Erasmus University Rotterdam

Erasmus School of Economics

Master Thesis

Non stationary extreme value theory applied to market risk in the U.S.

Thierry Volker

575020

April 26, 2022

Supervisor: prof. dr. C. Zhou

Second assessor: dr. A.J. Koning

Abstract

This research investigates a non stationary EVT approach, applied to market risk. The excess probability and the conditional excess distribution are modelled by covariates, using a LASSO for regularization. In addition, the excess probability is modelled by a Markov switching model. I test different model setups with a traditional backtest and a comparative backtest. While most comparative tests are statistically inconclusive, the results suggest that the non stationary EVT approach as it is defined in this study is a promising extension of stationary EVT. Also, there is a significant difference in performance when the market is in a calm state. The VIX is a crucial covariate that captures market uncertainty and volatility. A benchmark GARCH approach and the Markov switching model confirm the importance of modelling market uncertainty. One configuration of the non stationary EVT approach could serve as a more stable alternative for the GARCH-based model, which is favorable from a practitioner's point of view.

Erasmus Universiteit Rotterdam

Contents

1 Introduction

Financial markets have received significant attention from researchers for many decades. Since 'Black Monday' in 1987, market crashes are also studied in detail because of the impact of such events. The risk of a market crash falls under market risk, which is one of the main drivers of risk for financial institutions. Large market losses apply pressure on financial institutions, which sometimes causes them to fail. As a result, stock market crises are often accompanied with an economic downturn, including budget cuts, lost jobs and pensions, etc. Hence, severe stock market losses are outcomes that touch upon the lives of all kinds of people, including the average Joe. However, they also seem to be inherent to the way our society and economic policies are structured. Therefore, it is important to be well prepared when a crisis hits. At the same time, institutions should not be too conservative in their decisions, since this implies a waste of resources.

Risk management focuses on balancing the risks of highly unfavorable outcomes while using resources efficiently. To do so, the key is to have an accurate understanding of what the risks are. In the case of market risk, this entails that researchers try to estimate the probability that outcomes in the right tail of the loss distribution materialize. Based on those estimates, financial institutions save capital to overcome market downturns. This is not only important for the health and endurance of the institution itself; it is also mandatory. Regulators verify if financial institutions abide regulatory capital requirements and if the model they use for estimating market risk is adequate.

This research focuses on using extreme value theory (EVT) to study market risk. I investigate whether current EVT methods for assessing market risk can be improved by using non stationary models, which would imply better estimates of the risk with all its benefits. Using the peak-over-threshold (POT) method, the probability that a high loss occurs is split into the probability that a loss exceeds some threshold and the probability that certain losses above that threshold are observed, given that the threshold is exceeded. I refer to the first probability as the excess probability (or probability of exceedance). The second probability is characterized by the conditional excess distribution. Together, they are used to estimate the distribution of high losses. Gencay and Selcuk (2004) use the POT method in the context of market risk. Both the excess probability and the conditional excess distribution are assumed to be static in their research. Hambuckers et al. (2018a) model the parameters of the conditional excess distribution with covariates, applying regularization in the process. However, their area of application is operational risk.

This study adds to the current literature in two ways. Firstly, I extend the non stationary approach of Hambuckers et al. (2018a) to market risk by using covariates to estimate the parameters of the conditional excess distribution. Second, I estimate the excess probability using covariates via a logistic model and a Markov switching model. Together, the approach is a non stationary version of the standard EVT method for estimating market risk (e.g. in Gencay and Selcuk (2004)), hence the name 'non stationary EVT'. This approach reveals the role that covariates play in estimating market risk. In addition, the information that those covariates hold is used to possibly improve the estimates of the likelihood of high losses. To the best of my knowledge, non stationary EVT has not been applied to market risk at this point. I apply the analysis to U.S. stock market and use 48 covariates of various kinds, with data spanning from August 2001 to October 2021.

The extensions that non stationary EVT adds are tested using combinations of stationary and non stationary models, such that the impact of each extension isolated. Furthermore, additional models that use extensions for both the excess probability and the conditional excess distribution are estimated to examine performance when they are combined. For the test procedure, I use the methodology in Nolde and Ziegel (2017). Their test compares methods, such that it can be determined if one approach significantly outperforms another.

The results imply that all EVT approaches are adequate risk measurement procedure. In addition, the comparative backtests show that there is no significant difference in performance between the models, except when a subset of the test data is used that is considered to be generated in a calm state of the market. In this subcase, the GARCH and two non stationary EVT models perform the best. Even though the differences are not significant in most of the tests, the non stationary EVT approach shows promise over the stationary EVT model. The VIX is a crucial covariate that captures market uncertainty and volatility. Including the VIX results in similar results to the GARCH approach in McNeil and Frey (2000), which models market volatility in a different manner. The model that use a Markov switching model for the excess probability and a stationary conditional excess distribution shows good performance in the comparative tests, while being relatively stable in comparison to the other well performing models. Hence, it could serve as a more practical alternative for the GARCH-based model.

The rest of the paper is structured as follows. I review literature that is closely related to this research in Section 2. In section 3, a brief data description is provided. I give an outline in section 4 on the methodology. In section 5, the results of the research are given and provided with interpretation.

2 Literature review

Using the POT method in a market risk setting has been studied in the literature. Gencay and Selcuk (2004) use EVT by assuming a stationary generalized Pareto distribution (GPD) as the conditional excess distribution. McNeil and Frey (2000) and Singh et al. (2013) introduce dynamics in the POT model via a GARCH$(1, 1)$. After fitting the GARCH model to the data, a GPD is fitted to the residuals. Their 'dynamic EVT' accounts for conditional heteroskedasticity in financial time series. This approach shows promise over the standard EVT approach in Gencay and Selcuk (2004).

However, in market risk studies that apply EVT, the parameters of the GPD (the conditional excess distribution) and the probability of exceedance are considered to be stationary in all aforementioned studies. The stationary GPD can be extended to a non stationary GPD, which falls under the generalized additive model of location, scale, and shape (GAMLSS) in Rigby and Stasinopoulos (2005). Non stationary GPD models are used in other fields of risk management. Chavez-Demoulin et al. (2016) fit a non stationary GPD to operational losses, in the sense that the parameters of the GPD are functions of a small number of covariates. Hambuckers et al. (2018a) extend the approach of Chavez-Demoulin et al. (2016) by adding more covariates to the model. They use regularization in the form of a LASSO (Tibshirani, 1996) to counter overfitting. Hence, their model selects the most informative covariates for estimating the GPD parameters. In contrast, Hambuckers et al. (2018b) add a hidden Markov model to the methodology of Chavez-Demoulin et al. (2016), thereby accounting for differing behavior of losses in prosperous and crisis states of the world. However, the setting of operational risk is different from market risk, since the data frequency is usually quarterly in the field of operational risk, with multiple losses in each quarter. The data frequency for market risk studies is often higher (mostly daily). If a loss at a certain day is not large enough to 'peak over the threshold', then there is no observation for that day in the sample of extreme losses. Therefore, using a Markov switching element for the GPD estimation in a POT setting while studying market risk is not sound, since the Markov chain would not be equally spaced. Hence, I follow Hambuckers et al. (2018a), and refrain from using a hidden Markov model for the conditional excess distribution.

In the setting of Hambuckers et al. (2018a), other forms of regularization could be considered. There are various extensions of the original LASSO. The adaptive LASSO (AdLASSO) of Zou (2006) applies a coefficient-specific weighting to the regularization penalty, which could help with achieving more stable estimates. Also, AdLASSO can help reduce the bias that comes with the use of a LASSO term. The elastic net of Zou and Hastie (2005) is a combination of ridge regression and LASSO, which is shown to select highly correlated variables as a group. In contrast, the LASSO tends to choose one of the correlated variables and discards the rest. Also, the elastic net is expected to behave more stable then the LASSO in the situation of highly correlated variables. However, I choose the original LASSO over the elastic net since the parsimony of the LASSO is desired in this study. Many of the included covariates are highly related, some even are constructs of each other. The fact that the LASSO chooses the most informative one and discards the rest yields a more informative model than when it would include three versions of the same variable. Also, the LASSO requires less tuning of hyperparameters. LASSO is also preferred over AdLASSO since the approach is simpler and the differences with LASSO are expected to be minor and not necessarily in favour of the AdLASSO (Hambuckers et al., 2018a).

3 Data

This study investigates the left tail behavior of the S&P 500 index (SPX). The daily returns of the SPX are given in Figure 1. The data spans from August 2$^{\text{nd}}$ 2001 to October 1$^{\text{st}}$ 2021, eventually yielding a sample size of 5072 trading days. I choose this specific time span since one explanatory variable is not available before August 2$^{\text{nd}}$ 2001. The origin of the entire data set is Bloomberg.

I use a set of 48 covariates to model the SPX. Table 9 in Appendix A shows cohesive list of all the 49 variables (including the SPX) that are included in the analysis. It consists of other world major indices (with their volume for the US indices), technical variables, the VIX, natural resources, exchange rates, several types of interest rates, and spreads between interest rates. The selection of the covariates is largely based on Zhong and Enke (2019), who gather covariates that are used in various previous articles that study daily stock returns. The aim in Zhong and Enke (2019) is to try to predict returns, which includes gains. In contrast, I look exclusive at large losses. However, I reason that variables that are possibly related to stock returns are also likely to be useful in predicting large losses. That being said, the set of covariates in Table 9 is slightly different from the set of variables in Zhong and Enke (2019). These differences are motivated by the difference in nature between their study and this one and also reflect my own reasoning. Details on why certain variables are added or left out can be found in Appendix A.

The raw data has several missing values because of mismatching trading day schedules in different countries and randomly missing periods. As a solution, I take the SPX time series as the starting point of the cleaning process, in the sense that all the periods that correspond to the missing values in the SPX series are removed. This is motivated by the fact that the SPX is the endogenous variable. After this round of cleaning, the data only consists of observations where the SPX is observed. This however leaves several missing values in the series of the explanatory variables, which are resolved by linear interpolation. The vast majority of the gaps that are interpolated consist of one or two missing observations. For most variables, only one to four percent of the observations is interpolated.

After the cleaning process, the data is transformed. The SPX index is converted to negative log returns. The same transformation is applied to all the explanatory variables with exception of the the VIX, interest rates and interest rate spreads. In addition, all variables that are not the SPX or lags of the SPX are lagged, since outcomes of covariates should explain the SPX behavior of the next day. By choosing log differences for most of the explanatory variables, I assume that shocks explain shocks— that is, that a shock in the SPX return is explained by a shock in the explanatory variables, instead of the levels of those variables. Also, many covariates are clearly non stationary. Hence, considering levels of those variables would be uninformative.

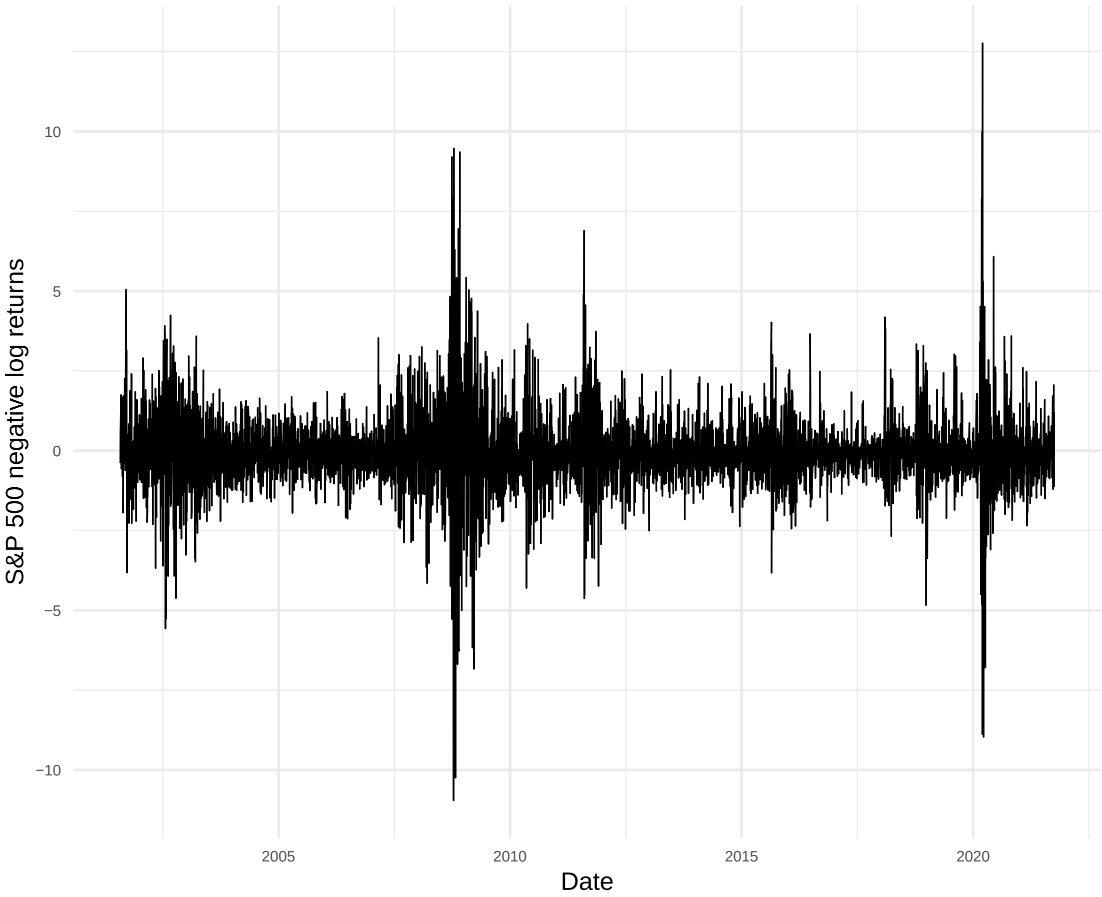


Figure 1: The S&P 500 negative log returns over the entire period spanned by the data.

I use different transformations for interest rates and the VIX. Firstly, the interest rates are converted to differences instead of log returns. Interest rates can be interpreted as a return, so considering differences instead of log returns is sensible. Second, the VIX is used in levels instead of (log) differences. Since the VIX and SPX returns are expected to be heavily connected in real time (they react to each other fast), a significant loss on a certain day is usually accompanied by a high positive change in the VIX on that same day. However, lags of the explanatory variables are used in this research, and the relative change of the VIX is not expected to have a significant effect on the relative change of the SPX on the next day. Therefore, I choose to use the VIX in levels, since the level of this variable holds information in itself. Time spans with high VIX levels are considered as uncertain times, which is expected to have an impact on stock returns.

4 Methodology

The analysis focuses on the loss of the SPX, denoted X with an unknown distribution $F(x)$. The aim is to estimate the value at risk (VaR), $\mathrm{VaR}_\alpha = F^{-1}(\alpha)$, which is a quantile of the distribution of X. The focus lies on the right tail of the loss distribution, so α is close to one. The tail can be modeled using extreme value theory (EVT). In fact, with EVT one considers only the tail and is not attempting to find a complete distribution for the loss X. Instead, the tail, which is the part that is of interest to risk managers, is modelled exclusively and therefore potentially more accurately.

There are two main approaches in the EVT framework: the block-maxima and the peak-over-threshold (POT) methods. In this article, the peak-over-threshold (POT) method is considered. The more traditional block-maxima method is found to make inefficient use of the data (Singh et al., 2013). Also, the POT method has a straightforward extension to computing the VaR, which is the end goal of modelling the tail of the distribution of X. In the classic POT framework where stationarity of the parameters is assumed, a high threshold value is fixed after which the outcomes of X that exceed this threshold are modelled. I assume that for X above a certain threshold u, the excess loss $Y = X - u$ follows a generalized Pareto distribution (GPD):

$$\mathrm{GPD}(y; \xi, \beta) = \begin{cases} 1 - (1 + \xi \frac{y}{\sigma})^{-\frac{1}{\xi}} & \text{if } \xi \neq 0 \\ 1 - \exp(-\frac{y}{\sigma}) & \text{if } \xi = 0, \end{cases} \tag{1}$$

with $y \geq 0$, $\xi \in \mathbb{R}$, and $\sigma > 0$. If $\xi < 0$, then $0 < y < -\sigma/\xi$. In the case of $\xi = 0$, the GPD takes an exponential form. The assumption that the conditional excess distribution Y (approximately) follows a GPD stems from the Gnedenko and Pickands–Balkema–De Haan theorems (Gnedenko (1943); Balkema and De Haan (1974); Pickands III (1975)). For the SPX, $\xi > 0$ is considered—that is, the case where the data is heavy tailed (Chavez-Demoulin et al., 2016).

The relevance of equation (1) follows after the distribution of X is rewritten as

$$\begin{aligned} F(x) &= 1 - P(X > x) \\ &= 1 - P(X > u)P(X - u > x - u | X > u) \\ &= 1 - \bar{F}(u)\bar{F}_u(y) \\ &= 1 - \bar{F}(u)(1 - F_u(y)), \end{aligned}$$

for $X > u$, where $y = x - u$, $\bar{F}(u) = P(X > u)$ is the probability of exceeding u and $F_u(y) = P(X - u \leq y | x > u)$ is the conditional excess distribution. As mentioned above, it is assumed that $F_u(y) = \text{GPD}(y; \xi, \sigma)$. Hence, if follows that

$$F(x) = 1 - \bar{F}(u)(1 - \text{GPD}(y; \xi, \sigma)), \tag{2}$$

and, after inverting $F(x)$ to find the VaR, that

$$\text{VaR}_\alpha = u + \frac{\sigma}{\xi}\left(\left(\frac{1-\alpha}{\bar{F}(u)}\right)^{-\xi} - 1\right). \tag{3}$$

Estimating the VaR requires an estimate of ξ, σ and $\bar{F}(u)$, which is done via maximum likelihood estimation. Similar to Chavez-Demoulin et al. (2016), the likelihoods for estimating the probability of exceedance and the parameters in the conditional excess distribution are estimated separately, since they are assumed to be independent after conditioning on an information set, e.g. covariates.

4.1 Estimation of the probability of exceedance

For ease of notation, define $\pi_u = \bar{F}(u) = P(X > u)$ as the excess probability. The simplest approach for estimating π_u is to use the empirical distribution of X (McNeil and Frey (2000); Gencay and Selcuk (2004)). Let T be the size of entire sample, which is split into a training and test set. If τ is the length of the training set, E is the set of samples in the training set that exceed the threshold u, and $N = |E|$ is the cardinality of the set E (its size), then the empirical distribution estimate is simply the ratio

$$\hat{\pi}_u^{\text{emp}} = \frac{N}{\tau}. \tag{4}$$

Since π_u appears in equation (3), it influences the VaR estimate directly. Thus, if estimates of π_u could be improved, estimates of the VaR would improve as well. I attempt to do that by considering a dynamic π_u.

4.1.1 Conditioning π_u on covariates

To achieve a dynamic probability, the first method that I consider is to condition π_u on a set of covariates. This results in $\pi_u(\boldsymbol{z}_t)$, $t \in \{1, ..., T\}$, with $\boldsymbol{Z} = (\boldsymbol{z}_1, ..., \boldsymbol{z}_T)'$ as the $T \times (M+1)$ matrix where $\boldsymbol{z}_t = (1, z_{1,t}, ..., z_{M,t})'$ consists of a one for including a constant and M covariates. $\boldsymbol{z}_t$ stores the values of the covariates at time $t-1$, which are used to model X_t. This allows for using the model in practice. The reasoning behind using $\pi_u(\boldsymbol{z}_t)$ is that movements or levels of certain variables could have an impact on the excess probability, and that accounting for this should result in more accurate estimates.

To estimate $\pi_u(\boldsymbol{z}_t)$, the problem can be reformulated as a classification problem. If $B_t \sim \text{Bern}(\pi_u(\boldsymbol{z}_t))$ is a Bernoulli random variable such that

$$B_t = \begin{cases} 1 & \text{if } X_t > u \\ 0 & \text{if } X_t \leq u, \end{cases}$$

then $\pi_u(\boldsymbol{z}_t)$ can be estimated via various methods. In this study, a logistic model is used for its estimation. Defining $\boldsymbol{\beta} = (\beta_0, ..., \beta_M)'$ as the vector of coefficients gives

$$\pi_u(\boldsymbol{z}_t \mid \boldsymbol{\beta}) = \frac{exp(\boldsymbol{\beta}'\boldsymbol{z}_t)}{1 + exp(\boldsymbol{\beta}'\boldsymbol{z}_t)}. \tag{5}$$

$\boldsymbol{\beta}$ can be estimated via maximum likelihood estimation. From Cameron and Trivedi (2005), the likelihood function of a Bernoulli random variable $B_t \sim \text{Bern}(\pi_u(\boldsymbol{z}_t \mid \boldsymbol{\beta}))$ for $t = \{1, ..., \tau\}$ is given by

$$L(\boldsymbol{\beta}|\boldsymbol{b}, \boldsymbol{Z}) = \prod_{t=1}^{\tau} (\pi_u(\boldsymbol{z}_t \mid \boldsymbol{\beta})^{b_t} (1 - \pi_u(\boldsymbol{z}_t \mid \boldsymbol{\beta}))^{1-b_t}), \tag{6}$$

where $\boldsymbol{b}$ is the τ-vector of realisations of B_t. The log-likelihood then follows as

$$l(\boldsymbol{\beta}|\boldsymbol{b}, \boldsymbol{Z}) = \sum_{t=1}^{\tau} \big\{ b_t \ln \pi_u(\boldsymbol{z}_t \mid \boldsymbol{\beta}) + (1 - b_t) \ln(1 - \pi_u(\boldsymbol{z}_t \mid \boldsymbol{\beta})) \big\}. \tag{7}$$

Maximizing equation (7) with respect to $\boldsymbol{\beta}$ yields the maximum likelihood estimates $\hat{\boldsymbol{\beta}}$.

4.1.2 Regularization

Since the number of covariates that can be considered for modelling stock return dynamics is high, I use regularization to avoid overfitting. In most applications, the aim of using regularization is to achieve two things. Firstly, the estimates of the coefficients should shrink, which should provide better out-of-sample estimation. Second, regularization should provide an interpretable model via forcing coefficients to exactly zero, where the coefficients of relatively uninformative variables are set to zero first. If a coefficient is zero, the variable corresponding to it is essentially excluded from the model. Hence, regularization can also perform variable selection.

Not all regularization methods achieve both things. For example, ridge regression is excellent at shrinking coefficients towards zero but fails to set them at exactly zero. I use the least absolute shrinkage and selection operator (LASSO), which was first proposed by Tibshirani (1996). In contrast to ridge regression, LASSO does force coefficients to zero, which is of importance in this study since there are many variables to consider. Defining $\mathcal{L}_{\pi_u}(\cdot)$ as the penalized log-likelihood for estimating π_u gives

$$\mathcal{L}_{\pi_u}(\boldsymbol{\beta}|\lambda, \boldsymbol{b}, \boldsymbol{Z}) = l(\boldsymbol{\beta}|\boldsymbol{b}, \boldsymbol{Z}) - \lambda \sum_{m=1}^{M} |\beta_m|, \tag{8}$$

where $\boldsymbol{Z}$ is standardized and $\lambda > 0$ is a hyperparameter that needs to be tuned. The LASSO term in equation (8) ensures that a variable selection takes place, such that only the (most) informative covariates will have a nonzero coefficient. In addition, it shrinks the nonzero coefficients such that out-of-sample variance is reduced. This process of shrinking introduces a bias however, as is always the case with shrinkage. The hyperparameter λ controls the intensity of the regularization, so choosing λ amounts to choosing a bias-variance trade-off. Furthermore, increasing λ drives more coefficients to zero. Hence, λ also determines the level of parsimony of the model. I select λ by using the Bayesian information criterion (BIC). For a given λ, the BIC is given by

$$\mathrm{BIC}(\lambda) = -2\mathcal{L}_{\pi_u}(\hat{\boldsymbol{\beta}}|\lambda, \boldsymbol{b}, \boldsymbol{Z}) + \ln(\tau)\mathrm{df}(\lambda), \tag{9}$$

where df(λ) is the estimated degrees of freedom of the model that follows from using λ. The value of λ that minimizes equation (9) is selected. The maximum likelihood estimates $\hat{\boldsymbol{\beta}}$ are found by maximizing the penalized likelihood:

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \ \mathcal{L}_{\pi_u}(\boldsymbol{\beta}|\lambda, \boldsymbol{y}, \boldsymbol{Z}). \tag{10}$$

df(λ) is estimated by the number of nonzero coefficients in the final solution of the LASSO, as in Zou et al. (2007). I prefer the BIC over the Akaike information criterion (AIC) since it generally yields more parsimonious models (Hambuckers et al., 2018a). Cross-validation techniques could also be used but are not considered due to long computation times.

4.1.3 Markov switching

Inspecting Figure 1 suggests that there are periods were high losses are documented relatively often, usually accompanied by high volatility in the market. In contrast, there are also period were the market is more calm with relatively few high losses. I use a Markov switching model to account for this difference in the behavior of the market. With a Markov switching model, one assumes that the process under investigation can be in one of multiple states, or regimes. In this application, I assume that there are two state: state 0, a state of expansion, and state 1, a state of crisis. More states could be considered if desired. A first order Markov process governs the transition between the states. The transition matrix

$$\boldsymbol{P} = \begin{pmatrix} p_{00} & p_{01} \\ \\ p_{10} & p_{11}, \end{pmatrix} \tag{11}$$

with $p_{ij} = p_{i|j} = P(S_t = i \mid S_{t-1} = j)$, where S is the unobserved state process and $i, j = \{0, 1\}$. I use an expectation-maximization (EM) algorithm to estimate the parameters of the model. The complete likelihood function is given by

$$L(\boldsymbol{\beta}|y_t, S_t) = \prod_{t=1}^{\tau} \prod_{i,j=0}^{1} \{p_{ij}(\pi_u^{(i)})^{y_t}(1-\pi_u^{(i)})^{1-y_t})\}^{\delta_{ij}(t)}, \tag{12}$$

where $\delta_{ij}(t)$ is an indicator variable that equals 1 if $S_t = i$ and $S_{t-1} = j$ and $\pi_u^{(i)}$ is the probability of exceedance in state i. The log likelihood follows as

$$l(\boldsymbol{\beta}|y_t, S_t) = \sum_{t=1}^{\tau} \sum_{i,j=0}^{1} \Big\{\delta_{ij}(t)\Big(\ln p_{ij} + y_t \ln \pi_u^{(i)} + (1-y_t)\ln(1-\pi_u^{(i)})\Big)\Big\}. \tag{13}$$

However, one does not observe S for any t. Using $\tilde{E}[\cdot]$, which is defined as an expectation over the delta's conditional on the information set $\mathcal{I}_\tau$, gives

$$\tilde{E}[\mathcal{L}(\boldsymbol{\beta}|y_t)] = \sum_{t=1}^{\tau} \sum_{i,j=0}^{1} \Big\{p_{ij}^*(t)\Big(\ln p_{ij} + y_t \ln \pi_u^{(i)} + (1-y_t)\ln(1-\pi_u^{(i)})\Big)\Big\}, \tag{14}$$

where $p_{ij}^* = P(S_t = i, S_{t-1} = j|\mathcal{I}_\tau)$. After finding the p_{ij}^* probabilities (E-step), one can maximize equation (14) with respect to $\pi_u^{(0)}$ and $\pi_u^{(1)}$ (M-step). These steps iterate until convergence. Further details can be found in Appendix B.

4.2 Estimating the conditional excess distribution

In market risk studies applying EVT, the parameters of the GPD have been considered to be stationary (McNeil and Frey (2000); Gencay and Selcuk (2004)). Following the methodology of Chavez-Demoulin et al. (2016) and Hambuckers et al. (2018a), this study extends the non stationary GPD to market risk. The excesses are modelled by a GPD as in equation (1), but with parameters that depend on covariates. This yields the excess variable $Y_n \sim \text{GPD}(y; \xi_n, \sigma_n)$ with Y_n, ξ_n, $\sigma_n > 0$, and $n \in E$ with E as defined in section 4.1. As in Chavez-Demoulin et al. (2016), I reparameterize σ_n, to create orthogonality with respect to the Fisher information metric. This avoids problems with the fitting procedure. The reparameterization is given by

$$\sigma_n = \frac{\exp(\kappa_n)}{1+\xi_n} \tag{15}$$

such that the $Y_n \sim \text{GPD}(y; \xi_n, \exp(\kappa_n)/(1+\xi_n))$. The reparameterization holds for $\xi_n > -1$ since $\sigma_n > 0$ for those values of ξ_n. This is a weak assumption since in general $\xi_n > 0$ holds for financial data. In this study, $\xi_n > 0$ is assumed. $\xi_n = \xi(\boldsymbol{z}_n)$ and $\kappa_n = \kappa(\boldsymbol{z}_n)$ are functions of covariates that vary over time. These functions can be specified as

$$\begin{aligned} \ln(\xi(\boldsymbol{z}_n)) &= a_0^{\xi} + \sum_{m=1}^{M} h_{\xi,j}(z_{m,n}^{\xi}) \\ \kappa(\boldsymbol{z}_n) &= a_0^{\kappa} + \sum_{m=1}^{M} h_{\kappa,j}(z_{m,n}^{\kappa}) \end{aligned} \tag{16}$$

where the log link function is applied in the first line to ensure that $\xi_n > 0$, and $h(\cdot)$ is a general function of the covariates that can be specified by the researcher. Chavez-Demoulin et al. (2016) use spline functions to characterise the effect of the covariates on the parameters. However, I follow Hambuckers et al. (2018a) in choosing h to be a linear function, for simplification purposes, turning the model into a generalized linear model for location, scale, and shape (GLMLSS) (Groll et al., 2019). Also, one could use distinct covariates for estimating for both ξ_n and κ_n, but since there is no apparent reason to do so in this study, I choose to set $\boldsymbol{z}_n^{\xi} = \boldsymbol{z}_n^{\kappa}$. The resulting GPD parameter equations are

$$\begin{aligned} \ln(\xi(\boldsymbol{z}_t)) &= a_0^{\xi} + \sum_{m=1}^{M} a_m^{\xi} z_{m,t} \\ \kappa(\boldsymbol{z}_t) &= a_0^{\kappa} + \sum_{m=1}^{M} a_m^{\kappa} z_{m,t}. \end{aligned} \tag{17}$$

Maximum likelihood estimation is used to estimate the parameters $\boldsymbol{a}^\xi = (a_0^\xi, ..., a_M^\xi)$ and $\boldsymbol{a}^\kappa = (a_0^\kappa, ..., a_M^\kappa)$. If $\boldsymbol{\theta} = (\boldsymbol{a}^\xi, \boldsymbol{a}^\kappa)'$ is the vector of all parameters to be estimated, then the log likelihood function is given by

$$l(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{Z}) = \sum_{n=1}^{N} \ln\big(\text{gpd}(y_n; \xi_n, \exp(\kappa_n)/(1+\xi_n))\big), \tag{18}$$

with gpd$(\cdot)$ being the probability density function of the GPD, $N = |E|$ as defined in section 4.1, and $\boldsymbol{y} = (y_1, ..., y_N)$ being the vector of realisations of the excesses. Similar to equation (8), define $\mathcal{L}_{\text{GPD}}$ as the penalized log likelihood function such that

$$\mathcal{L}_{\text{GPD}}(\boldsymbol{\theta} \mid \boldsymbol{\nu}, \boldsymbol{y}, Z) = l(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{Z}) - \nu^\xi \sum_{m=1}^{M} |a_m^\xi| - \nu^\kappa \sum_{m=1}^{M} |a_m^\kappa|. \tag{19}$$

where the constants a_0^ξ and a_0^κ are assumed not to be subject to regularization (Hambuckers et al., 2018a). The hyperparameters $\boldsymbol{\nu} = (\nu^\xi, \nu^\kappa)$ allow for a different power of the regularization for ξ_n and κ_n, and need to be tuned. This is done similarly to the procedure in section 4.1.2, with a two-dimensional grid search for the penalty parameters. The penalized maximum likelihood estimates are given by

$$\hat{\theta} = \arg\max_{\boldsymbol{\theta}} \ \mathcal{L}_{\text{GPD}}(\boldsymbol{\theta} \mid \boldsymbol{\nu}, \boldsymbol{y}, \boldsymbol{Z}) \tag{20}$$

For optimizing equation (20), I follow Hambuckers et al. (2018a) and use a quadratic approximation for the L_1-penalty terms, such that the maximization problem can be linearized(Oelker and Tutz, 2017). Further details on the estimation algorithm can be found in Appendix C

4.3 Threshold selection

So far, the threshold u has been assumed to be given. However, there is no exact rule that always determines the appropriate u. The selection of the threshold determines the amount of data points that are considered to be 'in the tail'. Here one considers a bias-variance trade-off— if too few points are included, the bias will be low since all points selected are safely regarded as extremes. In contrast, the variance of the estimates will be high because of a smaller sample, which can result in poor out-of-sample estimates. The converse is true when selecting too many points, which could also yield poor out-of-sample estimation due to a high bias. Because of the importance of the threshold selection, Gencay and Selcuk (2004) advise to use a combination of techniques.

I construct a Q-Q plot to gain a general understanding of the fat-tailedness of the loss data.

In addition, a mean-excess function (MEF) and plot give insight in where the data is heavy-tailed and equivalently where a GPD approximation is good fit. The MEF for a threshold v is given by

$$e(v) = \frac{\sum_{t=1}^{\tau} \max(X_t - v, 0)}{\sum_{t=1}^{\tau} I_{\{X_t > v\}}}, \tag{21}$$

where v can be varied and $I_{\{.\}}$ is an indicator function. The mean excess plot evaluates at the sequence $\{X_{(t)}, e(X_{(t)}) : 2 \leq t \leq \tau\}$, where $X_{(t)}$ is the t_{th} order statistic of X with $X_{(1)} \leq ... \leq X_{(\tau)}$. If the plot shows an upward linear trend above some threshold v then a GPD approximation should be accurate above v (Singh et al., 2013).

Lastly, I use a Hill plot, yielding more evidence on which threshold u is suitable to use. Hill (1975) suggests an estimator for the tail index γ, which is the inverse of ξ for the case where $\xi > 0$, i.e. the fat-tailed case. Given the order statistics $X_{(t)}$, the Hill-estimator is given by

$$\gamma^{\text{Hill}} = \big[\frac{1}{k}\sum_{i=1}^{k} \ln(X_{(t-i+1)}) - \ln(X_{(t-k)})\big]^{-1}, \tag{22}$$

where k is the pre-specified number of observations that exceed u. Varying k provides the Hill plot, which can be used for selecting the appropriate k. The aim is to select a k where the bias is not yet too much present, while trying to include as many points in the excess sample as possible to reduce variance. With k, u equivalently follows.

4.4 Model testing

I consider the following models to estimate the VaR in equation (3):

- Model 0: the dynamic GARCH$(1, 1)$ EVT approach in McNeil and Frey (2000).
- Model 1: estimate π_u with the empirical distribution and (ξ, σ) with the stationary GPD (equations 4 and 1, Gencay and Selcuk (2004)).
- Model 2: estimate π_u by conditioning on covariates and (ξ, σ) with the stationary GPD (equations 10 and 1).
- Model 3: estimate π_u with a Markov switching model and (ξ, σ) with the stationary GPD (section 4.1.3 and equation (1)).
- Model 4: estimate π_u with the empirical distribution and (ξ, σ) with the non stationary GPD (equations 4 and 20).

- Model 5: estimate π_u by conditioning on covariates and (ξ, σ) with the non stationary GPD (equations 10 and 20).
- Model 6: estimate π_u with a Markov switching model and (ξ, σ) with the non stationary GPD (section 4.1.3 and equation (20)).

For ease of notation, I define $\mathcal{M}_i$ as model i, $i = \{0, ..., 6\}$. $\mathcal{M}_0$ is used as a benchmark, since it is a well-known EVT approach to estimate the VaR. $\mathcal{M}_0$ fits a GARCH$(1, 1)$ model (Bollerslev, 1986) via quasi-maximum likelihood estimation. EVT is then applied to the residuals implied by the GARCH model. Together with the GARCH parameter estimates, this yields the VaR for the next day. For further details see Singh et al. (2013)).

Comparing $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$ to each other gives information on how much the VaR estimates are influenced by using more involved methods to estimate π_u. Additionally, comparing these three models to $\mathcal{M}_4$, $\mathcal{M}_5$ and $\mathcal{M}_6$ provides information on the effect of using covariates to model the GPD parameters. Lastly, comparing $\mathcal{M}_j$ for $j = \{2, ..., 6\}$ to $\mathcal{M}_0$ and $\mathcal{M}_1$ shows the effectiveness of the extensions that are considered in this study.

4.4.1 Test procedure

To test the performance of all the models, the data is split into a training set with observations $\{1, ..., \tau\}$ and a test set with observations $\{\tau + 1, ..., T\}$. The last 500 trading days of the sample are selected as the test set. On each day of the test set, the models are re-estimated using an expanding window. Hence, for making a prediction for the VaR at day $t \in \{\tau + 1, ..., T\}$, the information up and until day $t - 1$ is included. This approach makes use of the all the newly available data at a point in time.

I split the test procedure in two stages, following the suggestion in Nolde and Ziegel (2017). In the first stage I test whether models $\mathcal{M}_i$, $i = \{0, ..., 6\}$, can be rejected or not. Backtests that deal with this kind of question are referred to as 'traditional backtests'. The models that pass the test in the first stage advance to the second stage, where the performance of those models is compared. This gives the ability to conclude whether certain models perform significantly better than others.

I use the test proposed by Christoffersen (1998) as a traditional backtest, which is closely related to the first stage tests in Nolde and Ziegel (2017). Traditional backtests evaluate the following null hypothesis (Nolde and Ziegel, 2017):

$$\text{H}_0\text{:}\quad \text{“The risk measurement procedure } \mathcal{M}_i \text{ is correct.”}$$

Hence, these tests only provide judgement on whether a model is adequate or not. Still, the outcome of traditional backtests gives valuable information on the overall quality of a risk measurement procedure.

On each day of the test set, a one day ahead prediction of the VaR for next day is made, which is denote by r_t. After observing the outcome, it is evaluated if the loss realization x_t exceeded r_t. V_t is defined as an indicator variable of whether or not a violation occurred at day t:

$$V_t = \begin{cases} 1 & \text{if } x_t > r_t \\ 0 & \text{if } x_t \leq r_t. \end{cases}$$

$V = \sum_{t=\tau+1}^{T} V_t$ is the total number of violations in the test set. Evaluating the violations allows for testing the accuracy of the VaR estimates.

The test of Christoffersen (1998) uses likelihood ratio tests to test for unconditional coverage and independence over time. Given a confidence level α for VaR_α, unconditional coverage implies that the chosen coverage rate $(1-\alpha)$ is equal to the expectation of a violation, i.e. $E[V_t = 1] = 1-\alpha$ for all t. The property of independence demands that the VaR violations in the test set are independent from each other. Using a Markov chain, the impact of the outcome of V_{t-1} on V_t is tested. If being state 1 means that the last observation was a violation and state 0 means no violation, the transition matrix is given by

$$\boldsymbol{Q} = \begin{pmatrix} q_{00} & q_{01} \\ q_{10} & q_{11} \end{pmatrix}, \tag{23}$$

where $q_{ij} = q_{i|j} = P(V_t = i \mid V_{t-1} = j)$. The maximum likelihood estimate of Q is given by

$$\hat{\boldsymbol{Q}} = \begin{pmatrix} \hat{q}_{00} & \hat{q}_{01} \\ \hat{q}_{10} & \hat{q}_{11} \end{pmatrix} = \begin{pmatrix} \frac{n_{00}}{n_{00}+n_{10}} & \frac{n_{01}}{n_{01}+n_{11}} \\ \frac{n_{10}}{n_{00}+n_{10}} & \frac{n_{11}}{n_{01}+n_{11}} \end{pmatrix}, \tag{24}$$

where $n_{ij} = n_{i|j}$ is the number of times state j is followed by state i. Equation (24) allows for testing the independence property. Combining the tests for unconditional coverage and independence (for details see Christoffersen (1998)) yields the conditional coverage (CC) test statistic

$$\mathrm{LR}_{\mathrm{cc}} = -2\ln\left(\frac{\alpha^{T^\star - V}(1-\alpha)^V}{(\hat{q}_{00})^{n_{00}}(\hat{q}_{01})^{n_{01}}(\hat{q}_{10})^{n_{10}}(\hat{q}_{11})^{n_{11}}}\right) \sim \chi^2(2), \tag{25}$$

where $T^\star$ is the amount of observations in the test set. After evaluating the test statistic, H_0 is rejected or not. If H_0, it does not mean that the risk measure actually is correct, since the type I error of the test can not be controlled (Nolde and Ziegel, 2017).

In the second stage of the testing procedure, D is defined as the set of models that pass the traditional backtest. The aim of comparative testing is to assess if $\mathcal{M}_i$ performs significantly better than $\mathcal{M}_j$, with $i, j \in D, i \neq j$. The following null hypotheses are therefore tested (Nolde and Ziegel, 2017):

H_0^-: "The risk measurement procedure $\mathcal{M}_i$ performs at least as well as $\mathcal{M}_j$."

H_0^+: "The risk measurement procedure $\mathcal{M}_i$ performs at most as well as $\mathcal{M}_j$."

If H_0^- is rejected, then $\mathcal{M}_i$ performs worse than $\mathcal{M}_j$. If H_0^+ is rejected, then $\mathcal{M}_i$ performs better than $\mathcal{M}_j$. It is also possible that both H_0^- and H_0^+ cannot be rejected. In that case, the comparative tests cannot detect a significant difference between the performance of $\mathcal{M}_i$ and $\mathcal{M}_j$.

To test H_0^- and H_0^+, I follow Nolde and Ziegel (2017). Their test is based on the concept of elicitability and uses score functions that evaluate the quality of a forecast r. A risk measure (for example VaR) is elicitable if there exists a strictly consistent score function for it. A score function C is strictly consistent for the VaR if

$$E(C(\mathrm{VaR}^*, X)) < E(C(r, X)), \tag{26}$$

where r is the forecast estimate of the VaR, VaR^* is the VaR of the true unknown distribution F(x), and $r \neq \mathrm{VaR}^*$. A version for multidimensional risk measures can be found in Nolde and Ziegel (2017).

For the VaR measure, strictly consistent score functions are of the type

$$C(r, x) = (1 - \alpha - \mathbb{1}\{x > r\})G(r) + \mathbb{1}\{x > r\}G(x), \tag{27}$$

for an increasing function G (Nolde and Ziegel, 2017). Hence, the VaR is an elicitable risk measure, which allows for comparative backtesting. Following Nolde and Ziegel (2017), I choose G as the identity function. As a result, the score function becomes a hinge loss function:

$$C(r, x) = (1 - \alpha - \mathbb{1}\{x > r\})r + \mathbb{1}\{x > r\}x. \tag{28}$$

Given a scoring function C, it is defined that $\mathcal{M}_i$ C-dominates $\mathcal{M}_j$ on average if

$$E(C(r_{i,t}, X_t) - C(r_{j,t}, X_t)) \leq 0, \quad \text{for all } t. \tag{29}$$

Furthermore, it is assumed that a quantity $\omega \in \mathbb{R}$ exists and is given by

$$\omega = \lim_{T\to\infty} \frac{1}{T} \sum_{t=\tau+1}^{T} E(C(r_{i,t}, X_t) - C(r_{j,t}, X_t)), \tag{30}$$

where $r_{i,t}$ and $r_{j,t}$ are the forecasts of $\mathcal{M}_i$ and $\mathcal{M}_j$ respectively, with i and j defined as before. Under the mild assumption that the sequence $\{C(r_{i,t}, x_t) - C(r_{j,t}, x_t)\}_t$ is stationary, it follows that $\omega \leq 0$ if and only if $\mathcal{M}_i$ C-dominates $\mathcal{M}_j$. Conversely, $\omega \geq 0$ if and only if $\mathcal{M}_j$ C-dominates $\mathcal{M}_i$.

The hypotheses H_0^- and H_0^+ can be reformulated in terms of ω and therefore C-dominance as

$$\begin{aligned} \mathrm{H}_0^-: &\quad \omega \leq 0. \\ \mathrm{H}_0^+: &\quad \omega \geq 0. \end{aligned}$$

The hypotheses are tested with the test statistic

$$\Gamma = \frac{\Delta_T \overline{C}}{\hat{\Sigma}_T/\sqrt{T}}, \tag{31}$$

where $\hat{\Sigma}_T$ is an HAC estimator of the asymptotic variance $\Sigma = \mathrm{Var}(\sqrt{T}\Delta_T\overline{C})$ (e.g. Newey and West (1987); Andrews (1991)) and $\Delta_T\overline{C}$ is given by

$$\Delta_T \overline{C} = \frac{1}{T} \sum_{t=\tau+1}^{T} C(r_{i,t}, x_t) - C(r_{j,t}, x_t). \tag{32}$$

Under assumptions listed in Giacomini and White (2006), Γ is asymptotically standard normal. Hence, given a significance level η, H_0^- can be rejected if $1 - \Phi(\Gamma) \leq \eta$ and H_0^+ can be rejected if

$\Phi(\Gamma) \leq \eta$. If $\eta < \Phi(\Gamma) < 1-\eta$, then both H_0^- and H_0^+ cannot be rejected. This structure lends for a traffic light system. If H_0^+ is rejected and thus $\mathcal{M}_i$ passes the comparative backtest with respect to $\mathcal{M}_j$, then this is denoted by a green sign. If H_0^- is rejected and therefore the comparative backtest is failed by $\mathcal{M}_i$ with respect to $\mathcal{M}_j$, then this is denoted by a red sign. If $\eta < \Phi(\Gamma) < 1-\eta$, then $\mathcal{M}_i$ does not perform significantly better or worse than $\mathcal{M}_j$. This is denoted by an orange sign.

5 Results

This section consists of two parts. The first part discusses the estimation of the models in sections 4.1 and 4.2 on the training set. Also, various in-sample insights are highlighted. The second part is dedicated to testing the models on the test set as discussed in section 4.4.

5.1 In-sample results

The EVT analysis starts with the selection of a threshold. A threshold allows for labelling the data as an excess or no excess, which is the basis for estimating $\mathcal{M}_i$, $i = \{1, ..., 6\}$. The Q-Q plot in Figure 2 demonstrates that the data is heavy tailed. The mean-excess plot and the Hill plot in Figure 3 are used to determine a suitable threshold u. In the mean-excess plot, an upward sloping linear trajectory indicates heavy-tailedness that would be estimated well with a GPD. The mean-excess plot shows multiple instances of upward sloping linear parts of the mean-excess function. Those parts of the plot are ended abruptly by inconclusive or even downward sloping areas. This indicates that there are local areas that are consistent with a GPD configuration, but that there is a distortion (bias) between those configurations. The Hill plot confirms this. A stable horizontal part of the Hill plot indicates a stable tail index for the data, which is a measure for the heaviness of the tails. An upward sloping Hill plot (with ξ on the vertical axis) indicates an increasing bias. Figure 3b implies that there are two suitable candidates for a threshold. The first option u_1 is deep in the tail and is located at the point where the Hill plot stabilizes, making it a very suitable threshold. The second option u_2 is in the upward sloping part of the Hill plot in the grand scheme of things, but resides in a stable horizontal part of the trajectory. This means that u_2 is biased, but also representative for a part of the right tail of the sample. The Hill plot matches with Figure 3a, since u_1 and u_2 both mark the beginning of a stable upward sloping part of the mean-excess plot.

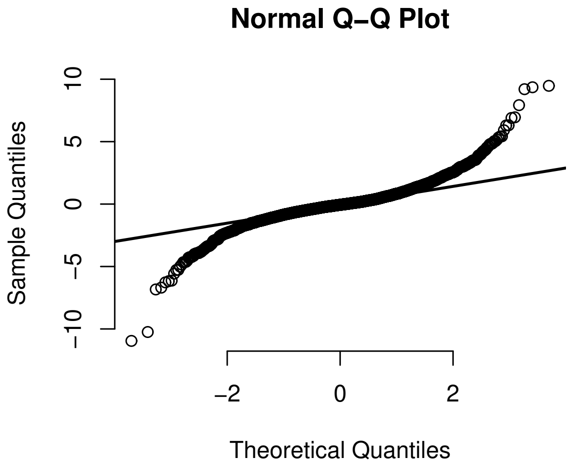
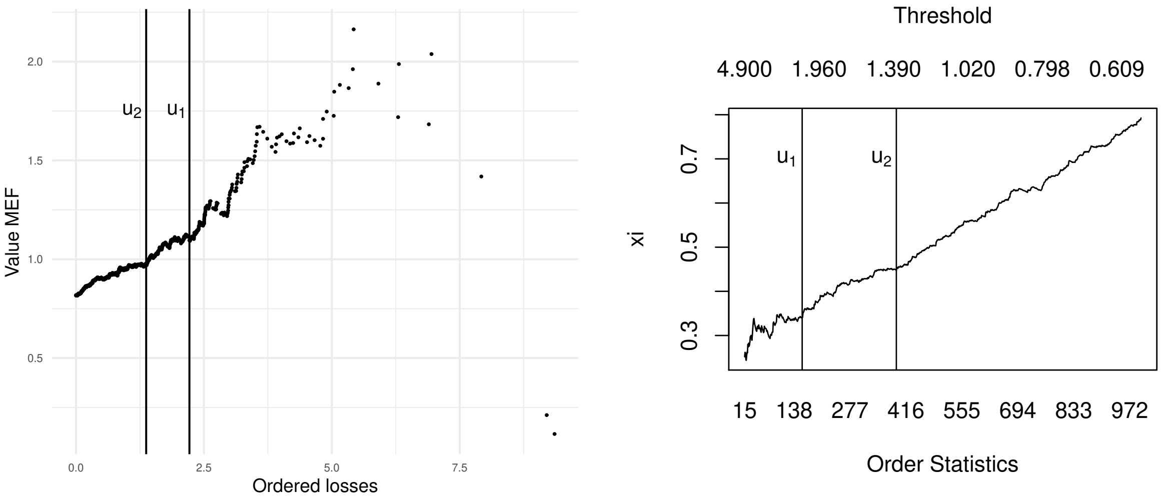


Figure 2: A Q-Q plot, where the log-returns of the SPX are plotted against the quantiles of a normal distribution.

Choosing between u_1 and u_2 is making a bias-variance trade-off. A bias is obviously undesirable in estimation procedures. However, including more data points in the excess sample results in better estimates in the sense of stability and precision. Also, the non stationary GPD is a challenging distribution to estimate and large samples, preferably still significantly larger than selecting u_2 would generate, are necessary (Hambuckers et al. (2018a); Groll et al. (2019)). Hence, I select u_2 as the threshold in this study, which yields an extreme sample of 392 observations. This is also supported by the relatively small difference in bias between u_1 and u_2. Going forwards, u_2 is referred to by u.



(a)

(b)

Figure 3: Figure 3a shows the mean-excess plot. The vertical axis gives the values of the mean-excess function against the ordered losses on the horizontal axis. The values of the ordered losses can also be interpreted as thresholds. Figure 3b shows the Hill plot. The vertical axis gives the tail index ξ, the bottom horizontal axis gives the amount of points included in the tail, and the top horizontal axis displays the corresponding threshold. Two possible options for a threshold, u_1 and u_2, are highlighted in both graphs.

Given the threshold selection, the models for the probability of excess and the conditional excess distribution are estimated for the training set. First, the results for the probability of excess models (section 4.1) are discussed. Searching over a grid for λ, the regularized conditional model for the excess probability of section 4.1.1 selects only the VIX. Figure 4 gives the path of the coefficients over a grid for λ. The estimates of the Markov switching probability of excess model of section 4.1.3 are given in Table 1. The states are persistent, with the 'good state' being more persistent than the 'bad state'. The high estimates of p_{11} and p_{22} indicate that periods of stability and uncertainty have a long expected duration, which coincides with perceived behavior of the world. Figure 5 gives the in-sample paths of the excess probability estimates, which highlights three interesting results.

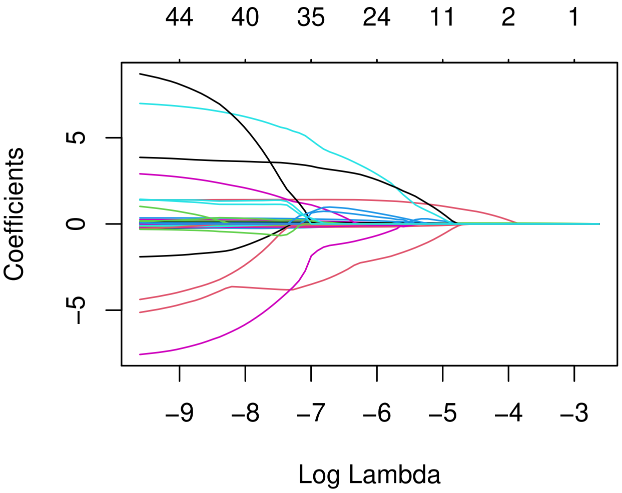


Figure 4: The values of the coefficients (vertical axis) are plotted against the values of the penalty term (bottom horizontal axis). The top horizontal axis shows the number of variables that are unequal to zero at the corresponding value of λ.

$\hat{p}_{11}$	$\hat{p}_{22}$	$\hat{\pi}_u^{(0)}$	$\hat{\pi}_u^{(1)}$
0.995	0.987	0.032	0.224

Table 1: The estimated parameters of the Markov switching model.

First, the conditional probability of excess and the Markov switching probability of excess estimates agree in their movements most of the time. This indicates that the VIX and the states structure in the Markov switching model capture a similar feature. One could label this feature as a measure of uncertainty and volatility. Second, the π_u model based on covariates is able to exceed the Markov switching π_u in its estimates. This is due to the structure of both models. The Markov model is bounded by $\hat{\pi}_u^{(1)}$, while the covariates model does not have this restriction. Therefore, the conditional model is more flexible than the Markov model. Third, the Markov model is less refined than the covariates model. The Markov model predicts its maximum and minimum value frequently, but it rarely attains intermediate values. This is caused by the Bernoulli density that is used in the Hamilton filter (Hamilton, 1989).

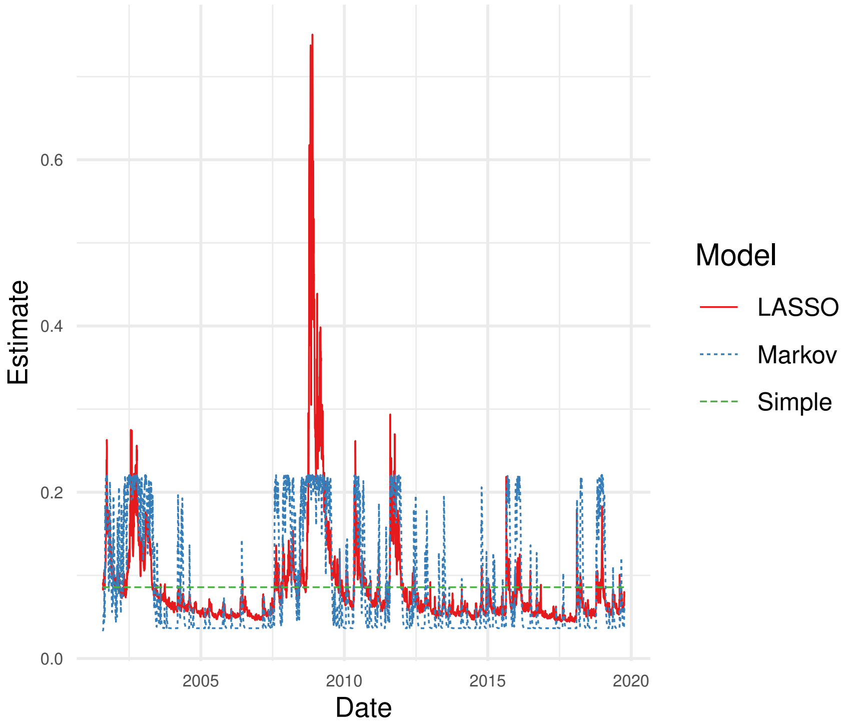


Figure 5: A comparison of the different methods for estimating the probability of excess, applied to the training set. Hence, these are in-sample estimates.

Filtering the data for observations that exceed u allows for the estimation of the conditional excess distribution (equation (1) and section 4.2). Trial runs indicate that the estimation procedure of equation (20) is unstable if (a) many variables are included and (b) variables that are highly correlated or constructs of each other are included, introducing a strong collinearity. Therefore, to ensure that such issues are largely avoided, I use a subset of covariates $\boldsymbol{z}^{\xi} = \boldsymbol{z}^{\kappa} \in \boldsymbol{z}$ before estimating equation (20). In terms of the notation of section 4.2, this corresponds to setting multiple coefficients equal to zero a priori. I investigate two sets of covariates, which are listed in Table 2. With selections A and B, I attempt to preserve as much information as possible by including variables of all kinds while also featuring variables that connect to different regions in the world. The difference between selections A and B is that selection B features the VIX. Including the VIX causes convergence issues for penalty terms on the lower end of the spectrum. In addition, the model converges less rapidly when it does converge, signaling that the model struggles to find the optimal parameter set. However, the VIX is expected to be an important factor in the estimation. Therefore, I also estimate a model that includes the VIX.

A	B
-	VIX
TSX	TSX
FTSE	FTSE
HSI	HSI
Oil	Oil
Gold	Gold
USDYEN	USDYEN
CAAA	CAAA
TS4	TS4

Table 2: Selected variables for the GPD estimation, before the estimation of (20) is performed. A constant is included in the models as well. For more information on the listed variables see Table 9 in Appendix A.

Table 3 shows the results of the penalized estimation. Details on the estimation procedure and the grid search for selection B are given in Appendix C. Following the idea in Hambuckers et al. (2018a), the coefficients are rounded such that variables with insignificant coefficients can be excluded. This is necessary because the LASSO is approximated and therefore it cannot set coefficients to exactly zero, which is what the LASSO is intended for. Also, the number of nonzero coefficients influences the degrees of freedom estimate in equation (9).

For selection A, the BIC selects a model with only constants, excluding all variables from the model. This results in the stationary GPD of equation (1). The added value of the covariates for maximizing the log likelihood does not outweigh the increased penalty of including them. The BIC therefore selects the outcomes that corresponds to high values of ν_ξ and ν_κ, which force the coefficients of the covariates to zero. The results for selection B are similar, with the distinction that the VIX and the CAAA rates are selected for the estimation of κ. The model prefers a very low ξ with a volatile κ, and hence a volatile σ. The VIX is the dominant force in the movements of σ, given its relatively high coefficient. In addition, the inclusion of the CAAA rates suggests that an increase in the rates of triple-A corporate bonds is linked to volatile periods with large losses, which is sensible. Figure 6 shows the path of σ for model configuration B, applied to the training sample. The graph is compared to the σ of the stationary GPD. As is also the case in Figure 5, the financial crisis of 2008 is prominently visible in the graph.

	A		B	
	$\boldsymbol{a}_\xi$	$\boldsymbol{a}_\kappa$	$\boldsymbol{a}_\xi$	$\boldsymbol{a}_\kappa$
Intercept	-1.758	-0.051	-30.193	-0.171
VIX	-	-	0.000	**0.473**
TSX	0.000	0.000	0.000	0.000
FTSE	0.000	0.000	0.000	0.000
HSI	0.000	0.000	0.000	0.000
Oil	0.000	0.000	0.000	0.000
Gold	0.000	0.000	0.000	0.000
USDYEN	0.000	0.000	0.000	0.000
CAAA	0.000	0.000	0.000	**0.066**
TS4	0.000	0.000	0.000	0.000

Table 3: The outcomes of the estimation of the regularized non stationary GPD for variable sets A and B. The bold faced numbers indicate covariates that are included in the model. The coefficients are rounded to three decimals, since they cannot be set exactly to zero by the LASSO as the penalty term is approximated. Also, the coefficients correspond to standardized variables.

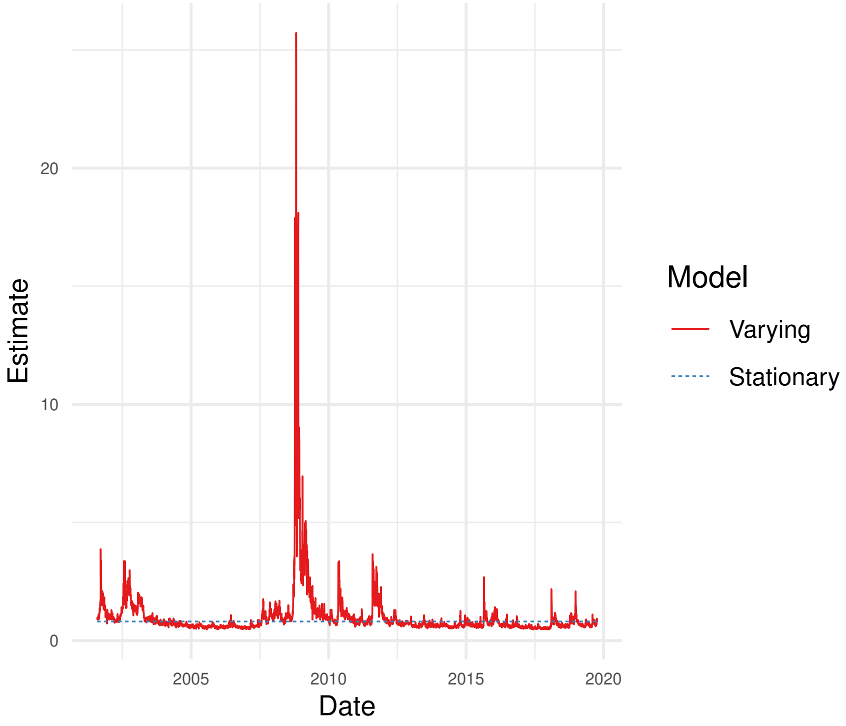


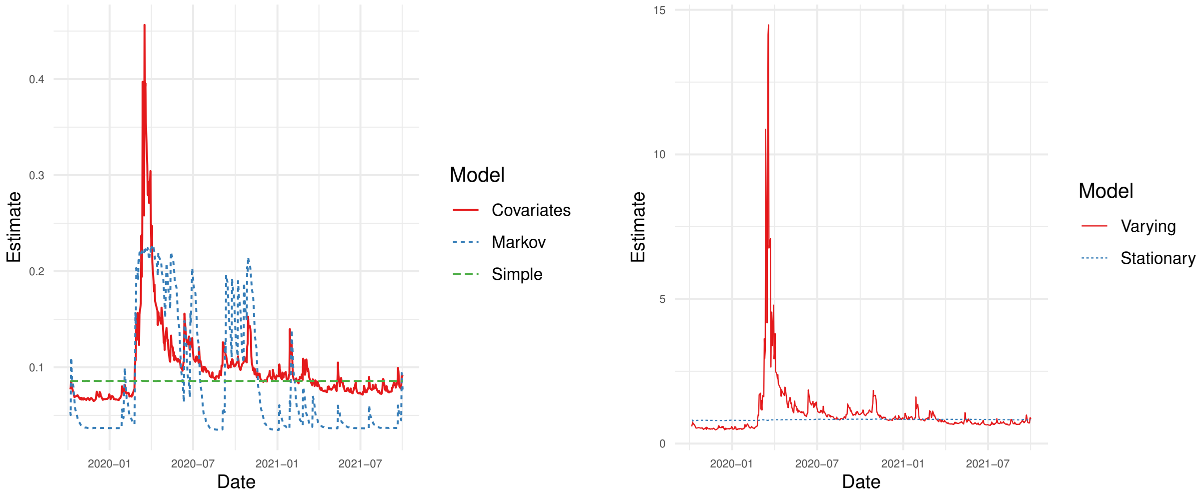
Figure 6: A comparison on the training set of the stationary σ and the σ that varies with covariates.

5.2 Out-of-sample results

Following the procedure as outlined in section 4.4, the models are tested for their predictive performance. The threshold is re-evaluated on each day of the test, using an expanding window. As a result, the estimates of the empirical π_u and the stationary ξ and σ vary slightly as the training set expands. Since it is infeasible to visually determine a suitable threshold via plots for the entire test set, the value of u is determined automatically. One can choose to fix the value of u or the value of the cumulative distribution function at u for the original training set. The difference between the options should be minor. I choose the latter, since the value of the cumulative distribution function has a conceptual meaning in the sense of where the extreme sample starts in general, fixing the proportion of extreme observations. The original u is just an instance of that mechanism. The models for the excess probability are also re-estimated on every day of the test set, as is $\mathcal{M}_0$. The model for the conditional excess distribution is re-estimated three times, once every 125 days.

Figure 7a shows the results of the estimated π_u for the test set. The graphs are similar to Figure 5, with agreeing movements of the covariates and the Markov model, a higher reach of the covariates model, and a Markov model that is less refined than the covariates model. As is the case for the training set, the LASSO selects only the VIX as a covariate of importance for the entire test set.

The re-estimated GPDs in the test set select only the VIX, thus dropping the CAAA variable. Figure 7b shows the results of the estimated σ for the stationary and the non stationary model. Interestingly, the graph of the non stationary σ shows very similar movements to Figure 7a, with the main difference being the very high value of σ at the emergence of the corona crisis. The similarities are not surprising however, since the VIX is the (dominating) selected variable in both models.

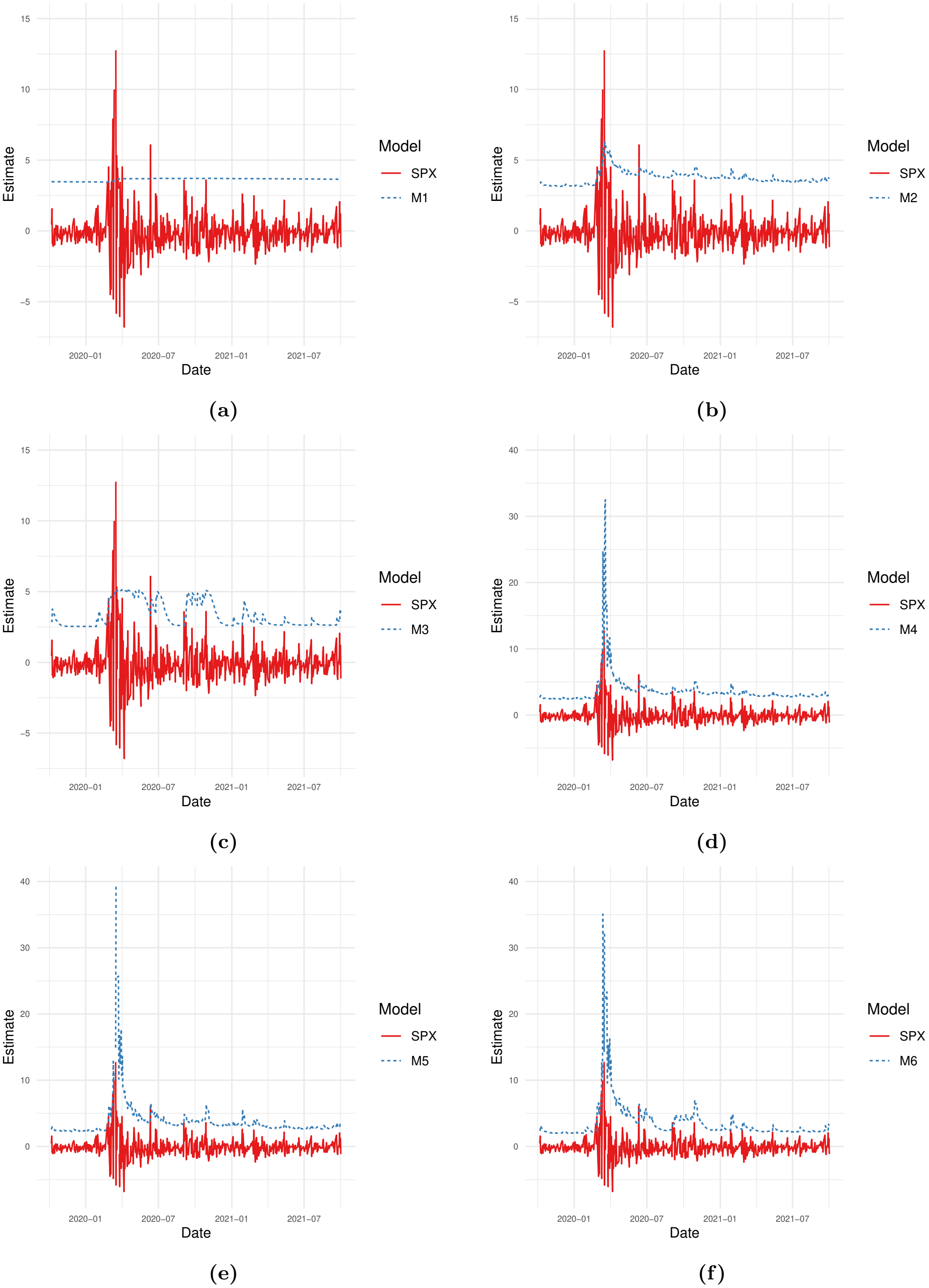


(a)

(b)

Figure 7: Figure 7a shows a comparison of the different methods for estimating the probability of excess, applied to the test set. Figure 7b shows the estimated σ for the stationary and the non stationary model, also applied to the test set.

The estimates of the $\text{VaR}_{0.99}$ by $\mathcal{M}_i$, $i = 1, ..., 6$, are given in Figure 8. Unsurprisingly, Figures 8a-8c show similar dynamics as Figure 7a, since π_u is the only parameter that varies over time for those models in the calculation of the VaR (equation (3)). Figures 8d-8f are clearly also influenced by the dynamics of σ (Figure 7b). As a result, the estimates of $\mathcal{M}_4$, $\mathcal{M}_5$, and $\mathcal{M}_6$ highly exceed the losses around March 2020. In that aspect, the graph of $\mathcal{M}_0$ in Figure 9 follows the path of the SPX more closely. The GARCH structure of the model allows it to be highly flexible, both in attaining high and low estimates. Also, $\mathcal{M}_0$ often agrees with the non stationary EVT models, which is expected since all models measure the level of uncertainty in the market and its participants, but in a different manner. The weakness of the GARCH model seems to be the strong influence of a high loss on subsequent estimates over a relatively long period of time. $\mathcal{M}_3$ and $\mathcal{M}_6$ share this feature to some degree, since the estimated state depends on the state of the previous period. The rest of the non stationary EVT models suffer less from the persistence of a high loss in subsequent estimates, since they do not have an explicit autoregressive structure.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 8: A comparison between the VaR estimates of models $\mathcal{M}_i$, $i = 1, ..., 6$ on the test set, combined with the corresponding graph of the SPX. Figures 8a-8c have the same scale, which is different from the scale of Figures 8d-8f.

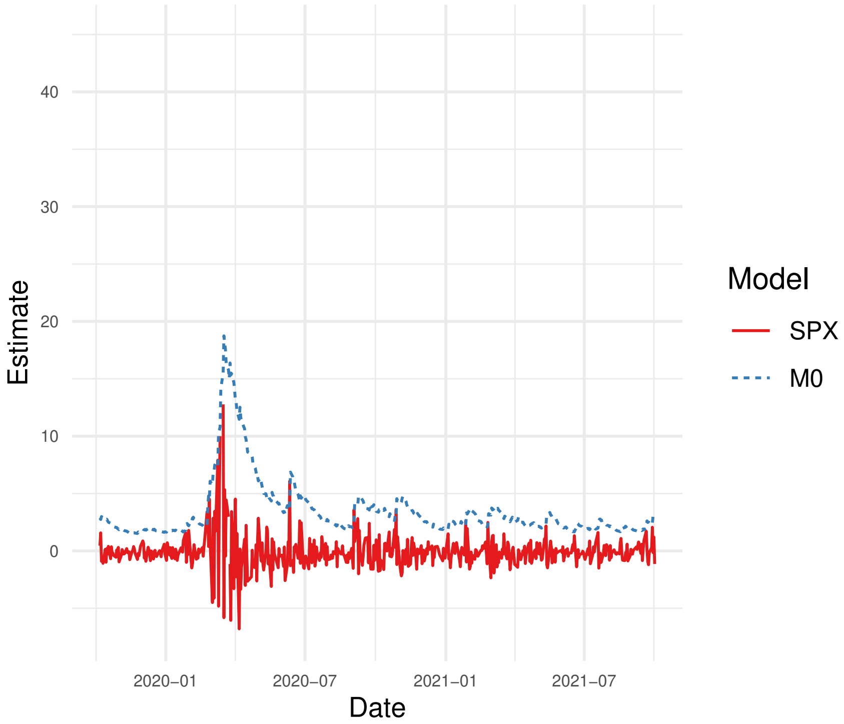


Figure 9: The estimation results of the $\text{VaR}_{0.99}$ for $\mathcal{M}_0$ and the path of the SPX for the test set, given on the scale of Figures 8d-8f.

The results of the traditional backtest are given in Table 4. All models cannot be rejected at a 5% level. However, $\mathcal{M}_1$, the stationary model, and $\mathcal{M}_3$, the model where π_u is modelled by a Markov switching model, can be rejected at a 10% level. This implies that the stationary model might not be adequate. Also, adding dynamics in π_u with a Markov switching model does not improve on $\mathcal{M}_1$ in the context of a traditional backtest. The Markov model may not be flexible enough since it is capped at a certain value, as Figure 8c demonstrates. Furthermore, $\mathcal{M}_4$ and $\mathcal{M}_6$ cannot be rejected at a 10% level, suggesting that modelling the GPD parameters with covariates has merit over the stationary GPD approach. Nonetheless, at a 5% level, all models are considered as adequate risk measurement procedures.

	$\mathcal{M}_0$	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$	$\mathcal{M}_5$	$\mathcal{M}_6$
Violations	7	9	7	9	7	4	6
P-value	0.632	0.094	0.149	0.094	0.632	0.845	0.845

Table 4: The results of the test in Christoffersen (1998), used in this research as a traditional backtest. Since the test set has a length of 500 trading days and $\alpha = 0.99$, the expected number of violations is 5.

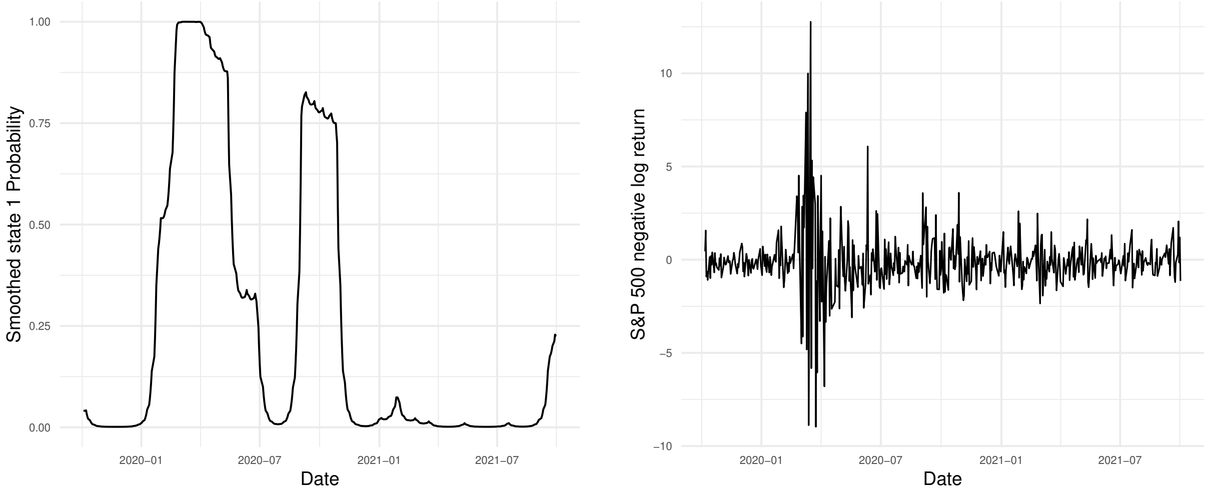
Since all models pass the traditional backtest, the comparative backtest of Nolde and Ziegel (2017) is performed to investigated whether certain models significantly outperform others. Given the loss function in equation (28), well performing models should follow the graph of the SPX closely while limiting the amount of violations. Table 5 show the results of the test, which are inconclusive. Some outcomes of the test statistic Γ approach levels where either H_0^+ and H_0^- can be rejected, but the difference between the models is not significant.

		Internal Model						
		$\mathcal{M}_0$	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$	$\mathcal{M}_5$	$\mathcal{M}_6$
Standard Model	$\mathcal{M}_0$	-	0.860	0.855	0.851	0.801	0.751	0.775
	$\mathcal{M}_1$	0.140	-	0.135	0.130	0.120	0.129	0.134
	$\mathcal{M}_2$	0.145	0.865	-	0.469	0.108	0.126	0.135
	$\mathcal{M}_3$	0.149	0.870	0.531	-	0.118	0.133	0.139
	$\mathcal{M}_4$	0.199	0.880	0.892	0.882	-	0.171	0.193
	$\mathcal{M}_5$	0.249	0.871	0.874	0.867	0.829	-	0.368
	$\mathcal{M}_6$	0.225	0.866	0.865	0.861	0.807	0.632	-

Table 5: The table shows the values of $\Phi(\Gamma)$ (section 4.4.1). The internal model outperforms the standard model if $\Phi(\Gamma) \leq 0.05$, and the internal model is outperformed by the standard model if $\Phi(\Gamma) \geq 0.95$. The results are inconclusive if $0.05 < \Phi(\Gamma) < 0.95$. The results follow from applying the test in Nolde and Ziegel (2017) on the entire test set.

I also investigate two subsets of the test set. The formation of the subsets is based on the state of the market, with intuition that is similar to section 4.1.3. As in section 4.1.3, the possible states are a volatile (crisis) and a calm (prosperous) state. I estimate the state using the smoothed conditional state probabilities in $\boldsymbol{P}^*(t)$, for $t = \tau + 1, ..., T$ (Appendix B). Figure 10a shows the results. Eyeballing Figure 10b, the estimated state probabilities match the SPX behavior well. I label periods as being in the volatile state when the estimated probability of being in the volatile state exceeds 0.10. Figure 10a indicates that this value of the estimated probability marks the start and end of a volatile period. With this selection rule, 181 of the 500 test points are in the volatile state and 319 observations are labeled as being in the calm state.

The mean scores of the models for the subsets and for the entire test set are given in Table 6. $\mathcal{M}_0$ performs best across the board, as it has the (shared) lowest score in each row of the table. $\mathcal{M}_5$ and $\mathcal{M}_6$ have similar test scores, suggesting that the models are comparable to $\mathcal{M}_0$ in performance. $\mathcal{M}_1$ consistently has the highest score, which is a reason to question its performance. Unfortunately, the statistical test in Nolde and Ziegel (2017) cannot be applied to the entire subsets, since the test uses a HAC estimator of the variance, which is conceptually incorrect for disjoint periods. To be able to perform the test on data that is exclusively from either state, I choose the first peak in Figure 10a as test data for the volatile state, yielding 115 data points, and the period after the second peak until the graph increases at the end as test data for the calm state, yielding 213 data points. I label these subperiods as SP_1 and SP_0, respectively.



(a)

(b)

Figure 10: Figure 10a shows the smoothed conditional state probabilities in $\boldsymbol{P}^*(t)$, for $t = \tau + 1, ..., T$ (Appendix B). State 1 is referred to as the volatile state. Figure 10b shows the trajectory of the SPX, which is used for validating the movements of Figure 10a.

The results for SP_1 (the volatile state) are given in Table 7. Again, the results are inconclusive. A possible explanation is that a sample size of 115 observations is too small for the test in Nolde and Ziegel (2017) to yield conclusive results. In contrast, the result for SP_0 (the calm state), which are given in Figure 11 and Table 8, show a significant difference in performance. Models $\mathcal{M}_0$ and $\mathcal{M}_6$ perform best, while $\mathcal{M}_3$ is only outperformed by $\mathcal{M}_6$. Figure 8 and Figure 9 show that the GARCH and the Markov Switching based models can achieve lower VaR estimates than the other models. This explains their strong perform in the calm state, since the score function also punishes the difference between market outcomes and VaR estimates when there is no excess. Low VaR estimates are convenient in a calm state of the market, since less capital is 'wasted' by keeping it in reserve. However, both $\mathcal{M}_0$ and $\mathcal{M}_6$ are highly volatile models. This is unfavorable, since changing capital buffers can be costly and even infeasible if the required changes are significant (Gencay and Selcuk, 2004). Although Figure 8c indicates that $\mathcal{M}_3$ is also volatile, it shows that periods of high and low VaR estimates stretch for a relatively long period of time, due to the persistence of the states (Table 1). This is convenient, since it provide an opportunity for less frequent alterations in capital reserves. Hence, from a practitioner's point of view, $\mathcal{M}_3$ could be the preferred choice.

	$\mathcal{M}_0$	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$	$\mathcal{M}_5$	$\mathcal{M}_6$
Total set	**0.048**	0.093	0.077	0.077	0.060	0.052	0.051
Volatile state	**0.087**	0.193	0.150	0.162	0.111	0.091	0.096
Calm state	**0.026**	0.036	0.036	0.029	0.030	0.030	**0.026**

Table 6: The mean scores for the score function in equation (28). The boldfaced number are the lowest in their row.

		Internal Model						
		$\mathcal{M}_0$	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$	$\mathcal{M}_5$	$\mathcal{M}_6$
Standard Model	$\mathcal{M}_0$	-	0.853	0.822	0.846	0.825	0.798	0.880
	$\mathcal{M}_1$	0.147		0.096	0.138	0.138	0.141	0.151
	$\mathcal{M}_2$	0.178	0.904	-	0.949	0.185	0.175	0.191
	$\mathcal{M}_3$	0.154	0.862	0.051	-	0.143	0.147	0.162
	$\mathcal{M}_4$	0.175	0.862	0.815	0.857	-	0.163	0.204
	$\mathcal{M}_5$	0.203	0.859	0.825	0.853	0.837	-	0.475
	$\mathcal{M}_6$	0.120	0.849	0.809	0.838	0.796	0.525	-

Table 7: The table shows the values of $\Phi(\Gamma)$ (section 4.4.1). The internal model outperforms the standard model if $\Phi(\Gamma) \leq 0.05$, and the internal model is outperformed by the standard model if $\Phi(\Gamma) \geq 0.95$. The results are inconclusive if $0.05 < \Phi(\Gamma) < 0.95$. The results follow from applying the test in Nolde and Ziegel (2017) to SP_1.

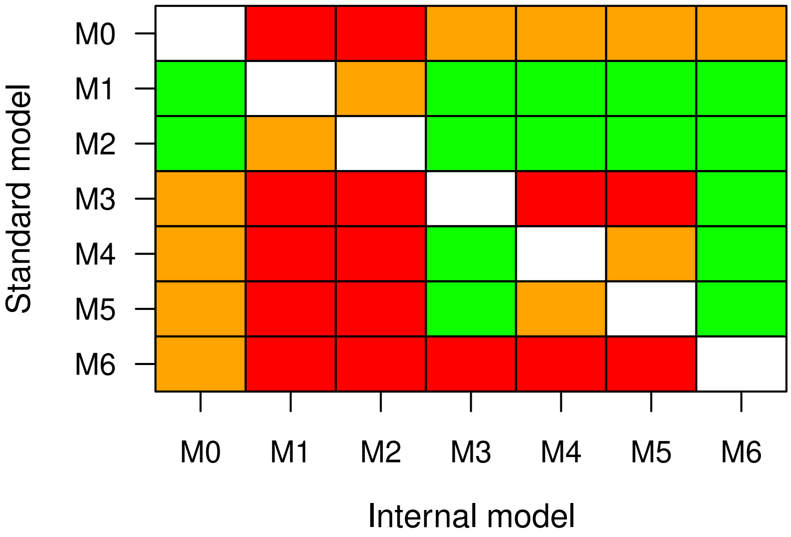


Figure 11: The results of the comparative backtest of Nolde and Ziegel (2017) for models $\mathcal{M}_i$, $i = 0, 1, ..., 6$. The green cells stand for the internal model outperforming the standard model, the red cells for the standard model outperforming the internal model, and the orange cells for inconclusive test results. The results follow from applying the test in Nolde and Ziegel (2017) to SP_0.

		Internal Model						
		$\mathcal{M}_0$	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$	$\mathcal{M}_5$	$\mathcal{M}_6$
Standard Model	$\mathcal{M}_0$	-	0.999	0.999	0.656	0.903	0.915	0.239
	$\mathcal{M}_1$	0.001	-	0.196	0.000	0.000	0.000	0.000
	$\mathcal{M}_2$	0.001	0.804	-	0.000	0.000	0.000	0.000
	$\mathcal{M}_3$	0.344	1.000	1.000	-	1.000	0.999	0.000
	$\mathcal{M}_4$	0.097	1.000	1.000	0.000	-	0.403	0.000
	$\mathcal{M}_5$	0.085	1.000	1.000	0.001	0.597	-	0.000
	$\mathcal{M}_6$	0.761	1.000	1.000	1.000	1.000	1.000	-

Table 8: The table shows the values of $\Phi(\Gamma)$ (section 4.4.1). The internal model outperforms the standard model if $\Phi(\Gamma) \leq 0.05$, and the internal model is outperformed by the standard model if $\Phi(\Gamma) \geq 0.95$. The results are inconclusive if $0.05 < \Phi(\Gamma) < 0.95$. The results follow from applying the test in Nolde and Ziegel (2017) to SP_0.

6 Conclusion

In this research, I investigate a non stationary EVT approach, applied to market risk. The excess probability and the conditional excess distribution are modelled using a broad selection of covariates. Both models are regularized to avoid overfitting and to gain insight in which covariates have a significant impact on performance. In addition, the excess probability is modelled by a Markov switching model, which is motivated by the clusters in which large losses occur. I test the different model setups with a traditional backtest and a comparative backtest. The traditional backtest examines if models are adequate methods for modelling the VaR, while the comparative backtest compares models in their performance.

The estimation of the excess probability with covariates yields that the VIX is the sole significant explanatory variable. Furthermore, the covariates model often shows consensus with the Markov switching model. This implies that the VIX in the covariates model and the good and bad states in the Markov model roughly capture the same feature of uncertainty in the market. However, since the structure of both models differs, their behavior does to. The covariates model is more flexible and more refined than the Markov model while the Markov model is more persistent. The VIX also features in the non stationary GPD model for the scale parameter, as well as the triple-A corporate interest rates. The model for the shape parameter only includes a constant. A configuration of the model without the VIX results in the BIC selecting the stationary model. Hence, the VIX plays an important role in the estimation of the non stationary GPD. Considering the importance of the VIX in both extensions that non stationary EVT provides, risk managers could make use of the VIX in their models.

The one day ahead forecasts imply that all models considered in this research are adequate risk measurement procedures. The stationary model and the model with a Markov model for the excess probability and a stationary GPD are rejected at a 10% level, but they pass the test at a 5% level. The comparative backtests on the entire test set are inconclusive, meaning that no model significantly outperforms any other model. When applying the comparative backtests to subperiods of the test set that correspond to either a volatile or a calm state of the market, the results give additional insights. The results are again inconclusive in the volatile state, which could be caused by a small test set. In the calm state, the tests indicate a significant difference in performance, with the GARCH-based model and two non stationary EVT models performing best. Both non stationary EVT models use a Markov model for the excess probability, while one uses a stationary

GPD and the other uses a non stationary GPD. This superior performance with respect to other models is due to the ability to achieve lower VaR estimates when there are no signs of additional risk. This shows potential to limit the amount of capital reserves when the market is in a calm state. Since all three models also pass the traditional backtest and therefore are valid risk measurement procedures, they are interesting options for practitioners. Out of the three best performing models in the calm state, the GARCH model and the model with the non stationary GPD are the most flexible, and they both outperform the model with a stationary GPD. However, the disadvantage of this flexibility is that the required capital changes frequently and significantly over time, which is a combination that could be a burden for risk managers. Changing capital allocations is costly and has its constraints. In contrast, the model with a Markov model for the excess probability and a stationary GPD is much more persistent in its estimates, which reduces the need for capital reserve alterations while still having great performance. Therefore, this model could be suitable option for a practitioner.

Several extensions can be made to further investigate the potential of non stationary EVT approaches. First, a (simulation) study on the selection of the threshold might help to improve performance. The threshold influences the estimation of the excess probability and the conditional excess distribution and therefore has a significant impact on the VaR estimates. In this research, I choose a threshold that follows from inspecting mean-excess and Hill plots. However, it could be the case that accepting a higher bias enhances the estimation of both parts of the non stationary EVT approach to such an extend that the forecasts are improved. Second, using different types of penalties could result in better performance. An example is the elastic net, although it has the disadvantage that it is computationally costly (Hambuckers et al., 2018a). Third, applying the tests on different and perhaps longer periods of data can yield additional insights. The test results rely heavily on the test data that is fed to the models. Hence, using different test sets could show more differences between the models. Finally, a combination of the covariates model and the Markov model for estimating the excess probability might provide better forecasts. The idea behind this approach is that the probability of excess is potentially explained by different covariates in different states of the world. It could be the case that certain covariates have a strong explanatory power in the prosperous state while the crisis state is better determined by other covariates (Hambuckers et al., 2018b). This would, however, require new methodology on state-wise regularization in a Markov switching setting (for example a LASSO-type penalty for each state), which is unexplored to the best of my knowledge.

References

Andrews, D. W. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 59(3):817–858.

Balkema, A. A. and De Haan, L. (1974). Residual life time at great age. *Annals of Probability*, 2(5):792–804.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327.

Cameron, A. C. and Trivedi, P. K. (2005). *Microeconometrics: methods and applications*. Cambridge university press.

Chavez-Demoulin, V., Embrechts, P., and Hofert, M. (2016). An extreme value approach for modeling operational risk losses depending on covariates. *Journal of Risk and Insurance*, 83(3):735–776.

Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review*, 39(4):841–862.

Gencay, R. and Selcuk, F. (2004). Extreme value theory and Value-at-Risk: Relative performance in emerging markets. *International Journal of Forecasting*, 20(2):287–303.

Giacomini, R. and White, H. (2006). Tests of conditional predictive ability. *Econometrica*, 74(6):1545–1578.

Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une serie aleatoire. *Annals of Mathematics*, 44(3):423–453.

Groll, A., Hambuckers, J., Kneib, T., and Umlauf, N. (2019). LASSO-type penalization in the framework of generalized additive models for location, scale and shape. *Computational Statistics & Data Analysis*, 140:59–73.

Hambuckers, J., Groll, A., and Kneib, T. (2018a). Understanding the economic determinants of the severity of operational losses: A regularized generalized Pareto regression approach. *Journal of Applied Econometrics*, 33(6):898–935.

Hambuckers, J., Kneib, T., Langrock, R., and Silbersdorff, A. (2018b). A Markov-switching generalized additive model for compound Poisson processes, with applications to operational loss models. *Quantitative Finance*, 18(10):1679–1698.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2):357–384.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *Annals of Statistics*, 3(5):1163–1174.

Kim, C.-J. (1994). Dynamic linear models with Markov-switching. *Journal of Econometrics*, 60(1-2):1–22.

McNeil, A. J. and Frey, R. (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. *Journal of Empirical Finance*, 7(3-4):271–300.

Newey, W. K. and West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708.

Nolde, N. and Ziegel, J. F. (2017). Elicitability and backtesting: Perspectives for banking regulation. *Annals of Applied Statistics*, 11(4):1833–1874.

Oelker, M.-R. and Tutz, G. (2017). A uniform framework for the combination of penalties in generalized structured models. *Advances in Data Analysis and Classification*, 11(1):97–120.

Pickands III, J. (1975). Statistical inference using extreme order statistics. *Annals of Statistics*, 3(1):119–131.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):507–554.

Singh, A. K., Allen, D. E., and Robert, P. J. (2013). Extreme market risk and extreme value theory. *Mathematics and Computers in Simulation*, 94:310–328.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Zhong, X. and Enke, D. (2019). Predicting the daily return direction of the stock market using hybrid machine learning algorithms. *Financial Innovation*, 5(1):1–20.

Zou, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society: Series B (statistical methodology)*, 67(2):301–320.

Zou, H., Hastie, T., and Tibshirani, R. (2007). On the “degrees of freedom” of the LASSO. *Annals of Statistics*, 35(5):2173–2192.

A Data details

The selection of covariates that I use is expected to be cohesive and is very similar to the one Zhong and Enke (2019). The main difference between both sets of variables is the inclusion of the VIX. The reason I include the VIX is that it is known as the fear thermometer of Wall Street, or differently put, a measure of uncertainty. Therefore, the VIX is expected to play a significant role in predicting large losses, which is the focus of this research.

There are also several smaller differences between the variables sets. First, in contrast to Zhong and Enke (2019), I include simple moving averages (SMAs) in the models instead of exponential moving averages. The SMA places an equal weight on each return and therefore has a longer memory than the exponential version, which gives a higher weight to more recent returns. I prefer SMAs since short term lags of the SPX are also included in the models. Therefore, SMAs are expected to add more information to the models. Furthermore, I included the LIBOR rates in the model while Zhong and Enke (2019) does not. This is motivated by the fact that London plays a central role in the financial world. Hence, the rates there could influence the markets in the U.S. as well, and probably in a different way than a T-Bill rate might do. Also, there are marginal differences in all the other types of variables. This is more of a personal choice, but I do not expect these differences to make a real impact on the results since they are very subtle. Finally, some variables in Zhong and Enke (2019) are left out, since I do not expect them to add value to the model, also keeping in mind that limiting the number of variables can be beneficial in the estimation procedure. The list of variables used in this research is given in Table 9.

Type	Name	Description
Technical variables		
	SPX	The log returns of the S&P 500 index. Note that this is not the SPY, which is an actual tradable fund. They are (almost) one-to-one.
	SPX_1	The log returns of the S&P 500 index, lagged once.
	SPX_2	The log returns of the S&P 500 index, lagged twice.
	SPX_3	The log returns of the S&P 500 index, lagged three times.

Type	Name	Description
	SMA10	The simple moving average of the SPX log returns of the previous 10 days.
	SMA20	The simple moving average of the SPX log returns of the previous 20 days.
	SMA50	The simple moving average of the SPX log returns of the previous 50 days.
	SMA200	The simple moving average of the SPX log returns of the previous 200 days.
Financial variables		
	VIX	The Chicago Board Options Exchange (CBOE) volatility index.
	SPXV	The log differences of the S&P 500 24-hour volume.
	DJIV	The log differences of the DJI 24-hour volume.
	IXICV	The log differences of the IXIC 24-hour volume.
Major world indices		
	DJI	The log returns of the Dow Jones Industry Average index.
	IXIC	The log returns of the NASDAQ Composite.
	TSX	The log returns of the S&P/TSX Composite.
	FTSE	The log returns of the FTSE 100 index.
	DAX	The log returns of the DAX.
	CAC	The log returns of the CAC 40 index.
	AEX	The log returns of the AEX.
	HSI	The log returns of the HSI.

Type	Name	Description
	NIKKEI	The log returns of the NIKKEI 225 index.
Natural resources		
	NG	The log differences of the natural gas future price.
	Oil	The log differences of the crude oil future prices.
	Gold	The log differences of the gold future price.
Exchange rates		
	USDGPB	The log differences of exchange rate between the dollar and the pound.
	USDEUR	The log differences of exchange rate between the dollar and the euro.
	USDCND	The log differences of exchange rate between the dollar and the Canadian dollar.
	USDYEN	The log differences of exchange rate between the dollar and the yen.
	USDCNY	The log differences of exchange rate between the dollar and the Chinese yuan.
Interest rates		
	CAAA	Change in Moody's seasoned AAA corporate bond yields.
	CBAA	Change in Moody's seasoned BAA corporate bond yields.
	LIBOR1M	Change in the LIBOR for a 1-month maturity.
	LIBOR3M	Change in the LIBOR for a 3-month maturity.
	LIBOR6M	Change in the LIBOR for a 6-month maturity.

Type	Name	Description
	LIBOR12M	Change in the LIBOR for a 12-month maturity.
	TB1M	Change in the rate on T-bills with a 1-month maturity.
	TB3M	Change in the rate on T-bills with a 3-month maturity.
	TB6M	Change in the rate on T-bills with a 6-month maturity.
	TB12M	Change in the rate on T-bills with a 12-month maturity.
	TB60M	Change in the rate on T-bills with a 60-month maturity.
	TB120M	Change in the rate on T-bills with a 120-month maturity.
	TB240M	Change in the rate on T-bills with a 240-month maturity.
Spreads between interest rates		
	TS1	Change in the spread between the rates on 3-month and 1-month T-bills.
	TS2	Change in the spread between the rates on 6-month and 1-month T-bills.
	TS3	Change in the spread between the rates on 120-month and 1-month T-bills.
	TS4	Change in the spread between the rates on 120-month and 3-month T-bills.
	TS5	Change in the spread between the rates on 120-month and 6-month T-bills.
	DS1	Change in the spread between Moody's rates for BAA and AAA corporate bonds.

Type	Name	Description
	DS2	Change in the spread between Moody's BAA rate and the 1-month T-bill rate.
	DS3	Change in the spread between Moody's BAA rate and the 3-month T-bill rate.
	DS4	Change in the spread between Moody's BAA rate and the 6-month T-bill rate.
	DS5	Change in the spread between Moody's BAA rate and the 60-month T-bill rate.
	DS6	Change in the spread between Moody's BAA rate and the 240-month T-bill rate.

Table 9: The list of all variables included in the analysis. In the columns from left to right, the type of the variables, their name in this study and a description is given. The source of all the data is Bloomberg.

B EM Algorithm

The EM algorithm in this research consists of a prediction, updating en smoothing step (Hamilton (1989); Kim (1994)). The Hamilton prediction step is

$$\hat{\boldsymbol{\zeta}}_{t+1|t} = \boldsymbol{P}\hat{\boldsymbol{\zeta}}_{t|t},$$

where $\hat{\boldsymbol{\zeta}}_{i|j}$ is the predicted state vector (with two entries) of time i, estimated at time j. The Hamilton updating step is

$$\hat{\boldsymbol{\zeta}}_{t|t} = \frac{\begin{bmatrix} f(y_t \mid S_t = 1) \\ f(y_t \mid S_t = 2) \end{bmatrix} \bullet \hat{\boldsymbol{\zeta}}_{t|t-1}}{\begin{bmatrix} f(y_t \mid S_t = 1) & f(y_t \mid S_t = 2) \end{bmatrix} \hat{\boldsymbol{\zeta}}_{t|t-1}},$$

where the sign $\bullet$ is the element wise multiplication of the vectors. The Kim smoother is given by

$$\begin{aligned}\hat{\boldsymbol{\zeta}}_{t|T} &= \hat{\boldsymbol{\zeta}}_{t|t} \bullet \boldsymbol{P}^{'}(\hat{\boldsymbol{\zeta}}_{t+1|T}/\hat{\boldsymbol{\zeta}}_{t+1|t}) \\ \boldsymbol{P}^{*}(t) &= \boldsymbol{P} \bullet (\hat{\boldsymbol{\zeta}}_{t|T}\hat{\boldsymbol{\zeta}}_{t-1|t-1})^{'}/(\hat{\boldsymbol{\zeta}}_{t|t-1}\begin{bmatrix}1 & 1\end{bmatrix}),\end{aligned}$$

where the / sign is a element wise division of the matrices, and $\boldsymbol{P}^{*}(t)$ is the matrix containing the probabilities $p_{ij}^{*}(t)$, $i, j = 0, 1$.

The algorithm is initially initialized in state 0, based the on the graph of the SPX and trial runs, with reasonable parameters for the state probabilities $\hat{\pi}^{(0)}$ and $\hat{\pi}^{(1)}$. In the test procedure, the algorithm is initialized at the results on the previous day, including the smoothed estimate of the initialization. Doing so allows for using less iterations. I use 6 iterations, based on trials in which the precision did not increase significantly since only one data point is added at a time.

C GPD estimation algorithm

I use a combination of the estimation procedure as in Hambuckers et al. (2018a) and Chavez-Demoulin et al. (2016). The optimal coefficients are found by using a penalized iterative reweighed least squares (PIRLS) algorithm, with a LASSO penalty. The penalty is approximated as in Oelker and Tutz (2017):

$$|\theta_j| \approx \sqrt{\theta_j^2 + c},$$

with $\theta_j \in \boldsymbol{\theta}$. Using this approximation allows for differentiating the penalty. Following Hambuckers et al. (2018a), I use $c = 10^{-7}$ and round the coefficients. This is necessary since the LASSO penalty is approximated. As in Hambuckers et al. (2018a), I round to three decimals since trials show that the third decimal can have a slight impact on the estimated distribution parameters, while the impact of the fourth decimal is negligible. Since I use reparameterized parameters, the estimation can be split for ξ and κ as in Chavez-Demoulin et al. (2016). One can find the coefficients in the equation for ξ via the recursive algorithm

$$\hat{\boldsymbol{a}}_{\xi}^{(k)} = \hat{\boldsymbol{a}}_{\xi}^{(k-1)} - v\boldsymbol{H}_{\text{pen}}^{-1}(\hat{\boldsymbol{a}}_{\xi}^{(k-1)})\boldsymbol{s}_{\text{pen}}(\hat{\boldsymbol{a}}_{\xi}^{(k-1)}).$$

In the algorithm, the penalized Hessian $\boldsymbol{H}_{\text{pen}}^{-1}$ and the penalized score function $\boldsymbol{s}_{\text{pen}}$ (both for the coefficients that feature in the model for ξ) are approximated as in Oelker and Tutz (2017), k

denotes the next iteration and v is the step length. I use $v = 0.5$, with the purpose of stabilizing the algorithm. The algorithm is identical for estimating the parameters in the model for κ. The algorithm is stopped when the relative change $||\hat{\boldsymbol{\theta}}_{(k)} - \hat{\boldsymbol{\theta}}_{(k-1)}||/||\hat{\boldsymbol{\theta}}_{(k-1)}||$ is smaller than 10^{-6} or after 200 iterations.

Instead of using the build in functions for the Hessian and gradient in R, I programmed them by hand. The derivation of the Hessian and the gradient follow from the chain rule. For the gradient of the (log) density f of observation n, this is

$$\frac{\partial f_n(\theta_j)}{\partial \theta_j} = \frac{\partial f_n(\theta_j)}{\partial \xi_n}\frac{\partial \xi_n}{\partial \theta_j} + \frac{\partial f_n(\theta_j)}{\partial \kappa_n}\frac{\partial \kappa_n}{\partial \theta_j},$$

of which either the left of the right term equals zero. For a θ_j that is part of the model for ξ_n (the terms for the case of κ_n are derived similarly), the Hessian follows from

$$\frac{\partial}{\partial \theta_l}\frac{\partial f_n(\theta_j)}{\partial \theta_j} = \frac{\partial^2 f_n(\theta_j)}{\partial \xi_n \partial \xi_n}\frac{\partial \xi_n}{\partial \theta_l}\frac{\partial \xi_n}{\partial \theta_j} + \frac{\partial^2 f_n(\theta_j)}{\partial \kappa_n \partial \xi_n}\frac{\partial \kappa_n}{\partial \theta_l}\frac{\partial \xi_n}{\partial \theta_j} + \frac{\partial f_n(\theta_j)}{\partial \xi_n}\frac{\partial^2 \xi_n}{\partial \theta_l \partial \theta_j},$$

where one or multiple terms are zero, with $\theta_l \in \boldsymbol{\theta}$. Using the exact gradient and Hessian increases the speed of the algorithm by a least 100 times. Also, the estimates are more precise since no approximations are necessary.

I choose starting values in the neighborhood of the stationary GPD, hence with the coefficients of the covariates being close to zero. A two-dimensional grid search is executed for ν_ξ and ν_κ to determine the optimal configuration of the model. However, in the case of selection B, the algorithm behave erratically. Due to convergence issues and singular Hessians, a proper grid search is not possible for low values of the penalty terms. This is also hinted at in Hambuckers et al. (2018a). Hence, I performed the grid search by hand. For configurations that might yield a low BIC, I approach the point that an extra variable is added to the model to achieve a low BIC for that model configuration. This is not ideal since it requires considerable manual labor, but it yields better solutions than the option of keeping penalty terms high to achieve easy convergence. Then, I compare the BIC values for the different sets of included variables. During this grid search, the value of ν_ξ is found to be irrelevant as long as it sets the coefficients in the equation of ξ equal to zero. Hence, only the grid for ν_κ is investigated extensively for selection B.