ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Master Thesis - MSc Quantitative Finance

Variable Selection in Tail Risk Modeling of Equity Returns

Author:

Hoxha, Xhulia

422839

Supervisor:

Zhou, C.

Co-reader:

Naghi, A. A.

Abstract

This paper incorporates machine learning techniques to study the behavior of the distribution of extreme values. Variable selection and model estimation is performed using gradient boosting methods together with the flexible additive GAMLSS models which incorporate covariates into the parameters of the extreme distribution. A simulation study is performed to test the performance of the proposed method under high-dimensional and high-correlation scenarios. We find that the model performance is quite satisfactory in terms of both variable selection and estimation of covariate effects. For a real-data application, we explain the behaviour of the S&P 500 market index during extreme events by incorporating the effect of 37 potential covariates. To evaluate the adequacy of our model, we compute 95% confidence intervals using the quantiles of the fitted conditional distribution, predict the Value-at-Risk conditional on covariates as well as compare the performance with other benchmark models.

April 30, 2021

Contents

1 Introduction

Extreme value theory (EVT) studies the impact of extremal events and stress scenarios by modelling the behavior of the tail distribution of a particular variable of interest. This theory has been widely applied in various fields such as finance (to analyze unusually large insurance losses), environmental science (to study extreme flood events, temperature levels or earthquakes) as well as in healthcare, engineering and even astronomy. Classical EVT faces multiple challenges with extremal data often being scarce and heavy-tailed, consequently affecting the implementation of an appropriate model for parameter estimation. Financial data, for example, typically show serial dependence, trends or clustering over time. In attempt to explain some of this non-stationary behavior, recent studies propose (semi-)parametric and non-parametric models that incorporate covariates or additional variables to the parameters describing the tail of a distribution. One major difficulty for incorporating covariates in extremes is the large number of potential covariates. This could be considered either as a variable selection problem, where we select only a subset of the most explanatory covariates, or as a dimension reduction problem, where we transform the data into a lower-dimension setting of constructed factors which help explain the tail risk more efficiently. In this paper, we restrict our focus on the former by incorporating machine learning boosting techniques for variable selection in tail risk modeling.

Introduced by Rigby and Stasinopoulos (2005), generalized additive models for location, scale and shape (GAMLSS) are regression and smoothing models where the parameters of the underlying distribution of the response variable are modelled as additive functions of a set of explanatory variables. One important feature of GAMLSS is that the relationship with the covariates can be modelled as parametric, semi-parametric or non-parametric, giving these models a high level of flexibility. Furthermore, the distribution for the response variable in GAMLSS can be selected from any family of distributions including highly skewed or kurtotic distributions, which may be more appropriate as compared to the restrictive Gaussian or exponential distribution. GAMLSS use a backfitting Gauss-Newton algorithm to iteratively fit the additive predictors. This allows the simultaneous fitting of all distribution parameters given a fixed set of covariates, using relatively small computational requirements. However, GAMLSS alone do not address the major issue of variable selection. In high-dimensional data settings, with a large number of covariates and/or distribution parameters, the estimation becomes cumbersome or, most of the time, infeasible and a selection of the most informative covariates is necessary and often required.

Mayr et al. (2012) develop an alternative approach to the classical estimation of GAMLSS which applies machine learning boosting techniques for model fitting and variable selection in the GAMLSS framework. Their *gamboostLSS* algorithm adapts component-wise gradient boosting as a new fitting method to estimate and select predictors iteratively and simultaneously across all distribution parameters. Their algorithm is an efficient data-driven variable selection technique that overcomes the limitations of the GAMLSS fitting procedure of Rigby and Stasinopoulos (2005) under high-dimensional data scenarios where the number of covariates is larger than the number of observations. Moreover, it avoids the potential shortcomings of using information criteria, such as the generalized Akaike information criterion (GAIC), for model selection.

Although GAMLSS can be applied to an abundance of distributions, from discrete to con-

tinuous, with up to four (or more) parameters, research on the performance of GAMLSS on extreme value distributions has been limited. Chavez-Demoulin and Davison (2005) proposed a first step to adapting the GAM framework to an EVT setting by taking into account the effect of covariates on the tail behaviour. Chavez-Demoulin et al. (2016) also present a general methodology for modeling loss data depending on covariates, such as business line and time. In both papers, inference and model fitting is done by penalized maximum likelihood estimation. However, to the best of our knowledge, the variable selection problem remains untackled in an EVT setting. This paper contributes to the current literature by incorporating machine learning techniques to study the behavior of the distribution of extreme values. Our methodology ties together extreme value analysis for the fundamentals of the tail distribution, with the flexible GAMLSS to incorporate covariates into the parameters of the extreme distribution, and gradient boosting techniques for an efficient and careful selection of the most prominent covariates that help explain this tail behaviour.

We perform a simulation study to evaluate the performance of the proposed method. The results show that the model is able to accurately capture the effect of the most prominent covariates on the scale parameter of the extreme value distribution. As widely recognized in the current literature, the shape parameter is very difficult to estimate (Coles, 2001; Park and Kim, 2016). This is also evidenced in our study. For this reason, we suggest to incorporate a GAMLSS covariate model to the scale parameter alone. A sensitivity analysis is conducted to test the robustness of the simulation results when certain assumptions of the data-generation process are changed. We consider a low-dimensional and a high-dimensional setting for both the number of observations as well as the number of covariates introduced to the model. Furthermore, we investigate whether high correlation among covariates has an impact on the results.

To test our methodology on a real-data application, an empirical analysis is conducted stock returns. Financial return time series exhibit non-stationarity, trends and volatility clustering and often they are subject to dependencies with other financial series such as stock indices or macroeconomic variables. We look into extreme events in the daily S&P 500 market index returns during the last 20 years. Then, we apply the *gamboostLSS* algorithm to select among 37 potential covariates, including important market indices (such as NASDAQ, Dow Jones), volatility indices (VIX), bond yields (treasury bills with maturities varying from 3 months to 20 years), commodities (crude oil, gold, silver), foreign exchange (FX) rates, Fama-French factors and other variables that might explain the tail behavior of stock losses (negative returns). We estimate their predictor effect. Finally, to evaluate the performance of our methodology, we compare how the estimation of risk measures such as Value-at-Risk (VaR) can be adequately captured by the model and how much it can be improved as compared to classical EVT models which does not include covariates to model the tail distribution parameters, but rather assumes they are time-constant.

This paper is structured as follows. A review of the relevant theoretical framework can be found in Section 2. The Methodology section begins with a detailed overview of the classical extreme value theory and the generalized additive models of location, scale and shape. We tie together these two fields, in order to extend the available literature and give a solution to the variable selection problem using the *gamboostLSS* algorithm. The gradient boosting technique,

the inherence and parameter estimation of the models used are explained in detail in Section 3. A simulation study is conducted in Section 4 to evaluate the accuracy and the ability of the gradient boosting algorithm to perform data-driven variable selection in an extreme value setting. In Section 5, we conduct an empirical analysis on S&P 500 loss data where we apply our methodology to select among a large set of explanatory variables that explain best extreme loss distribution characteristics. Section 6 briefly summarizes our main results as well as gives some final remarks and recommendations for further research.

2 Literature Review

Extreme value theory (EVT) studies the behavior of the tail distribution of a random variable. There are two main approaches in EVT, namely the block maxima (BM) and the peaks-over-threshold (POT) approach. The BM method divides the data into several equally-sized, non-overlapping blocks and restricts the focus only on the largest observations within each block. These extreme data points, the block maxima, approximately follow a generalized extreme value (GEV) distribution (Gumbel, 1958). With extreme values being scarce, the BM method may be considered as wasteful as it disregards a lot of the information found in other extreme values besides the block maxima. The POT approach, on the other hand, focuses its attention to extreme values above a certain threshold. The corresponding exceedances follow a generalized Pareto (GP) distribution (Pickands, 1975). The choice of the threshold is of crucial importance in this approach. Figure 1 gives a graphical visualisation of both EVT methods. In this paper, we restrict our attention solely on the POT approach, described in detail in Section 3.1.

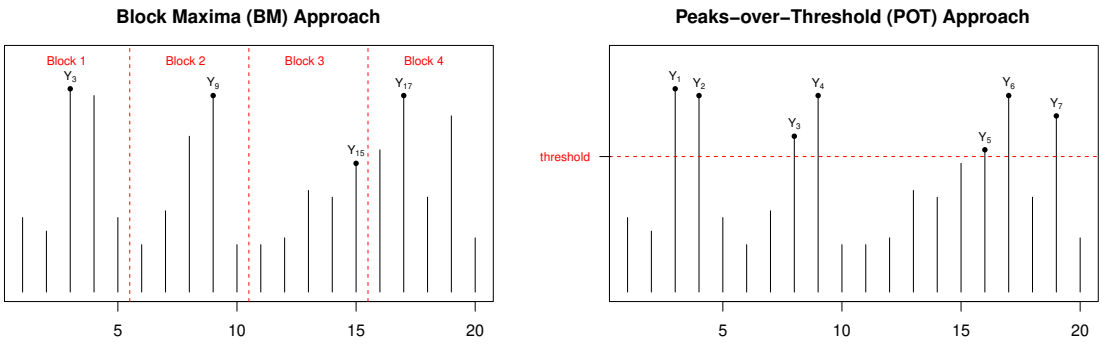

Figure 1: The BM vs. POT Approach

Classical EVT, however, assumes the distribution parameters are constant over time and that the realizations are identically and independently distributed. In practice, this assumption is often violated for extreme values as they usually exhibit temporal dependence and clustering. Dynamic EVT allows the distribution parameters to be varying by modelling their dependence on additional explanatory variables, such as time (Coles, 2001).

Generalized additive models for location, scale and shape (GAMLSS) are flexible regression and smoothing models where the parameters of the underlying distribution of the response variable are modelled as additive functions of a set of predictor variables. Introduced by Rigby and Stasinopoulos (2005), GAMLSS are an extension of the classical generalized linear models (GLM) by Nelder and Wedderburn (1972) and generalized additive models (GAM) by Hastie

and Tibshirani (1990). In comparison, GAMLSS simultaneously model all the parameters of the underlying distribution, not only the conditional mean, that is, not only the location, but also the scale and other shape parameters. Furthermore, the distribution for the response variable in GAMLSS can be selected from any family of distributions including highly skewed or kurtotic distributions, which may be more appropriate for modeling the variable of interest as compared to the restrictive Gaussian or exponential distributions. Similarly to the traditional GAMs, the relationship between the distribution parameters and the covariates is modelled through additive predictor functions which can be parametric, semi-parametric or non-parametric in form. Moreover, the inclusion of random effect terms is also possible. The estimation of the additive predictors in GAMLSS is based on maximum (penalized) likelihood maximization through Newton-Raphson or Fisher-scoring algorithms. In high-dimensional settings, however, as more covariates or distribution parameters are included, these fitting methods have been proved to be highly unstable and a selection of the most informative covariates is necessary.

Variable selection when it comes to GAMLSS is problematic for several reasons. First, as all distribution parameters are fitted (not only the location paramater as in the traditional GAM setting, but also the scale and shape parameters), the complexity and computational requirements increase, especially when using large datasets. Secondly, the high level of flexibility that GAMLSS offer requires efficient methods for variable selection that avoid overfitting. To select the most informative covariates and avoid overfitting, Rigby and Stasinopoulos (2005) compare between GAMLSS models using the generalized Akaike information criterion (GAIC). As the GAIC is a generalization of the AIC and BIC, it inherits the same properties and, hence, shortcomings. Variable selection using these information criteria has often been criticized as being unstable and to include a large number of uninformative covariates (Ripley, 2004). Moreover, they have shown substantial bias when comparing between linear and non-linear models (Greven and Kneib, 2010).

Mayr et al. (2012) introduce machine learning techniques into the GAMLSS framework to address the issue of variable selection. Their *gamboostLSS* algorithm adapts component-wise gradient boosting as a new fitting method to estimate and select predictor effects iteratively. The main idea of gradient boosting is to minimize an expected loss function along the steepest gradient descent (Friedman, 2001). Boosting generally results in an additive prediction function composed of a set of base-learners. Component-wise gradient boosting fits these base-learners iteratively to the negative gradient vector of the loss criterion and updates only the best fitting base-learner every iteration (see Bühlmann and Hothorn, 2007). Thus, in a high-dimensional data setting with an extensive number of potential covariates, component-wise boosting performs variable selection inherently, as non-informative covariates are never selected and excluded from the model. The choice of base-learners, which carry out the fitting of the gradient vectors using the covariates, is crucial for the application of the *gamboostLSS* algorithm, as they define the type of effect that covariates will have on the predictors of the GAMLSS distribution parameters. Examples of base-learners include simple linear models, penalized regression splines (for non-linear effects) as well as classification or regression trees and neural networks.

An alternative method for variable selection in GAMLSS is introduced more recently by Hambuckers et al. (2018), where LASSO-type penalization is incorporated in the estimation.

This strategy, however, is only entailed for linear covariate effects.

In this paper, we bridge the extensive theory of EVT with machine learning boosting algorithms used in GAMLSS for variable selection. Chavez-Demoulin and Davison (2005) proposed a first step to adapting the GAM framework to an EVT setting by taking into account the effect of covariates on the tail behaviour. Chavez-Demoulin et al. (2016) use additive models for modeling operational loss data depending on covariates such as business line and time. However, research on variable selection strategies to model tail behaviour using covariates has been very scarce.

3 Methodology

In this section, we present the methods and estimation techniques required to perform variable selection in tail risk modeling. Section 3.1 introduces the classical POT approach from EVT, the generalized Pareto distribution and its properties as well as some discussion points which emphasize the necessity of moving from the classical EVT to a dynamic one where the distribution parameters are not constant, but varying. Section 3.2 describes in detail GAMLSS models which assume an additive functional relationship between distribution paramaters and some explanatory variables, and how we adapt this to an EVT context. Finally, Section 3.3 gives a solution to the variable selection problem by introducing gradient boosting techniques to model extreme value distribution parameters while selecting informative covariates that can explain their behaviour.

3.1 The Classical Peaks-Over-Threshold (POT) Approach

Suppose $X_1, \dots, X_T$ are a sequence of identically and independently distributed (i.i.d) observations from an unknown distribution F. For a given finite threshold $u \geq 0$, let $\{X_{t_1}, \dots, X_{t_N}\} \subseteq \{X_1, \dots, X_T\}$ where $\{t_1, \dots, t_N\} = \{i : X_i > u\}$ and $Y_i = X_{t_i} - u$, $i \in \{1, \dots, N\}$, the corresponding excesses. We follow the notation in Chapter 3 of Embrechts et al. (1997) to assume that

1. the number of exceedances N over a time interval $[0, T]$ approximately follows a Poisson distribution with rate λ, hence $N \sim \text{Poi}(\lambda T)$;

2. the excesses $Y_1, \dots, Y_N$ over the threshold u approximately follow (independently of N) a generalized Pareto distribution (GPD), denoted by $\text{GPD}(\xi, \beta)$ with shape parameter $\xi \in \mathbb{R}$, scale parameter $\beta > 0$ and distribution function

$$G_{\xi,\beta}(x) = \begin{cases} 1 - (1 + \xi x/\beta)^{(-1/\xi)}, & \text{if } \xi \neq 0, \\ 1 - \exp(-x/\beta), & \text{if } \xi = 0, \end{cases}$$

for $x \in D(\xi, \beta)$ such that

$$D(\xi, \beta) = \begin{cases} [0, \infty) & \text{if } \xi \geq 0, \\ [0, -\beta/\xi] & \text{if } \xi < 0, \end{cases}$$

Depending on the tail characteristics, or equivalently, on the shape parameter ξ, the following distribution classifications can be distinguished. A positive ξ entails a heavy-tailed (or power-tailed) distribution such as Pareto, Student t or Fréchet. The case $\xi = 0$ corresponds to distributions whose tail essentially decays exponentially like the Normal, Gamma or Lognormal distribution, while $\xi < 0$ shapes short-tailed distributions with a finite right endpoint, such as Weibull, Uniform or Beta distributions. These three cases are graphically shown in Figure 2. In financial applications, heavy-tailed distributions typically fit best return series, hence we restrict our focus entirely on the $\xi \geq 0$ scenario.

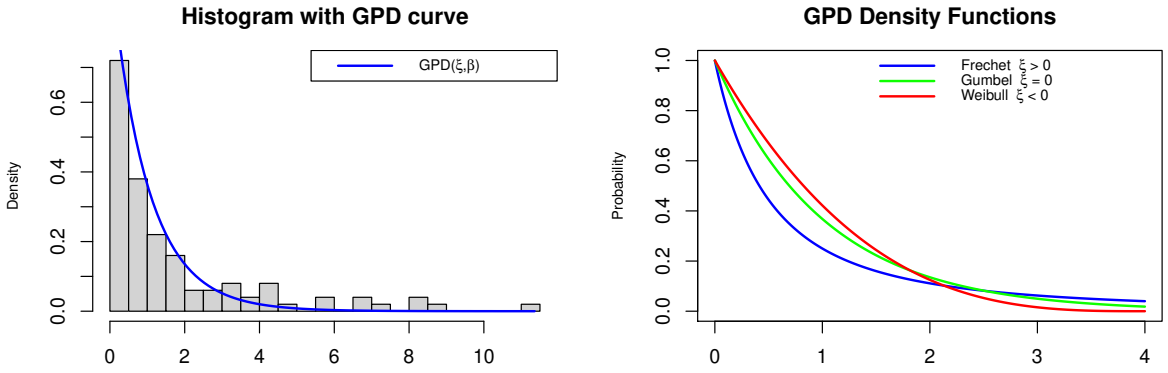

Figure 2: The Generalized Pareto Distribution (GPD)

There are three noteworthy points of discussion here. First, the choice of the threshold is of crucial important in the POT approach because it implies a trade-off between bias and variance. For smaller values of u, more observations are used for inference, hence resulting in a smaller variance. Larger values, on the other hand, result in a smaller bias. In general, choosing a threshold such that around 10% of the data are included as exceedances is recommended as an initial choice. However, a sensitivity analysis across different values of the threshold u is necessary to make a better informed decision regarding this choice.

Secondly, the Poisson process rate λ and the GPD parameters (ξ, β) can be estimated separately and independently of each other. The independence condition between the number of exceedances N and the excesses $Y_1, \ldots, Y_N$ results in the following likelihood function

$$L(\mathrm{Y}; \lambda, \xi, \beta) = \frac{(\lambda T)^N}{N!} \exp(-\lambda T) \prod_{i=1}^{N} g_{\xi,\beta}(Y_i),$$

where $\mathrm{Y} = (Y_1, \ldots, Y_N)'$ and $g_{\xi,\beta}$ is the density of $G_{\xi,\beta}$. The log-likelihood function then splits into two parts

$$l(\mathrm{Y}; \lambda, \xi, \beta) = l(\mathrm{Y}; \lambda) + l(\mathrm{Y}; \xi, \beta),$$

where

$$l(\mathrm{Y}; \lambda) = -\lambda T + N \log(\lambda) + \log\left(\frac{T^N}{N!}\right) \quad \text{and} \quad l(\mathrm{Y}; \xi, \beta) = \sum_{i=1}^{N} l(Y_i; \xi, \beta)$$

with

$$l(Y_i;\xi,\beta) = \log g_{\xi,\beta}(Y_i)$$
$$= \begin{cases} -\log(\beta) - (1+1/\xi)\log(1+\xi Y_i/\beta), & \text{if } \xi > 0, Y_i \geq 0 \text{ or } \xi < 0, Y_i \in [0, -\beta/\xi] \\ -\log(\beta) - Y_i/\beta, & \text{if } \xi = 0 \\ -\infty, & \text{otherwise.} \end{cases}$$

Thus, the maximization for the two estimation problems can be carried out separately and independently of each other. One common risk measure is the Value-at-Risk (VaR), which computes the maximum loss expected to occur at a $100\alpha\%$ probability level. Given estimates of the shape and scale parameters for the GP distribution, this measure is calculated as

$$\widehat{\mathrm{VaR}}_\alpha = u + \frac{\hat{\beta}}{\hat{\xi}}\left[\left(\frac{1-\alpha}{N/T}\right)^{-\hat{\xi}} - 1\right]. \tag{3.1}$$

Finally, the classical POT approach makes the assumption of identically and independently distributed observations. In practice, this is violated with financial data that typically show serial dependence, trends or clustering over time. Furthermore, financial losses might depend on covariates or additional variables that are possibly predictive for the dataset. Such covariates can be categorical factors, economic factors or even time. The non-stationary in time series can be handled either by filtering the data resulting in a stationary series of residuals or by using parametric, semi-parametric or non-parametric models to model the parameters of the Poisson and the GPD distribution (see Davison and Smith, 1990; Chavez-Demoulin and Davison, 2005).

3.2 Generalized Additive Models for Location, Scale and Shape (GAMLSS)

Extending the assumption of the classical POT approach into a dynamic setting where the parameters of the GPD depend on covariates, we introduce generalized additive models for location, scale and shape (GAMLSS) to model the relation between extreme values and the additional explanatory variables. Similar to Section 3.1, we assume threshold excesses follow a generalized Pareto distribution, $Y_i \sim \mathrm{GPD}(\xi_i, \beta_i)$ for $i \in \{1, \ldots, N\}$, where the shape parameter ξ and the scale parameter β are not constant but varying for every observation. Given p potential covariates $\mathrm{X} = (x_1, \ldots, x_p)$, GAMLSS specify each distribution parameter vector as

$$\begin{aligned} \eta_\xi(\mathrm{X}) = \eta_\xi = g_\xi(\xi) = \xi_0 + \sum_{j=1}^{p} h_{\xi,j}(x_j) \\ \eta_\beta(\mathrm{X}) = \eta_\beta = g_\beta(\beta) = \beta_0 + \sum_{j=1}^{p} h_{\beta,j}(x_j) \end{aligned} \tag{3.2}$$

where η (hereby, a generalized notation for either η_ξ or η_β) is the additive predictor vector which is related to the parameter through a link function g that guarantees that parameter estimates remain within the appropriate range depending on the underlying distribution. The shape parameter does not have a restriction on its domain, therefore an identity link function is appropriate, $\eta_\xi = g_\xi(\xi) = \xi$. The scale parameter, on the other hand, only takes positive values,

hence a logarithmic transformation function is necessary, $\eta_\beta = g_\beta(\beta) = \log(\beta)$. The effect of the covariates x_j on the additive predictors is specified through some base-learners h_j for $j \in \{1, \ldots, p\}$. The idea is to let the data determine the relationship between the distribution parameters and the explanatory variables. The base-learners, also known as smoothers, can be simple (parametric) linear regression models, $h_j^{linear}(x_j) = x_j\beta_j$, (non-parametric) penalized regression splines, $h_j^{smooth}(x_j)$, as well as classification or regression trees, thus making GAMLSS highly flexible. Each distribution parameter can have a subset of non-informative covariates. In that case, the effect will be estimated as $h_j(x_j) = 0$. Finally, ξ_0 and β_0 denote the intercept for each additive model.

GAMLSS models are fitted using penalized maximum likelihood using two possible algorithms. The first fitting procedure is based on Cole and Green (1992), the CG algorithm, and it uses the first derivatives and the expected values of the second and cross derivatives of the likelihood function with respect to the distribution parameters. The second algorithm is by Rigby and Stasinopoulos (1996a, 1996b), the RB algorithm, which, in contrast, assumes that the expected values of the cross derivatives are approximately zero. The estimation of the GAMLSS parameters is then done iteratively using a backfitting Gauss-Newton algorithm. This procedure is implemented in the freely available R package `gamlss` (Stasinopoulos and Rigby, 2007).

However, GAMLSS alone do not address the major issue of variable selection. As it happens often in practical applications, the variable of interest can have an extensive number of potential candidates as explanatory variables. Especially in high-dimensional data settings, a selection of the most informative covariates is necessary and often required. Rigby and Stasinopoulos (2005) use information criteria (such as AIC and GAIC) to select the best-performing model which contains only the most influential covariates. As previously mentioned, this approach has various shortcomings. In the next section, we discuss an alternative to the classical estimation of GAMLSS which applies machine learning boosting techniques for model fitting and variable selection.

3.3 Variable Selection in GAMLSS

Let $\mathrm{Y} = (Y_1, \ldots, Y_N)$ be the response variable where $Y_i \sim GPD(\xi_i, \beta_i)$ for $i \in \{1, \ldots, N\}$ and $\mathrm{X} = (x_1, \ldots, x_p)$ an $(N \times p)$ matrix of covariates. The relationship between the response variable and the covariates is modelled using GAMLSS as specified in Equation (3.2). In a high-dimensional setting, where the number of potential explanatory variables can be quite extensive, a selection of the most informative covariates is necessary. To solve this variable selection problem, we use component-wise gradient boosting as in Mayr et al. (2012). The fitting of the parameters of the conditional distribution is done by minimizing the expected value of a specified loss function. Hence, to model the distribution parameters we minimize

$$(\hat{\eta}_\xi, \hat{\eta}_\beta) = \underset{\eta_\xi, \eta_\beta}{\operatorname{argmin}} \Big\{ \mathbb{E}_{Y,X} \Big[\mathrm{Loss}\big(\mathrm{Y}, \eta_\xi(\mathrm{X}), \eta_\beta(\mathrm{X})\big) \Big] \Big\}.$$

Given that this expectation is unknown in practice, gradient boosting methods estimate

predictors by minimizing the *empirical risk*

$$\sum_{i=1}^{N} \text{Loss}\big(\text{Y}_i, \eta_\xi(\text{X}_i), \eta_\beta(\text{X}_i)\big) = -\sum_{i=1}^{N} l(\text{Y}_i; \xi_i, \beta_i) \tag{3.3}$$

which is the total loss across observations. In our case, the loss function is the negative log-likelihood of the distribution of the response variable. Component-wise gradient boosting fits the base-learners $h_j(x_j)$ to the negative gradient vector of the loss function one at a time per iteration. Hence, in every boosting iteration m, we compute the partial derivative of the log-likelihood of Y_i with respect to each of the additive predictors. This is done iteratively, first for the shape parameter ξ and then for the scale parameter β. In a general notation representing each of the parameters, the negative gradient of the loss function

$$u_i^{[m]} = \frac{\partial}{\partial \eta} l(Y_i; \xi, \beta)$$

is evaluated at $(\xi, \beta) = (\hat{\xi}_i^{[m-1]}, \hat{\beta}_i^{[m-1]}) = \Big(g_\xi^{-1}(\hat{\eta}_{\xi,i}^{[m-1]}), g_\beta^{-1}(\hat{\eta}_{\beta,i}^{[m-1]})\Big)$ for $i = 1, \ldots, N$. The gradient descent approach then includes only the best-fitting base-learner to the negative gradient (corresponding to the best-fitting covariate x_{j^*}) which minimizes the least squares criterion

$$j^* = \underset{1 \le j \le p}{\text{argmin}} \sum_{i=1}^{N} (u_i^{[m]} - h_j(x_{j,i}))^2.$$

The choice of the smoother $h_j(\cdot)$ is crucial here. Examples of base-learners include simple linear models, penalized regression splines as well as classification or regression trees. As these base-learners define the type of effect that covariates will have on the predictors of the GAMLSS distribution parameters, we decide to use penalized regression splines (P-splines) to account for non-linear effects. More details on how the fitting of the base-learners is performed in boosting algorithms can be found in Bühlmann and Hothorn (2007). Next, the boosting algorithm updates additive predictors only by a small step-length (typically a value of $\gamma = 0.1$ is used) of the current fit of the base-learner

$$\hat{\eta}^{[m]} = \hat{\eta}^{[m-1]} = \hat{\eta}^{[m-1]} + \gamma \cdot \hat{h}_{j^*}^{[m]}(x_{j^*}).$$

At the end of the iteration, the additive predictor is updated and the algorithm circles among the different distribution parameters. A schematic overview of this updating process for the GPD distribution parameters in iteration m is

$$\begin{aligned} \frac{\partial}{\partial \eta_\xi} l(Y_i; \hat{\xi}^{[m-1]}, \hat{\beta}^{[m-1]}) &\xrightarrow{\text{update}} \hat{\eta}_\xi^{[m]} \Longrightarrow \hat{\xi}^{[m]}, \\ \frac{\partial}{\partial \eta_\beta} l(Y_i; \hat{\xi}^{[m]}, \hat{\beta}^{[m-1]}) &\xrightarrow{\text{update}} \hat{\eta}_\beta^{[m]} \Longrightarrow \hat{\beta}^{[m]} \end{aligned}$$

The stopping criterion $m_{stop} = (m_{\text{stop},1}, m_{\text{stop},2})$ plays an important role in the boosting algorithm as it decides the number of iterations the algorithm performs for each distribution parameter, which ultimately controls the variable selection procedure. The additive predictors are updated only while the number of boosting iterations has not exceeded the stopping criterion.

If the stopping criterion is reached for one of the distribution parameters, the corresponding additive predictor stops being updated and the algorithm moves on to the next distribution parameter. In each boosting iteration, the best-fitting covariate is selected and added to the GAMLSS model. Once the stopping criterion is reached, all the covariates that are not selected yet are assumed to be non-informative and hence are excluded. For the implementation and estimation of the *gamboostLSS* algorithm, we use the R package `gamboostLSS` (Hofner et al., 2018) designed for variable selection in high-dimensional GAMLSS settings. The pseudocode for the gradient boosting algorithm to fit GAMLSS is given in Algorithm 1.

3.3.1 Hyperparameter Tuning

The *gamboostLSS* algorithm with its iterative gradient boosting procedure requires the following parameters:

- The stopping criterion $m_{stop} = (m_{stop,1}, m_{stop,2})$ setting the maximum number of iterations for the algorithm. This hyperparameter is crucial for the *gamboostLSS* algorithm. For small datasets, it may be convenient to let the algorithm run until convergence. In high-dimensional scenarios, however, early stopping is necessary to prevent overfitting. The choice of $m_{stop,k}$ controls for both the amount of shrinkage applied and the complexity of the model. Early stopping reduces the number of covariates selected for each additive predictor and shrinks their predictor functions to zero, which directly addresses our variable selection problem. The optimal $\boldsymbol{m}_{stop}$ can be determined using one-dimensional ($m_{stop} = m_{stop,1} = m_{stop,2}$) or multi-dimensional cross-validation ($m_{stop,1} \neq m_{stop,2}$).
- The learning rate for the gradient descent, or the step-length, $\gamma_{sl} = (\gamma_{sl,1}, \gamma_{sl,2})$ for each of the distribution parameters. The tuning of this hyperparameter has shown not to affect the results. In general, a value of 0.01 or 0.1 is used.

4 Simulation Study

In this section, we conduct a simulation study to evaluate the accuracy of the variable selection as well as model estimation adequacy of our proposed model. Section 4.1 gives the data-generation process of our simulation as well as the main results. Section 4.2 performs a sensitivity analysis by considering low- and high-dimensional scenarios for both the number of observations and the number of covariates. Furthermore, we investigate how correlation amongst covariates affects the simulation results in Section 4.3.

4.1 Data-Generation Process and Simulation Results

We simulate $N = 1000$ extreme observations from a Generalized Pareto distribution, in line with the EVT theory of Section 3.1.

$$Y_i \sim GPD(\xi_i, \beta_i), \quad i = 1, \ldots, N.$$

Furthermore, we take into consideration 30 covariates, $\mathrm{X} = (x_1, \ldots, x_{30})$, independently and identically simulated from a standard uniform distribution, $U(0, 1)$. Note that using the log-

Algorithm 1: Component-wise Gradient Boosting for Variable Selection in GAMLSS

Input:

The response variable $\mathrm{Y} = (Y_1, \ldots, Y_N)$ where $Y_i \sim GPD(\xi_i, \beta_i)$ for $i \in \{1, \ldots, N\}$,
The covariates $\mathrm{X} = (x_1, \ldots, x_p)$,
The stopping criterion $m_{stop} = (m_{stop,1}, m_{stop,2})$ for each distribution parameter,
The learning rate for the gradient descent $\gamma = (\gamma_1, \gamma_2)$.

Output:

The fitted GPD distribution parameters $\hat{\boldsymbol{\theta}} = (\hat{\xi}, \hat{\beta}) = (\hat{\eta}_\xi, \exp(\hat{\eta}_\beta))$,
The selected predictors x_{j^*} for $j^* \in \mathcal{J}$, the set of the indices of the selected predictors.

1 Initialize the additive predictors $\hat{\eta}_\xi^{[0]} = 0$, $\hat{\eta}_\beta^{[0]} = 0$, or simply, $\hat{\eta}_{\theta_k}^{[0]} = 0$ for $k = 1, 2$ corresponding to each distribution parameter and set $m = 0$.

2 Specify a set of base-learners $h_{k,1}(\cdot), \ldots, h_{k,p_k}(\cdot)$, where p_k is the cardinality of the set of base-learners specified for θ_k.

3 **for** $m = 1, \ldots, \max(m_{stop,1}, m_{stop,2})$ **do**

4 **for** $k = 1, 2$ **do**

5 **if** $m > m_{stop,k}$ **then**

6 Set $\eta_{\theta_k}^{[m]} := \eta_{\theta_k}^{[m-1]}$ and skip this iteration.

7 **else**

8 Compute the negative partial derivative by plugging in the current estimates $\hat{\boldsymbol{\theta}}^{[m-1]} = \left(\hat{\xi}^{[m-1]}, \hat{\beta}^{[m-1]}\right) = \left(\hat{\eta}_\xi^{[m-1]}, \exp(\hat{\eta}_\beta^{[m-1]})\right)$:

$$u_{k,i}^{[m]} = \frac{\partial}{\partial \eta_{\theta_k}} l(Y_i; \boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_i^{[m-1]}}, \quad i = 1, \ldots, N$$

9 Fit each of the base-learners specified for the parameter θ_k in Step 2 to the negative gradient vector $\boldsymbol{u}_k^{[m]}$.

$$\left(x_j, \boldsymbol{u}_k^{[m]}\right) \xrightarrow{\text{base-learner}} \hat{h}_{k,j}^{[m]}(x_j), \quad \text{for } j = 1, \ldots, p_k.$$

10 Select only the covariate j^* that best fits the negative partial-derivative vector according to the least squares criterion:

$$j^* = \underset{1 \le j \le p_k}{\operatorname{argmin}} \sum_{i=1}^{n} (u_{k,i}^{[m]} - h_{k,j}(x_j))^2.$$

11 Update the current additive predictor η_{θ_k} as follows:

$$\hat{\eta}_{\theta_k}^{[m-1]} := \hat{\eta}_{\theta_k}^{[m-1]} + \gamma_k \cdot h_{k,j*}(x_{j^*}),$$

where γ_k is the learning rate ($0 \le \gamma_k \ll 1$). Also, update the additive predictor for the next iteration

$$\hat{\eta}_{\theta_k}^{[m]} := \hat{\eta}_{\theta_k}^{[m-1]}.$$

arithmic link function to ensure that the GPD parameters are strictly positive makes the parameters susceptible to outliers. The assumption of standard uniformly distributed covariates limits that. Next, we assume that the additive predictors are dependent on the covariates in the following functional form

$$\begin{aligned} \log(\xi) = \eta_\xi &= x_1^2 + \sin(x_2) - 3\exp(x_3) - x_4 \\ \log(\beta) = \eta_\beta &= x_3^2 + \sqrt{x_4} + \exp(x_5) - x_6 \end{aligned} \tag{4.1}$$

In this setting, there are six informative covariates and 24 uninformative ones. As the relationship between the additive predictors and the covariates is non-linear, the base-learners we choose for the GAMLSS fit are smooth P-splines, with 20 knots, a second-order difference penalty and 4 degrees of freedom as control parameters. In this study, we perform 100 simulation runs. The "true" model of Equation 4.1 returns values of ξ with mean around 0.02 (standard deviation of 0.024) and values of β with mean 0.38 (0.279). Figure 3 gives a plot of the simulated

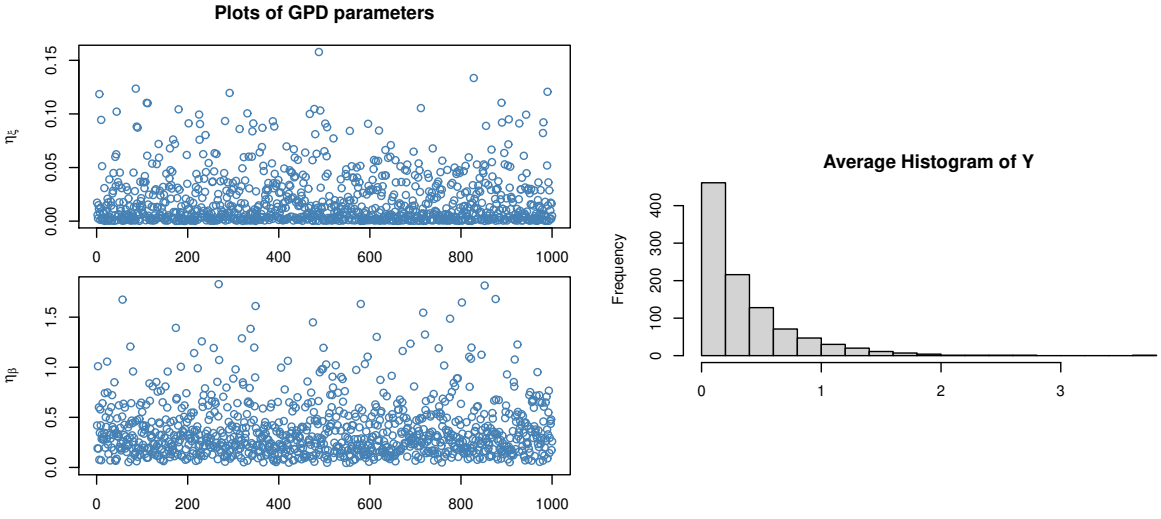

Figure 3: Plots of the additive predictors for GPD parameters (η_ξ, η_β) simulated using Equation 4.1 as the true-underlying model (left) and a histogram of the simulated Y values using average parameter values and 1000 observations (right)

GPD parameters and a histogram of the simulated Y values using average parameter values and 1000 observations. Note that this is just to give an impression of what the histogram of Y looks on average, not to set our focus on the unconditional distribution of Y.

As previously mentioned, the stopping iteration m_{stop} plays a crucial role in determining the number of variables selected by the model to estimate the additive predictors. Although the incorporation of additional variables to the model results in a lower empirical risk, there is overfitting of the model. To avoid this, early-stopping is necessary, meaning we should not let the algorithm run until convergence. With early-stopping, the effects of less informative covariates shrink to zero as the gradient boosting algorithm performs a data-driven variable selection and updating. In our simulation study, the true model is based on four explanatory variables for the shape parameter, x_1, x_2, x_3 and x_4 and four for the scale parameter, namely x_3, x_4, x_5 and x_6. Therefore, we should pick an appropriate number of stopping iterations that selects four

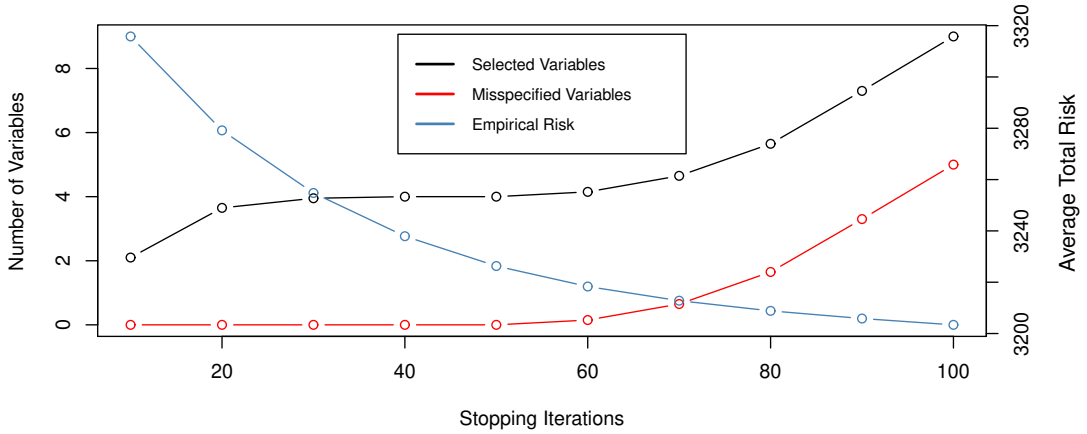

Figure 4: Number of selected variables (black), number of incorrectly selected variables (red) and average total empirical risk (blue) per stopping criterion (ranging from 10 to 100 iterations)

covariates for each additive predictor. Figure 4 shows how this number varies per different stopping criteria ranging from 10 to 100, when using a one-dimensional stopping criterion. On average, throughout the simulations that we conducted, the average number of selected variables (black line) is around four where $m_{stop} = 40$ iterations. The number of incorrectly specified variables (red line) also remains at zero until this point. Note that, by (in-)correctly selected variables, we refer to the (non-)informative variables that the model selects. For lower numbers of stopping iterations, the model is missing out on important informative covariates, while for higher stopping criteria the probability of incorrectly selecting non-informative covariates increases. The average empirical risk (blue line), which is the empirical risk as computed in Equation (3.3) averaged accross the 100 simulation runs, decreases with the stopping iterations, as more explanatory variables are added to the model. However, this comes at the cost of possible overfitting.

Table 1: Results from 100 simulation runs

	# Covariates = 30
	$N = 1000$
Selection Rates (Informative, ξ)	31.6%
Selection Rates (Non-Informative, ξ)	6.7%
Selection Rates (Informative, β)	100%
Selection Rates (Non-Informative, β)	0.1%
Total Risk	3234
Running Time	70.26 sec

The results of the simulation are summarized in Table 1. Throughout the 100 simulation runs, the informative covariates were selected 100% of the time for the scale parameter, while only 31.6% for the shape parameter. Moreover, the average selection rate for the other non-informative covariates remains quite low at 0.1% and 6.7%, respectively. The boosting algorithm works extremely well for the scale parameter, shrinking the effect of the less informative variables to zero while carefully selecting the ones with the most significant impact on the additive

predictors. Given the more complex nature of the non-linear functional form between the explanatory variables and the GPD parameters, the importance of these results is amplified even further. The results are not as satisfactory for the shape parameter. The model is not able to capture the true effect for all of the informative variables. However, the challenges arising with the modelling of the shape parameter using covariates are not surprising; the shape parameter is difficult to estimate and this is well-known in the current literature (Coles, 2001).

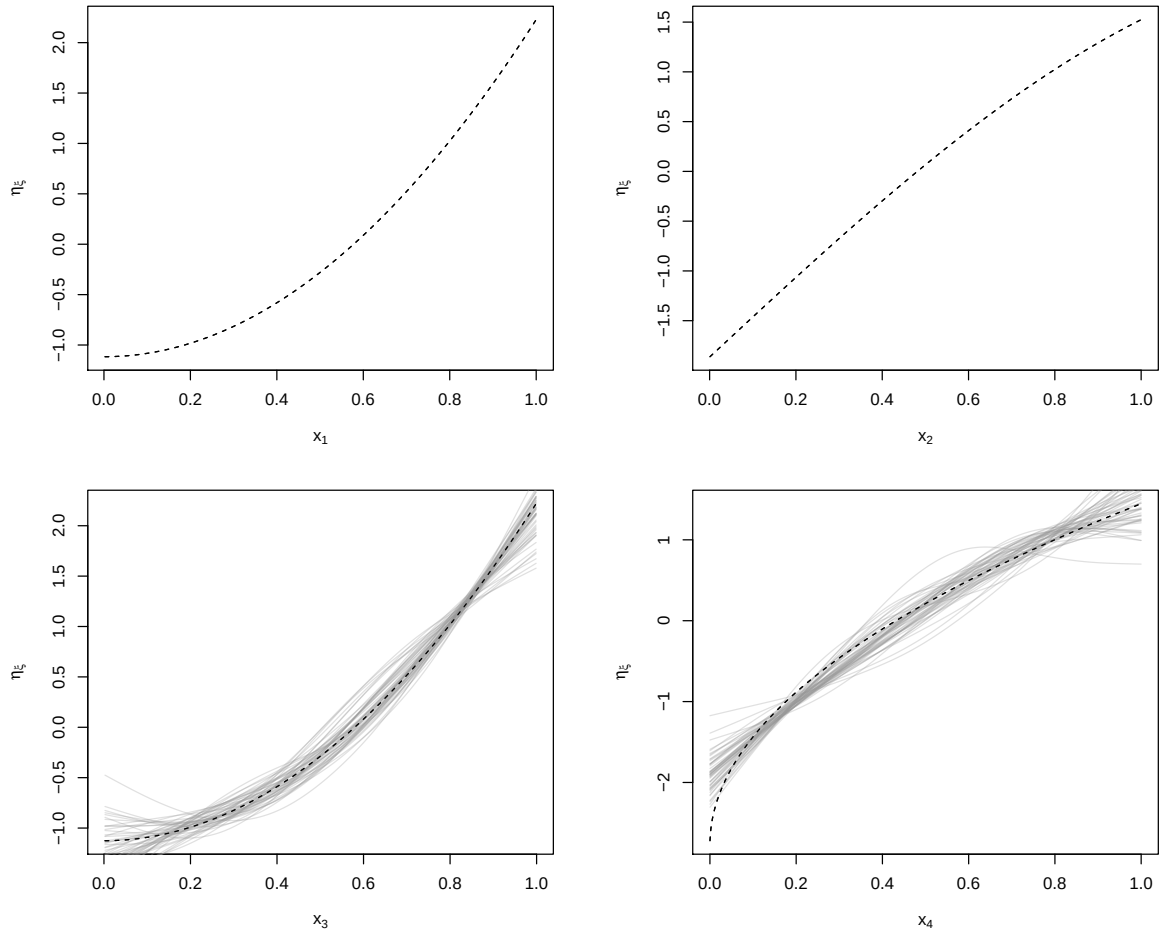

Figure 5: Estimated non-linear effects (———) for the shape parameter ξ (using 100 simulations and $m_{stop} = 40$ iterations) against the true-underlying effect (- - - -)

The estimated effects of the informative covariates on the additive predictors η_ξ and η_β for 100 simulation runs, are plotted against the underlying functions in Figures 5 and 6. These results are in line with the ones in Table 1. For the shape parameter, the model is only able to capture the effect of two of the four informative covariates, x_3 and x_4. Although these covariates are not selected in every simulation, the estimated effect is very close to the true effect when they do. In contrast, the results for the scale parameter are very adequate. The algorithm not only selects the informative covariates correctly, but the approximations of the functional form of the covariates are extremely close to the true model. Given these results, the overall performance of our model could be potentially improved if we keep the shape parameter constant. This is investigated further later. Note that the simulation is conducted under the certain assumptions which might be perceived as favourable. First, the number of observations can be considered

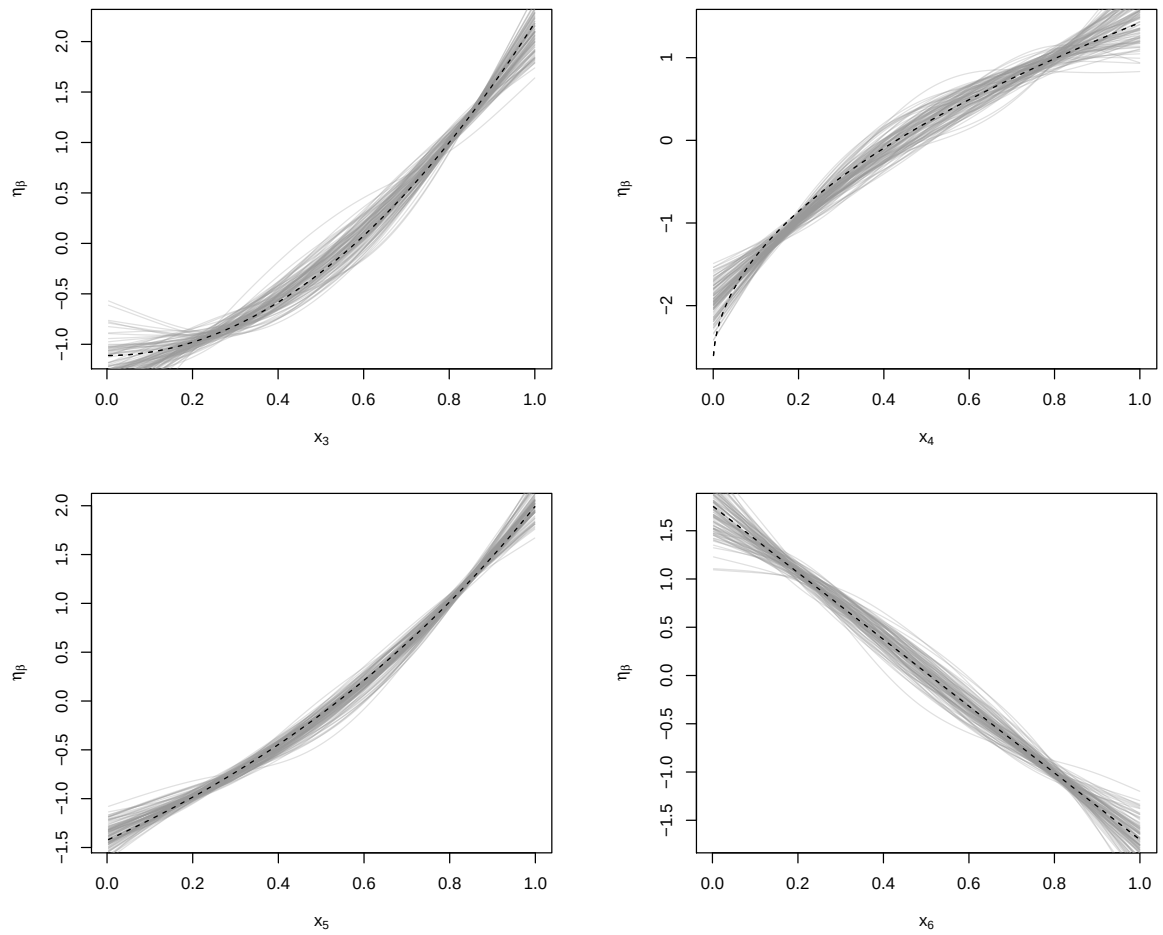

Figure 6: Estimated non-linear effects (———) for the scale parameter β (using 100 simulations and $m_{stop} = 40$ iterations) against the true-underlying effect (- - - -)

high. In practice, this assumption could be unrealistic as extreme values are scarce by nature. Second, the number of covariates is lower than the number of observations. In a high-dimensional setting, where the number of explanatory variables exceeds the number of the response variables, it would be of great interest to see whether these results continue to hold. Last, the covariates have no correlation with one another. High-correlation between covariates might also have an impact on our conclusions. A further investigation for these points of concern is conducted in the later sections.

As the true-underlying distribution of the response variable is known in the simulation, the performance of the model can be evaluated by comparing the true and estimated values for the risk measure Value-at-Risk (VaR) as calculated in Equation 3.1. Table 2 shows the Mean Squared Error (MSE), calculated as the mean squared difference between the actual and the fitted values for VaR at a 95%, 97.5% and 99% level. Our covariate-dependant model with varying distribution parameters is compared with two simplified models: one model where only the shape parameter is kept constant (keeping the scale parameter varying) and one where both the shape and the scale parameters are constant (in line with the classical POT approach). The results indicate two things. First, introducing covariates to the distribution parameters increases performance significantly compared to the model with constant parameters. This confirms once

again the importance of modeling distribution parameters that are varying and transitioning from the classical EVT to a dynamic one. Secondly, setting the shape parameter as constant does not have a large impact on the model performance. The model with a constant shape parameter outperforms our model when we look at the 95% VaR, while the opposite is observed for higher levels of probability. For the 97.5% VaR the difference between the two models is relatively small. Overall, our model is able to capture the more extreme observations better than the other two simplified models.

Table 2: Mean Squared Error (MSE) values for the VaR at a 95%, 97.5% and 99% level

	$\text{MSE}(\text{VaR}_{0.95})$	$\text{MSE}(\text{VaR}_{0.975})$	$\text{MSE}(\text{VaR}_{0.99})$
Varying ξ and β	0.352	0.884	3.064
Constant ξ and varying β	0.313	0.893	3.391
Constant ξ and β	0.913	1.816	4.732

So far, we have assumed a one-dimensional stopping criterion, meaning the stopping criterion is the same for both distribution parameters. An alternative would be to conduct the simulation using a two-dimensional stopping criterion, where the number of iterations varies per parameter. However, since the algorithm is not able to adequately capture the effect of the shape parameter, we do not expect that varying the stopping criterion for ξ, would improve the ability of the algorithm for accurately modelling β. Figure 7 indicates that neither the number of covariates selected, nor the total empirical risk, are dependent on the stopping criterion for the shape parameter. Furthermore, the selection rates also remain unchanged, hence confirming our initial expectations.

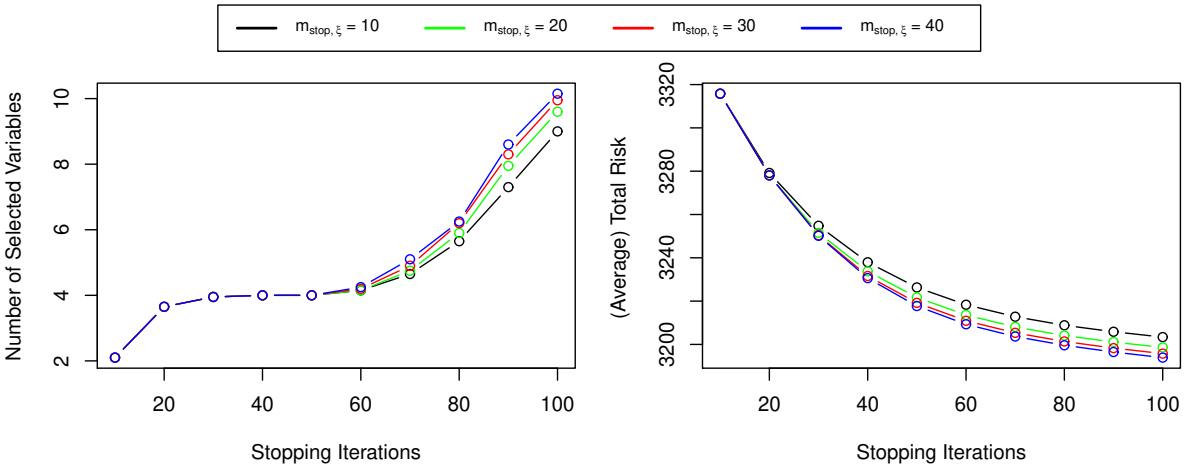

Figure 7: Number of selected variables and average total empirical risk using a two-dimensional stopping criterion

4.2 Sensitivity Analysis

In this section, we investigate further if the results from the simulation study remain valid if the initial assumptions are changed. Table 3 gives the selection rates for the informative and non-informative covariates, as well as the total risk and the running time for $N = \{150, 500, 800\}$

observations and $p = \{50, 250, 500, 1000\}$ covariates. To avoid that any random effect influences the simulation results of the *gamboostLSS* algorithm, the covariates (and, hence, the GPD parameters) are set to remain constant across the simulations. First, in the low-dimensional scenario where the number of observations is small, we observe that the model performs the worst regardless of the number of covariates. The selection rates for the informative covariates are the lowest, while the selection rates for the non-informative covariates are highest, as compared to larger datasets. As the gradient boosting algorithm performs data-driven variable selection, it is common sense that it would perform better given a larger set of observations.

Table 3: Results from 100 simulation runs in several high-dimensional settings

	# Covariates = 50			# Covariates = 250		
	$N = 150$	$N = 500$	$N = 800$	$N = 150$	$N = 500$	$N = 800$
Selection Rates (Informative)	74.5%	98.5%	100%	68.1%	99.0%	100%
Selection Rates (Non-Informative)	12.1%	1.2%	0.4%	3.7%	0.5%	0.1%
Total Risk	475	1627	2600	490	1599	2595
Running Time	28.89 sec	39.74 sec	101.25 sec	138.64 sec	199.08 sec	534.54 sec
	# Covariates = 500			# Covariates = 1000		
	$N = 150$	$N = 500$	$N = 800$	$N = 150$	$N = 500$	$N = 800$
Selection Rates (Informative)	65.0%	98.5%	99.0%	57.5%	98%	100%
Selection Rates (Non-Informative)	2.3%	0.6%	0.4%	1.26%	0.8%	0.01%
Total Risk	492	1617	2620	473	1615	2620
Running Time	314.40 sec	552.35 sec	1204.18 sec	583.67 sec	872.40 sec	2178.71 sec

Second, the algorithm performs worse in high-dimensional settings where the number of covariates exceeds the number of obeservations. This can be observed in the decreasing selection rates for the significant covariates as the number of covariates increases, and it is most obvious on smaller datasets.

4.3 Highly-Correlated Covariates

We further examine the simulation results when there is correlation between covariates. Once again, consider the initial setting with 1000 extreme observations and 30 covariates. Assume that the covariates follow a multivariate normal distribution with zero mean, unit variance and pair-wise correlation ρ between variables. Hence, $x_j \sim N(0_{p\times 1}, \Sigma_{p\times p})$ for $j = 1, \ldots, p$, where diag(Σ) = 1 and the off-diagonal elements equal ρ. We consider different values for $\rho = \{0.5, 0.7, 0.8, 0.9\}$ and compare between the various correlation scenarios.

Table 4: Results with correlated covariates from 100 simulation runs

	$N = 1000$ and # Covariates = 30				
	$\rho = 0$	$\rho = 0.5$	$\rho = 0.7$	$\rho = 0.8$	$\rho = 0.9$
Selection Rates (Informative)	100%	100%	75.8%	73.5%	67.5%
Selection Rates (Non-Informative)	0%	1.2%	3.7%	5.1%	5.6%
Stopping Criterion ($m_{stop,\beta}$)	40	32	30	27	25
Total Risk	3234	3226	3219	3223	3206
Running Time	70.26 sec	64.82 sec	62.93 sec	58.25 sec	56.06 sec

Table 4 summarizes the main results from 100 simulation runs in high-correlation settings.

As expected, introducing correlation to the covariates distorts the performance of the model on two fronts. The selection rates for the informative covariates decrease with the increase in the correlation, while the selection rates for the non-informative covariates increase. Note that the "wrongful" selection of a variable that is highly-correlated to an informative covariate does not necessarily mean the model performance is poor, as all correlated explanatory variables will have similar effects on the additive predictors. In high-correlation cases, the algorithm cannot distinguish between the different covariates. To make sure the selection rates are comparable, the stopping criterions are set such that under each scenario only four covariates are selected by the model. As the correlation amongst covariates becomes higher, less iterations are needed to achieve this.

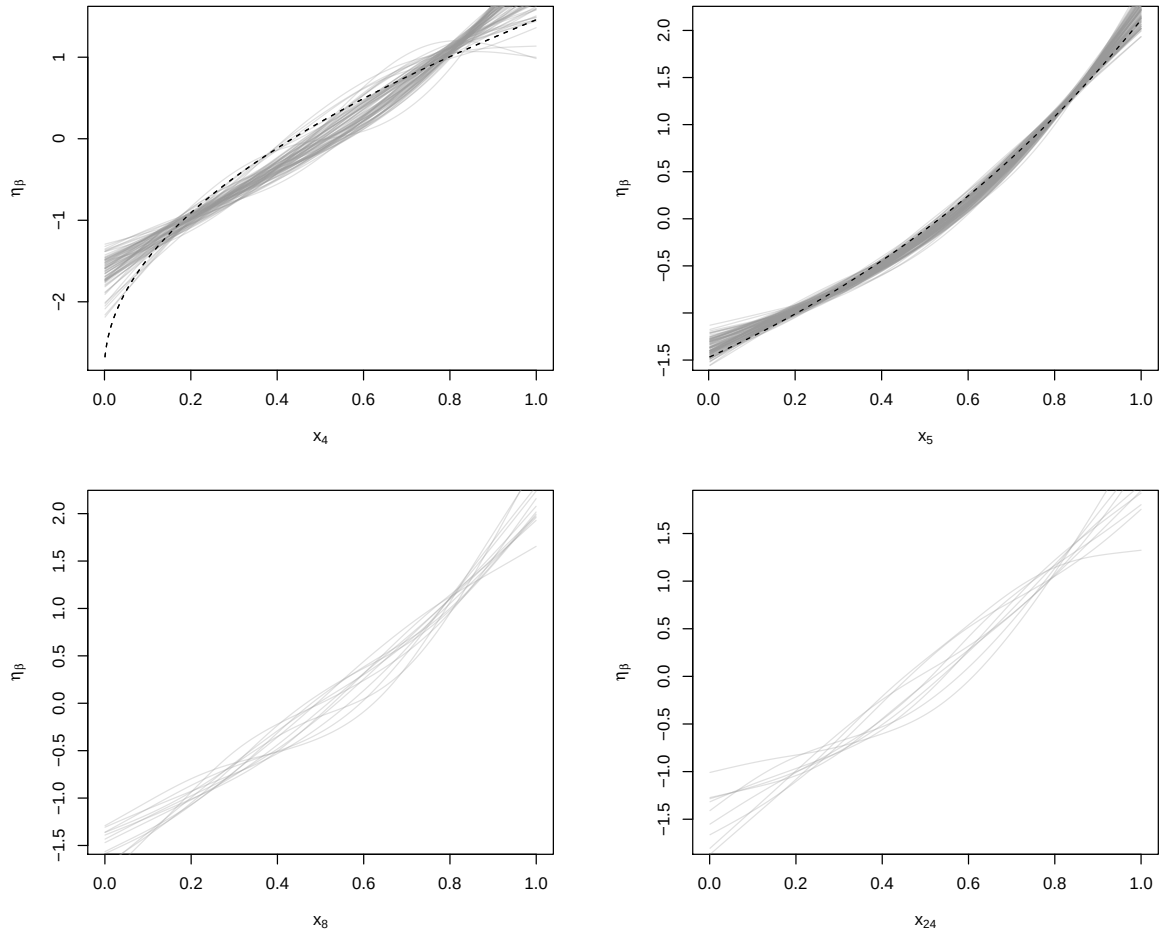

Figure 8: Estimated non-linear effects (———) for the scale parameter β (using 100 simulations and $m_{stop} = 25$ iterations) against the true-underlying effect (- - - -) when correlation is high among covariates ($\rho = 0.9$). The informative covariates are plotted at the top, non-informative at the bottom.

Figure 8 gives the estimated covariate effects for the scale paramater for two of the informative covariates x_4 and x_5 and the two most-selected non-informative covariates within 100 simulations in a high-correlation setting where $\rho = 0.9$. In this case, the informative covariates have selection rates of 74% and 100%, while the non-informative ones are selected 12% and 9% of the time, respectively. Comparing with Figure 6, the estimated effects of x_5 are once

again very close to the true effects, whereas the effects of x_4 seem to be less accurate, especially closer to the bounds. On the other hand, the true-underlying relationship between the additive predictors and the non-informative covariates is unknown. However, there is a clear functional form, mostly resembling the exponential or polynomial, which is similar to the true effect of some of the informative covariates. This indicates that model selects non-informative explanatory variables that are highly correlated to the informative ones. The main conclusion that we derive from this analysis is that correlation among explanatory variables largely affects model performance, however this is compensated by the fact that highly-correlated covariates have similar relationships with the response variable.

5 Application

In the previous section, we show that our methodology is adequate not only under optimal assumptions, but also under high-dimensional settings with a large number of covariates and (to a lesser extent) when covariates were highly-correlated with each other. Testing assumptions is crucial to evaluate the performance of the model under different scenarios which are ideally as close to reality as possible. In this section of the paper, we test the model with real data. Section 5.1 gives a detailed description of the response variable and the covariates we use as well as the data sources, summary statistics and the main literature supporting the selection of the covariates. The optimal hyper-parameters and the optimal threshold are selected prior to estimating the model in Section 5.2 and 5.3. Finally, the main results on variable selection and covariate effects are explained in Section 5.4 and on the predictive performance of the model in Section 5.5.

5.1 Data Description

For a real-data application, we look into the extreme events in the S&P 500 market index during the last 20 years. Our dataset consists of daily losses (negative returns) from August 31st, 2000 until January 29th, 2021, amounting to 4496 observations. This timeframe includes important global events and financial crises such as the (post-peak) dot-com bubble, 9/11 attacks, the financial crisis of 2007-2009, the Euro area debt crisis and the most recent global pandemic of COVID-19. Using the POT approach, we restrict our focus to extreme observations, defined as the data points that are above a certain threshold. Given this threshold, the original data is transformed into excess losses by subtracting the threshold from the extremes. After performing a model sensitivity analysis to carefully choose the threshold (more details can be found in Section 5.3), we only include the 899 observations that are above the 80-th percentile.

Financial return time-series often exhibit non-stationary behaviour, such as heteroskedasticity and volatility clustering. This is also evident in our excess loss data, which is plotted in Figure 9. To account for the non-stationarity, we use GAMLSS to model distribution parameters that are time-varying by introducing covariates that might trigger some of this behaviour. For this reason, a selection of 37 potential covariates is made including important market indices (such as NASDAQ, Dow Jones, FTSE100), volatility indices (VIX), bond yields (treasury bills with maturities varying from 3 months to 20 years), commodities (crude oil, gold, silver), for-

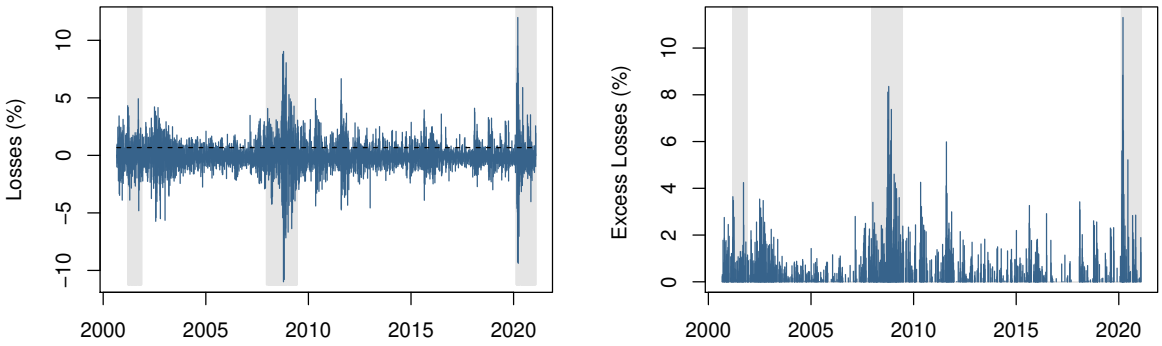

Figure 9: S&P 500 losses (left) and excess losses above the 80-th percentile (right). The grey shaded regions mark the NBER-based recession indicators.

eign exchange (FX) rates, Fama-French factors and other variables that might explain the tail behavior of S&P 500 losses. Our data is retrieved from several databases such as Center for Research in Security Prices (CRSP) and Compustat databases accessible via the Wharton Research Data Services (WRDS) website, Federal Reserve Economic Data (FRED) by the Federal Reserve Bank of St. Louis and the Federal Reserve Board (FRB).

The list of explanatory variables can be quite extensive, some of them being highly informative and some to a lesser extent. The selection of the covariates is made largely based on the current literature, intuition and data availability at a daily frequency. Hilliard (1979) investigates the relationship among the world's largest stock exchanges. They conclude that most inter-continental exchanges have a tendency to move simultaneously, while intra-continental ones are not as highly related (with the exception of the New York and Amsterdam stock exchanges). Hence, our excess S&P 500 losses should be more closely related to other U.S. market indices than, for example, European or Asian ones. Giot (2005) show that VIX, which is an implied volatility index derived from S&P 500 option prices, has a significant negative effect on the S&P 500 index. Moreover, this effect is asymmetric, meaning the impact on negative index returns is greater than on positive ones.

Fogler et al. (1981) give evidence to how stock market returns are related to government bond returns and corporate bonds with default risk and how changes in maturity affect this relationship. Therefore, we include U.S. Treasury bills with different maturities of 5, 10 and 30 years as covariates. Flannery and James (1984) shows that stock returns of financial institutions are very sensitive to interest rate movements. As a benchmark for short-term interest rates, we consider the over-night, 1-month, 3-month, 6-month and 1-year LIBOR.

Kilian and Park (2009) study the relationship between crude oil price shocks and U.S. equity returns. Except for crude oil, we also include other commodities, such as gold, silver and platinum, as potential explanatory variables for movements in extreme losses in the S&P 500 market index. In general, commodities are found to be positively correlated to inflation and negatively correlated to stock returns. Nieh and Lee (2001) explore the potential response of

equity prices to changes in daily foreign exchange rates during 1993 until 1996 for the G-7 countries. They find no significant correlation in the U.S. market, however the current literature is divided on this.

Finally, the Fama and French (2015) 5-factor model constructs value-weighted portfolios based on size, value, operating profitability and investment pattern. We include four of the factors, namely the SMB (Small-Minus-Big) which measures the difference in average returns between small and large firms, HML (High-Minus-Low book-to-market ratio firms) measuring the difference in average returns between growth and value firms, RMW (Robust-Minus-Weak operating profitability) and CMA (Conservative-Minus-Aggresive investment portfolios), excluding only the market premium factor.

Table 5: Descriptive statistics on the daily S&P 500 losses, extreme returns above the 80th percentile and the corresponding excesses (in percentage terms) from August 31st 2000 until March 30th, 2019.

	Mean	St. Dev.	Minimum	20^{th} Percentile	80^{th} Percentile	Maximum	Skewness	Nr. Obs.
S&P500 losses	-0.029	1.314	-10.986	-0.760	0.677	11.984	0.207	4496
Extreme losses	1.694	1.187	0.679	0.880	2.216	11.984	3.106	899
Excess losses	1.017	1.187	0.002	0.203	1.538	11.307	3.106	899

The descriptive statistics for the S&P 500 losses (negative returns), the extreme losses above the 80th percentile (threshold) and the corresponding excess losses (extreme losses − threshold) are summarized in Table 5. The S&P 500 losses average around -0.03% and take the maximum of almost 11.98% in the beginning of 2020 with the start of the pandemic. The 80th percentile of the S&P 500 losses, which marks the threshold, is at 0.68%. That means that our extreme or excess loss data composes 20% of the entire dataset, that is, 899 observations.

The list of the potential covariates as well as the summary statistics are presented in Table 6. We include 11 market and volatility indices, 3 government bonds, 4 short-term interest rates (less than 1 year), 3 commodities, 11 foreign exchange rates (with USD as one of the currencies) and 4 Fama-French factors from their 5-factor model (excluding the market premium), totalling up to 37 explanatory variables. Note that, for all the covariate data (that are not returns), we compute the percentage change. This way, we make sure the variables are similarly scaled and the interpretation is easier. Furthermore, there is missing data in our financial series. We decide to exclude these data points. Overall, most covariates have a positive average return. Exceptions include the interest rates, the 10- and 30-year treasury bills, crude oil and some exchange rates. The most volatile variables include crude oil, which has experienced losses as high as 300%, and volatility index VIX, with the highest return of 115%. The market indices exhibit similar properties with each other and their returns range within an interval of -14% and 14%. Some of the exchange rates are the least volatile covariates, together with the Fama-French factors. For all the covariates, the Jarque-Bera statistics indicate that they are not normally distributed.

Table 6: Descriptive statistics of the covariate data (expressed in percentage change terms) including the mean, standard deviation, minimum, maximum, skewness, kurtosis, the Jarque-Bera test statistic for normality and the corresponding probability value in parenthesis.

	Mean	St. Dev.	Minimum	Maximum	Skewness	Kurtosis	JB Test	p-Value
Market Indices								
AEX	0.008	1.501	-10.753	10.277	0.066	7.372	10197	(0.000)
CAC40	0.006	1.542	-12.277	14.230	0.144	7.348	10143	(0.000)
DAX	0.026	1.582	-12.239	14.411	0.111	6.765	8593	(0.000)
DJA	0.030	1.265	-12.927	11.365	-0.215	12.083	27418	(0.000)
EURONEXT100	0.008	1.418	-11.972	13.177	0.080	7.671	11041	(0.000)
FTSE100	0.007	1.266	-10.874	11.753	0.048	8.983	15135	(0.000)
HANGSENG	0.026	1.520	-13.667	14.347	0.180	10.630	21217	(0.000)
NASDAQ	0.038	1.603	-12.321	10.477	-0.094	5.445	5568	(0.000)
NIKKEI225	0.025	1.573	-12.124	14.150	-0.297	6.933	9083	(0.000)
RUSSELL2000	0.043	1.621	-14.272	9.391	-0.405	6.413	7837	(0.000)
VIX	0.293	7.913	-39.129	115.598	2.138	17.190	58843	(0.000)
Government Bonds								
TBILL5Y	0.008	3.668	-39.043	43.124	0.615	16.195	49467	(0.000)
TBILL10Y	-0.002	2.704	-31.725	49.900	1.624	48.634	445488	(0.000)
TBILL30Y	-0.008	1.899	-22.881	30.096	0.490	35.964	242719	(0.000)
Interest Rates								
LIBOR1M	-0.055	1.694	-34.229	22.623	-3.557	88.680	1484055	(0.000)
LIBOR3M	-0.047	1.370	-23.863	18.271	-1.643	53.254	533796	(0.000)
LIBOR6M	-0.055	1.326	-21.016	12.597	-1.722	36.824	256499	(0.000)
LIBOR12M	-0.047	1.582	-22.328	16.041	-0.520	25.076	118115	(0.000)
Commodities								
CRUDEOIL	-0.028	5.804	-305.966	90.843	-33.960	1.778.568	593985443	(0.000)
GOLD	0.049	1.182	-9.354	9.028	-0.144	5.647	5997	(0.000)
SILVER	0.059	2.109	-25.751	13.872	-0.864	11.157	23904	(0.000)
FX Rates								
USDEUR	0.007	0.601	-2.958	4.729	0.093	2.732	1407	(0.000)
USDGBP	0.002	0.609	-7.845	4.535	-0.634	10.524	21073	(0.000)
USDAUD	0.006	0.810	-7.884	8.008	-0.509	11.951	26979	(0.000)
CNYUSD	-0.006	0.157	-1.998	1.833	0.029	23.009	99281	(0.000)
JPYUSD	0.000	0.623	-5.082	3.109	-0.334	4.224	3430	(0.000)
CADUSD	0.005	0.566	-4.945	3.880	0.120	4.872	4463	(0.000)
BRLUSD	0.026	1.056	-9.216	9.054	0.193	8.701	14228	(0.000)
INRUSD	0.005	0.451	-3.686	4.017	0.302	10.742	21710	(0.000)
CHFUSD	-0.014	0.675	-12.210	9.298	-0.851	33.839	215264	(0.000)
HKDUSD	0.000	0.034	-0.441	0.335	-0.814	21.765	89326	(0.000)
SGDUSD	-0.007	0.331	-2.357	2.732	0.040	5.161	4998	(0.000)
Fama-French Factors								
SMB	0.017	0.603	-4.580	5.730	0.224	5.041	4805	(0.000)
HML	0.005	0.730	-4.890	6.700	0.443	9.791	18127	(0.000)
RMW	0.021	0.482	-3.020	3.270	0.092	4.261	3413	(0.000)
CMA	0.013	0.402	-3.380	2.430	-0.183	7.409	10323	(0.000)

Section 4.3 discusses on the importance of the dependence structure between covariates and how it affects the model performance. The higher the correlation between explanatory variables, the lower the selection rate for informative covariates and the higher the probability to incorrectly select non-informative ones. Figure 10 gives the correlation matrix between the variables in the form of a heatmap. We also include the S&P 500 excess losses, which are extremely (negatively) correlated with the other U.S. market index returns, such as Dow Jones, NASDAQ and Russell 2000. All other market indices are also closely related, except for the ones from Asian stock exchanges. Crude oil is amongst the least correlated covariates to our response variable, together with gold, the 3, 6 and 12-month LIBOR rates and some exchange rates like HKD/USD. The VIX is positively correlated to S&P 500 losses.

The correlations between the covariates range between -0.85 and 0.98, with only 5% of the

values being higher than 0.5. Hence, our algorithm should be reliable enough to make a good selection of the most informative covariates for our S&P 500 excess loss dataset. The most highly correlated explanatory variables are among the market indices, Treasury bills, LIBOR rates and some of the exchange rates, hence among the same asset class. Gold and silver prices are also very positively correlated.

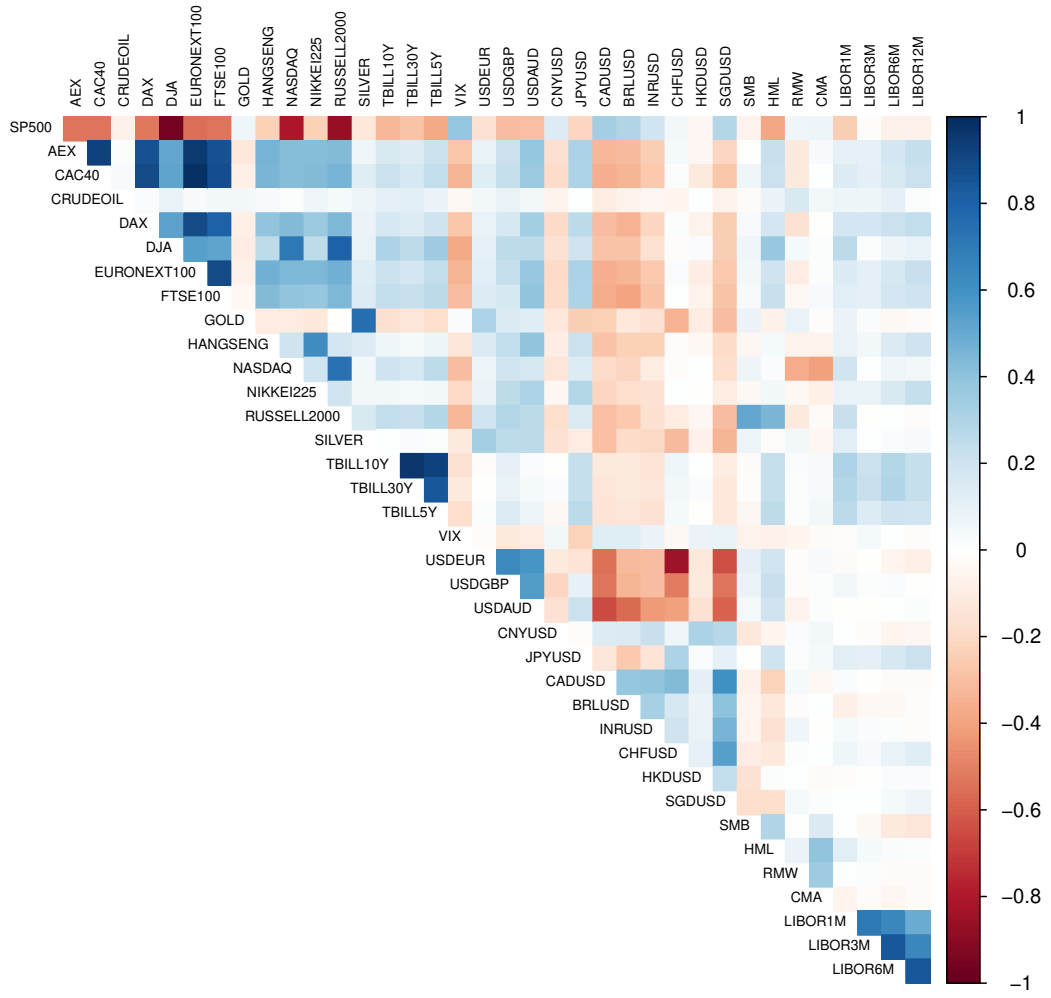

Figure 10: Correlation matrix of the covariates in the form of a heatmap. Positively correlated variables are marked in blue, uncorrelated in white and negatively correlated in red.

Figure 11 plots the S&P 500 excess losses against a selection of covariates. There is a clear negative relation between the S&P 500 excess losses and the percentage changes in the Dow Jones, NASDAQ and RUSSELL 2000 indices. If their price is decreasing, S&P 500 losses increase, and this effect amplifies as the change in price is higher. The percentage change in the VIX is mostly positive for the excess S&P 500 losses, meaning that if the volatility increases in the market, the index has a tendency to suffer losses. Furthermore, the more volatility there is in the market the greater these losses are. Looking at the Fama-French factors, when the returns of the SMB and HML factors are largely positive or negative, the S&P 500 index experiences the largest losses. This effect is not as obvious with the LIBOR rates. Finally, crude oil seems to have no relation to the excess losses.

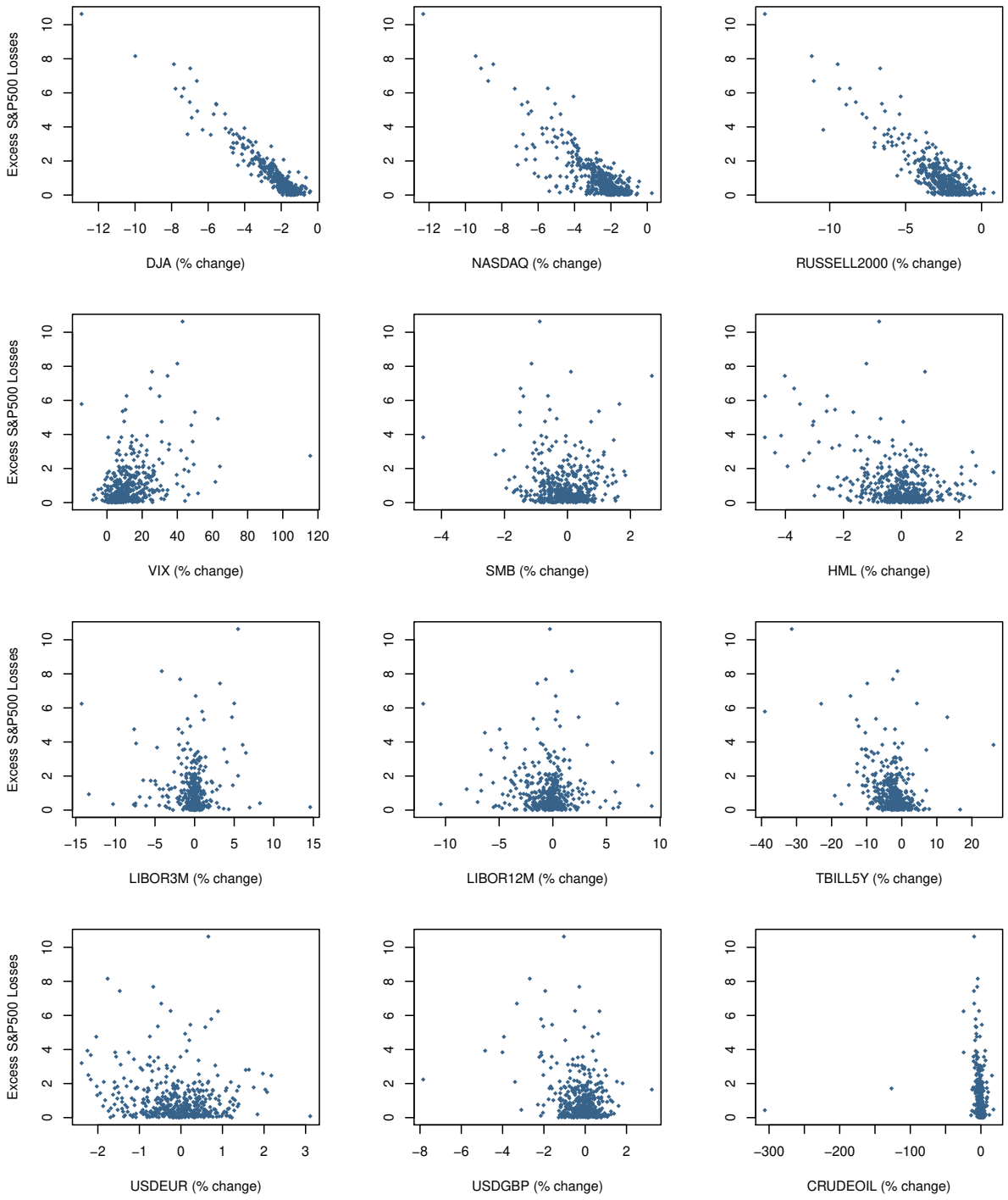

Figure 11: Scatterplot of S&P 500 extreme losses against covariates

5.2 Choosing the optimal stopping criterion

Before estimating the model, we need to choose the optimal hyper-parameters for the algorithm. Fine-tuning the hyper-parameters is of crucial importance for the model to provide adequate and reliable results. In Section 3.3.1, we explain how the stopping criterion for the boosting algorithm determines the amount of shrinkage and the number of selected covariates. For large datasets,

early stopping is necessary to prevent overfitting. Hence, to determine the optimal number of stopping iterations m_{stop}, we conduct a 10-fold cross validation. As there are two distribution parameters, we use a two-dimensional grid for different values of the stopping criterion for each parameter. Then, we select the optimal stopping criterion which minimizes the predictive empirical risk.

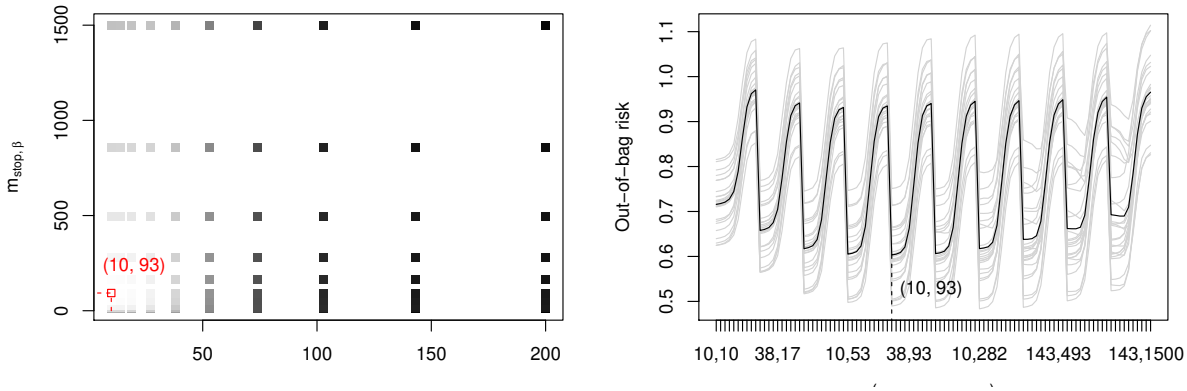

Figure 12: 10-fold cross validation. Heatmap (right) and out-of-bag risk plot (left). The optimal number of boosting iterations is $m_{stop} = (m_{stop,\xi}, m_{stop,\beta}) = (10, 93)$

Figure 12 gives the results of the 10-fold cross validation. Once again the results confirm that the shape parameter cannot be estimated accurately, thus a larger stopping criterion for ξ does not improve the performance of the model. Hence, we focus on the stopping criterion for the scale parameter β and use the minimum number of iterations for ξ. The predictive empirical risk is minimized at $m_{stop,\beta} = 93$ boosting iterations.

5.3 Model sensitivity with respect to the threshold

The POT approach, explained in detail in Section 3.1, assumes that the excesses over a threshold approximately follow a generalized Pareto distribution with shape parameter ξ and scale parameter β. In general, an initial choice of a threshold such that 10% of the data is included as exceedances is recommended. The choice of the threshold is important as it implies a trade-off between bias and variance. Hence, to make an informed decision, we look into different values for the threshold u. We test the goodness-of-fit of the generalized Pareto distribution using the fitted values for the parameters when we include exceedances above the 80-th, 85-th and 90-th percentile of the S&P 500 losses.

For a GPD setting where the data is assumed to be identically and independently distributed, meaning the distribution parameters are constant, Choulakian and Stephens (2001) use Q-Q plots to evaluate the goodness-of-fit. Bee et al. (2019) make modifications to the former approach to adapt the methodology to a dynamic setting with varying GPD paramaters. Given a threshold u and the observations $X_{t_1}, \ldots, X_{t_N}$, the fitted exceedances are $Y_i = X_{t_i} - u \sim GPD(\hat{\xi}_i, \hat{\beta}_i)$, where $\hat{\xi}_i$ and $\hat{\beta}_i$ denote the estimated shape and scale parameters for $i = 1, \ldots, N$. To get exponentially distributed variables from GPD ones, the exceedances Y_i are transformed to

$$\tilde{Y}_i = \frac{1}{\hat{\xi}_i} \log\left(1 + \hat{\xi}_i \frac{Y_i}{\hat{\beta}_i}\right).$$

Let the sorted exponentially distributed variables $\tilde{Y}_i$ be denoted as $\tilde{y}_{(1)}, \ldots, \tilde{y}_{(N)}$ Then, the Q-Q plot is obtained by plotting the pairs $\left\{\left(\tilde{y}_{(i)}, -\log(1 - i/(N+1))\right), i = 1, \ldots N\right\}$. Figure 13 gives the Q-Q plots for fitted S&P 500 excess losses using the 80-th, 85-th and 90-th percentile as a threshold. Overall, the fitted values are very close to the theoretical ones, confirming that our model estimates are quite adequate. Comparing between the different thresholds, we observe that the more observations are included as exceedances, the more accurate the estimates are after fitting the GAMLSS. However, the more observations, the higher the bias. To statistically

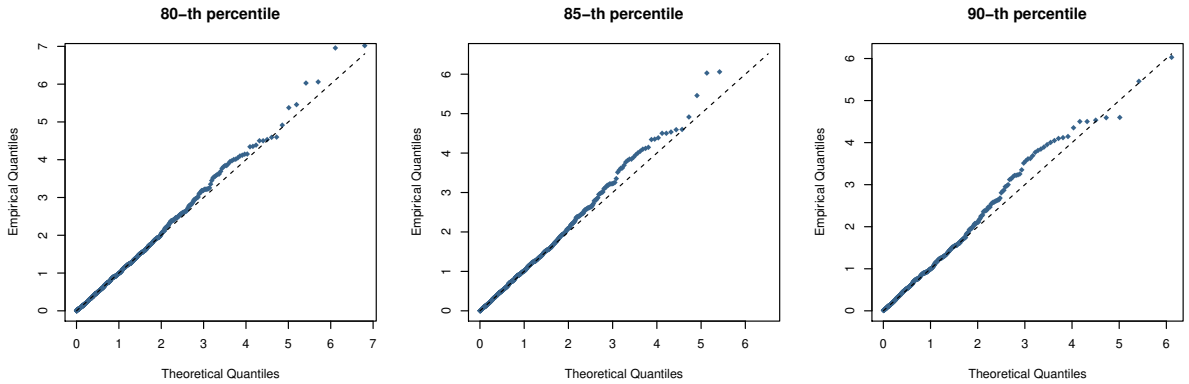

Figure 13: Q-Q plots using the 80-th, 85-th and 90-th percentile as a threshold for the S&P 500 excess losses. The 45-degree line (- - - -) plots the theoretical quantile of a unit-rate exponential distribution. The empirical quantiles of the excess losses transformed to exponential are plotted on the y-axis.

test whether the empirical distribution (fitted) is significantly different from the theoretical one (observed), we perform the Kolmogorov-Smirnov (KS) test. Furthermore, a comparison between the KS statistics evaluates the goodness-of-fit for the three cases. The KS test statistics and probability values for the 80-th, 85-th and 90-th percentile threshold are found in Table 7. The KS statistic is at its lowest using the 80-th percentile as a threshold, confirming that the model fits best the data at this threshold level.

Table 7: Kolmogorov-Smirnov (KS) test for different thresholds

Threshold	KS test statistic	p-value
80th-percentile	0.018	(0.998)
85th-percentile	0.022	(0.996)
90th-percentile	0.031	(0.981)

5.4 Covariate selection and estimated effects on the GPD scale parameter

Using the 80-th percentile of the S&P 500 losses as threshold, $(m_{stop,\xi}, m_{stop,\beta}) = (10, 93)$ stopping iterations for our algorithm, as well as smooth P-spline base-learners, 20 knots, a second-order difference penalty and 4 degrees of freedom as control parameters, we fit GAMLSS on the scale parameter of the tail distribution of the excess losses. The variable selection algorithm distinguishes and selects the following covariates as informative: Dow Jones, NASDAQ, Russell 2000, the Fama-French factors HML and SMB, the USD/EUR and CNY/USD exhange rates,

gold and the volatility index VIX.

It is of crucial importance to perform an in-depth analysis of which variables are selected by the algorithm as well as their estimated effects. Our simulation study shows that the algorithm is able to perform variable selection and model estimation quite adequately given certain assumptions. Its performance is somewhat challenged when high-dimensionality or correlation is introduced in the model. Therefore, the validity of the results should always be tested and covariates effects which are not in line with economic theory or intuition need further inspection and may also be ignored if no evidence can be found in their support.

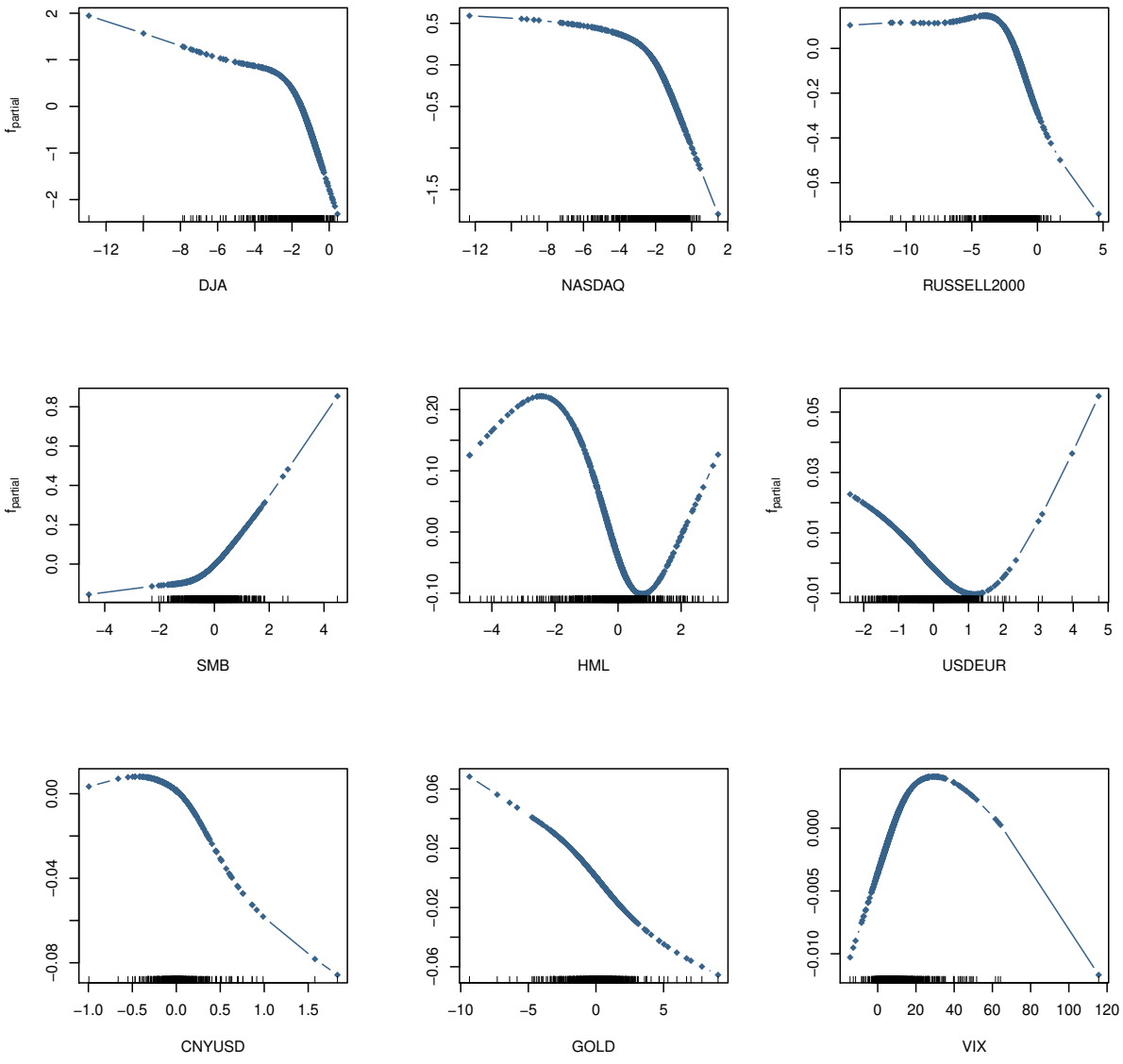

Figure 14: Partial covariate effects on the scale parameter. The covariates (in percentage change units) are plotted against the additive predictor $\eta(\beta)$, hence are in a logarithmic scale. Only the covariates selected by the estimated model are included.

Figure 14 plots the estimated partial covariate effects on the additive predictor $\eta(\beta)$. As β is the scale parameter of a GP distribution, a higher β value would result in a higher tail risk. This means that there is a higher chance to have extreme values, or that the magnitude of the

potential extreme losses would be higher. Note that, the partial effect ignores all other model parameters, and thus we cannot make an interpretation on the size but only on the functional form of the covariate effect. Recall once again that covariates are expressed in percentage change or return terms.

The selected covariates can be categorized into three groups according to the magnitude of their effect. First, the U.S. stock indices, Dow Jones, NASDAQ and Russell 2000, are the most influential covariates as they have the largest effect on the predictor. In practice, these indices are indicative of the overall market performance and behave quite similarly. Hence, it comes as no surprise that the partial effect for all three is also very similar. The scale parameter for S&P 500 extreme losses is the highest when the other indices perform badly. This negative relation is more prominent when the index returns fall from 0% to -4%, then the marginal effect becomes smaller as the values are more extreme, resulting in a cliff-shaped effect.

Second, other important covariates selected by the algorithm that are slightly less influential when looking at the magnitude are the Fama-French SMB and HML factors and the USD/EUR and CNY/USD exchange rates. The SMB factor has a positive effect on the scale parameter, meaning that the chance of having more extreme losses becomes higher when smaller firms outperform larger ones. This is to be expected as the S&P 500 stock index is composed of the stocks of 500 large-cap publicly traded firms, and a large return on the SMB factor could be due to the poor performance of firms with a high market capitalization. There seems to be no effect when the SMB becomes negative, which is also the case when Russell 2000 performs poorly. Russell 2000 follows the stocks of 2000 small-cap U.S. companies which might explain the similar behavior with the SMB factor. For the HML factor, the effect is negative when the difference between high and low book-to-market ratio firms is small and positive when this difference is larger (in absolute terms). Decreasing HML returns translates into value firms underperforming relative to growth firms. Value companies are usually large-cap and are well-established firms, which can be an explanation as to why the risk of extreme S&P 500 losses is larger when HML decreases. On the other hand, among 11 exchange rates, only two were selected as explanatory variables, the USD/EUR and CNY/USD rates. When the USD/EUR exchange rate decreases, meaning the U.S. Dollar depreciates relative to the Euro, there are more extreme losses in the index. There seems to be the same effect when the the USD/EUR increases, however there are too few data points to draw a conclusion in this case. The CNY/USD exchange rate has a negative effect on the scale parameter of the S&P excess loss distribution, although the rate itself does not fluctuate a lot (most extreme observations are within a 0.5% absolute change in the exchange rate). Once again, as the U.S. dollar depreciates, the risk for larger S&P 500 losses is higher.

Finally, the volatility index and gold have very little impact on the scale parameter as compared to the other covariates. The corresponding graph shows that the scale parameter increases as VIX returns reach 30% and then slightly decreases for higher values. However, the magnitude of the effect that VIX has on the scale parameter is smaller than other covariates. This could be evidence that either the volatility index is not as prominent of an explanatory variable or the model is not able to adequately capture its effect and hence it may be ignored. Gold appears to have a negative relationship with the S&P 500 tail risk parameter. This is to be

expected as gold and stock returns are usually negatively correlated to each other. Furthermore, we have seen the gold price rise through some of the biggest stock market crashes in history, making it an appealing hedging instrument against turmoil periods.

5.5 Performance evaluation

In this section, we evaluate the performance of the model. After fitting the model to our data and selecting the informative covariates, we obtain estimates for the distribution parameters which are dependent on the covariate effects. Figure 15 plots the fitted values for the scale and the shape parameter. On average, the scale and the shape parameter have a mean of 1.039 and 0.133, respectively. During turmoil periods with a higher probability of having extreme losses, the scale parameter is the highest while the shape parameter at its lowest. Note that for the shape parameter, the estimates may not as reliable as for the scale parameter.

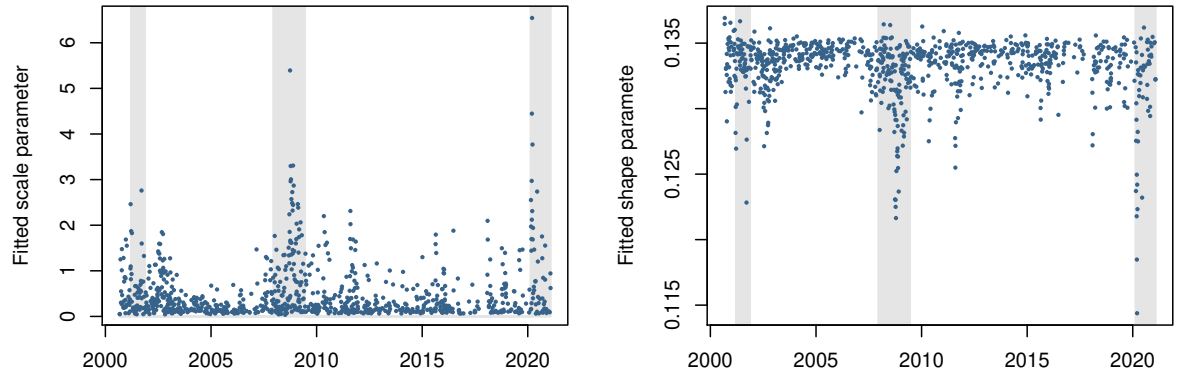

Figure 15: Fitted values for the scale and the shape parameters given the estimated covariate effects. The grey shaded regions mark the NBER-based recession indicators.

One way to assess the performance of the model, is to compute the 95% confidence intervals for the S&P 500 excess losses. Given the covariates, the fitted values of the response variable can be computed from the conditional distribution, $\hat{Y}_i \sim GPD(\hat{\xi}_i, \hat{\beta}_i)$ for $i = 1, \ldots, 899$. We repeat this step 100 times to obtain the 95% confidence intervals which are computed from the quantiles of the estimated conditional distribution, as follows

$$\text{CI}_{0.95}(\hat{Y}_i) = \left[Q_{0.025}(\hat{Y}_i),\ Q_{0.975}(\hat{Y}_i)\right]$$

where Q_α is the $100\alpha\%$ quantile. Figure 16 plots the S&P 500 excess loss data and the 95% confidence intervals derived from the fitted conditional distribution of excess losses given the covariates. The coverage ratio, computed as the probability that the observed S&P 500 excess losses are included in the 95% confidence interval, is 94.89%. In both plots, it is clear that the model adequately captures the heteroskedasticity in the loss data. For greater excess losses, the confidence intervals become wider as a result of a higher variability in the data that is captured by introducing the covariates.

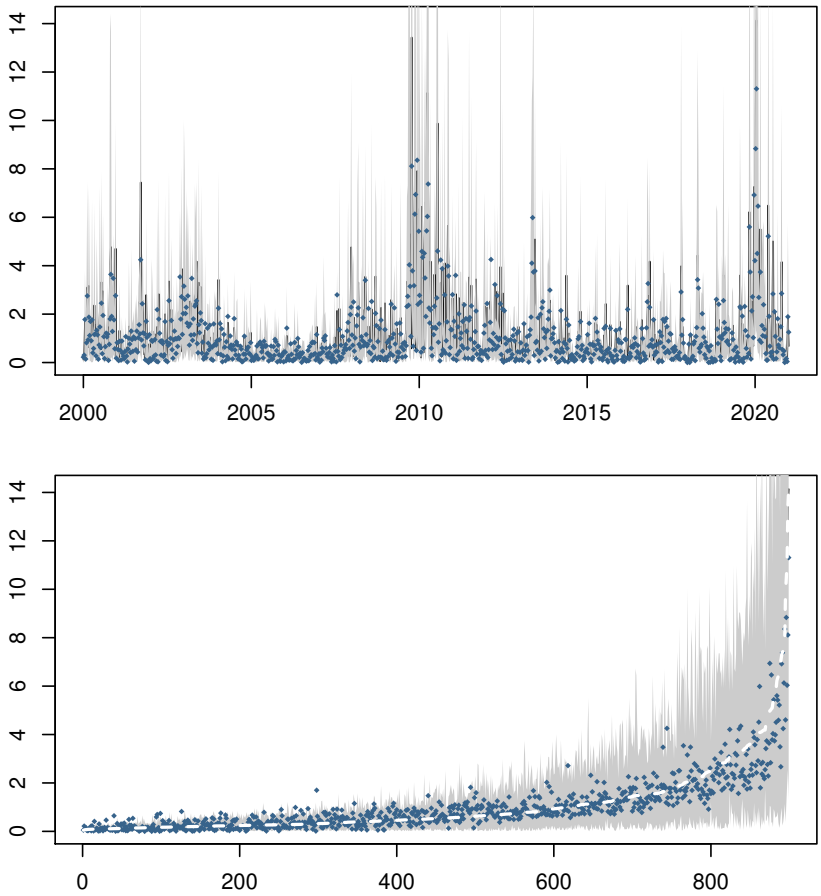

Figure 16: 95% confidence intervals for S&P 500 excess losses based on the estimated conditional GPD parameters from GAMLSS. • S&P 500 observed excess losses, ■ confidence intervals from the quantiles of the estimated conditional distribution and (white, dashed line) the conditional mean of the fitted S&P 500 excess losses. The data is sorted in the bottom figure.

In the dynamic setting where the effect of covariates is taken into account, the risk measure VaR cannot be computed as in Equation 3.1 as that would correspond to the unconditional VaR. Given the estimated distribution parameters, the conditional VaR is calculated as

$$\widehat{\mathrm{VaR}}_\alpha = u + \frac{\hat{\beta}}{\hat{\xi}} \left[\left(\frac{1-\alpha}{\hat{p}} \right)^{-\hat{\xi}} - 1 \right].$$

where $\hat{p} = \hat{\mathrm{P}}\left[X > u \mid (x_1, \ldots, x_p)\right]$ is an estimator of the conditional probability that the S&P 500 losses are above the threshold, given the set of covariates. The conditional probability differs from the unconditional one which we previously set to be 20%. To find an estimator for p, this problem can be approached as a classification problem where the response variable takes a value of 1 if the corresponding loss exceeds the threshold and 0 otherwise. The conditional probability can then be modelled using a logistic regression or Support Vector Machine (SVM).

For both classification methods, the full dataset is scaled and split into training (75%) and test data (15%) used for fitting and prediction, respectively. Furthermore, a linear relationship between the binary response variable and selected covariates is assumed. The included covariates

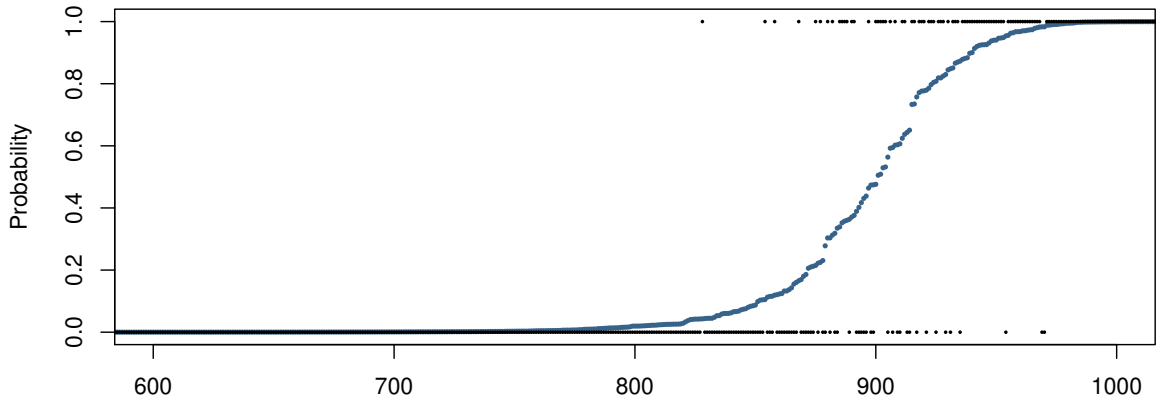

Figure 17: Logistic regression. • Conditional probability that S&P 500 losses exceed the threshold, • observed test data. The first 600 observations are ignored in the plot for better visualization of the results.

are the Dow Jones, NASDAQ, Russell 2000, SMB, HML and VIX which were chosen as influential explanatory variables by our algorithm. We exclude the exchange rates and gold which are less significant. Using a logistic regression model, the estimated value of the conditional probability is 20.24%. Figure 17 plots the predicted conditional probability together with the actual values of the binary response variable for the test sample. For the SVM with a linear kernel, the confusion matrix is given in Table 8. The conditional probability, in this case, is computed as the number of predicted exceedances (regardless whether correctly identified) over the number of observations in the test sample, hence $\hat{p} = (16 + 208)/1124 = 20.02\%$. From the confusion matrix, we can also derive the prediction accuracy of the SVM, which is 97.4%.

Table 8: Confusion matrix for linear SVM

		Prediction		
		0	**1**	**Total**
Actual	**0**	883	16	899
	1	17	208	225
	Total	900	224	**1124**

Moreover, we consider incorporating forward step-wise selection in the logistic regression model which we can compare with our initial model. Forward step-wise selection using the Akaike information criterion (AIC) adds one covariate at each step and stops when the algorithm no longer improves by adding another variable. Starting from a model without any covariates to potentially including all covariates (informative and non-informative), the final step-wise

regression proposes including the following explanatory variables: Dow Jones, NASDAQ, Russell 2000, SMB, VIX, DAX, Nikkei 225, Libor 12M and HML. The four latter covariates are found to be the least informative with a significance level of more than 10%. The conditional probability is 20.13% in this case. Overall, this is in line with the selection procedure performed by our model, as the step-wise regression selects a very similar set of covariates.

Using the estimated conditional probability from the logistic regression using our proposed set of covariates, we compute the $\text{VaR}_{0.95}$ of S&P 500 losses. We want to look at the percentage of violations, that is, the percentage of observed data that exceeds the conditional $\text{VaR}_{0.95}$. This number should be close to 95%. If the observed number of exceedances is much higher than the expected number, the model underestimates the tail risk. Otherwise, in case the observed number of violations is smaller, tail risk is overestimated. Figure 18 plots the conditional $\text{VaR}_{0.95}$ against the actual S&P 500 loss data. The percentage of violations is found to be 97.3%. To test whether this difference is significantly large, we use a binomial test which is a standard back-testing method for VaR evaluation. The test statistic does not indicate that there is a major difference. However, this is not the case when we evaluate VaR at higher probability levels such as 97.5% and 99%. This means that the model has a tendency to overestimate the risk level of the S&P 500 losses by estimating values of VaR that are higher than the actual ones, especially at higher probability levels.

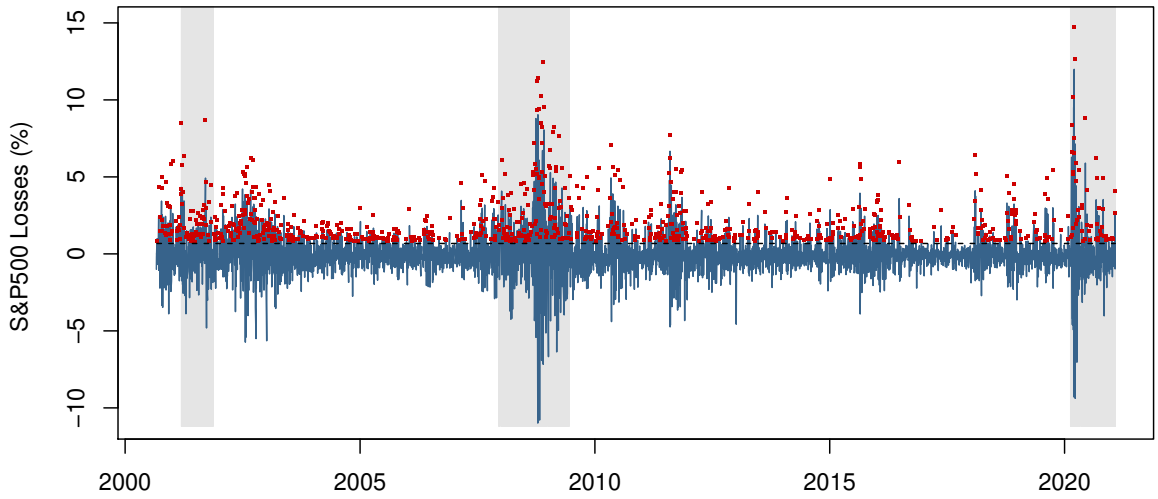

Figure 18: Plot of the the actual S&P 500 losses (blue, solid line), • conditional $\text{VaR}_{0.95}$ given covariates and the threshold (black, dashed line).

To test the goodness-of-fit of the model in the application, we use Q-Q plots and the KS test to compare our model with two simplified versions of it. In contrast to our model which has varying distribution parameters, the second model assumes the shape parameters is constant while the scale parameter varies, and the third one assumes both are constant. From Figure 19, we can observe that the first two models are not that different from each other and perform quite well. The assumption of constant distribution parameters, however, is quite strict and fails to capture the variation in the S&P 500 excess losses. These results evidence the existing relation between the distribution parameters and the covariates, especially for the scale parameter.

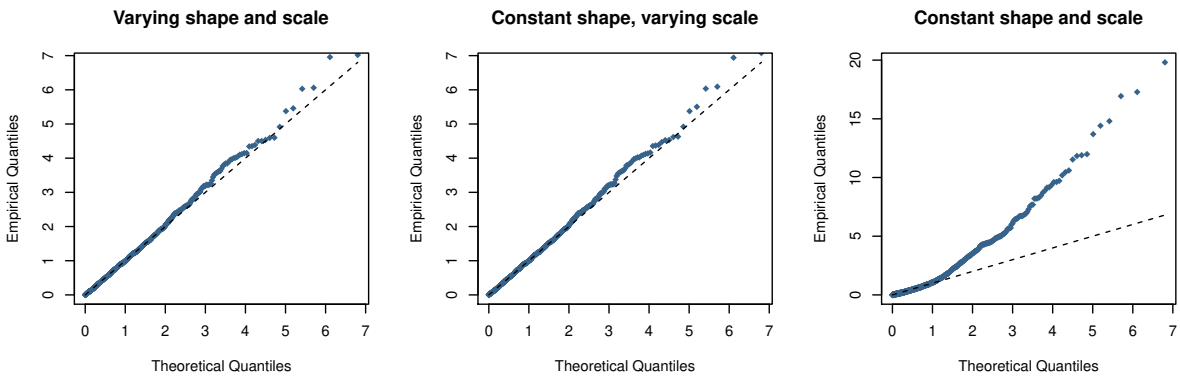

Figure 19: Q-Q plots for three different models: both shape and scale parameters are varying (left), constant shape but varying scale (middle) and constant shape and scale (right). The 45-degree line (- - - -) plots the observed S&P 500 excess losses of a unit-rate exponential distribution. The fitted values transformed to exponential are plotted on the y-axis.

Table 9: Kolmogorov-Smirnov (KS) test for different models

Model	KS test statistic	p-value
Varying ξ and β	0.018	(0.998)
Constant ξ and varying β	0.017	(0.999)
Constant ξ and β	0.121	(0.000)

In Table 9, the KS statistics for our dynamic model with varying parameters, the model with only a constant shape parameter and one where both distribution parameters are constant are given. Again, only in the latter case, where the scale and shape paramaters are constant, the KS statistics rejects the hypothesis under which the distributions of the fitted and observed values for the S&P 500 excess losses are the same.

6 Conclusion

This paper combines together extreme value theory, the GAMLSS framework and machine learning techniques to perform data-driven variable selection in tail risk modelling. In our simulation study, the performance of the model was satisfactory in several fronts. First, the selection rates for the informative variables are quite high, while the probability of selecting non-informative covariates remains quite low. This remains true under high-dimensional settings, where the number of covariates exceeds the number of observations. In the case of highly-correlated covariates, the selection rates for informative explanatory variables decrease, however this does not mean the model is performing worse. Covariates that are highly-correlated have similar effects on the distribution parameters, hence non-informative covariates are also informative in a way. Secondly, the model is able to capture the covariate effects for the scale parameter quite adequately. The non-parametric functional form of the GAMLSS models gives a lot of flexibility. One drawback of this model is that it is not able to capture the behavior of the shape parameter. The estimation of the shape parameter remains cumbersome because of the sensitive nature of this parameter and how susceptible it is to outliers.

We test our methodology on S&P 500 extreme losses and a list of 37 covariates. The U.S. market indices Dow Jones, NASDAQ, Russell 2000, the SMB and HML Fama-French factors, VIX, USD/EUR, CNY/EUR and gold are selected as the informative covariates. Using 95% confidence intervals, we show that our model is able to capture quite well the heteroskedasticity in loss data due to the inclusion of explanatory variables. Using classification models, such as the logistic regression and SVM, we compute the conditional VaR at a 95% probability level and show that the model overall captures tail risk quite accurately with a slight tendency for overestimating risk. Finally, using a KS test we show that our covariate-dependent model is better able to capture the distribution of the extreme variables than a model assuming constant scale and shape parameters. Introducing covariates, and hence variation in the distribution parameters, improves the model estimation as well as variable selection.

There are some limitations to our methodology which can be explored in future research. The estimation of the shape parameter in the same context of variable selection using GAMLSS can be a potential continuation of our research. Furthermore, EVT faces multiple challenges when dealing with scarce data. Training the algorithm using larger datasets as well as comparing its predictive performance with other benchmark models will give more insights on the model performance.

A Simulation Codes

The necessary R packages and codes used for the simulation study in Section 4 are described in detail here.

`gamboostLSS-package`	This package includes multiple functions that use gradient boosting techniques to fit GAMLSS models. It is publicly available and created by Mayr et al. (2012).
`GPDdist.R`	This code is a function that specifies the CDF, PDF and quantiles of the GP distribution. It is used to include GPD in the family of distributions that the `gamboostLSS` package is able to recognize.
`simStudy.R`	This code is a function which uses as inputs the desired number of observations, number of covariates, number of simulations and number of boosting iterations, conducts simulations using functions from the package `gamboostLSS` and returns a list of results per simulation, such as the fitted values for the GPD distribution parameters, selected covariates, empirical risk, etc.
`resultsSimStudy.R`	This code is used to plot the results retrieved from calling the function `simStudy.R`. It creates plots and histograms for the fitted distribution parameters as well as plots for the marginal effects for the covariates.

B Application Codes

The necessary R packages and codes used for the real-data application in Section 5 are described in detail here.

`mainSP500.R`	This code is used for the application part of this paper and is composed of several sections. First, it reads the S&P 500 and covariate data and performs the gradient descent algorithm to GAMLSS models for the GPD distribution variables of the excess losses. Additionally, it performs 10-fold cross-validation to find the optimal number of iterations, it computes the descriptive statistics, the correlation between covariates, and makes several plots of the results. Finally, it performs logistic regression and SVM to compute the conditional probability and VaR.

References

Bee, M., Dupuis, D. J., and Trapin, L. (2019). Realized Peaks over Threshold: A Time-Varying Extreme Value Approach with High-Frequency-Based Measures*. *Journal of Financial Econometrics*, Vol. 17(2): 254–283.

Bühlmann, P. and Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, Vol. 22(4): 477–505.

Chavez-Demoulin, V. and Davison, A. C. (2005). Generalized additive modelling of sample extremes. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, Vol. 54(1): 207–222.

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

Choulakian, V. and Stephens, M. (2001). Goodness-of-fit tests for the generalized pareto distribution. *Technometrics*, Vol. 43: 478–484.

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

Davison, A. C. and Smith, R. L. (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 52(3): 393–442.

Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*. Springer-Verlag, Berlin.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, Vol. 116(1): 1–22.

Flannery, M. J. and James, C. M. (1984). The effect of interest rate changes on the common stock returns of financial institutions. *The Journal of Finance*, Vol. 39(4): 1141–1153.

Fogler, H. R., John, R., and Tipton, J. (1981). Three factors, interest rate differentials and stock groups. *Journal of Finance*, Vol. 36(2): 323–335.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, Vol. 29(5): 1189–1232.

Giot, P. (2005). Relationships between implied volatility indexes and stock index returns. *The Journal of Portfolio Management*, Vol. 31(3): 92–100.

Greven, S. and Kneib, T. (2010). On the behaviour of marginal and conditional AIC in linear mixed models. *Biometrika*, Vol. 97(4): 773–789.

Gumbel, E. J. (1958). *Statistics of Extremes*. Columbia University Press, New York.

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

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman and Hall, London.

Hilliard, J. E. (1979). The relationship between equity indices on world exchanges. *The Journal of Finance*, Vol. 34(1): 103–114.

Hofner, B., Mayr, A., Fenske, N., and Schmid, M. (2018). ***gamboostLSS**: Boosting Methods for GAMLSS Models*. R package version 2.0-1.1.

Kilian, L. and Park, C. (2009). The impact of oil price shocks on the U.S. stock market. *International Economic Review*, Vol. 50(4): 1267–1287.

Mayr, A., Fenske, N., Hofner, B., Kneib, T., and Schmid, M. (2012). Generalized additive models for location, scale and shape for high dimensional data – a flexible approach based on boosting. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, Vol. 61(3): 403–427.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized Linear models. *Journal of the Royal Statistical Society. Series A (General)*, Vol. 135(3): 370–384.

Nieh, C. C. and Lee, C. F. (2001). Dynamic relationship between stock prices and exchange rates for G-7 countries. *The Quarterly Review of Economics and Finance*, Vol. 41(4): 477–490.

Park, M. H. and Kim, J. H. T. (2016). Estimating extreme tail risk measures with generalized Pareto distribution. *Computational Statistics & Data Analysis*, Vol. 98: 91–104.

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

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized Additive Models for Location, Scale and Shape. *Applied Statistics*, Vol. 54(3): 507–554.

Ripley, B. D. (2004). Selecting amongst large classes of models. In *Methods and Models in Statistics*, pages 155–170. Imperial College Press, London.

Stasinopoulos, D. M. and Rigby, R. A. (2007). Generalized Additive Models for Location Scale and Shape (GAMLSS) in R. *Journal of Statistical Software*, Vol. 23(7).