Erasmus University Rotterdam

Erasmus School of Economics

Master Thesis Quantitative Finance

Implementing Covariate Selection Techniques in a Dynamic Peak-Over-Threshold Approach

D.R.A. Lemmen

445526dl

Abstract

In this paper, we extend the existing dynamic Peak-Over-Threshold methodology with the following covariate selection techniques: best subset selection (BeSS), Lasso, and Relaxed Lasso. We then test all three covariate selection techniques using a simulation study. We find that all three methods work approximately equally well for the loss frequency. For loss severity, we find that Relaxed Lasso performs best in selecting covariates, while BeSS tends to select too few covariates and Lasso too many. Finally, we apply our methodology on a market risk dataset. We find that the S&P Volatility Index is selected the most often by all three covariate selection techniques. The S&P Volatility Index has a positive relation with both frequency of extreme losses and the expected loss in case of such an extreme event.

Supervisor:

prof.dr. C. Zhou

Second assessor:

M.F.O. Welz

July 6, 2021

Contents

1 Introduction

Modeling the distribution of extreme losses can provide quantitative insights into risks, which are often required by regulators. Moreover, it might also lead to insights into the drivers behind the risk exposure. These insights can be very useful for reducing risk exposure.

The standard methods for modeling the distribution of extreme losses include the Block Maxima approach and the Peak-over-Threshold (POT) approach (Gumbel, 1958; Pickands, 1975). In the Block Maxima approach, the maximum in each time period is modelled using a Generalized Extreme Value (GEV) distribution. The POT approach models the frequency and severity of exceedances of a (high) threshold. The frequency and severity are modeled as two separate and independent distributions. This gives the POT approach the advantage of being more flexible in modeling and the ability to incorporate more data. In this 'classic' POT approach, the distributions of both the frequency and severity are assumed to be stationary. However, Jagannathan and Wang (1996) observe non-stationarity for risk exposure of stocks. Therefore, Chavez-Demoulin, Embrechts, and Hofert (2016) propose a dynamic POT approach that lets the distribution parameters depend on covariates.

As the availability of data has grown rapidly in recent years, many potential covariates have become available. However, this leads to a practical problem when implementing the dynamic POT approach, including all covariates in the current methodology could lead to several issues. Coefficients in the model could become unidentified, or the model could be overfitted to the data. Moreover, the current methodology does not select a subset of covariates that significantly drive the risk exposure.

This paper focuses on applying covariate selection in the dynamic POT approach. There exist different methods for covariate selection. Hastie, Tibshirani, and Tibshirani (2020) provide an extensive comparison of best subset selection and Lasso regularization (L_1) and conclude that neither of them dominates the other in all cases. However, Relaxed Lasso, a variation on Lasso, is found to perform best. Therefore, we extend the dynamic POT approach with these three covariate selection techniques. We test the methodology extensively using a simulation study, where we test performance on both providing the best distribution parameter estimate and the ability to select the correct covariates. Finally, we apply the proposed methodology on a market risk dataset in combination with a large number of potential covariates.

The structure of this paper is as follows. In Section 2, we discuss the current literature and methods that are relevant to our research. Continuing, in Section 3, we describe the proposed methodology. Thereafter, in Section 4, we discuss the setup and results of the simulation study. Next, in Section 5, we elaborate on the application to a market risk dataset and present the results. In Section 6, we present our findings and draw conclusions about our results. Lastly, in Section 7, we discuss the limitations of our research and provide suggestions for further research.

2 Literature Review

In this section, the current literature is discussed in detail. First, we discuss the 'classic' POT approach. Secondly, we discuss an extension of this approach, the dynamic POT approach. Lastly, we discuss the three covariate selection techniques that are considered.

2.1 Peak-Over-Threshold (POT) approach

The POT approach is used for modeling losses above a (high) threshold, u. In this method, two distributions are fitted: a distribution that models the number of exceedances of the threshold, N_t, and a distribution that models the severity of such an exceedance, L_i. Where t denotes the time period and i denotes the index of the exceedance.

Embrechts, Klüppelberg, and Mikosch (1997) suggest the following model to approximate the frequency and severity of extreme losses:

- The frequency of a high threshold exceedance approximately follows a Poisson process, i.e. $N_t \sim \text{Pois}(\lambda)$ with the rate parameter $\lambda > 0$.

- The severity of a loss above a high threshold approximately follows a generalized Pareto distribution (GPD) independent of N_t, i.e. $L_i - u = Y_i \sim \text{GPD}(\xi, \sigma)$ with the shape parameter $\xi \in \mathbb{R}$ and scale parameter $\sigma > 0$. The pdf and cdf of the GPD are defined as follows,

$$g_{\xi,\sigma}(y) = \begin{cases} \frac{1}{\sigma}\left(1 + y\frac{\xi}{\sigma}\right)^{-\frac{\xi+1}{\xi}}, & \xi \neq 0, \\ \frac{1}{\sigma}\exp\left(-\frac{y}{\sigma}\right), & \xi = 0, \end{cases} \tag{1}$$

$$G_{\xi,\sigma}(y) = \begin{cases} 1 - \left(1 + y\frac{\xi}{\sigma}\right)^{-1/\xi}, & \xi \neq 0, \\ 1 - \exp(-y/\sigma), & \xi = 0. \end{cases} \tag{2}$$

When $\xi < 0$, y is bounded between 0 and $-\sigma/\xi$. For $\xi = 0$, the GPD corresponds to the Exponential distribution. Most often, ξ is therefore restricted in the following way: $\xi > 0$. This restriction corresponds to the heavy-tail case. The GPD has infinite variance when $\xi \geq 0.5$, and no finite moment when $\xi \geq 1$ (Moscadelli, 2011).

The likelihood function of the POT approach takes the following form,

$$L(\lambda, \xi, \sigma; \mathbf{Y}) = \frac{\lambda^n}{n!} \exp(-\lambda) \prod_{i=1}^{n} g_{\xi,\sigma}(Y_i), \tag{3}$$

where $\mathbf{Y} = (Y_1, ..., Y_n)$, and $g_{\xi,\sigma}(\cdot)$ corresponds to the pdf of the GPD.

As the frequency and severity are asymptotically independent, the log likelihood of equation (3) can be split into two parts:

$$\ell(\lambda, \xi, \beta; \mathbf{Y}) = \ell(\lambda; \mathbf{Y}) + \ell(\xi, \sigma; \mathbf{Y}), \tag{4}$$

where

$$\ell(\lambda; \mathbf{Y}) = n \ln(\lambda) - \ln(n!) - \lambda, \tag{5}$$

$$\ell(\xi, \sigma; \mathbf{Y}) = -n \ln(\sigma) - (1 + 1/\xi) \sum_{i=1}^{n} \ln\left(1 + \xi \frac{Y_i}{\sigma}\right). \tag{6}$$

The split of the log-likelihood in equation (4) separates the frequency and severity part of the log-likelihood. This implies that the estimation of the frequency and severity distributions can be done separately.

2.2 Dynamic POT approach

The ‘classic’ POT approach assumes that the distribution parameters are equal for all observations and thus do not depend on any covariates or time. In practice, these stationarity assumptions are often violated. The loss distributions may depend on several covariates. For example, loss distributions might change with the business line or economic conditions.

Therefore, Chavez-Demoulin et al. (2016) extend the ‘classic’ POT approach into a dynamic POT approach which allows its parameters to depend on covariates. In this dynamic POT approach, λ is replaced by

$$\lambda(x, t) = \exp(f_\lambda(x) + h_\lambda(t)), \tag{7}$$

where $f_\lambda(\cdot)$ maps the covariates to correspondingly many constants, and $h_\lambda(t)$ is a non-parametric smoothing function regarding time t.

As the parameters of the Poisson distribution and GPD can be estimated separately, the estimation of the Poisson distribution becomes a standard generalized additive model (GAM). Wood (2017) describes an algorithm to estimate such a model.

In the dynamic POT approach, the severity of a loss becomes dependent on covariates as well. Letting ξ and σ directly depend on the covariates may lead to statistical identification issues during estimation (Chavez-Demoulin & Embrechts, 2004). By replacing ξ or σ with so-called orthogonal parameters solves this potential issue. Two reparameterizations are possible, however the reparameterization on σ is easier to compute and more stable. Cox and Reid (1987) provide the following reparameterization,

$$\nu = \ln((1+\xi)\sigma). \tag{8}$$

For the dependence on the covariates, the parameters ξ and ν are replaced by

$$\xi(x,t) = f_\xi(x) + h_\xi(t), \tag{9}$$

$$\nu(x,t) = f_\nu(x) + h_\nu(t), \tag{10}$$

where $f_p(\cdot)$ maps the covariates to correspondingly many constants, and $h_p(t)$ is a non-parametric smoothing function regarding time t with $p \in \{\xi, \nu\}$.

Chavez-Demoulin et al. (2016) propose an iterative procedure to estimate ξ and ν simultaneously. After estimation of the model, Chavez-Demoulin et al. (2016) perform a graphical goodness-of-fit test. If $y_{(t,i)}$ (approximately independently) follow GPD(ξ_t, σ_t), then $R_{(t,i)} = 1 - G_{\xi_t,\sigma_t}(y_{(t,i)})$ approximately forms a random sample from a standard uniform distribution. Therefore, we can check using a Q-Q plot whether,

$$r_{(t,i)} = -\ln(1 - G_{\xi_t,\sigma_t}(y_{(t,i)})), \tag{11}$$

are distributed approximately as independent standard exponential variables. Where $y_{(t,i)}$ denotes the ith exceedance of the threshold in time period t, i.e. $y_{(t,i)} = l_{(t,i)} - u$, with

$l_{(t,i)}$ defined as the ith loss in time period t and u denoting the threshold.

2.3 Best subset selection

Best subset selection maximizes the likelihood function under the restriction that the number of non-zero coefficients is equal to or lower than a predetermined hypertuning parameter k (Hocking & Leslie, 1967), i.e.,

$$\max_{\beta} \ell(\beta) \quad \text{subject to} \quad ||\beta||_0 \leq k, \tag{12}$$

where $||\beta||_0$ is the L_0 norm of β, i.e. the number of nonzero elements in β.

A major drawback of this technique is the optimization difficulties that arise when the number of covariates grows. Bertsimas, King, and Mazumder (2016) propose a mixed-integer formulation for this optimization problem. Using this formulation, this difficulty can be overcome. However, this mixed-integer formulation is only valid for linear least squares regressions. Another optimization method, the primal-dual active set (PDAS) method, is proposed by Ito and Kunisch (2014). This method iteratively updates the active set (i.e. the selected covariates) through the use of primal and dual variables. The PDAS algorithm is generalized by Wen, Zhang, Wang, and Quan (2020) for general convex loss functions. The basic idea of this algorithm is to iteratively estimate the model with the active set, denoted as A_i in the ith iteration, and then update the active set to the covariates that have the most effect on the log-likelihood, measured by $\left(\Delta_i\right)_j$ in the ith iteration for the jth covariate, until the active set converges. We use this algorithm to estimate our models with best subset selection for a given k, which is given in Algorithm 1 below.

Algorithm 1: Primal-dual active set (PDAS) algorithm

```
$\mathcal{A}_0 = \{1,\ldots,k\}$ ;
$\mathcal{A}_1 = \emptyset$;
$i = 1$ ;
while $i < i_{max}$ and $\mathcal{A}_{i-1} \neq \mathcal{A}_i$ do
    $\beta_i = \arg\max_{\beta} \ell(\beta) \quad \text{subject to} \quad \beta_{\mathcal{A}^c_{i-1}} = \mathbf{0}$ ;
    $\left(\Delta_i\right)_j = \frac{1}{2}\left(h_i\right)_j \left(\left(\beta_i\right)_j - \frac{\left(g_i\right)_j}{\left(h_i\right)_j}\right)^2, \quad j = 1,\ldots,p$ ;
    $\mathcal{A}_i = \left\{j : \left(\Delta_i\right)_j \geq \left[\Delta_i\right]_{(k)}\right\}$;
    $i = i + 1$;
end
return $\beta_i$;
```

where g_i denotes the gradient of $\ell(\beta)$ evaluated at β_i, h_i the diagonal elements of the Hessian of $\ell(\beta)$ evaluated at β_i, $\left[\Delta_i\right]_{(k)}$ the kth order statistic of Δ_i and $i_{\max}$ the maximum number of iterations.

There are several options to determine the value for k. One option is Cross-Validation which optimizes the prediction performance. However, using Cross-Validation can be time-consuming especially with high-dimensional data. Wen et al. (2020) propose an alternative way of determining k, the sequential primal-dual active set (SPDAS) algorithm. This algorithm performs the best subset selection algorithm for an increasing k and chooses k such that some criteria is optimized. Suggested criteria include the Akaike information criterion (AIC), Bayesian information criterion (BIC), and extended Bayesian information criterion (Akaike, 1974; Schwarz, 1978; Chen & Chen, 2008).

Wen et al. (2020) find in their application that the optimal k is found at the so-called elbow point, i.e. the point where increasing k does not lead to a large improvement in the likelihood. The elbow heuristic is used in determining the optimal number of clusters (Kodinariya & Makwana, 2013). Delgado, Anguera, Fredouille, and Serrano (2015) describe an algorithm to identify this elbow point. We use this algorithm in combination with SPDAS to determine the optimal k.

2.4 Lasso (L_1) regularization

The Lasso regularization technique as proposed by Tibshirani (1996) is very popular. Lasso subtracts an L_1 penalty from the maximum likelihood function, leading to the following maximization problem,

$$\beta^{lasso}(\kappa) = \arg\max_{\beta} \ell(\beta) - \kappa ||\beta||_1, \tag{13}$$

where $||\beta||_1$ is the L_1 norm of β, i.e. the sum of absolute values of the elements, and κ is a predetermined hypertuning parameter.

Solving this maximization problem using standard numerical optimization algorithms does not lead to a sparse solution. Therefore, Fu (1998) propose the shooting algorithm. This algorithm calculates a sparse solution of Lasso regularized linear regressions. The idea behind this algorithm is to solve the optimization problem in Equation (13) and then iteratively set coefficients to zero whose effect of becoming zero is small until the coefficients converge. Alternatives for obtaining a sparse solution for linear regressions have been developed since, such as Least Angle Regression (Efron et al., 2004) and pathwise coordinate descent (Friedman, Hastie, Höfling, & Tibshirani, 2007). These algorithms converge very fast and work extremely well even when the number of parameters is large (Wu & Lange, 2008). Note, however, that there are no algorithms in the current literature that provide a sparse solution for the optimization problem in Equation (13) when $\ell(\beta)$ has a more general form. Therefore, we adapt the shooting algorithm of Fu (1998) in such a way that it works for general log-likelihood specifications.

Lasso's hyperparameter, κ, is commonly chosen by the application of Cross-Validation (Cawley & Talbot, 2010; Mosteller & Tukey, 1968). We follow this and choose κ using Cross-Validation.

2.5 Relaxed Lasso

Adding Lasso regularization has two effects on the solution: model selection and parameter shrinkage. Lasso sets a subset of the coefficients to zero, and these covariates are thus excluded from the model. On the other hand, the parameters in the model that are not set to zero are shrunk towards zero. The covariates corresponding to the parameters that are not set to zero are called the active set, denoted by A_κ. In the standard Lasso formulation, both effects are controlled with one hyperparameter, and it is not possible to limit the second effect. Therefore, Meinshausen (2007) propose the Relaxed Lasso estimator. The Relaxed Lasso estimator maximizes the following optimization problem,

$$\beta^{\text{relax}} = \arg\max_{\beta} \ell(\beta^{\kappa}) - \varphi\kappa\|\beta\|_1, \tag{14}$$

with $\kappa \in [0, \infty)$ and $\varphi \in (0, 1]$

In Equation (14), β^κ is defined as the parameter vector where the parameters that are not contained in the Lasso solution for a given κ are set to zero. That is,

$$\beta_k^{\kappa} = \begin{cases} 0 & \text{if} \quad \beta_k^{\text{lasso}}(\kappa) = 0, \\ \beta_k & \text{else.} \end{cases} \tag{15}$$

Hastie et al. (2020) propose a simplified version of Relaxed Lasso which can be estimated more easily. Their formulation of the Relaxed Lasso is as follows,

$$\beta^{\text{relax}} = \varphi\beta^{\text{lasso}}(\kappa) + (1-\varphi)\beta^{\text{LS}}(\kappa), \tag{16}$$

where $\beta^{\text{LS}}(\kappa)$ is defined as a full-length vector (same dimensions as β) with the coefficients of the least squares regression on the covariates of the active set, A_κ, and zeros for the coefficients corresponding to the covariates not in the active set.

This formulation of Relaxed Lasso is only applicable for regressions. As we have a more general optimization problem, we adapt the formulation to also work for general log-likelihood specifications.

In the literature the two hyperparameters, κ and φ, are often chosen by k-fold Cross-Validation (Hastie et al., 2020; Cawley & Talbot, 2010; Mosteller & Tukey, 1968). We use this method to determine both hypertuning parameters.

3 Methodology

The proposed methodology is based on the POT approach and thus split into two parts. First, we estimate the loss frequency distribution. Secondly, we estimate the distribution of loss severity. As the frequency and severity are independent the estimation of the two parts can be done separately. Both parts are subject to various covariate selection techniques which we discuss in Subsection 3.4. The proposed methodology produces a distribution for both the frequency and severity of losses in a certain time period, e.g. one month. The assumption is made that both distributions stay unchanged within such a time period. Therefore, the covariates need to have the same frequency as this set time period. We also assume that the extreme losses above the threshold in a given time period are i.i.d. draws from the distribution.

3.1 Loss frequency

Like the dynamic POT approach, as suggested by Chavez-Demoulin et al. (2016), the loss frequency is modeled using a non-homogeneous Poisson distribution with a time-varying rate λ_t. Unlike the methodology of Chavez-Demoulin et al. (2016), λ_t does not directly depend on time, and thus does not include any smoothing functions. As the proposed methodology is able to include many covariates the effect of the time variable is assumed to be captured by covariates. The number of extreme losses in a given time period t is denoted by N_t, leading to $N_t \sim \text{Pois}(\lambda_t)$. Where λ_t is depended on covariates in the following way,

$$\ln(\lambda_t) = c_{0,\lambda} + \sum_{i=1}^{p} c_{i,\lambda} x_{i,t}, \tag{17}$$

where $c_{0,\lambda}$ corresponds to the intercept, $c_{i,\lambda}$ to the ith coefficient and $x_{i,t}$ to the ith covariate for time period t.

Note that this is equivalent to a Poisson regression of the covariates on the number of extreme losses. Therefore, the log-likelihood of equation (5) becomes the following,

$$\ell(\boldsymbol{c}_\lambda;\boldsymbol{Y}) = \sum_t \Big(\ln(\lambda_t)N_t - \lambda_t - \ln(N_t!)\Big), \tag{18}$$

where $\boldsymbol{c}_\lambda = \{c_{0,\lambda}, c_{1,\lambda}, \dots, c_{p,\lambda}\}$ and $\boldsymbol{Y} = \{N_t, y_{(t,1)}, \dots, y_{(t,N_t)} | \forall t\}$.

As optimization is only done over $\boldsymbol{c}_\lambda$, Equation (18) can be simplified to,

$$\ell(\boldsymbol{c}_\lambda;\boldsymbol{Y}) = \sum_t \Big(\ln(\lambda_t)N_t - \lambda_t\Big). \tag{19}$$

3.2 Loss severity

The loss severity is modeled with a non-stationary generalized Pareto distribution (GPD). Similar to the rate parameter of the loss frequency, we loosen the assumption of stationary shape and scale parameters of the GPD. Moreover, the shape and scale parameters are assumed to depend on covariates. To prevent identification issues the scale parameter is chosen to be transformed in the following way,

$$\nu_t = \ln((1+\xi_t)\sigma_t). \tag{20}$$

We discuss the details of this transformation in Section 2.2.

To ensure that the domain of the extreme losses is not bounded from above, ξ_t is restricted to the heavy tail case, i.e. $\xi_t > 0$. This restrictions is enforced by making the dependence structure of ξ_t log-linear. The transformation of Equation (20) already enforces the restrictions that $\sigma_t > 0$. Therefore, the dependence structure of ν_t is kept linear. This leads to the following relation with respect to the covariates for the shape and transformed scale parameters,

$$\ln(\xi_t) = c_{0,\xi} + \sum_{i=1}^{p} c_{i,\xi} x_{i,t}, \tag{21}$$

$$\nu_t = c_{0,\nu} + \sum_{i=1}^{p} c_{i,\nu} x_{i,t}, \tag{22}$$

where $c_{0,\cdot}$ corresponds to the intercept and $c_{i,\cdot}$ to the ith coefficient.

The log likelihood of Equation (6) becomes the following after accounting for the non-stationarity and parameter transformation,

$$\begin{aligned}\ell(\boldsymbol{c}_\xi, \boldsymbol{c}_\nu; \boldsymbol{Y}) = \sum_t & N_t \ln(1+\xi_t) - N_t \nu_t \\ & - \frac{\xi_t + 1}{\xi_t} \sum_{i=1}^{N_t} \ln\left(1 + \frac{\xi_t(1+\xi_t)}{\exp(\nu_t)} y_{(t,i)}\right),\end{aligned} \tag{23}$$

where $\boldsymbol{c}_\xi = \{c_{0,\xi}, c_{1,\xi}, \ldots, c_{p,\xi}\}$ and $\boldsymbol{c}_\nu = \{c_{0,\nu}, c_{1,\nu}, \ldots, c_{p,\nu}\}$.

3.3 Non-linear optimization

For the loss frequency, estimation techniques for the used covariate selection methods are already implemented in the software package R. For the loss severity, this is not the case. Therefore, direct maximization of the log-likelihood is required. This maximization is done using the BFGS (Broyden, 1970; Flecther, 1970; Goldfarb, 1970; Shanno, 1970). The BFGS method is a quasi-Newton optimization technique that iteratively updates the inverse Hessian. We calculate the gradients of $\ell(\boldsymbol{c}_\xi, \boldsymbol{c}_\nu)$ with respect to $\boldsymbol{c}_\xi$ and $\boldsymbol{c}_\nu$ separately. The gradients can be found in equations (24) and (25) respectively. The partial derivatives with respect to $\ln(\xi_t)$ and ν_t can be found in Appendix Section 8.1.1.

$$\nabla_\xi \ell(\boldsymbol{c}_\xi, \boldsymbol{c}_\nu) = \sum_t \sum_{i=1}^{N_t} \frac{\partial \ell(\boldsymbol{c}_\xi, \boldsymbol{c}_\nu)}{\partial \ln(\xi_t)}\bigg|_{\xi_t=\xi_t, \nu_t=\nu_t, y=y_{(t,i)}} \boldsymbol{x}_t, \tag{24}$$

$$\nabla_\nu \ell(\boldsymbol{c}_\xi, \boldsymbol{c}_\nu) = \sum_t \sum_{i=1}^{N_t} \frac{\partial^2 \ell(\boldsymbol{c}_\xi, \boldsymbol{c}_\nu)}{\partial \nu_t}\bigg|_{\xi_t=\xi_t, \nu_t=\nu_t, y=y_{(t,i)}} \boldsymbol{x}_t, \tag{25}$$

where $\boldsymbol{x}_t' = (1, x_{1,t}, \ldots, x_{p,t})$.

3.4 Covariate selection

For both distributions of the proposed POT approach, we apply three covariate selection techniques.

3.4.1 Best subset selection

First of all, we apply the classic method best subset selection. For the loss frequency, we maximize the log-likelihood of equation (19) with the restriction on the L_0 norm of all coefficients (except the intercept).

Estimation is done using the PDAS algorithm as described in Section 2.3. For computation of Δ in the algorithm the gradient and the diagonal elements of the Hessian of $\ell(\boldsymbol{c}_\lambda)$ are necessary. These have the following analytical form,

$$\nabla\ell(\boldsymbol{c}_\lambda) = \sum_t (N_t - \lambda_t)\boldsymbol{x}_t, \tag{26}$$

$$\nabla^2\ell(\boldsymbol{c}_\lambda) = -\sum_t \lambda_t \boldsymbol{x}_t \boldsymbol{x}_t'. \tag{27}$$

For the loss severity, the log-likelihood of equation (23) is maximized with the restriction on the L_0 norm of all coefficients (except the intercept). Similar to the loss frequency, estimation is done using the PDAS algorithm. As only the diagonal elements of the Hessian are used in the algorithm, both the gradient and Hessian of $\ell(\boldsymbol{c}_\xi, \boldsymbol{c}_\nu)$ are derived in block forms with respect to $\boldsymbol{c}_\xi$ and $\boldsymbol{c}_\nu$. The Hessian of $\ell(\boldsymbol{c}_\xi, \boldsymbol{c}_\nu)$ with respect to $\boldsymbol{c}_\xi$ and $\boldsymbol{c}_\nu$ can be found in Equations 28 and 29 respectively. The second partial derivatives with respect to $\ln(\xi_t)$ and ν_t can be found in Appendix Section 8.1.1. For completeness, we also give the mixed second partial derivative in the Appendix.

$$\nabla_\xi^2\ell(\boldsymbol{c}_\xi, \boldsymbol{c}_\nu) = \sum_t \sum_{i=1}^{N_t} \left.\frac{\partial^2\ell(\boldsymbol{c}_\xi, \boldsymbol{c}_\nu)}{\partial \ln(\xi_t)^2}\right|_{\xi_t=\xi_t, \nu_t=\nu_t, y=y_{(t,i)}} \boldsymbol{x}_t \boldsymbol{x}_t', \tag{28}$$

$$\nabla_\nu^2\ell(\boldsymbol{c}_\xi, \boldsymbol{c}_\nu) = \sum_t \sum_{i=1}^{N_t} \left.\frac{\partial^2\ell(\boldsymbol{c}_\xi, \boldsymbol{c}_\nu)}{\partial \nu_t^2}\right|_{\xi_t=\xi_t, \nu_t=\nu_t, y=y_{(t,i)}} \boldsymbol{x}_t \boldsymbol{x}_t'. \tag{29}$$

In the PDAS algorithm as discussed in Section 2.3 the assumption is made that each covariate has one coefficient that influences the log likelihood. However, the GPD has two parameters which both depend on the covariates and therefore each covariate has two coefficients that influence the log-likelihood. Therefore, the influence on the log likelihood of the GPD is relatively measured as follows,

$$\Delta_j = h_{j,\xi}\left(c_{j,\xi} - \frac{g_{j,\xi}}{h_{j,\xi}}\right)^2 + h_{j,\nu}\left(c_{j,\nu} - \frac{g_{j,\nu}}{h_{j,\nu}}\right)^2, \quad j = 1, \ldots, p, \tag{30}$$

where $g_{j,\cdot}$ denotes the jth element of the gradient, $\nabla_\cdot \ell(\boldsymbol{c}_\xi, \boldsymbol{c}_\nu)$, and $h_{j,\cdot}$ denotes the jth diagonal element of the block Hessian, $\nabla_\cdot^2 \ell(\boldsymbol{c}_\xi, \boldsymbol{c}_\nu)$.

For both distributions, the hyperparameter k is chosen by the SPDAS algorithm. The choice of k is the elbow point of the log-likelihoods, identified by the algorithm of Delgado et al. (2015).

3.4.2 Lasso

Secondly, Lasso regularization is applied to the log-likelihood functions of both distributions. The penalization term is the L_1 norm of all coefficients (except the intercept).

As the estimation of the loss frequency becomes a Lasso regularized Poisson regression, the algorithm by Friedman, Hastie, and Tibshirani (2010) can be used to obtain a sparse solution. However, the estimation of the loss severity does not collapse to a Lasso regularized GLM. Unfortunately, numerically solving the maximization problem of Equation 13 using the standard numerical optimization algorithms does not lead to a sparse solution. Therefore, an algorithm is proposed to obtain a sparse solution for the loss severity. The algorithm is based on the shooting algorithm (Fu, 1998). The main difference between the proposed algorithm and the shooting algorithm is the measurement of the effect of a coefficient becoming zero, c. This is measured by the derivative with respect to β of log-likelihood instead of the derivative with respect to β of the sum of squared residuals. The proposed algorithm puts two restrictions on the form of the log-likelihood. The log-likelihood, $\ell(\beta)$, needs to be concave and at least once differentiable.

Algorithm 2: Shooting Algorithm for Lasso with Maximum Likelihood

```
Result: $\beta^{\text{lasso}}(\kappa)$
$\beta^{\text{lasso}}(\kappa) = \arg\max_{\beta} \ell(\beta) - \kappa||\beta||_1$ (numerically solved, i.e. non-sparse Lasso
  solution);
$\varepsilon_c = \infty$ ;
while $\varepsilon_c > \varepsilon$ do
    $\beta^{\text{prev}}(\kappa) = \beta^{\text{lasso}}(\kappa)$ ;
    $c_i = \frac{\partial \ell(\beta)}{\partial \beta_i}\Big|_{\beta=\beta^{\text{prev}}_{-i}(\kappa)}$ ;
    $\beta^{\text{lasso}}_i(\kappa) = \begin{cases} 0 & \text{if} \quad |c_i| \leq \kappa \\ \beta^{\text{prev}}_i(\kappa) & \text{else} \end{cases}$ ;
    $\varepsilon_c = (\beta^{\text{lasso}}(\kappa) - \beta^{\text{prev}}(\kappa))'(\beta^{\text{lasso}}(\kappa) - \beta^{\text{prev}}(\kappa))$ ;
end
where $\beta^{\text{prev}}_{-i}(\kappa)$ is equal to the vector $\beta^{\text{prev}}(\kappa)$ with the $i$th element set to zero and $\varepsilon$
  the predefined convergence tolerance.
```

The hypertuning parameter κ is chosen by ten-fold Cross-Validation. The grid for κ is chosen to be an exponentially spaced grid between κ_{min} and κ_{max} with

$$\kappa_{\text{max}} = \frac{\ell_{\text{NR}} - \ell_{\text{naive}}}{\sum_t N_t}, \tag{31}$$

$$\kappa_{\text{min}} = \kappa_{\text{max}} \times 10^{-4}, \tag{32}$$

where ℓ_{naive} is the maximized log likelihood when all covariates are included and ℓ_{naive} is the maximized log likelihood when no covariates are included and thus only the intercept is optimized. This grid selection is based on Friedman et al. (2010) and adjusted for log-likelihood optimizations.

3.4.3 Relaxed Lasso

Lastly, Relaxed Lasso is applied to both the loss frequency and severity. For ease of estimation, the simplified version of Relaxed Lasso is used. The application to the loss frequency is relatively straightforward as the parameters can be obtained using Poisson regression.

As the loss severity does not fall into the class of GLM the application of the simplified version of Relaxed Lasso is not directly possible. It is possible to apply Relaxed Lasso by solving the optimization problem of equation (14). However, this optimization is not trivial in practice. Therefore, the simplified version of Relaxed Lasso is adapted in the following way such that it no longer depends on a regression solution.

$$\beta^{\text{relax}} = \varphi\beta^{\text{lasso}}(\kappa) + (1-\varphi)\beta^{\text{NR}}(\kappa), \tag{33}$$

where $\beta^{\text{NR}}(\kappa)$ is defined as a $p \times 1$ vector with the not regularized solution when only the covariates of the active set, A_κ, are included and zeros in the places of the other covariates.

The two hyperparameters κ and φ are chosen by ten-fold Cross-Validation. For the hyperparameter κ the grid as described in Section 3.4.2 is used. For the hyperparameter φ, Friedman et al. (2010) use $\varphi \in \{0, 0.25, 0.5, 0.75, 1\}$. However, as $\varphi = 1$ equals the Lasso solution, we use $\varphi \in \{0, 0.25, 0.5, 0.75\}$.

3.5 Goodness-of-fit test

We use the graphical goodness-of-fit test as described in Section 2.2 to check whether we set the threshold u correctly. We check the Q-Q plots for signs of misspecification. When this is the case, we perform the graphical goodness-of-fit test for different values for the threshold u and use the results of the threshold where the Q-Q plots show no or the least signs of misspecification.

The graphical goodness-of-fit test requires visual inspection of each Q-Q plot. This becomes impossible when the number of analyses is large. Therefore, we use the Kolmogorov-Smirnov (KS) test as described by Massey (1951) to quantify the distance between the theoretical and empirical distribution. In the KS-test we compare $r_{(t,i)}$ as described in Equation 11 with the standard exponential distribution. Therefore, the null hypothesis of the KS-test is that the losses are a draw from the estimated distribution. Thus, rejection of the KS-test means that we reject that the losses are a draw from the estimated distribution.

4 Simulation study

To assess the performance of the proposed methodology in selecting the covariates used in the data generating process (DGP) out of many covariates, a simulation study is conducted. In the simulation study, we simulate the frequency and severity of extreme losses from a Poisson and generalized Pareto distribution respectively. We simulate non-extreme losses from a normal distribution. The parameters of these distributions depend on a subset of the simulated covariates. The generation of a single simulation is performed as follows.

First, we set the threshold u to a quantile between 80% and 95% of the standard normal distribution. Second, we generate 50 covariates for each time period t by drawing from a i.i.d. standard normal distribution, i.e. $x_{i,t} \overset{iid}{\sim} N(0,1)$, $\forall(i,t)$. From this set of covariates a subset, denoted by A, with size $k \in \{3,4,5,6\}$ is taken which will be used in the DGP.

As the methodology estimates the coefficients for ν_t instead of σ_t, we let ν_t depend on the covariates such that the coefficients can be compared. This leads to the following (log-)linear relations,

$$\ln(\lambda_t) = c_{0,\lambda} + \sum_{i\in A} c_{i,\lambda} x_{i,t}, \qquad \ln(\xi_t) = c_{0,\xi} + \sum_{i\in A} c_{i,\xi} x_{i,t}, \qquad \nu_t = c_{0,\nu} + \sum_{i\in A} c_{i,\nu} x_{i,t}, \tag{34}$$

where $c_{0,\cdot}$ denotes the intercept corresponding to the distribution parameter and $c_{\cdot,\cdot}$ denotes the coefficient corresponding to the distribution parameter and the covariate.

Chavez-Demoulin, Embrechts, and Nešlehová (2006) find values for ξ that range between 0.2 and 0.8 and values for σ that range between 1 and 2. Therefore, we choose the coefficients for ξ_t and ν_t in such a way that the resulting values are in the same magnitude. To ensure sufficient losses are available for each time period we choose the coefficients for λ_t in such a way that the resulting values are between 10 and 100. Appendix Table 13 describes the distributions the intercepts and coefficients are (independently) drawn from.

Using these resulting distribution parameters we simulate the frequency and severity of the extreme losses. First, we draw the number of extreme losses for each time period (denoted with N_t) independently from a non-homogeneous Poisson distribution with rate λ_t, i.e. $N_t \sim \text{Pois}(\lambda_t)$, $\forall t$. Secondly, for each time period N_t, we draw the excess losses from a non-homogeneous generalized Pareto distribution (GPD) with parameters ξ_t and $\sigma_t = \frac{\exp(\nu_t)}{1+\xi_t}$, i.e. $y_{(t,i)} \sim \text{GPD}(\xi_t, \sigma_t)$, $\forall i \in \{1, \ldots, N_t\}$ and $\forall t$. The inverse CDF of the GPD is defined as follows,

$$F^{-1}(y) = \frac{\sigma((1-y)^{-\xi} - 1)}{\xi}. \tag{35}$$

Therefore, a random draw, $y_{(t,i)} \sim \text{GPD}(\xi_t, \sigma_t)$, can be simulated using a random draw from a standard uniform distribution, $U \sim \text{Uni}(0,1)$, using the following transformation,

$$y_{(t,i)} = \frac{\sigma_t(U^{-\xi_t} - 1)}{\xi_t}. \tag{36}$$

Next, we transform the excess losses into extreme losses by adding the threshold u. Lastly, we simulate the non-extreme losses. For each time period, we simulate $500 - N_t$ non-extreme losses to get a total of 500 losses per time period. These losses are independent draws from a standard normal distribution with the restriction that the value is lower than u.

This procedure leads to a dataset that contains 500 losses (of which N_t are extreme) and 50 covariates for each time point t. After the generation of a simulation, we apply the methodology. The chosen threshold, selected covariates, and coefficients are not available for the estimation process.

4.1 Correlation in covariates

We also test the performance of the proposed methodology when the potential covariates are correlated. Instead of drawing $x_{i,t}$ from a i.i.d. standard normal distribution, $x_t = \{x_{1,t}, \ldots, x_{50,t}\}$ is drawn from a multivariate normal distribution, i.e. $x_t \overset{iid}{\sim} N(\mathbf{0}, \Sigma)$.

Joe (2006) propose a method to simulate a random correlation matrix. Joe (2006) parameterize the correlation matrix in correlations and partial correlations and draw these from a Beta distribution. In their method, they use one parameter, α, which influences the distribution of the correlation matrix. We choose $\alpha = 1$ which leads to a uniform distribution over the space of positive definite correlation matrices. We choose the variance of x_t to be 1. Therefore, the covariance matrix, Σ, is equal to the correlation matrix.

4.2 Performance measures

We use the following performance measure to assess the performance of the proposed methods in the simulation study. First, we consider the percentage of the selected covariates that are identified as a significant covariate (TP, true positive). Secondly, we consider the percentage of the not-selected covariates that are correctly identified as not significant (TN, true negative). We calculate these statistics using the following formulas,

$$\mathtt{TP} = \frac{1}{|A|} \sum_{i \in A} I_{\hat{\beta}_i \neq 0}, \tag{37}$$

$$\mathtt{TN} = \frac{1}{|A^c|} \sum_{i \in A^c} I_{\hat{\beta}_i = 0}. \tag{38}$$

Next, we also consider the mean squared error (MSE) of the estimated coefficients.

$$\mathtt{MSE} = \frac{1}{51} \sum_{i=0}^{50} (\beta_i - \hat{\beta}_i)^2. \tag{39}$$

4.3 Results

First, we perform 100 simulations with no correlation in the covariates. Table 1 shows the mean of the performance measures with the threshold, u, chosen at the 90% quantile of the losses. For the loss frequency, we find that, on average, BeSS has the highest true positive rate. The average true negative rate for Relaxed Lasso is 0.99 and the true positive rate is lower, 0.87, than the other two. This indicates that Relaxed Lasso might be too restrictive in selecting covariates. BeSS produces the lowest MSE, indicating that on average the error in the estimated coefficients is the lowest with BeSS. BeSS and Lasso correctly select all covariates in the active set (i.e. $\mathtt{TP} = 1$) 81 and 75 times, respectively, and Relaxed Lasso does only this 58 times. However, Relaxed Lasso is able to select only the correct covariates (i.e. TP = TN = 1) 50 times. BeSS and Lasso are able to do this only 26 and 24 times respectively.

Table 1: Mean over 100 simulation of the performance measures for the simulation study with no correlation in the covariates. Threshold, u, is chosen at 90% quantile.

	Loss frequency			Loss severity		
Method	TP	TN	MSE	TP	TN	MSE
BeSS	**0.94**	0.98	**2.65**	0.91	**1.00**	13.53
Lasso	0.92	0.96	2.76	**0.96**	0.49	**12.05**
Relaxed Lasso	0.87	**0.99**	2.69	**0.96**	0.87	14.06

The reported MSE is scaled by 10^2.

For the loss severity, we find that Lasso and Relaxed Lasso produce the highest true positive rate, 0.96. For Lasso, we observe a relatively low true negative rate, 0.49, indicating that Lasso selects too many covariates. BeSS has a true negative rate of 1 and a lower true positive rate, indicating that BeSS is on the restrictive side when selecting covariates.

The number of times all covariates of the active set are selected is in line with this. Namely, BeSS does this 61 times, Lasso 83 times, and Relaxed Lasso 84 times. Moreover, the number of times only the covariates of the active set are selected by BeSS is 57, by Relaxed Lasso 31 and Lasso is never able to do this. Regarding the MSE, we observe high values compared to the median, which ranges between 3.29-3.34. This is due to two simulations for which all three methods produce an MSE which is two orders of magnitude greater than the MSE's for other simulations. BeSS produces the lowest median MSE, closely followed by Relaxed Lasso and Lasso. The median of the performance measures can be found in Appendix Table 14.

Appendix Tables 15 and 16 show the mean of the performance measures with the threshold, u, chosen at the 80% and 95% quantile of the losses respectively. For the loss frequency, the results with these different thresholds are similar to the results shown in Table 1. The true positive rates are slightly lower when the threshold is chosen at the 80% quantile. For the MSE, the values are higher when the 80% quantile is used, and lower when the 95% quantile is used. This indicates that selecting a too high value for the threshold still leads to accurate estimates. For the loss severity, we find lower true positive and negative rates when the threshold is chosen at the 80% quantile. The MSE for all three methods is higher when the 80% quantile is used. When the 95% quantile is used we find similar true positive and negative rates as with the 90% quantile. However, the MSE is lower than the mean and median MSE of the 90% quantile. Thus, for the loss severity, we find the same pattern that a higher threshold leads to better performance.

Table 2: Count table with significance levels of the KS-statistics for all simulations without correlation in the covariates for each threshold and method.

	80% quantile			90% quantile			95% quantile		
Significance	BeSS	Lasso	RL	BeSS	Lasso	RL	BeSS	Lasso	RL
$< 1\%$	100	100	100	77	77	76	46	45	44
$< 5\%$	100	100	100	79	79	78	50	50	51
$< 10\%$	100	100	100	80	81	80	51	56	51
$\geq 10\%$	0	0	0	20	19	20	49	44	49

RL denotes Relaxed Lasso

Table 2 contains the number of simulations for which we reject the KS-test at certain significance levels. When the threshold is chosen at the 80% quantile we reject the KS-test at the 1% level for all simulations and for each method. For the 90% quantile, we reject the KS-test at the 1% level 77 times for BeSS and Lasso and 76 times for Relaxed Lasso. We do not reject the KS-test at the 10% level 19 times for Lasso and 20 times for BeSS and Relaxed Lasso. For the 95% quantile, the number of times we do not reject the KS-test increases to 44-49 times. Thus, we observe a trend where the fit becomes better when the threshold is chosen at a higher value. However, it should be noted that the number of points that we fit decreases when the threshold is chosen at a higher value.

Figure 1 contains the Q-Q plots using both the actual and estimated parameter values for the first four simulations as examples. For the first simulation, we find that the right skewness is underestimated by all three methods. We also reject the KS-test at the 1% level for this simulation. For simulations 2, 3, and 4 the Q-Q plots indicate a good fit by all methods. However, for simulation 3 we also reject the KS-test at the 1% level for all methods. For simulations 2 and 4 we do not reject the KS-test at the 10% level for all methods. For simulation 1, the inaccuracy is large due to the inaccurate estimates for ξ where the MSE is above average for all three methods. For simulation 3, the inaccuracy is due to the threshold. The estimated threshold for this simulation is lower than the threshold used in the DGP.

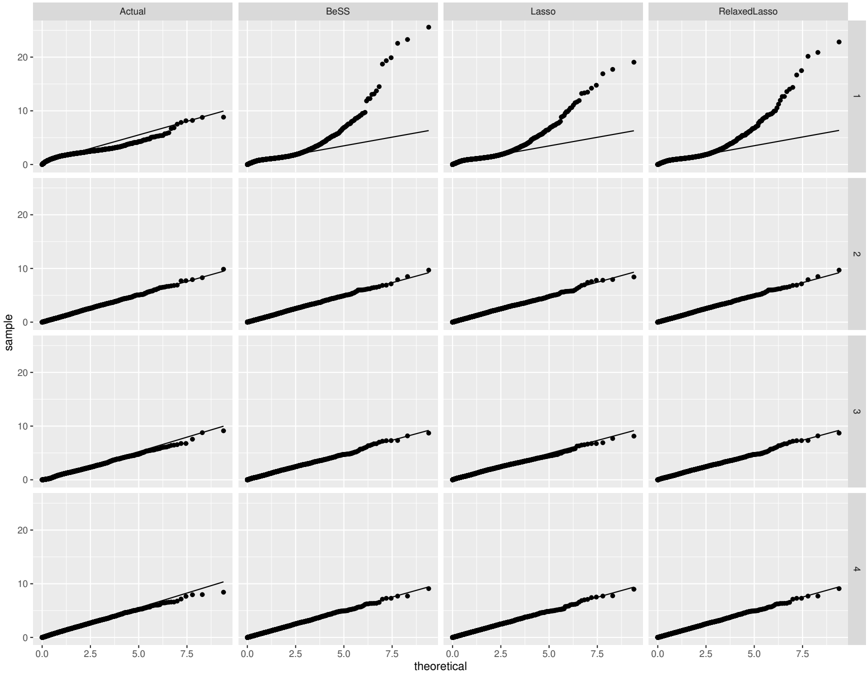

Figure 1: Q-Q plots of the losses above the 90% quantile using the actual parameters of the distribution and estimated parameters of the distribution by each method for the first four simulations.

Next, we drop the independence of the covariates and perform 100 simulations with correlation in the covariates. For each simulation, we simulate a random covariance matrix as described in Section 4.1. Table 3 shows the mean of the performance measures with the threshold, u, chosen at the 90% quantile of the losses. For the loss frequency, we find that BeSS outperforms both Lasso and Relaxed Lasso in terms of true positive rate and MSE. Relaxed Lasso produces a slightly higher true negative rate. BeSS and Lasso are able to correctly select all covariates of the active set (i.e. $\texttt{TP} = 1$) 79 and 76 times respectively. Relaxed Lasso does this 64 times. Moreover, BeSS and Relaxed Lasso are able to select only the correct covariates (i.e. $\texttt{TP} = \texttt{TN} = 1$) 41 times, Lasso does this only 15 times.

Table 3: Mean over 100 simulation of the performance measures for the simulation study with correlation in the covariates. Threshold, u, is chosen at 90% quantile.

	Loss frequency			Loss severity		
Method	TP	TN	MSE	TP	TN	MSE
BeSS	**0.93**	0.98	**2.30**	0.88	**1.00**	15.14
Lasso	0.92	0.94	2.41	0.95	0.53	**4.31**
Relaxed Lasso	0.88	**0.99**	2.33	**0.96**	0.85	7.69

The reported MSE is scaled by 10^2.

For the loss severity, BeSS has a relatively low true positive ratio and a true negative rate of 1, indicating that BeSS selects fewer covariates than in the DGP. This might be caused by the correlation in the covariates. Lasso, on the other hand, has a relatively high true positive ratio and a low true negative ratio. This indicates that Lasso does the opposite and selects too many covariates. Relaxed Lasso has a slightly higher true positive rate than Lasso, however, its true negative rate is in between BeSS and Lasso. Lasso and Relaxed Lasso are able to select all covariates in the active set 76 and 81 times respectively and BeSS only 52 times. However, Lasso never selects only the covariates in the active set. BeSS and Relaxed Lasso do this 47 and 11 times respectively. In terms of MSE, Lasso outperforms both BeSS and Relaxed Lasso. For BeSS and Relaxed Lasso, we observe relatively high values for the MSE compared to the median. For both methods, this is due to two simulations that produce a large MSE. For both simulations, the 90% quantile is below the threshold used in the DGP. The median MSE leads to the same conclusion as the mean. All median values are reported in Appendix Table 14.

Appendix Tables 18 and 19 show the results with the threshold, u, chosen at the 80% and 95% quantile respectively. We find a similar pattern as in the case of no covariate correlation. For both the loss frequency and severity, we observe that the MSE is higher when the threshold is set at 80% and lower when the threshold is set at 95%. Notably, for the loss severity, we find that both BeSS and Relaxed Lasso outperform Lasso in terms of MSE when the threshold is chosen at the 95% quantile.

Table 4: Count table with significance levels of the KS-statistics for all simulations with correlation in the covariates for each threshold and method.

	80% quantile			90% quantile			95% quantile		
Significance	BeSS	Lasso	RL	BeSS	Lasso	RL	BeSS	Lasso	RL
< 1%	100	100	100	74	73	74	41	41	41
< 5%	100	100	100	74	74	74	42	44	43
< 10%	100	100	100	74	74	74	44	47	45

RL denotes Relaxed Lasso

Table 4 contains the number of simulations for which we reject the KS-test at certain significance levels. When the threshold is chosen at the 80% quantile we reject the KS-test at the 1% level for all simulations and for each method. For the 90% quantile, we reject the KS-test at the 1% level 74 times for BeSS and Relaxed Lasso and 73 times for Lasso. We do not reject the KS-test at the 10% level 26 times for all three methods. For the 95% quantile, the number of times we do not reject the KS-test increases to 53-56 times. We observe the same trend as in the simulations with no correlation in the covariates where the fit becomes better when the threshold is chosen at a higher value.

Figure 2 contains the Q-Q plots using both the actual and estimated parameter values for the first four simulations as examples. We observe that the Q-Q plots indicate a good fit for simulations 1 and 2 by all methods. For both of these simulations, we do not reject the KS-test at the 10% level for all three methods. However, for simulations 3 and 4 we find that the right skewness is underestimated by all three methods. In line with this, we reject the KS-test for these simulations at the 1% level for all three methods. For both these simulations, we find that the 90% quantile of the data is below the real threshold and therefore non-extreme losses are used as extreme losses.

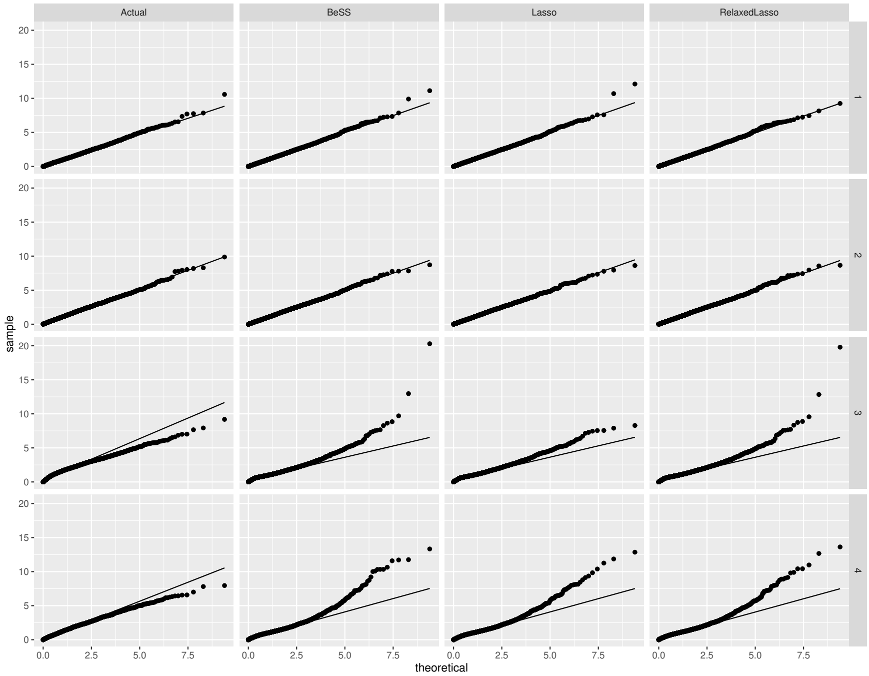

Figure 2: Q-Q plots of the losses above the 90% quantile using the actual parameters of the distribution and estimated parameters of the distribution by each method for the first four simulations.

5 Application to market risk

We apply the proposed methodology to a market risk dataset. This application is split into three parts. First, we analyze the daily loss returns of the S&P 500. Secondly, we analyze the loss returns of the 500 companies in the S&P500 per company, i.e. we perform 500 analyses. Lastly, we analyze the loss returns of each sector. In all of these analyses the time period, denoted in the methodology by t, is set to be one month. That is, we fit distributions for the loss frequency and severity per month and the covariates have a monthly frequency.

We obtain the closing prices of the S&P 500 index from Standard and Poor's website, and we download the adjusted closing prices[1] of the 500 companies in this index from Yahoo Finance. All closing prices are obtained for the years 2011-2020, i.e. 2,517 trading days. We convert The (adjusted) closing prices into loss returns.

For all three analyses, we use the large macro-economic dataset with a monthly frequency, FRED-MD. FRED-MD is developed by the Research Division of the Federal Reserve Bank of St. Louis and contains over 120 macro-economic variables. All variables in the dataset have either a monthly frequency or have been transformed to have a monthly frequency. Some variables are adjusted for the use of statistical analysis (McCracken & Ng, 2016). All transformations can be found in the FRED-MD Updated Appendix. We standardize all variables for our methods to work correctly. The lagged values of the covariates are used in the analysis such that the fitted distributions are forecasts.

5.1 S&P500

First, we analyze the loss returns of the S&P500 index. As each month has around 21 trading days, each month in this analysis has around 21 loss returns. To have sufficient extreme losses to train the models the threshold, u, is set to the 80% quantile. This means that on average each month has around 4 extreme losses. Appendix Table 20 shows the

[1] Yahoo provides adjusted closing prices which are the closing prices adjusted for dividends and stock splits following the Center for Research in Security Prices (CRSP) standards.

number of exceedances per month. Seven of the 120 months contain no extreme losses and the most extreme losses (12) occurred in March 2020. Appendix Figure 7 contains the histogram of the daily loss returns of the S&P500 index.

5.1.1 Results

First, we perform BeSS, Lasso, and Relaxed Lasso on the loss frequency. Both Lasso and Relaxed Lasso select no covariates indicating that there is no significant relation between the covariates and frequency of extreme losses. However, BeSS selects three covariates, namely, the S&P Dividend Yield, U.S./U.K. Foreign Exchange Rate and the S&P Volatility Index. For the S&P Dividend Yield the relation is negative, indicating that an increase in dividend yield leads to a decrease in the frequency of extreme losses. The U.S./U.K. Foreign Exchange Rate and the S&P Volatility Index both have a positive relation.

Table 5: Estimated coefficients by all three methods for the S&P500 index. The threshold is chosen at 80%.

	BeSS			Lasso			Relaxed Lasso		
Coefficients	λ	ξ	ν	λ	ξ	ν	λ	ξ	ν
(Intercept)	1.85	-6.11	-0.24	1.43	-65.95	-0.19	1.43	-14.80	-0.23
IPDCONGD	0.00	3.60	0.03	0.00	0.00	0.00	0.00	0.00	0.00
IPNCONGD	0.00	0.00	0.00	0.00	0.00	0.05	0.00	-0.58	0.11
IPBUSEQ	0.00	0.00	0.00	0.00	0.00	-0.02	0.00	0.00	0.00
USGOVT	0.00	0.00	0.00	0.00	0.00	0.09	0.00	6.92	0.00
CES0600000007	0.00	0.00	0.00	0.00	0.00	-0.03	0.00	0.00	0.00
S.P.div.yield	-1.21	-1.10	-0.19	0.00	0.00	-0.03	0.00	0.00	0.00
T5YFFM	0.00	0.00	0.00	0.00	0.00	-0.03	0.00	0.00	0.00
EXSZUSx	0.00	0.00	0.00	0.00	0.00	-0.05	0.00	0.00	0.00
EXUSUKx	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
OILPRICEx	0.00	0.00	0.00	0.00	0.00	-0.02	0.00	0.00	0.00
CPIAPPSL	0.00	0.00	0.00	0.00	0.00	-0.03	0.00	0.00	0.00
CUSR0000SA0L2	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
VXOCLSx	0.03	-0.70	0.33	0.00	0.00	0.07	0.00	-0.32	0.20

A description of the variables can be found in the FRED-MD Appendix.

Next, we perform BeSS, Lasso, and Relaxed Lasso on the loss severity. BeSS once again selects three covariates: namely, the Industrial Production Durable Consumer Goods, the S&P Dividend Yield, and the S&P Volatility Index. Lasso selects 10 covariates, which can be seen in Table 5. Relaxed Lasso also selects three covariates: namely, the Industrial Production Non-Durable Consumer Goods, the number of U.S. government employees, and the S&P Volatility Index.

To assess the estimated fit we perform the KS-test for all three methods. For all three methods, we do not reject the KS-test at the 10% significance level. Figure 3 contains the Q-Q plots for all three methods. We do not observe any structural deviations in the Q-Q plots. Thus, all three methods lead to a good fit of the losses.

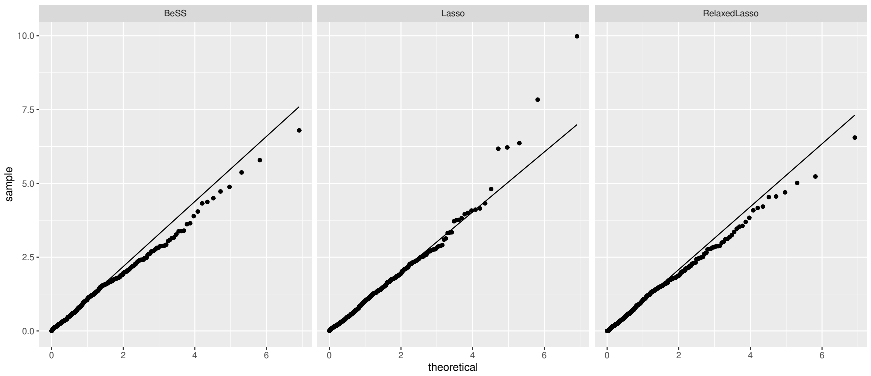

Figure 3: Q-Q plots of the losses above the 80% quantile using the estimated parameters of the distribution by each method.

5.2 Single stock analysis

For each of the 500 companies in the S&P500 index, an analysis is performed on the loss returns. The threshold, u, is chosen at the 80% quantile. Two of the 500 companies have more than ten months without extreme losses.

5.2.1 Results

First, we perform BeSS, Lasso, and Relaxed Lasso on the loss frequency. Lasso and Relaxed Lasso select no covariates 57% and 55% of the time respectively. Therefore, the median of the number of selected covariates is zero for both. For BeSS, this median is 3 with a mean of 3.3. The means for Lasso and Relaxed Lasso are 2.2 and 2 respectively. As can be seen in Figure 4, Lasso and Relaxed Lasso select more than 20 covariates a small number of times whereas BeSS always selects less than 15 covariates.

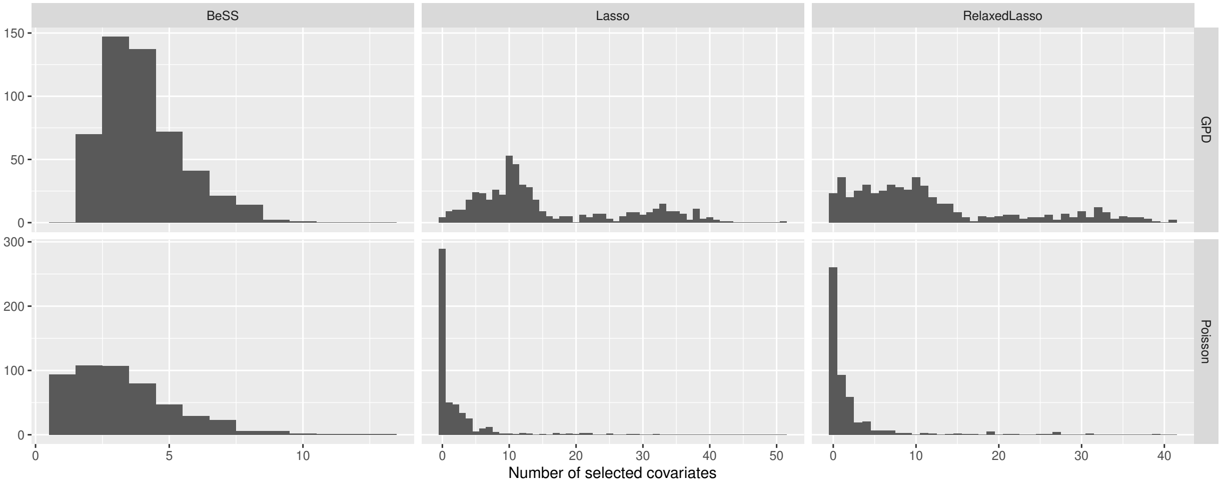

Figure 4: Histograms of the number of selected covariates per method for loss frequency and severity.

Table 6 shows the 10 most chosen variables per method. All three methods select the S&P Volatility Index the most often and for all three methods the median of the corresponding coefficients is positive. The S&P 500 Dividend Yield, Switzerland/U.S. Foreign Exchange Rate and Industrial Production index are also chosen often by all three methods. With the median of the corresponding coefficients being positive for the S&P 500 Dividend Yield and Switzerland/U.S. Foreign Exchange Rate and negative for the Industrial Production index. Lasso and Relaxed Lasso often choose the Consumer Price Index with a negative median of the corresponding coefficient.

Table 6: The 10 most chosen variables for the loss frequency per method.

BeSS		Lasso		Relaxed Lasso	
Median	3		0		0
Mean	3.3		2.2		2.0
Variable	%	Variable	%	Variable	%
VXOCLSx	52.5	VXOCLSx	33.9	VXOCLSx	34.1
S.P.div.yield	28.7	CPIAPPSL	12.7	CPIAPPSL	11.3
EXSZUSx	16.8	S.P.div.yield	9.3	IPNCONGD	8.7
IPNCONGD	16.0	EXSZUSx	9.3	EXSZUSx	8.5
IPBUSEQ	12.7	IPNCONGD	6.9	S.P.div.yield	8.1
IPFUELS	10.5	ANDENOx	6.9	IPB51222S	6.1
TOTRESNS	9.3	IPBUSEQ	5.9	TOTRESNS	6.1
ANDENOx	9.1	IPFUELS	5.5	ANDENOx	5.9
BAA	8.9	TOTRESNS	5.5	IPFUELS	5.5
PPICMM	8.9	IPB51222S	5.3	IPBUSEQ	5.3

A description of the variables can be found in the FRED-MD Appendix.

Next, we perform all three methods on the loss severity. In contrast to the loss frequency, Lasso and Relaxed Lasso select, on average, more than double the number of covariates than BeSS. This can also be seen in Figure 4. BeSS always selects 10 covariates or less, whereas Lasso and Relaxed Lasso sometimes select over 40 covariates.

Table 7 contains the most chosen variables for the loss severity per method. Again, for all three methods, the S&P Volatility Index is the most selected covariate. For all three methods, the median of the corresponding coefficients is negative with respect to ξ_t and positive with respect to ν_t. It is therefore difficult to state which effect an increase in the Volatility Index has on the expected loss. Lasso and Relaxed Lasso often select the number of U.S. government employees. This might be due to the co-movement of the number of U.S. government employees and large (economic) events including market crashes.

Table 7: The 10 most chosen variables for the loss severity per method.

BeSS		Lasso		Relaxed Lasso	
Median	4		11		9
Mean	4.0		15.4		11.7
Variable	%	Variable	%	Variable	%
VXOCLSx	52.9	VXOCLSx	89.3	VXOCLSx	83.8
IPDCONGD	38.0	USGOVT	77.2	USGOVT	66.7
USGOVT	36.6	IPBUSEQ	55.6	IPBUSEQ	48.5
EXSZUSx	23.8	EXSZUSx	51.3	T5YFFM	42.6
S.P.div.yield	18.8	S.P.div.yield	49.9	EXSZUSx	39.6
IPNCONGD	15.4	T5YFFM	49.3	S.P.div.yield	39.4
DTCTHFNM	14.7	IPNCONGD	44.8	IPNCONGD	35.0
PERMITNE	13.1	PERMITNE	41.0	DTCTHFNM	32.1
T5YFFM	13.1	CES0600000007	39.4	CES0600000007	31.3
HOUSTNE	12.5	OILPRICEx	39.0	PERMITNE	30.9

A description of the variables can be found in the FRED-MD Appendix.

To assess the provided fit by each method we perform KS-tests. Table 8 contains the number of analyses for which we reject the KS-test at certain significance levels. We find that we reject the KS-tests the most often for Lasso indicating that the fit provided by Lasso is worse. At the 1% significance level, we reject the KS-tests the least often for the fits provided by Relaxed Lasso. For the 5% and 10% significance levels, we reject the least often for the fits provided by BeSS. As the number of rejections at the 10% significance level is almost double for Relaxed Lasso compared to BeSS we conclude that BeSS provides the best fit.

Table 8: Count table with significance levels of the KS-statistics for all analyses for each method and the number of expected rejections based on the significance levels.

Significance	BeSS	Lasso	Relaxed Lasso	Expected rejections
< 1%	32	53	19	5
< 5%	42	132	57	25
< 10%	58	203	103	50

5.3 Sector analysis

The 500 companies in the S&P500 index can be split into 11 sectors[2]. For each of the 11 sectors, an analysis on the loss returns is performed. In these analyses, we assume that the coefficients for stocks within the same sector are equal. Such an assumption, despite being arbitrary, has the advantage to allow more loss returns in each month. Correspondingly, the threshold, u, can be chosen at a higher value while maintaining sufficient extreme losses for training. However, this analysis violates an assumption made in the methodology. The methodology assumes that all extreme losses are drawn independently from a GPD. Empirically, stock returns are cross-sectionally correlated and thus not independent draws from a distribution, especially when the absolute stock returns are large (Cizeau, Potters, & Bouchaud, 2001). We keep this violation in mind when interpreting the results.

The threshold, u, is chosen to be the 95% quantile. On average 46 losses exceed this threshold per month per sector, within sector average vary between 23 and 75. Appendix Table 21 contains the summary statistics of the number of exceedances per month.

[2] The division into these sectors is according to the Global Industry Classification Standard (GICS).

5.3.1 Results

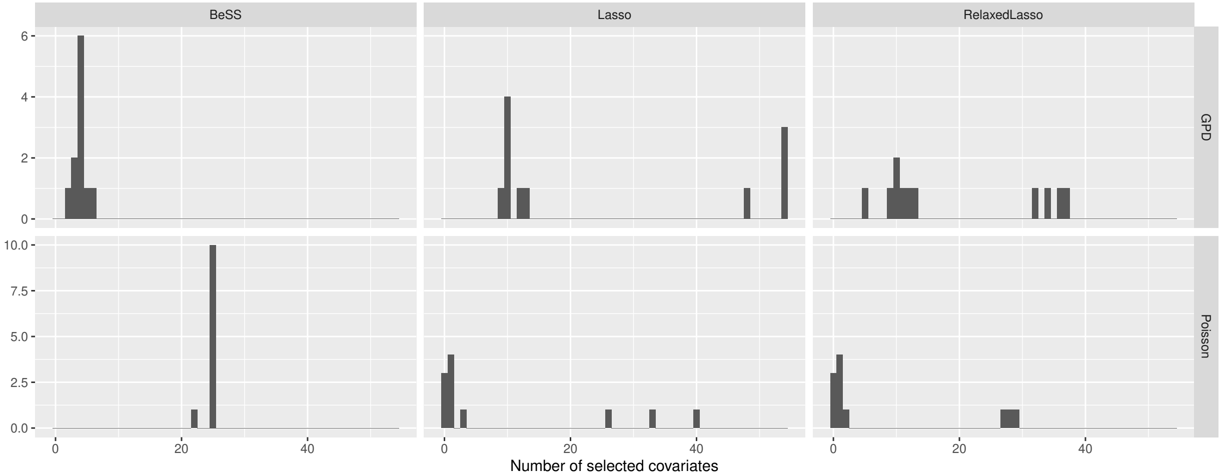

Figure 5: Histograms of the number of selected covariates per method for loss frequency and severity.

First, we analyze the loss frequency of the 11 sectors. As can be seen in Figure 5, BeSS always chooses over 20 covariates. Lasso and Relaxed Lasso choose either a small number of covariates (0-3) or more than 25 covariates. Lasso chooses more than 25 covariates for the sectors Consumer Staples, Industrials, and Materials. Relaxed Lasso does this for the sectors Communication Services, Materials, and Utilities.

Table 9 shows the 10 most chosen covariates per method. For all three methods, the S&P Volatility Index is the most chosen variable and the relation is always positive. We observe that the covariates selected by BeSS always include a set of 5 variables and the relations are always in the same direction: the S&P Volatility Index (positive relation), the number of new Housing Units (positive relation), the number of new Housing Units authorized (negative relation), the S&P 500 Dividend Yield (negative relation) and the number of filled unemployed claims (negative relation).

Table 9: The 10 most chosen variables for the loss frequency per method.

BeSS		Lasso		Relaxed Lasso	
Median	25		1		1
Mean	24.7		9.6		8.2
Variable	%	Variable	%	Variable	%
VXOCLSx	100.0	VXOCLSx	72.7	VXOCLSx	63.6
HOUSTNE	100.0	CPIAPPSL	36.4	IPNCONGD	36.4
PERMITNE	100.0	IPNCONGD	27.3	IPNMAT	36.4
S.P.div.yield	100.0	IPBUSEQ	27.3	UEMP5TO14	36.4
CLAIMSx	100.0	UEMPLT5	27.3	CLAIMSx	36.4
IPB51222S	90.9	UEMP5TO14	27.3	USGOVT	36.4
BAA	90.9	USGOVT	27.3	CES0600000007	36.4
DNDGRG3M086SBEA	90.9	CES0600000007	27.3	HOUSTNE	36.4
IPBUSEQ	81.8	HOUSTNE	27.3	HOUSTMW	36.4
USGOVT	81.8	HOUSTMW	27.3	PERMITNE	36.4

A description of the variables can be found in the FRED-MD Appendix.

Next, we analyze the loss severity. We observe that BeSS always chooses less than 10 covariates. For both Lasso and Relaxed Lasso we observe that both choose around 10 covariates for most sectors. However, for Energy, Materials, Real Estate, and Utilities both choose over 30 covariates. For all three methods, we find that the number of U.S. government employees is the most chosen covariate. All three methods estimate the relation to be negative with respect to ξ_t and positive with respect to ν_t. Lasso also always chooses the number of unemployed for 15-26 weeks with a positive relation with respect to ξ_t and negative with respect to ν_t.

Table 10: The 10 most chosen variables for the loss severity per method.

BeSS		Lasso		Relaxed Lasso	
Median	4		12		12
Mean	3.9		25.8		19
Variable	%	Variable	%	Variable	%
USGOVT	72.7	USGOVT	100.0	USGOVT	100.0
IPBUSEQ	63.6	UEMP15T26	100.0	UEMP15T26	90.9
T5YFFM	45.5	TB6SMFFM	90.9	IPBUSEQ	81.8
UEMP15T26	27.3	T5YFFM	90.9	TB6SMFFM	81.8
UEMPLT5	18.2	IPBUSEQ	72.7	T5YFFM	72.7
ANDENOx	18.2	PERMITNE	72.7	S.P.div.yield	63.6
GS5	18.2	S.P.div.yield	72.7	GS5	63.6
EXSZUSx	18.2	GS5	72.7	UEMP5TO14	54.5
VXOCLSx	18.2	EXSZUSx	72.7	PERMITNE	54.5
IPDCONGD	9.1	BAA	63.6	TOTRESNS	54.5

A description of the variables can be found in the FRED-MD Appendix.

To assess the fit provided by each method we perform KS-tests. Table 11 contains the p-values for each KS-test corresponding to the sectors and methods. We never reject the KS-test when BeSS is used. For Lasso and Relaxed Lasso, we reject the KS-test for five and three sectors respectively at the 5% significance level. This indicates that BeSS provides the best fit.

Table 11: P-values (in %) of KS test for each all 11 sectors per estimation method. The threshold is chosen at the 95% quantile.

Sector	BeSS	Lasso	Relaxed Lasso
Communication Services	90.9	39.6	72.7
Consumer Discretionary	98.7	40.8	97.0
Consumer Staples	95.8	0.0***	89.5
Energy	22.8	27.8	25.8
Financials	27.0	0.0***	20.8
Health Care	55.1	0.4***	65.3
Industrials	13.1	0.0***	2.7**
Information Technology	45.9	5.9*	38.3
Materials	22.6	71.4	52.6
Real Estate	41.7	33.8	3.4**
Utilities	88.2	0.1***	0.3***

*** p-value < 1% ** p-value < 5% * p-value < 10%

Additionally, we perform a graphical goodness-of-fit test using the Q-Q plots. Figure

6 contains the Q-Q plots for all 11 sectors estimated using all three methods. Here we again observe that the fit provided by Lasso is the least accurate as the Q-Q plots contain relatively many points far from the theoretical line. From the Q-Q plots, we cannot clearly distinguish a difference in goodness-of-fit between BeSS and Relaxed Lasso.

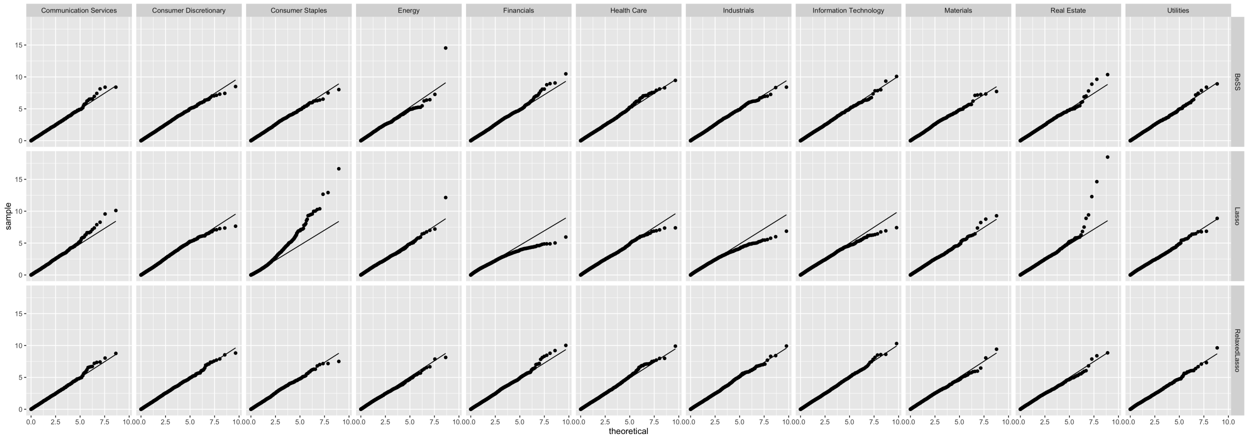

Figure 6: Q-Q plots of all 11 sectors estimated using all three methods. The threshold is chosen at the 95% quantile.

As BeSS provides the best fit we further inspect the selected covariates by BeSS for each sector. We observe that many of the relations with the selected covariates are ambiguous. This makes it difficult to interpret these relations. We observe the following notable relations: Communication Services and Consumer Discretionary are both negatively associated with the number of unemployed. This is a peculiar relationship, especially for Consumer Discretionary as it contains companies that produce non-essential consumer goods. For the sector Energy, we observe a negative relation with the price of Crude Oil indicating that an increase in the price of Crude Oil leads to a lower expected loss. Another notable relationship is seen for the sector Financials: we observe a negative relation with the total reserves of depository institutions, indicating that having more reserves leads to a lower expected loss for financial companies.

Table 12: Selected covariates by BeSS in the loss severity analysis for each sector. The relation is given in the parentheses.

Sector	Selected covariates
Communication Services	UEMP15T26 (-), USGOVT (+/-), T5YFFM (+/-), EXSZUSx (+/-)
Consumer Discretionary	IPBUSEQ (+/-), UEMP15T26 (-), USGOVT (+/-), ANDENOx (+), GS5 (-), DTCOLNVHFNM (+/-)
Consumer Staples	UEMP15T26 (+/-), S.P.div.yield (-), T5YFFM (-), VXOCLSx (+/-)
Energy	IPDCONGD (+/-), OILPRICEx (-)
Financials	IPBUSEQ (+/-), IPDMAT (+/-), ANDENOx (+/-), TOTRESNS (-)
Health Care	IPBUSEQ (+/-), UEMPLT5 (-), USGOVT (+/-), T5YFFM (-), DNDGRG3M086SBEA (+/-)
Industrials	IPBUSEQ (+/-), USGOVT (+/-), EXSZUSx (+/-)
Information Technology	IPBUSEQ (+/-), USGOVT (+/-), GS5 (-), VXOCLSx (-)
Materials	IPBUSEQ (+/-), UEMPLT5 (+/-), USGOVT (+)
Real Estate	IPBUSEQ (+/-), USGOVT (+), PERMITNE (+/-), T5YFFM (+/-)
Utilities	UEMP5TO14 (+), USGOVT (+), FEDFUNDS (+/-), T5YFFM (+/-)

A description of the variables can be found in the FRED-MD Appendix.
(+) denotes a positive relation.
(-) denotes a negative relation.
(+/-) denotes an ambiguous relation, i.e. the relations with ξ and ν are in opposite direction.

6 Conclusion

In this paper, we extend the dynamic POT approach with three covariate selection techniques: best subset selection, Lasso, and Relaxed Lasso. We test our methodology in a simulation study and assess the performance of all three covariate selection techniques. Finally, we apply our methodology on a market risk dataset. In this section, we present our main findings.

From the simulation study, we learn that all three proposed methods perform well in selecting covariates and estimating their coefficients for the loss frequency. We do not find that any methods are dominating or are being dominated by others. However, we find that performance drops when the threshold is chosen at a too low value. Moreover, when the threshold is chosen at a value above the actual threshold performance remains almost the same. The most likely reason for this is that when the threshold is chosen at a too low value non-extreme losses are used in the analysis as extreme losses. As the non-extreme losses come from a different DGP than the extreme losses the performance drops. When the threshold is chosen at a too high value information is lost as not all extreme losses are taken into account. However, there are no non-extreme losses used in the analysis. Therefore, the performance drop is little when there are still sufficient losses to analyze. Additionally, we find that even when the covariates are correlated the methods remain performing well in both selecting the correct covariates and estimating their coefficients.

For the loss severity, we find that Lasso tends to select too many covariates and BeSS tends to select a too sparse model. Relaxed Lasso tends to fit a model in between both, in terms of sparsity. For coefficient estimation accuracy we find that Lasso outperforms BeSS and Relaxed Lasso when the threshold is chosen at a value lower than the threshold in the DGP. When the threshold is chosen at a value above the threshold in the DGP performance in terms of coefficient estimation accuracy is very similar for all three methods. Again, we see that correlation in the covariates has little effect on the performance. For the threshold sensitivity, we see that both the estimation accuracy as the provided fit, assessed by the KS-test, drops dramatically when the threshold is chosen at a too low value. However, when the threshold is chosen at a value above the actual threshold both the estimation accuracy as the provided fit have little to no performance drop. Therefore, we conclude that it is better to choose the threshold at a relatively high value when the exact threshold is unknown. However, when the threshold is chosen at a too high value information is lost. Graphically inspecting the Q-Q plots and performing KS-tests for different thresholds should provide good insights.

In all of our applications, we find that the S&P Volatility Index is the most chosen covariate for both the loss frequency and severity. When inspecting the Q-Q plots for the loss severity and performing KS-tests we find that BeSS provides the best fits, closely followed by Relaxed Lasso. The most likely reason for this is that BeSS tends to select the sparsest model and Relaxed Lasso selects models that are sparser than the models selected by Lasso. From our applications, we conclude that the S&P Volatility Index has a positive relation with the frequency of extreme losses and the expected loss in case of such an extreme event.

7 Discussion and Further Research

The conclusions drawn in our paper are subject to some limitations. In the proposed methodology the threshold is assumed to remain constant in the whole analysis. Especially, when applying the methodology to a long time period this assumption might be violated. Therefore, in further research, the methodology could be extended by allowing for a (time-)varying threshold.

In the simulation study, the simulated covariates are drawn from a (multivariate) standard normal distribution. In practice, however, covariates may have distributions very different from the standard normal case. Moreover, our methods are only tested using continuously distributed covariates. We choose to simulate the non-extreme losses independently from a standard normal distribution (subject to being lower than the chosen threshold). For further research, it would be interesting to investigate the performance of the proposed methods when the non-extreme losses are also drawn from a non-homogeneous (dependent) distribution.

Chavez-Demoulin et al. (2016) developed their methodology for the use on operational risk data. As operational risk data is not readily available we choose to apply our methodology on market risk data. For further research, we would suggest applying our methodology on operational risk data if available.

References

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, *19*(6). doi: 10.1109/TAC.1974.1100705

Bertsimas, D., King, A., & Mazumder, R. (2016). Best subset selection via a modern optimization lens. *Annals of Statistics*, *44*(2). doi: 10.1214/15-AOS1388

Broyden, C. G. (1970). The convergence of a class of double-rank minimization algorithms 1. General considerations. *IMA Journal of Applied Mathematics (Institute of Mathematics and Its Applications)*, *6*(1). doi: 10.1093/imamat/6.1.76

Cawley, G. C., & Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, *11*.

Chavez-Demoulin, V., & Embrechts, P. (2004). Smooth extremal models in finance and insurance. *Journal of Risk and Insurance*, *71*(2). doi: 10.1111/j.0022-4367.2004.00085.x

Chavez-Demoulin, V., Embrechts, P., & Hofert, M. (2016). An Extreme Value Approach for Modeling Operational Risk Losses Depending on Covariates. *Journal of Risk and Insurance*, *83*(3). doi: 10.1111/jori.12059

Chavez-Demoulin, V., Embrechts, P., & Nešlehová, J. (2006). Quantitative models for operational risk: Extremes, dependence and aggregation. *Journal of Banking and Finance*, *30*(10). doi: 10.1016/j.jbankfin.2005.11.008

Chen, J., & Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, *95*(3). doi: 10.1093/biomet/asn034

Cizeau, P., Potters, M., & Bouchaud, J. P. (2001). Correlation structure of extreme stock returns. *Quantitative Finance*, *1*(2). doi: 10.1080/713665669

Cox, D. R., & Reid, N. (1987). Parameter Orthogonality and Approximate Conditional Inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, *49*(1). doi: 10.1111/j.2517-6161.1987.tb01422.x

Delgado, H., Anguera, X., Fredouille, C., & Serrano, J. (2015). Novel clustering selection criterion for fast binary key speaker diarization. In *Proceedings of the annual conference of the international speech communication association, interspeech* (Vol. 2015-Janua).

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., Ishwaran, H., Knight, K., … Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, *32*(2). doi: 10.1214/009053604000000067

Embrechts, P., Klüppelberg, C., & Mikosch, T. (1997). Modelling Extremal Events for Insurance and Finance. *ASTIN Bulletin*, *28*(2). doi: 10.2143/ast.28.2.519071

Flecther, R. (1970). New approach to variable metric algorithms. *Computer Journal*, *13*(3). doi: 10.1093/comjnl/13.3.317

Friedman, J., Hastie, T., Höfling, H., & Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, *1*(2). doi: 10.1214/07-aoas131

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1). doi: 10.18637/jss.v033.i01

Fu, W. J. (1998). Penalized regressions: The bridge versus the lasso? *Journal of Com-*

putational and Graphical Statistics, *7*(3), 397–416. doi: 10.1080/10618600.1998.10474784

Goldfarb, D. (1970). A family of variable-metric methods derived by variational means. *Mathematics of Computation*, *24*(109). doi: 10.1090/s0025-5718-1970-0258249-6

Gumbel, E. J. (1958). *Statistics of extremes*. Columbia university press.

Hastie, T., Tibshirani, R., & Tibshirani, R. (2020). Best Subset, Forward Stepwise or Lasso? Analysis and Recommendations Based on Extensive Comparisons. *Statistical Science*, *35*(4). doi: 10.1214/19-sts733

Hocking, R. R., & Leslie, R. N. (1967). Selection of the Best Subset in Regression Analysis. *Technometrics*, *9*(4). doi: 10.1080/00401706.1967.10490502

Ito, K., & Kunisch, K. (2014). A variational approach to sparsity optimization based on Lagrange multiplier theory. *Inverse Problems*, *30*(1). doi: 10.1088/0266-5611/30/1/015001

Jagannathan, R., & Wang, Z. (1996). The conditional CAPM and the cross-section of expected returns. *Journal of Finance*, *51*(1). doi: 10.1111/j.1540-6261.1996.tb05201.x

Joe, H. (2006). Generating random correlation matrices based on partial correlations. *Journal of Multivariate Analysis*, *97*(10). doi: 10.1016/j.jmva.2005.05.010

Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of Cluster in K-Means Clustering. *International Journal of Advance Research in Computer Science and Management Studies*, *1*(6).

Massey, F. J. (1951). The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*, *46*(253). doi: 10.1080/01621459.1951.10500769

McCracken, M. W., & Ng, S. (2016). FRED-MD: A Monthly Database for Macroeconomic Research. *Journal of Business and Economic Statistics*, *34*(4). doi: 10.1080/07350015.2015.1086655

Meinshausen, N. (2007). Relaxed Lasso. *Computational Statistics and Data Analysis*, *52*(1). doi: 10.1016/j.csda.2006.12.019

Moscadelli, M. (2011). The Modelling of Operational Risk: Experience with the Analysis of the Data Collected by the Basel Committee. *SSRN Electronic Journal*. doi: 10.2139/ssrn.557214

Mosteller, F., & Tukey, J. W. (1968). Data Analysis, Including Statistics. In *The handbook of social psychology: Vol. 2. research methods.*

Pickands, J. (1975). Statistical Inference using Extreme Order Statistics. *The Annals of Statistics*, *3*(1).

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2). doi: 10.2307/2958889

Shanno, D. F. (1970). Conditioning of quasi-Newton methods for function minimization. *Mathematics of Computation*, *24*(111). doi: 10.1090/s0025-5718-1970-0274029-x

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1). doi: 10.1111/j.2517-6161.1996.tb02080.x

Wen, C., Zhang, A., Wang, X., & Quan, S. (2020). Bess: An R package for best subset selection in linear, logistic and cox proportional hazards models. *Journal of Statistical Software*, *94*. doi: 10.18637/jss.v094.i04

Wood, S. N. (2017). *Generalized additive models: An introduction with R, second edition*. doi: 10.1201/9781315370279

Wu, T. T., & Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *Annals of Applied Statistics*, *2*(1). doi: 10.1214/07-AOAS147

8 Appendix

8.1 Methodology

8.1.1 Partial derivatives

$$\begin{aligned}\frac{\partial \ell(\boldsymbol{c}_\xi, \boldsymbol{c}_\nu)}{\partial \xi_t} &= 1/(1+\xi_t) + \ln(1+\xi_t(1+\xi_t)\exp(-\nu_t)y)/\xi_t^2 \\ &\quad - (1+1/\xi_t)\frac{(1+2\xi_t)\exp(-\nu_t)y}{1+\xi_t(1+\xi_t)\exp(-\nu_t)y} \qquad (40)\end{aligned}$$

$$\frac{\partial \ell(\boldsymbol{c}_\xi, \boldsymbol{c}_\nu)}{\partial \ln(\xi_t)} = \frac{\partial \ell(\boldsymbol{c}_\xi, \boldsymbol{c}_\nu)}{\partial \xi_t}\xi_t \qquad (41)$$

$$\begin{aligned}\frac{\partial^2 \ell(\boldsymbol{c}_\xi, \boldsymbol{c}_\nu)}{\partial \xi_t^2} &= -1/(1+\xi_t)^2 - 2\ln(1+\xi_t(1+\xi_t)\exp(-\nu_t)y)/\xi_t^3 \\ &\quad + \frac{2(1+2\xi_t)\exp(-\nu_t)y}{\xi_t^2(1+\xi_t(1+\xi_t)\exp(-\nu_t)y)} \\ &\quad - (1+1/\xi_t)\exp(-\nu_t)y\frac{2(1+\xi_t(1+\xi_t)\exp(-\nu_t)y) - (1+2\xi_t)^2\exp(-\nu_t)y}{(1+\xi_t(1+\xi_t)\exp(-\nu_t)y)^2} \qquad (42)\end{aligned}$$

$$\frac{\partial^2 \ell(\boldsymbol{c}_\xi, \boldsymbol{c}_\nu)}{\partial \ln(\xi_t)^2} = \frac{\partial^2 \ell(\boldsymbol{c}_\xi, \boldsymbol{c}_\nu)}{\partial \xi_t^2}\xi_t^2 + \frac{\partial \ell(\boldsymbol{c}_\xi, \boldsymbol{c}_\nu)}{\partial \xi_t}\xi_t \qquad (43)$$

$$\frac{\partial \ell(\boldsymbol{c}_\xi, \boldsymbol{c}_\nu)}{\partial \nu_t} = \frac{-1+(1+\xi_t)\exp(-\nu_t)y}{1+\xi_t(1+\xi_t)\exp(-\nu_t)y} \qquad (44)$$

$$\frac{\partial^2 \ell(\boldsymbol{c}_\xi, \boldsymbol{c}_\nu)}{\partial \nu_t^2} = \frac{-(1+\xi_t)^2\exp(-\nu_t)y}{(1+\xi_t(1+\xi_t)\exp(-\nu_t)y)^2} \qquad (45)$$

$$\begin{aligned}\frac{\partial^2 \ell(\boldsymbol{c}_\xi, \boldsymbol{c}_\nu)}{\partial \xi_t \nu_t} &= (1+1/\xi_t)\frac{(1+2\xi_t)y\exp(\nu_t)}{(\xi_t(1+\xi_t)y+\exp(\nu_t))^2} \\ &\quad - \frac{(1+\xi_t)y/\xi_t}{\xi_t(1+\xi_t)y+\exp(\nu_t)} \qquad (46)\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 \ell(\boldsymbol{c}_\xi, \boldsymbol{c}_\nu)}{\partial \ln(\xi_t)\nu_t} &= (\xi_t+1)\frac{(1+2\xi_t)y\exp(\nu_t)}{(\xi_t(1+\xi_t)y+\exp(\nu_t))^2} \\ &\quad - \frac{(1+\xi_t)y}{\xi_t(1+\xi_t)y+\exp(\nu_t)} \qquad (47)\end{aligned}$$

8.2 Simulation study

Table 13: Distributions of intercepts and coefficients

Coefficient	Distribution
$c_{0,\lambda}$	Uni$(2,4)$
$c_{0,\xi}$	Uni$(-2,0)$
$c_{0,\nu}$	Uni$(-2,1)$
$c_{\cdot,\cdot}$	$2(B-0.5)\times U$ where $U \sim$ Uni$(0.1,0.5)$ and $B \sim$ Bernoulli(0.5)

8.2.1 Results

Table 14: Median over 100 simulation of the performance measures for the simulation study with no correlation in the covariates. Threshold, u, is chosen at 90% quantile.

	Loss frequency			Loss severity		
Method	TP	TN	MSE	TP	TN	MSE
BeSS	**1.00**	0.98	**1.72**	**1.00**	**1.00**	**3.29**
Lasso	**1.00**	0.98	1.83	**1.00**	0.49	3.36
Relaxed Lasso	**1.00**	**1.00**	1.73	**1.00**	0.98	3.34

The reported MSE is scaled by 10^2.

Table 15: Mean over 100 simulation of the performance measures for the simulation study with no correlation in the covariates. Threshold, u, is chosen at 80% quantile.

	Loss frequency			Loss severity		
Method	TP	TN	MSE	TP	TN	MSE
BeSS	**0.86**	0.98	**6.32**	0.87	**1.00**	68.97
Lasso	0.82	0.97	6.43	**0.97**	0.40	**40.66**
Relaxed Lasso	0.78	**1.00**	6.36	0.95	0.77	48.10

The reported MSE is scaled by 10^2.

Table 16: Mean over 100 simulation of the performance measures for the simulation study with no correlation in the covariates. Threshold, u, is chosen at 95% quantile.

	Loss frequency			Loss severity		
Method	TP	TN	MSE	TP	TN	MSE
BeSS	**0.96**	0.98	1.23	0.92	**1.00**	2.48
Lasso	0.95	0.96	**1.22**	0.95	0.56	**2.41**
Relaxed Lasso	0.90	**1.00**	**1.22**	**0.97**	0.91	2.43

The reported MSE is scaled by 10^2.

Table 17: Median over 100 simulation of the performance measures for the simulation study with correlation in the covariates. Threshold, u, is chosen at 90% quantile.

	Loss frequency			Loss severity		
Method	TP	TN	MSE	TP	TN	MSE
BeSS	**1.00**	0.98	**1.12**	**1.00**	**1.00**	3.39
Lasso	**1.00**	0.96	1.26	**1.00**	0.55	**2.83**
Relaxed Lasso	**1.00**	**1.00**	1.17	**1.00**	0.91	2.87

The reported MSE is scaled by 10^2.

Table 18: Mean over 100 simulation of the performance measures for the simulation study with correlation in the covariates. Threshold, u, is chosen at 80% quantile.

	Loss frequency			Loss severity		
Method	TP	TN	MSE	TP	TN	MSE
BeSS	**0.85**	0.98	**5.89**	0.83	**0.99**	66.45
Lasso	0.82	0.96	6.02	**0.95**	0.46	**41.05**
Relaxed Lasso	0.77	**0.99**	5.93	0.94	0.71	45.97

The reported MSE is scaled by 10^2.

Table 19: Mean over 100 simulation of the performance measures for the simulation study with correlation in the covariates. Threshold, u, is chosen at 95% quantile.

	Loss frequency			Loss severity		
Method	TP	TN	MSE	TP	TN	MSE
BeSS	0.94	0.98	1.22	0.89	**1.00**	**2.59**
Lasso	**0.95**	0.93	**1.18**	0.95	0.57	2.73
Relaxed Lasso	0.90	**0.99**	1.21	**0.96**	0.82	2.67

The reported MSE is scaled by 10^2.

8.3 Application to market risk

8.3.1 S&P500 index

Table 20: Number of loss returns of the S&P500 index per month that exceeded the threshold, $u = 0.5\%$.

Year	January	February	March	April	May	June	July	August	September	October	November	December	Total
2011	2	2	7	2	7	7	7	8	10	6	8	6	72
2012	2	2	3	6	8	5	6	3	3	4	5	4	51
2013	0	4	2	4	5	8	0	6	2	4	1	1	37
2014	6	2	5	5	3	2	3	2	5	7	0	8	48
2015	10	0	6	2	4	6	4	8	6	2	3	10	61
2016	9	6	2	5	5	4	1	3	5	2	3	2	47
2017	1	0	2	1	1	3	1	2	1	0	1	1	14
2018	2	8	7	7	4	3	4	2	1	10	7	8	63
2019	4	1	3	1	9	1	4	8	4	3	0	3	41
2020	4	6	12	8	6	6	6	1	8	8	3	1	69

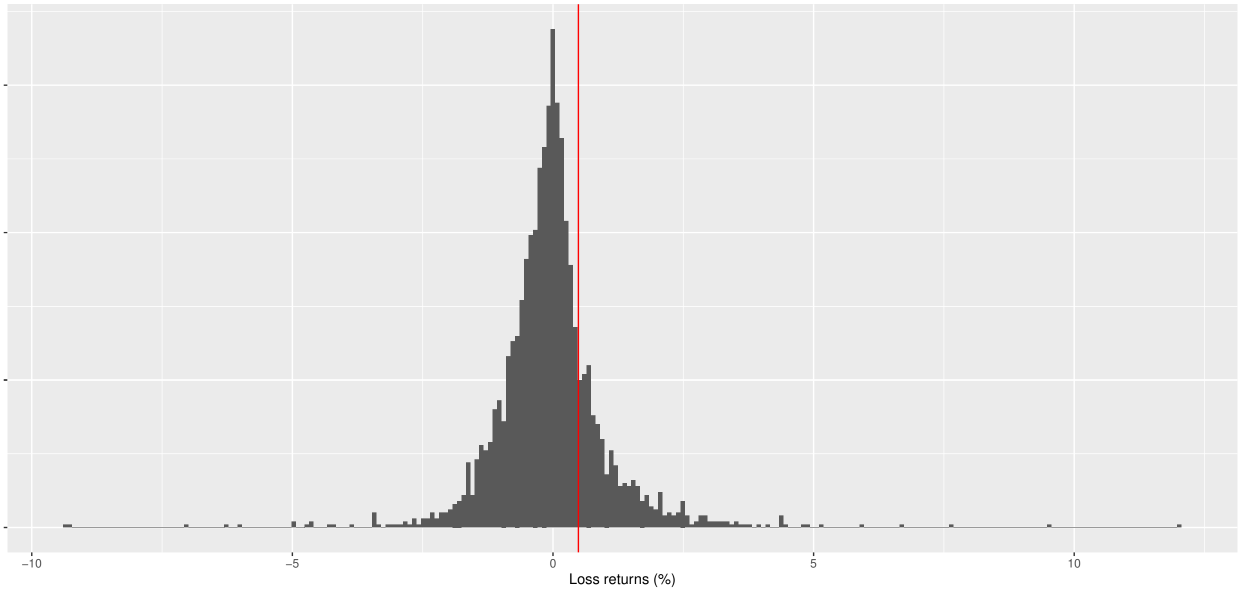

Figure 7: Histogram of the daily loss returns of the S&P500 index for the years 2011-2020. The threshold, $u = 0.5\%$, is marked in red.

8.3.2 Sector analysis

Table 21: Summary statistics of the number of losses exceeding the threshold per month.

Sector	Min	Median	Mean	Max
Communication Services	1	16.00	23.62	216
Consumer Discretionary	5	41.50	64.05	569
Consumer Staples	2	23.50	32.47	265
Energy	0	11.00	23.73	200
Financials	1	32.00	67.38	654
Health Care	14	43.00	64.98	509
Industrials	9	43.00	72.94	641
Information Technology	8	51.00	74.54	622
Materials	1	17.00	27.01	230
Real Estate	0	15.50	30.41	252
Utilities	1	20.00	29.36	252