Extreme value statistics using related variables

Ahmed, Hanan

DOI:
10.26116/a6ez-y955

Publication date:
2022

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Ahmed, H. (2022). *Extreme value statistics using related variables*. CentER, Center for Economic Research. https://doi.org/10.26116/a6ez-y955

CentER

Extreme value statistics using related variables

HANAN AHMED

Tilburg University

Extreme value statistics using related variables

Proefschrift

ter verkrijging van de graad van doctor aan Tilburg University op gezag van de rector magnificus, prof. dr. W.B.H.J. van de Donk, in het openbaar te verdedigen ten overstaan van een door het college voor promoties aangewezen commissie in de Aula van de Universiteit op vrijdag 10 juni 2022 om 10.00 uur door

Hanan Emad Galal Ahmed

geboren te Cairo, Egypte.

Promotores: Prof. dr. John H.J. Einmahl (Tilburg University)
Prof. dr. Chen Zhou (Erasmus University Rotterdam)

Overige Commissieleden: Prof. dr. Jan Beirlant (Katholieke Universiteit Leuven)
dr. Juan-Juan Cai (Vrije Universiteit)
Prof. dr. Johan Segers (Universite catholique de Louvain)
Prof. dr. Bas J.M. Werker (Tilburg University)

Dedicated to my mother

Amal Zohir..

Acknowledgments

Here comes the end of my Ph.D. journey. It was a big learning chapter in my life not only on an academic level but also on a personal level. I had the chance to learn many lessons, some of which were not easy. I was only able to do it with all the love and support of many people who were always around throughout to the end.

Four years ago when I decided to move to the Netherlands for my Ph.D., I had major support from the biggest figure in my life my mother Amal Zohir, however, how hard she felt to let me go. I want to dedicate this book to her as the strongest woman who is the reason and inspiration behind all the good parts of my life.

I want to start by thanking the main two pillars of this thesis, my supervisors Prof. dr. John H.J. Einmahl and Prof. dr. Chen Zhou. I am very thankful for all the help and support you have provided to me through the road. From day one when I arrived in the Netherlands and had John's help to carry my luggage to all the moments when John was very open to all kinds of discussions. John was a kind, thoughtful and supportive supervisor. Having Chen as a supervisor was a big addition to this thesis. I really appreciate all his advice and way of challenging ideas. I was also interested in Chen's discussions on a personal level which helped me in handling many situations. I'm grateful for all the experiences (including the hard ones) I had with my supervisors which were big learning moments for me.

It is a great honor for me to have Prof. dr. Jan Beirlant, dr. Juan-Juan Cai, Prof. dr. Johan Segers and Prof. dr. Bas J.M. Werker as my Ph.D. committee. Their comments and suggestions during my pre-defense were very helpful and led to this improved final version of my dissertation.

I am thankful to the Department of Econometrics and OR, Tilburg University for providing a great working environment and their great effort to keep maintaining that during the hard transition of the corona time. I want to thank Dr. Pavel Cizek for all his support, help, and effort till the end of my Ph.D. Samantha Koot from HR was very keen on all my requests, very supportive, and very willing to help.

I am very thankful for all my special office mates. Dr. Mánuel László Mágó thanks for all the very deep long discussions (they were really thoughtful as you are), all the help and support. Jorgo Goossens thanks for being around and supporting me many times. Xiaoyue Zhang thanks for being so kind and helpful. Mánuel, Jorgo, and Xiaoyue were great office mates, I was so lucky to have them. My office neighbor Emanuel Marcu who I am so grateful and thankful for all the moments he supported me and all things I have learned from him. I am also so grateful for all the fun times, events, and gathering with the rest of the department: Cliemans, Liz, Laura, Santiago, Ittai, Serhan, Suraj, Lieske, Kadircan, Lkhagvaa, Maciej. Maria Gantner, Rob Vermunt, Joris van Kempen, and Albert ten Have thanks so much for all the support through out and after my internship, it was such a great experience to work with you all.

I am so grateful and thankful for Prof. dr. Abd Elnasser Saad, who is a father to me. He never gave up or stopped supporting me from day one. He keeps encouraging me and pushing me forward. He is such an idol that I am always learning from and looking up to. To Marwa Sabry thanks for being such good friend no matter the distance. Thanks to all my Egyptian fellows in the Netherlands who were always helpful and around all times especially Enas and Haitham.

Finally, my deepest thanks and gratitude to my family: my parents (Amal and Emad), my siblings (Soha and Omar), Saber, Ahmed and my partner Wessam for always being by my side and supporting me no matter what I choose to do. I am so grateful for all the unconditional love and care you provide me with all the time.

Contents

Chapter 1

Introduction

Extreme value theory mainly handles data related to rare events, that are originated as a consequence of huge changes, such as earthquakes or large changes in asset pricing. Various fields have such applications where they use extreme value theory, for instance, meteorology (see de Haan (1990), Coles and Walshaw (1994) and Buishand et al. (2008)), health (Einmahl et al. (2019) and Thomas et al. (2016)), internet auctions (de Haan et al. (2009)), sports (Einmahl and Magnus (2008)), risk management in finance and insurance (Embrechts et al. (2013), Klüppelberg and Mikosch (1997) and Donnelly and Embrechts (2010)).

Extreme value theory is developed in parallel with the central limit theory, where extreme value theory is based on the limit behaviour of the sample maximum (or minimum) rather than the limit behaviour of the partial sums. Historically extreme value theory started in the 20th century, when Fisher and Tippett (1928) initiated the theory by formulating the limiting distribution of the sample maximum (or minimum). Through time extreme value theory becomes more specified in terms of the required conditions and characteristics of the limit distribution (see Von Mises (1936), Gnedenko (1943) and de Haan (1971)). Extreme value theory is introduced in the univariate and multivariate setting, see de Haan and Ferreira (2006) and Beirlant et al. (2004) for a comprehensive study of extreme value theory in both settings.

In this introduction we discuss some core ideas of extreme value theory in the univariate and multivariate setting, with focus on the notions needed in the thesis. We then preview the outline of the thesis and explain the main ideas in each chapter.

1.1 Univariate extreme value theory

Let $X_1, \dots, X_n$ be a random sample with distribution function F. Univariate extreme value theory encounters the case where there exists a sequence of positive numbers $a_n > 0$ and a sequence of real numbers b_n, such that as $n \to \infty$

$$\frac{\max\limits_{1 \le i \le n} X_i - b_n}{a_n} \xrightarrow{d} Z,$$

where Z has a non-degenerate distribution G. Equivalently, It considers that F belongs to the max-domain of attraction of an extreme value distribution, if as $n \to \infty$

$$\mathbb{P}\left(\frac{\max\limits_{1 \le i \le n} X_i - b_n}{a_n} \le x \right) = F^n(a_n x + b_n) \to G(x),$$

for every continuity point x of G, where G is a non-degenerate distribution and denoted as the extreme value distribution. Fisher and Tippett (1928) and Gnedenko (1943) first describe the extreme value distribution, that there exist $\gamma \in \mathbb{R}$, $a > 0$ and $b \in \mathbb{R}$ such that $G(x) = G_\gamma(ax + b)$, where

$$G_\gamma(x) = \begin{cases} \exp(-(1 + \gamma x)^{-1/\gamma}), & 1 + \gamma x > 0, \\ \exp(-e^{-x}), & \gamma = 0. \end{cases}$$

The parameter γ is known as the extreme value index, it describes the heaviness of the tail of the distribution. We distinguish between three types the tail of the distribution based on the sign of γ.

- If $\gamma > 0$, then the right endpoint of the distribution F is ∞ and that refers to a heavy tail distribution. Moments of order greater than $1/\gamma$ do not exist. The sequences a_n and b_n can be chosen such that G has a Fréchet$(1/\gamma)$ distribution. Examples of distributions in such domain of attraction are Pareto, Student and Cauchy distributions.

- If $\gamma = 0$, the right endpoint of the distribution F can be finite or infinite. The distribution has a light right tail and moments of any order exist. The sequences a_n and b_n can be chosen such that G has a Gumbel distribution. Examples of distributions in this domain of attraction are Exponential and Normal distributions.

- If $\gamma < 0$, the distribution has a finite endpoint. The sequences a_n and b_n can be chosen such that G has a reverse Weibull$(-1/\gamma)$ distribution. The uniform distribution is one of the distributions that belongs to such domain of attraction.

There are several estimators proposed for γ. For positive γ, the Hill (1975) estimator is the most used in the literatures. For $\gamma \leq 0$, there are other options, such as, the maximum likelihood estimator (Smith, 1987), the moment estimator (Dekkers et al., 1989), and the probability weighted moment estimator (Hosking and Wallis, 1987). Based on the estimator of the extreme value index, other important and widely used inferences can be obtained such as that for an extreme quantile: given a very small probability p, an extreme quantile is defined as $x_p = \inf\{x : 1 - F(x) \leq p\}$. For $\gamma \in \mathbb{R}$, the first order condition

$$\lim_{t\to\infty} \frac{U(tx) - U(t)}{a(t)} = \frac{x^\gamma - 1}{\gamma}, \tag{1.1}$$

where $U(.) = F^{-1}(1-1/.)$ is the tail quantile of the distribution function F, and $a(t) > 0$. The extreme quantile is derived based on (1.1), where for a small probability $p = p(n)$ such that $\lim\limits_{n\to\infty} p(n) = 0$,

$$\hat{x}_p = X_{n-k:n} + \hat{a}\left(\frac{n}{k}\right) \frac{(\frac{k}{np})^{\hat{\gamma}} - 1}{\hat{\gamma}},$$

where $X_{n-k:n}$ is the k^{th} order statistics of $\{X_1\}_{i=1}^n$, $k \in \{1, \ldots, n-1\}$, and $\hat{a}\left(\frac{n}{k}\right)$ is a proper estimator for $a\left(\frac{n}{k}\right)$.

1.2 Multivariate extreme value theory

Now we discuss the extension to the multiple dimension case. Multivariate extreme value theory determines the limit distribution of the componentwise maxima of a random vector. Let $(X_1, Y_{1,2}, \ldots, Y_{1,d}), \ldots, (X_n, Y_{n,2} \ldots, Y_{n,d})$ be a random sample from a $d-$variate distribution function F. Suppose there exist positive sequences $a_{n,1}, \ldots, a_{n,d}$ and sequences of real numbers $b_{n,1}, \ldots, b_{n,d}$, then F belongs to the max-domain of attraction, if as $n \to \infty$

$$F^n(a_{n,1}x_1 + b_{n,1}, a_{n,2}y_2 + b_{n,2} \ldots, a_{n,d}y_d + b_{n,d}) \to G(x),$$

for every continuity point $(x_1, y_2, \ldots, y_d)$ of G, which is known as the multivariate extreme value distribution.

Let $F_1, \ldots, F_d$ be the marginals of F. The multivariate max domain of attraction implies d univariate max-domain of attraction and the existence of the limit

$$\lim_{t \downarrow 0} t^{-1}\mathbb{P}(1-F_1(X_1) \leq tx_1 \text{ or } 1-F_d(Y_{1,2}) \leq ty_2 \ldots \text{ or } 1-F_d(Y_{1,d}) \leq ty_d) =: l(x, y_2, \ldots, y_d)$$

for all $(x, y_2, \ldots, y_d) \in [0, \infty)^d$. The function l is denoted as the tail dependence function, which describes the tail dependence structure. For a general review of the tail dependence, see Chapter 6 in de Haan and Ferreira (2006) and Huang (1992).

The d-variate distribution function F can be represented in terms of its marginal distribution functions and dependence structure, for instance, its copula C, then $F(x) = C(F_1(x), F_2(y) \ldots, F_d(y))$. The tail dependence function can then be

$$l(x, y_2, \ldots, y_d) = \lim_{t \downarrow 0} t^{-1}(1 - C(1 - tx, 1 - ty_2, \ldots, 1 - ty_d)),$$

Variable of Interest	X_1	X_2	.	.	X_n				
Covariate	Y_1	Y_2	.	.	Y_n	Y_{n+1}	.	.	Y_{n+m}

Table 1.1: Semi-supervised model

that shows the relation between the dependence structure and the tail dependence function. Unlike the univariate extreme value theory there is no unique way to define the extreme quantile for the multivariate distribution.

1.3 Outline of the thesis

The thesis bundles four chapters, where we propose novel methods to improve estimation of extreme inferences using related covariate(s). The thesis is motivated by the fact of the rarity of the extreme data, which significantly affects the quality of the parameters estimation.

In Chapter 2, 3 and 4 we use the semi-supervised model (SSM), which is represented as follows. The variable of interest (or target variable) is available for only n observations, known as labeled data. For the labeled data, we also observe the covariate(s) (or related variables). In addition, there are extra m observations for the covariate denoted as unlabeled data. Table 1.1 provides an illustration for the SSM in the bivariate case, i.e. one covariate. It can be extended to the multivariate case by adding more covariates

Chapter 2. Improved estimation of the extreme value index using related variables. A new improved estimator for the positive extreme value index ($\gamma > 0$) is introduced. In Chapter 2, we first consider the SSM in the bivariate case, where the variable of interest has a heavy tailed distribution with extreme value index γ_1, and the related variable similarly has a heavy tailed distribution with extreme value index γ_2. We assume that there exists tail dependence between the two variables, defined as

$$R(x,y) = \lim_{t \downarrow 0} \frac{1}{t} \mathbb{P}\left(1 - F_1\left(X_1\right) \leq tx, 1 - F_2\left(Y_1\right) \leq ty\right),$$

where $R(x,y) = x + y - l(x,y)$. We propose new adaptation for the extreme value index of the variable of interest as

$$\hat{\gamma}_{1,2} = \hat{\gamma}_1 + \frac{\hat{\gamma}_1}{\hat{\gamma}_{2+}} \left(\frac{\hat{R}(1,1) - \frac{k}{k_+}\hat{R}(1, \frac{k_+}{k}\frac{n}{n+m})}{1 + \frac{k}{k_+} - 2\frac{n}{n+m}} \right) (\hat{\gamma}_{2+} - \hat{\gamma}_2), \tag{1.2}$$

where $\hat{\gamma}_1, \hat{\gamma}_2$ and $\hat{\gamma}_{2+}$ are estimated using the Hill estimator, the tail copula R is estimated empirically (see Drees and Huang (1998)), $k \in \{1, \ldots, n-1\}$ and $k_+ \in \{k+1, \ldots, n+m-1\}$. We prove the asymptotic distribution of our adapted estimator, where the asymptotic variance is reduced substantially and the asymptotic bias can be kept at the same level as that of the Hill estimator. We observe that the main two key factors which contribute to improving the performance of our adapted estimator, especially in terms of the variance reduction, are: the availability of more observations of the related variable and the existence of high tail dependence between the variable of interest and related variable. Our results are extended to the multivariate data setting with proof for the asymptotic distribution of the adapted estimator. An extensive simulation study is conducted which confirms the improved performance of our adapted estimator. We finally present an application on the earthquake financial losses, where we show the use of the our adapted estimator and estimate the extreme quantile based on the adapted estimator.

Chapter 3 Extreme value statistics in semi-supervised models. We use the SSM to obtain the semi-supervised estimator (SSE) for the extreme value index when $\gamma > -\frac{1}{2}$. We assume that only the distribution of the variable of interest belongs to the max-domain of attraction with an extreme value index ($\gamma > -\frac{1}{2}$). We choose a number g that mimics the extreme value index for the covariate, then transform the labelled data of the covariate empirically based on the labelled and unlabelled data and g. Assuming that there is tail dependence between the variable of interest and the related covariate, denoted as $R(x, y)$, the SSE of the extreme value index

$$\hat{\gamma}_g = \hat{\gamma} - \frac{1+\hat{\gamma}}{1+g} \hat{R}_g(\hat{g} - g),$$

where $\hat{\gamma}$ and $\hat{g}$ are the pseudo maximum likelihood estimators (pseudo-MLEs) (Smith, 1987) for γ and g,

$$\hat{R}_g = \hat{R}(1,1) + \frac{g-\hat{\gamma}}{\hat{\gamma}+g+1} \left((2\hat{\gamma}+1) \int_0^1 \frac{\hat{R}(s,1)}{s^{1-\hat{\gamma}}} ds - (2g+1) \int_0^1 \frac{\hat{R}(1,t)}{t^{1-g}} dt \right),$$

and $\hat{R}$ is estimated empirically (see Drees and Huang (1998)).
Chapter 3 provides a general results for the tail quantile process of the variable of interest and the non-standard tail quantile process of the related covariate. Based on the tail quantile results, we proved the asymptotic distribution for the SSE of the extreme value index, where the asymptotic bias is kept the same as that of the pseudo-MLE. We then observe the performance of the SSE in terms of the asymptotic variance reduction, which depends on the known g and the unknown γ and R. Through a detailed simulation study we show the amount of variance reduction when using the SSE compared to the pseudo-MLE for the extreme value index. The asymptotic results are introduced and proved for the multivariate setting. An application to railfall in France is used to demonstrate the use of the SSE.

Chapter 4 Extreme quantile estimation in semi-supervised models. In this chapter we extend the use of the SSM to estimate the extreme quantile. Improving the estimation of the extreme quantile is very crucial in many applications as it affects the decision reliability. For $\gamma > -\frac{1}{2}$, Let x^* be the right endpoint of F_1, $x^* = \sup\{x : F_1(x) < 1\}$, and define the excess distribution function

$$F_{1t}(x) = \mathbb{P}(X \leq x + t | X > t) = \frac{F_1(t+x) - F_1(t)}{1 - F_1(t)}, x > 0.$$

The distribution function F_1 belongs to the max-domain of attraction ($F_1 \in D(G_\gamma)$) if

$$\lim_{t \to x^*} F_{1t}(x\sigma(t)) = H_\gamma(x), \tag{1.3}$$

where $\sigma(t)$ is a positive scale function and $H_\gamma(x)$ is the generalized Pareto distribution. Under the SSM in Chapter 3, we start by obtaining SSE for the scale parameter of the variable of interest $\sigma\left(U_1(\frac{n}{k})\right)$. Based on the transformation of the labelled data of the related covariate, as in Chapter 3, the number $\left(\frac{n}{k}\right)^g$ mimics the scale of the covariate. The SSE for the scale is

$$\hat{\sigma}_g = \hat{\sigma}\left(1 - \frac{\hat{S}_g}{1+(1+g)^2}\left(\frac{\tilde{\sigma}_g}{(\frac{n}{k})^g} - 1\right)\right),$$

where $\hat{\sigma}, \tilde{\sigma}_g$ are the pseudo-MLE for σ and $\left(\frac{n}{k}\right)^g$, and $\hat{S}_g$ is a function of γ, g, and R. Then the SSE for the extreme quantile is obtained, using the SSE for the extreme value index and the scale, as

$$\hat{x}_{p_g} = \hat{U}_1\left(\frac{n}{k}\right) + \hat{\sigma}_g \frac{(\frac{k}{np})^{\hat{\gamma}_g} - 1}{\hat{\gamma}_g}.$$

The asymptotic distribution for the SSE of the scale and the extreme quantile are proved using the tail quantile processes which introduced in Chapter 3. The asymptotic distributions have the same bias as the asymptotic distributions based on the pseudo-MLE, we observe the asymptotic variance to measure the performance of the SSEs. We present an extensive simulation study where we show the obtained variance reduction using the SSE for the scale and the extreme quantile. Additionally we show that using the SSE for the extreme value index in the extreme quantile is sometimes not enough to obtain variance reduction, and the use of the SSE for the scale is important in such cases. The results for the multivariate setting are provided for both the SSE for the scale and the extreme quantile.

Chapter 5 Insurance risk and machine learning: Estimating conditional Value-at-Risk using random forest. We extend the peak over threshold approach (POT) to include a large number of covariates in the parameters when modelling heavy tailed response variable. In this chapter, we consider a d-dimensional covariate Y and a response variable X that has a distribution function F, such that $(X, Y) \in \mathbb{R}^{d+1}$ is dependent random vector. We assume that F belongs to the max-domain of attraction with a positive extreme value index if

$$\lim_{t\to\infty} \mathbb{P}(X > tx|X > t, Y = y) = x^{-1/\gamma(y)},$$

where

$$\gamma(y) = \lim_{t\to\infty} E\left(\log\left(\frac{X}{t}\right)\middle| X > t, Y = y\right),$$

is a continuous function $\mathbb{R}^d \to \mathbb{R}^+$. We define the conditional VaR_α of X given a set of covariates $Y = y$, as $VaR_\alpha(y) := VaR_\alpha(X|Y = y)$, which satisfies $\mathbb{P}(X \geq VaR_\alpha(y)|Y = y) = 1 - \alpha$. For a high threshold u,

$$\frac{\mathbb{P}(X \geq VaR_\alpha(y)|Y = y)}{\mathbb{P}(X \geq u|Y = y)} \approx \left(\frac{VaR_\alpha(y)}{u}\right)^{-1/\gamma(y)}.$$

Define

$$g(y) = \mathbb{P}(X \geq u|Y = y) = E(I_{X>u}|Y = y), \tag{1.4}$$

where $I[.]$ is an indicator function that equals to 1 when $X > u$ and 0 otherwise and

$$\gamma(y) \approx E\left(\log\left(\frac{X}{u}\right)\middle| X > u, Y = y\right), \tag{1.5}$$

then

$$\widehat{VaR_\alpha(y)} = u\left(\frac{\hat{g}(y)}{1-\alpha}\right)^{\hat{\gamma}(y)},$$

where $\hat{g}(y)$ and $\hat{\gamma}(y)$ are proper estimators for $g(y)$ and $\gamma(y)$. Based on (1.4) and (1.5), we estimate $g(y)$ and $\gamma(y)$ using the random forest classification and regression models.

We asses the performance of the proposed methodology using an extensive simulation study, where the rooted mean squared error turns to be low for the estimators of $\gamma(y)$ and $g(y)$, and consequently $VaR_\alpha(y)$. We show the strong ability of the random forest models to detect the top important variables that affect model estimation. Finally we apply our methodology on a claim loss dataset from anonymized insurance company which contains a large number of covariates. The methodology performance is evaluated using two backtests, where we observe that using the top important covariates based on the random forest models substantially affects the backtesting results.

Chapter 2

Improved estimation of the extreme value index using related variables

[Based on joint work with John H.J. Einmahl, *Improved estimation of the extreme value index using related variables*, Extremes 2019, 22, 553-569.]

Abstract. Heavy tailed phenomena are naturally analyzed by extreme value statistics. A crucial step in such an analysis is the estimation of the extreme value index, which describes the tail heaviness of the underlying probability distribution. We consider the situation where we have next to the n observations of interest another $n+m$ observations of one or more related variables, like, e.g., financial losses due to earthquakes and the related amounts of energy released, for a longer period than that of the losses. Based on such a data set, we present an adapted version of the Hill estimator. For this adaptation the tail dependence between the variable of interest and the related variable(s) plays an important role. We establish the asymptotic normality of this estimator and find that it shows greatly improved behavior relative to the Hill estimator, in particular the asymptotic variance is substantially reduced, whereas in the natural setting the asymptotic bias remains unchanged. A simulation study confirms the substantially improved performance of our adapted estimator. We also present an application to the aforementioned earthquake losses.

Key words. Asymptotic normality, Heavy tail, Hill estimator, Tail dependence, Variance reduction

2.1 Introduction

Consider univariate extreme value theory for heavy tails, that is, the case where the extreme value index γ is positive. This index describes the tail heaviness of the underlying probability distribution, the larger γ, the heavier the tail. See de Haan and Ferreira (2006) or Beirlant et al. (2004) for a comprehensive introduction to univariate and multivariate extreme value theory and Gomes and Guillou (2015) for a more recent review of the univariate case. Given a random sample, we often estimate γ with the well-known Hill (1975) estimator. Such an estimate of γ is the crucial ingredient for estimating important tail functionals of the distribution, like very high quantiles, very small tail probabilities, but also the expected shortfall or an excess-of-loss reinsurance premium.

In this paper we first consider the semi-supervised model (SSM) in the bivariate case. For the data structure of the SSM in the bivariate case, see Section 1.3 and Table 1.1. We assume that the variable of interest has extreme value index $\gamma_1 > 0$ and the related variable is a heavy-tailed, with extreme value index $\gamma_2 > 0$, that should help to improve the estimation of γ_1. The m observations of the related variable are independent of the pairs and mutually independent. Such a situation occurs in, e.g., an insurance setting when we have recorded both variables for a certain period of time (2008-2017, say), but in addition have data for the second variable only, for an earlier period (1980-2007, say). We can think of financial losses as the variable of interest and some physical quantity (like wind speed, air pressure, earthquake magnitude, water height) as the related variable (see Section 2.5). The independence assumption between the n pairs and m earlier observations is then naturally fulfilled. Specifically, the situation with hurricane losses as variable of interest and (transformed) air pressures as related variable was brought to our attention by a reinsurance company. A related situation where our setup can occur is when in a certain period the related variable is measured more frequently than the variable of interest. Also in a cross-sectional context our setup can be relevant. E.g., in a medical setting it often happens that for a group of n patients a specific variable is measured together with one (or more) other variable(s), whereas for a (larger) group of m patients, due to cost constraints, the specific variable is not measured, only the related variable(s), see Chakrabortty and Cai (2018). We will also, as suggested in the medical setting, consider the situation where there is *more* than one related variable, the multivariate case, but in this introductory section we will focus on the bivariate case.

We can estimate γ_1 with the Hill estimator $\hat{\gamma}_1$ and γ_2 with the Hill estimators $\hat{\gamma}_2$, based on the n data, and $\hat{\gamma}_{2+}$, based on all $n+m$ data. The latter estimator is better than $\hat{\gamma}_2$, “hence” their difference can be used to update and improve $\hat{\gamma}_1$. For this updating the strength of the tail dependence between both variables is important and should be estimated. A detailed derivation of our adapted Hill estimator is presented in the next section. We will show that our estimator improves greatly on the Hill estimator, in particular the asymptotic variance is substantially reduced, whereas in the natural setting the asymptotic bias remains unchanged. To the best of our knowledge this approach is novel and there are no results of this type in the literature.

The remainder of this paper is organized as follows. In Section 2.2, for the clearness of the exposition, the bivariate case is treated as indicated above and the asymptotic normality of the adapted estimator is established and in Section 2.3 the corresponding results for the multivariate case are presented. In Section 2.4, the finite sample performance of our estimator is studied through a simulation study, which confirms the improved performance of the adapted Hill estimator. In Section 2.5, we present an application to earthquake damage amounts with the “amount of energy released” as related variable. The proofs of the results in Section 2.3 are deferred to Section 2.6. Since Section 2.3 generalizes Section

2.2, the proofs of Section 2.2 can be obtained by specializing those of Section 2.3, and are hence omitted.

2.2 Main Results: the bivariate case

Let F be a bivariate distribution function with marginals F_1 and F_2. Assume that F is in the bivariate max-domain of attraction (i.e., $F \in D(G)$) with both extreme value indices γ_1 and γ_2 *positive*, see Chapter 6 in de Haan and Ferreira (2006). Let $U_j(\cdot) = F_j^{-1}(1 - 1/\cdot)$ be the tail quantile corresponding to $F_j, j = 1, 2$. Then $F \in D(G)$ with positive extreme value indices implies that U_j is regularly varying with index γ_j, $j = 1, 2$, i.e., $\lim_{t\to\infty} U_j(tx)/U_j(t) = x^{\gamma_j}, x > 0$. Let (X_1, Y_1) have distribution function F. Then $F \in D(G)$ also implies the existence of the tail copula R defined by

$$R(x, y) = \lim_{t\downarrow 0} \frac{1}{t}\mathbb{P}\left(1 - F_1\left(X_1\right) \le tx, 1 - F_2\left(Y_1\right) \le ty\right), \tag{2.1}$$

$(x, y) \in [0, \infty]^2 \setminus \{(\infty, \infty)\}$. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a bivariate random sample from F, and let $Y_{n+1}, \ldots, Y_{n+m}$ be a univariate random sample from F_2, independent from the n pairs. Denote the order statistics of the $X_i, i = 1, \ldots, n$, with $X_{1,n} \le \ldots \le X_{n,n}$ and use similar notation for the order statistics of the $Y_i, i = 1, \ldots, n$, and also for the order statistics of all the $Y_i, i = 1, \ldots, n + m$. For $k \in \{1, \ldots, n - 1\}$ define the Hill (1975) estimator of γ_1 by

$$\hat{\gamma}_1 = \frac{1}{k}\sum_{i=0}^{k-1} \log X_{n-i,n} - \log X_{n-k,n}. \tag{2.2}$$

Define, using the same k, similarly the Hill estimator $\hat{\gamma}_2$ based on the $Y_i, i = 1, \ldots, n$:

$$\hat{\gamma}_2 = \frac{1}{k}\sum_{i=0}^{k-1} \log Y_{n-i,n} - \log Y_{n-k,n}. \tag{2.3}$$

Also, let $\hat{\gamma}_{2+}$ be the Hill estimator of all $Y_i, i = 1, \ldots, n + m$, with k replaced by $k_+ \in \{k + 1, \ldots, n + m\}$:

$$\hat{\gamma}_{2+} = \frac{1}{k_+}\sum_{i=0}^{k_+-1} \log Y_{n+m-i,n+m} - \log Y_{n+m-k_+,n+m}. \tag{2.4}$$

Throughout for the asymptotical theory we will assume that $m = m(n)$ and that

$$k \to \infty,\ \frac{k}{n} \to 0,\ \sqrt{\frac{k}{k_+}} \to \nu \in (0,1),\ \frac{n}{n+m}\frac{k_+}{k} \to \beta \in (0,1], \quad \text{as } n \to \infty. \tag{2.5}$$

Observe that we now also have $k_+ \to \infty$, $m \to \infty$, and $k_+/(n+m) \to 0$; actually $n/(n+m) \to \beta\nu^2 \in (0,1)$.

First we consider the joint asymptotic normality of the three Hill estimators $\hat{\gamma}_1 = \hat{\gamma}_1(k), \hat{\gamma}_2 = \hat{\gamma}_2(k)$, and $\hat{\gamma}_{2+} = \hat{\gamma}_{2+}(k_+)$. For this, we need the usual second order conditions, on F_1 and F_2: there exist positive or negative functions $A_j, j = 1,2$, with $\lim_{t\to\infty} A_j(t) = 0$, such that for $x > 0$

$$\lim_{t\to\infty} \frac{\frac{U_j(tx)}{U_j(t)} - x^{\gamma_j}}{A_j(t)} = x^{\gamma_j}\frac{x^{\rho_j}-1}{\rho_j}, \quad \text{for some } \rho_j \le 0,\ j = 1,2. \tag{2.6}$$

Proposition 2.2.1 *If $F \in D(G)$, conditions (2.5) and (2.6) hold, and $\sqrt{k}A_j(\frac{n}{k}) \to \lambda_j \in \mathbb{R}, j = 1,2,$ as $n \to \infty$, then*

$$\left(\sqrt{k}(\hat{\gamma}_1-\gamma_1), \sqrt{k}(\hat{\gamma}_2-\gamma_2), \sqrt{k_+}(\hat{\gamma}_{2+}-\gamma_2)\right) \xrightarrow{d} N\left(\left(\frac{\lambda_1}{1-\rho_1}, \frac{\lambda_2}{1-\rho_2}, \frac{\lambda_2\beta^{-\rho_2}}{\nu(1-\rho_2)}\right), \breve{\Sigma}\right), \tag{2.7}$$

with, see (2.1),

$$\breve{\Sigma} = \begin{bmatrix} \gamma_1^2 & R(1,1)\gamma_1\gamma_2 & \nu R(1,\beta)\gamma_1\gamma_2 \\ R(1,1)\gamma_1\gamma_2 & \gamma_2^2 & \nu\beta\gamma_2^2 \\ \nu R(1,\beta)\gamma_1\gamma_2 & \nu\beta\gamma_2^2 & \gamma_2^2 \end{bmatrix}.$$

Corollary 2.2.1 *Under the conditions of Proposition 2.2.1, as $n \to \infty$,*

$$\left(\sqrt{k}(\hat{\gamma}_1-\gamma_1), \sqrt{k}(\hat{\gamma}_{2+}-\hat{\gamma}_2)\right) \xrightarrow{d}$$
$$N\left(\left(\frac{\lambda_1}{1-\rho_1}, \frac{\lambda_2(\beta^{-\rho_2}-1)}{1-\rho_2}\right), \begin{bmatrix} \gamma_1^2 & \left(\nu^2R(1,\beta)-R(1,1)\right)\gamma_1\gamma_2 \\ \left(\nu^2R(1,\beta)-R(1,1)\right)\gamma_1\gamma_2 & \left(1+\nu^2-2\nu^2\beta\right)\gamma_2^2 \end{bmatrix}\right).$$

Corollary 2.2.1 is the basis for deriving our adapted Hill estimator. For this derivation only, take $\lambda_1 = \lambda_2 = 0$. The tail copula R is estimated as usual, cf. Drees and Huang (1998), by

$$\hat{R}(x,y) = \frac{1}{k}\sum_{i=1}^{n} 1_{[X_i \geq X_{n-[kx]+1,n}, Y_i \geq Y_{n-[ky]+1,n}]}, \quad x, y \geq 0. \tag{2.8}$$

Now consider $(\hat{\gamma}_1, \hat{\gamma}_{2+} - \hat{\gamma}_2)$ and its approximate bivariate normal distribution according to Corollary 2.2.1, with estimated covariance matrix:

$$N\left((\gamma_1, 0), \frac{1}{k}\begin{bmatrix} \hat{\gamma}_1^2 & (\frac{k}{k_+}\hat{R}(1, \frac{k_+}{k}\frac{n}{n+m}) - \hat{R}(1,1))\hat{\gamma}_1\hat{\gamma}_{2+} \\ (\frac{k}{k_+}\hat{R}(1, \frac{k_+}{k}\frac{n}{n+m}) - \hat{R}(1,1))\hat{\gamma}_1\hat{\gamma}_{2+} & \left(1 + \frac{k}{k_+} - 2\frac{n}{n+m}\right)\hat{\gamma}_{2+}^2 \end{bmatrix}\right). \tag{2.9}$$

Maximizing this approximate likelihood of the single observation $(\hat{\gamma}_1, \hat{\gamma}_{2+} - \hat{\gamma}_2)$ with respect to γ_1, we obtain our adapted estimator for γ_1:

$$\hat{\gamma}_{1,2} = \hat{\gamma}_1 + \frac{\hat{\gamma}_1}{\hat{\gamma}_{2+}}\left(\frac{\hat{R}(1,1) - \frac{k}{k_+}\hat{R}(1, \frac{k_+}{k}\frac{n}{n+m})}{1 + \frac{k}{k_+} - 2\frac{n}{n+m}}\right)(\hat{\gamma}_{2+} - \hat{\gamma}_2). \tag{2.10}$$

The main result of this section, the asymptotic normality of this estimator, shows that it improves substantially on the Hill estimator.

Theorem 2.2.1 *Under the conditions of Proposition 2.2.1, as $n \to \infty$,*

$$\sqrt{k}(\hat{\gamma}_{1,2} - \gamma_1) \xrightarrow{d}$$
$$N\left(\frac{\lambda_1}{1-\rho_1} + \frac{\gamma_1}{\gamma_2}\cdot\frac{R(1,1) - \nu^2 R(1,\beta)}{1 + \nu^2 - 2\nu^2\beta}\cdot\frac{\lambda_2(\beta^{-\rho_2} - 1)}{1-\rho_2}, \gamma_1^2\left[1 - \frac{(R(1,1) - \nu^2 R(1,\beta))^2}{1 + \nu^2 - 2\nu^2\beta}\right]\right).$$

Remark 1 Note that in case $\rho_1 \neq \rho_2$, we have, since $|A_j|$ is regularly varying at ∞ with index ρ_j, $j = 1, 2$, that $\lambda_1 = 0$ or $\lambda_2 = 0$. Hence in this case the expression for the asymptotic bias is simplified. In case $\lambda_2 = 0$ (which is implied by $\rho_1 > \rho_2$) or $\beta = 1$ or $\rho_2 = 0$, the Hill estimator and the adapted estimator have the same asymptotic bias $\lambda_1/(1 - \rho_1)$.

We highlight the natural choice $\beta = 1$ in the following corollary.

Corollary 2.2.2 *Under the conditions of Proposition 2.2.1 with $\beta = 1$, as $n \to \infty$,*

$$\sqrt{k}(\hat{\gamma}_{1,2} - \gamma_1) \xrightarrow{d} N\left(\frac{\lambda_1}{1-\rho_1}, \gamma_1^2\left[1 - (1-\nu^2)R^2(1,1)\right]\right).$$

Remark 2 Since the asymptotic biases of both estimators are the same now, we can in the comparison focus on the asymptotic variances. Clearly the asymptotic variance of the adapted Hill estimator never exceeds the γ_1^2 of the classical Hill estimator. The (relative) variance reduction is equal to $(1-\nu^2)R^2(1,1)$, which is positive in case of tail dependence, i.e., $R(1,1) > 0$. When, e.g., $m = n$ and $k_+ = 2k$, this becomes $\frac{1}{2}R^2(1,1)$. Then, depending on the value of $R(1,1) \in [0,1]$, the variance reduction can be as large as 50%. In case of tail independence ($R(1,1) = 0$), the estimators have the same asymptotic variances. In such a case a "better" related variable should be looked for.

Remark 3 It is well-known that choosing a good k is a difficult problem in extreme value theory. We will not address this problem here, but compare for many values of k our adapted estimator and the Hill estimator, see Remark 2, Remark 4, and the simulation section. On the other hand, there are many methods for choosing the k of the Hill estimator, see, e.g., Caeiro and Gomes (2015). If one of these methods *for the Hill estimator itself* is used, we can choose the same k for our adapted estimator and obtain the discussed improvements.

2.3 Main results: the multivariate case

Now we consider a d-variate distribution function F, with marginals $F_1, \dots, F_d$ and corresponding tail quantile functions U_j, $j = 1, \dots d$; write F_- for the distribution function of the last $d-1$ components of a random vector with distribution function F. We assume that F is in the multivariate max-domain of attraction, that is $F \in D(G)$, with all extreme value indices $\gamma_1, \dots, \gamma_d$ positive. Let R_{ij} be the tail copula of the i-th and the j-th component, $1 \le i, j \le d, i \ne j$, see (2.1).

Let $(X_1, Y_{1,2}, \ldots, Y_{1,d}), \ldots, (X_n, Y_{n,2}, \ldots, Y_{n,d})$, be a d-variate random sample from F and let

$(Y_{n+1,2}, \ldots, Y_{n+1,d}), \ldots, (Y_{n+m,2}, \ldots, Y_{n+m,d})$ be a $(d-1)$-variate random sample from F_{-}, independent of the d-variate random sample of size n. Let $\hat{\gamma}_1, \hat{\gamma}_j$, and $\hat{\gamma}_{j+}$ be the Hill estimators based on $X_1, \ldots, X_n$, $Y_{1,j}, \ldots, Y_{n,j}$, and $Y_{1,j}, \ldots, Y_{n+m,j}$, $j = 2, \ldots, d$, respectively, cf. (2.2), (2.3), and (2.4); here again we replace k with k_+ for $\hat{\gamma}_{j+}, j = 2, \ldots, d$. First we consider the joint asymptotic normality of all the $2d-1$ Hill estimators.

Proposition 2.3.1 *If $F \in D(G)$, condition (2.5) holds, condition (2.6) holds for $j = 1, \ldots, d$, and $\sqrt{k}A_j(\frac{n}{k}) \to \lambda_j \in \mathbb{R}, j = 1, \ldots, d$, as $n \to \infty$, then*

$$\Big(\sqrt{k}(\hat{\gamma}_1 - \gamma_1), \sqrt{k}(\hat{\gamma}_2 - \gamma_2), \sqrt{k_+}(\hat{\gamma}_{2+} - \gamma_2), \ldots, \sqrt{k}(\hat{\gamma}_d - \gamma_d), \sqrt{k_+}(\hat{\gamma}_{d+} - \gamma_d)\Big) \xrightarrow{d} N(\breve{\mu}_d, \breve{\Sigma}_d), \tag{2.11}$$

where

$$\breve{\mu}_d = \left(\frac{\lambda_1}{1-\rho_1}, \frac{\lambda_2}{1-\rho_2}, \frac{\lambda_2 \beta^{-\rho_2}}{\nu(1-\rho_2)}, \ldots, \frac{\lambda_d}{1-\rho_d}, \frac{\lambda_d \beta^{-\rho_d}}{\nu(1-\rho_d)}\right),$$

$$\breve{\Sigma}_d = \begin{bmatrix} \gamma_1^2 & R_{12}(1,1)\gamma_1\gamma_2 & \nu R_{12}(1,\beta)\gamma_1\gamma_2 & \cdots & R_{1d}(1,1)\gamma_1\gamma_d & \nu R_{1d}(1,\beta)\gamma_1\gamma_d \\ R_{12}(1,1)\gamma_1\gamma_2 & \gamma_2^2 & \nu\beta\gamma_2^2 & \cdots & R_{2d}(1,1)\gamma_2\gamma_d & \nu R_{2d}(1,\beta)\gamma_2\gamma_d \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ R_{1d}(1,1)\gamma_1\gamma_d & R_{2d}(1,1)\gamma_2\gamma_d & \nu R_{2d}(1,\beta)\gamma_2\gamma_d & \cdots & \gamma_d^2 & \nu\beta\gamma_d^2 \\ \nu R_{1d}(1,\beta)\gamma_1\gamma_d & \nu R_{2d}(1,\beta)\gamma_2\gamma_d & R_{2d}(1,1)\gamma_2\gamma_d & \cdots & \nu\beta\gamma_d^2 & \gamma_d^2 \end{bmatrix}.$$

Corollary 2.3.1 *Under the conditions of Proposition 3.3.1, as $n \to \infty$,*

$$\Big(\sqrt{k}(\hat{\gamma}_1 - \gamma_1), \sqrt{k}(\hat{\gamma}_{2+} - \hat{\gamma}_2), \ldots, \sqrt{k}(\hat{\gamma}_{d+} - \hat{\gamma}_d)\Big) \xrightarrow{d} N\Big(\mu_d, \Sigma_d\Big), \tag{2.12}$$

where $\mu_d = \left(\frac{\lambda_1}{1-\rho_1}, \frac{\lambda_2(\beta^{-\rho_2}-1)}{1-\rho_2}, \ldots, \frac{\lambda_d(\beta^{-\rho_d}-1)}{1-\rho_d}\right)$, $\Sigma_d = \Gamma\Gamma^T \circ H$ *("$\circ$" denotes the Hadamard* *or entrywise product), with*

$$H = \begin{bmatrix} 1 & h_{12} & . \; . \; . & h_{1d} \\ h_{12} & h & . \; . \; . & h_{2d} \\ . & . & & . \\ . & & . & . \\ . & & \;\;\; . & . \\ . & & \;\;\;\;\;\; . & . \\ . & & & . \\ h_{1d} & h_{2d} & . \; . \; . & h \end{bmatrix}, \; \Gamma = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ . \\ . \\ . \\ \gamma_d \end{bmatrix},$$

$h = 1 + \nu^2 - 2\nu^2\beta$, $h_{1i} = \nu^2 R_{1i}(1,\beta) - R_{1i}(1,1)$, *and* $h_{ij} = (1+\nu^2)R_{ij}(1,1) - \nu^2\big(R_{ij}(1,\beta) + R_{ij}(\beta,1)\big)$, $i = 2,\dots,d$, $j = i+1,\dots,d$.

Very similar to the the bivariate case we approximate, for $\lambda_j = 0, j = 1,\dots,d$, the d-variate normal limiting distribution of $(\hat{\gamma}_1, \hat{\gamma}_{2+} - \hat{\gamma}_2, \dots, \hat{\gamma}_{d+} - \hat{\gamma}_d)$, with mean vector $(\gamma_1, 0, \dots, 0)$, and estimate the approximated $\frac{1}{k}\Sigma_d$, where for the estimation of R_{ij}, $\hat{R}_{ij}$ is defined similarly as $\hat{R}$ in (2.8). The thus obtained approximated and estimated version of $\frac{1}{k}\Sigma_d$ is denoted by $\frac{1}{k}\hat{\Sigma}_d$. In this normal distribution the only unknown parameter is the first component of the mean: γ_1, cf. (2.9). Maximizing this approximate likelihood of the single observation $(\hat{\gamma}_1, \hat{\gamma}_{2+} - \hat{\gamma}_2, \dots, \hat{\gamma}_{d+} - \hat{\gamma}_d)$ with respect to γ_1, we obtain our adapted estimator for γ_1:

$$\hat{\gamma}_{1,d} = \hat{\gamma}_1 + \sum_{j=2}^{d} \frac{\hat{\Sigma}^{-1}_{1j}}{\hat{\Sigma}^{-1}_{11}} (\hat{\gamma}_{j+} - \hat{\gamma}_j),$$

where A^{-1}_{ij} denotes the entry in the i^{th} row and j^{th} column of the inverse of the matrix A. Using, in the obvious notation, $\hat{\Sigma}_d = \hat{\Gamma}\hat{\Gamma}^T \circ \hat{H}$ (see above), we can rewrite our adapted estimator as

$$\hat{\gamma}_{1,d} = \hat{\gamma}_1 + \sum_{j=2}^{d} \frac{\hat{\gamma}_1}{\hat{\gamma}_{j+}} \frac{\hat{H}^{-1}_{1j}}{\hat{H}^{-1}_{11}} (\hat{\gamma}_{j+} - \hat{\gamma}_j). \tag{2.13}$$

Theorem 2.3.1 *Assume H is invertible. Then under the conditions of Proposition 3.3.1,* *as $n \to \infty$,*

$$\sqrt{k}(\hat{\gamma}_{1,d} - \gamma_1) \xrightarrow{d} N\left(\frac{\lambda_1}{1-\rho_1} + \sum_{j=2}^{d} \frac{\gamma_1}{\gamma_j} \frac{H_{1j}^{-1}}{H_{11}^{-1}} \frac{\lambda_j(\beta^{-\rho_j} - 1)}{1-\rho_j}, \sigma^2\right), \tag{2.14}$$

where

$$\sigma^2 = \gamma_1^2 \Big(1 - \frac{1}{(H_{11}^{-1})^2} \Big[2H_{11}^{-1} \sum_{j=2}^{d} [R_{1j}(1,1) - \nu^2 R_{1j}(1,\beta)] H_{1j}^{-1} - [1 + \nu^2 - 2\nu^2\beta] \sum_{j=2}^{d} (H_{1j}^{-1})^2$$
$$- 2 \sum_{i=2}^{d} \sum_{j>i}^{d} [(1+\nu^2) R_{ij}(1,1) - \nu^2 \big(R_{ij}(1,\beta) + R_{ij}(\beta,1)\big)] H_{1i}^{-1} H_{1j}^{-1}\Big]\Big).$$

Corollary 2.3.2 *Under the conditions of Theorem 2.3.1 with $\beta = 1$, as $n \to \infty$,*

$$\sqrt{k}(\hat{\gamma}_{1,d} - \gamma_1) \xrightarrow{d} N\left(\frac{\lambda_1}{1-\rho_1}, \sigma^2\right),$$

where the asymptotic variance now simplifies to

$$\sigma^2 = \gamma_1^2 \Big(1 - \frac{1-\nu^2}{(H_{11}^{-1})^2} \Big[2H_{11}^{-1} \sum_{j=2}^{d} R_{1j}(1,1) H_{1j}^{-1} - \sum_{j=2}^{d} (H_{1j}^{-1})^2 - 2 \sum_{i=2}^{d} \sum_{j>i}^{d} R_{ij}(1,1) H_{1i}^{-1} H_{1j}^{-1}\Big]\Big),$$

where $\beta = 1$ also yields simplified entries for the matrix H.

Remark 4 We have seen that in the bivariate case for $\beta = 1$ and $\nu^2 = \frac{1}{2}$ the reduction in asymptotic variance is equal to $\frac{1}{2}R^2(1,1)$. For, e.g., $R(1,1) = 0.8$, this becomes 0.320. Now consider the trivariate case with the same values for β and ν^2 and with (also) $R_{12}(1,1) = R_{13}(1,1) = 0.8$, but $R_{23}(1,1) = 0.4$. Then the reduction in asymptotic variance, see the next section, becomes much larger: 0.457. In other words, adding a third variable that has the same (as the second variable) tail copula value at (1,1) with the variable of interest and does not have a high tail dependence with the second variable reduces the asymptotic variance much more than when using only one related variable.

2.4 Simulation study

In this section we will perform a simulation study in order to compare the finite sample behavior of the adapted estimator and the Hill estimator. We will consider 6 bivariate distributions and 8 trivariate distributions and 3 different pairs (n, m). Every setting is replicated 10,000 times.

To be precise, we consider the Cauchy distribution restricted to the first quadrant/octant in dimensions $d = 2$ and $d = 3$. This Cauchy density is proportional to

$$(1 + xS^{-1}x^T)^{-(1+d)/2},$$

where the 2×2 or 3×3 scale matrix S has 1 as diagonal elements and s as off-diagonal elements, but when $d = 3$ we take $S_{23} = S_{32} = r$. For s we take the values 0, 0.5, and 0.8, respectively. When $d = 3$ we take $r = s$, but for $s = 0.5$ and $s = 0.8$ we also take $r = 0$ and $r = 0.3$, respectively. Approximated $R(1, 1)$-values are given in Table 2.1. In the case $r < s$, two values are given; the lower one is $R_{23}(1, 1)$.

$d = 2$			$d = 3$				
$s = 0$	$s = 0.5$	$s = 0.8$	$s = 0$	$s = 0.5$	$s = 0.5$	$s = 0.8$	$s = 0.8$
			$r = 0$	$r = 0.5$	$r = 0$	$r = 0.8$	$r = 0.3$
0.59	0.67	0.76	0.59	0.68	0.69	0.77	0.81
					0.59		0.63

Table 2.1: $R(1, 1)$-values for the Cauchy distribution

We will also consider the bi- and trivariate logistic distribution function with standard Fréchet marginals:

$$F(x_1, \ldots, x_d) = \exp\left\{-\left(x_1^{-1/\theta} + \ldots + x_d^{-1/\theta}\right)^\theta\right\}, \quad x_1 > 0, \ldots, x_d > 0; \quad d = 2 \text{ or } d = 3.$$

For θ we take the values 0.1, 0.3, and 0.5, respectively. The corresponding $R(1, 1)$-values are 0.93, 0.77, and 0.59. All γ-values in the simulations are equal to 1.

We use the following values for n, m, and k:

- $n = 1000$, $m = 500$, and $k = 100$,
- $n = 1000$, $m = 1000$, and $k = 100$,
- $n = 500$, $m = 1000$, and $k = 50$.

Then we choose k_+ according to

$$\frac{k}{k_+} = \frac{n}{n+m}. \tag{2.15}$$

In case $d = 2$, using (2.15), our adapted estimator in (2.10) specializes to

$$\hat{\gamma}_{1,2} = \hat{\gamma}_1 + \frac{\hat{\gamma}_1}{\hat{\gamma}_{2+}} \hat{R}(1,1)(\hat{\gamma}_{2+} - \hat{\gamma}_2),$$

and the asymptotic variance in Theorem 2.2.1 becomes $\gamma_1^2 \Big(1 - (1-\nu^2)R^2(1,1)\Big)$. When $d = 3$, using (2.15), our adapted estimator in (2.13) can be rewritten as

$$\hat{\gamma}_{1,3} = \hat{\gamma}_1 + \frac{\hat{\gamma}_1}{\hat{\gamma}_{2+}} \frac{\hat{R}_{12}(1,1) - \hat{R}_{13}(1,1)\hat{R}_{23}(1,1)}{1 - \hat{R}_{23}^2(1,1)} (\hat{\gamma}_{2+} - \hat{\gamma}_2) + \frac{\hat{\gamma}_1}{\hat{\gamma}_{3+}} \frac{\hat{R}_{13}(1,1) - \hat{R}_{12}(1,1)\hat{R}_{23}(1,1)}{1 - \hat{R}_{23}^2(1,1)} (\hat{\gamma}_{3+} - \hat{\gamma}_3),$$

and the asymptotic variance in Theorem 2.3.1 specializes to

$$\sigma^2 = \gamma_1^2 \left(1 - (1-\nu^2)\left(\frac{R_{12}^2(1,1) + R_{13}^2(1,1) - 2R_{12}(1,1)R_{13}(1,1)R_{23}(1,1)}{1 - R_{23}^2(1,1)}\right)\right).$$

Tables 2.2 and 2.3 show the (empirical percentages of) variance reduction as discussed below Theorems 2.2.1 and 3.3.1, and above, based on the 10,000 estimates. We see that the variance reduction ranges from about 10% to more than 50%, that is, our adapted estimator yields much better results than the Hill estimator. A stronger tail dependence between the variable of interest and the related variable(s) yields a larger variance reduction. In case $d = 3$, due to the exchangeability of the components of the logistic distribution, a stronger tail dependence between the variable of interest and the related variables, yields also a stronger tail dependence between the two related variables and hence increasing the dimension from 2 to 3 does not help that much, but in case of the Cauchy distribution with $r < s$ we see a large improvement when adding the third variable. Comparing the numbers in the table with the (not presented) theoretical asymptotic reductions shows that the

	$d = 2$			$d = 3$				
	$s = 0$	$s = 0.5$	$s = 0.8$	$s = 0$	$s = 0.5$	$s = 0.5$	$s = 0.8$	$s = 0.8$
				$r = 0$	$r = 0.5$	$r = 0$	$r = 0.8$	$r = 0.3$
$n = 1000, m = 500$	10.5%	12.4%	17.3%	12.4%	17.9%	18.9%	21.2%	26.5%
$n = 1000, m = 1000$	15.5%	20.1%	28.9%	21.0%	27.3%	30.5%	31.8%	40.3%
$n = 500, m = 1000$	20.6%	27.7%	38.3%	27.6%	36.4%	38.9%	41.3%	51.4%

Table 2.2: Empirical variance reduction for the Cauchy distribution

	$d = 2$			$d = 3$		
	$\theta = 0.1$	$\theta = 0.3$	$\theta = 0.5$	$\theta = 0.1$	$\theta = 0.3$	$\theta = 0.5$
$n = 1000, m = 500$	26.8%	17.4%	8.8%	28.8%	20.7%	13.4%
$n = 1000, m = 1000$	41.1%	27.3%	14.4%	42.0%	31.4%	20.6%
$n = 500, m = 1000$	54.5%	37.4%	21.4%	58.9%	43.1%	27.8%

Table 2.3: Empirical variance reduction for the logistic distribution

empirical numbers are about the same but slightly smaller, partly due to the variability of the tail copula estimators, which does not show up in the asymptotic variance.

Although the asymptotic biases are the same (see Corollaries 2.2.2 and 2.3.2), we also present, in order to show the full behavior of the estimator, similar tables for the reduction in mean squared error (MSE). We see in Tables 2.4 and 2.5 that considering the MSE instead of the variance yields approximately the same reduction percentages. This shows that indeed our adapted estimator substantially outperforms the Hill estimator.

For every simulation setting we have only taken one value of k. It is of interest to investigate the sensitivity to the choice of k of the variance reduction in Tables 2.2 and 2.3. For the two tables below we doubled the value of k compared to the initial settings in Tables 2.2 and 2.3. For the choice of k_+ the formula in (2.15) is still used, i.e., k_+ is also doubled. Tables 2.6, 2.7 and 2.8 show that this large change in k leads to about the

	$d = 2$			$d = 3$				
	$s = 0$	$s = 0.5$	$s = 0.8$	$s = 0$	$s = 0.5$	$s = 0.5$	$s = 0.8$	$s = 0.8$
				$r = 0$	$r = 0.5$	$r = 0$	$r = 0.8$	$r = 0.3$
$n = 1000, m = 500$	10.6%	12.5%	17.3%	12.5%	18.1%	19.0%	21.3%	26.6%
$n = 1000, m = 1000$	15.7%	20.2%	28.9%	21.1%	27.6%	30.6%	31.9%	40.3%
$n = 500, m = 1000$	20.6%	27.8%	38.3%	27.8%	36.4%	38.9%	41.3%	51.3%

Table 2.4: Empirical MSE reduction for the Cauchy distribution

	$d = 2$			$d = 3$		
	$\theta = 0.1$	$\theta = 0.3$	$\theta = 0.5$	$\theta = 0.1$	$\theta = 0.3$	$\theta = 0.5$
$n = 1000, m = 500$	26.3%	17.5%	9.1%	28.4%	20.6%	13.4%
$n = 1000, m = 1000$	40.7%	27.0%	14.7%	41.3%	31.0%	20.6%
$n = 500, m = 1000$	55.1%	38.0%	22.6%	59.4%	43.9%	28.7%

Table 2.5: Empirical MSE reduction for the logistic distribution

same percentages of variance reduction, in other words, when comparing the adapted Hill estimator and the Hill estimator the choice of k is not so important.

	$s = 0$	$s = 0.5$	$s = 0.8$
$n = 1000, m = 500, k = 200$	10.1%	12.4%	17.6%
$n = 1000, m = 1000, k = 200$	15.2%	20.2%	26.0%
$n = 500, m = 1000, k = 100$	19.9%	27.3%	36.1%

Table 2.6: Empirical variance reduction for the $2d-$ Cauchy distribution

	$s = 0$	$s = 0.5$	$s = 0.5$	$s = 0.8$	$s = 0.8$
	$r = 0$	$r = 0.5$	$r = 0$	$r = 0.8$	$r = 0.3$
$n = 1000, m = 500, k = 200$	13.5%	17.3%	18.4%	21.4%	25.0%
$n = 1000, m = 1000, k = 200$	19.4%	25.7%	28.1%	33.5%	38.4%
$n = 500, m = 1000, k = 100$	27.6%	35.4%	36.0%	44.1%	52.7%

Table 2.7: Empirical variance reduction for the $3d-$ Cauchy distribution

	$d = 2$			$d = 3$		
	$\theta = 0.1$	$\theta = 0.3$	$\theta = 0.5$	$\theta = 0.1$	$\theta = 0.3$	$\theta = 0.5$
$n = 1000, m = 500, k = 200$	27.2%	19.4%	10.1%	28.5%	21.0%	14.6%
$n = 1000, m = 1000, k = 200$	39.6%	27.5%	14.5%	44.6%	30.8%	20.3%
$n = 500, m = 1000, k = 100$	55.3%	38.0%	21.8%	57.1%	43.6%	26.0%

Table 2.8: Empirical variance reduction for the logistic distribution

2.5 Application

We consider financial losses (in US$) due to earthquakes as variable of interest with the corresponding energy released as related variable. The aim of this application is to assess the tail heaviness of the loss distribution and also to estimate a very high quantile of the losses. We make use of the adapted Hill estimator, since the losses are influenced by the amounts of energy and hence these variables are expected to be tail dependent.

The earthquakes concern 29 countries[1]. The data are provided by the National Oceanic and Atmospheric Administration (NOAA). Ignoring tsunami losses, we consider the financial losses of categories at least “moderate” for the time period from 1993 through 2017.

[1] Algeria, Burma, Chile, China, Ecuador, El Salvador, Germany, Greece, Haiti, Iceland, India, Indonesia, Iran, Italy, Japan, Mexico, Morocco, Nepal, New Zealand, Nicaragua, Pakistan, Philippines, Russia, Taiwan, Tajikistan, Tanzania, Thailand, Turkey, United States.

We used linear regression analysis per country for imputation of missing loss values, with "number of deaths due to the earthquake" and "severity of the financial loss" (a categorical variable) as independent variables. We also corrected the financial losses for inflation. The highest loss in the data set is US\$ 36×10^9. We obtained the related Richter scale magnitude M of the earthquakes for the much longer period 1940 through 2017. (Note that also for the earlier period 1940-1992 the financial loss *categories* are available and again we used only magnitudes with losses at least "moderate", as for the period 1993-2017.) The energy E released by earthquakes (in megajoules) is given by $E = 2 \times 10^{1.5(M-1)}$; Lay and Wallace (1995).

We have $n = 330$ and $m = 512$. Figure 2.1 shows the log-log plot of the top 60 observations of the data with logarithm of the data rank in a descending order. We can observe a linear pattern in Figure 2.1, which indicates an empirical power law for the data. Figure 2.2 shows a plot of the adapted Hill estimator against k, with k_+ based on (2.15).

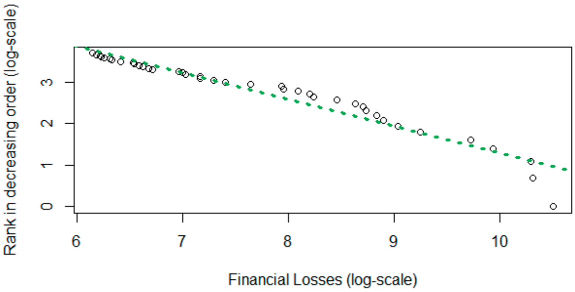


Figure 2.1: Log-Log plot of the financial losses

We take the average value of the estimates over the region $k = 40, \ldots, 60$. This yields the average Hill estimate $\hat{\gamma}_1 = 1.504$ and our final average estimate of γ_1, which is somewhat lower than the Hill estimate:

$$\hat{\gamma}_{1,2} = 1.465.$$

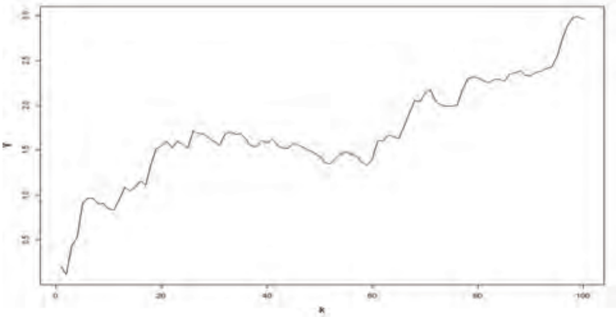


Figure 2.2: Adapted Hill estimator of the financial losses of the earthquakes

Both estimates indicate that the loss distribution has a very heavy right tail.

We also estimate the high quantile $F_1^{-1}(1-p)$ of the loss distribution for $p = \frac{1}{n} = \frac{1}{330}$. This high quantile is estimated as usual (see, e.g., page 138 of de Haan and Ferreira (2006)) with

$$X_{n-k,n}\left(\frac{k}{np}\right)^{\hat{\gamma}},$$

where $\hat{\gamma}$ is the Hill estimator or the adapted Hill estimator (and $k = 40, \ldots, 60$). This yields for the average high quantile estimate US\$ 130×10^9 when we use the Hill estimates and US\$ 113×10^9 when we use our estimates of γ_1, which is a reduction of 17 billion dollars. This shows that, from an insurer's perspective, improved (that is, less variable) estimation of the extreme value index can lead to huge changes in high quantiles, here the 25 year return level. The lower estimate we obtain indicates less risk for (re)insurers.

Acknowledgement We thank an associate editor and three referees for many insightful comments, which helped to improve this paper.

2.6 Proofs

Proof of Proposition 3.3.1: Let C be a copula corresponding to the distribution function of $(-X_1, -Y_{1,2}, \ldots, -Y_{1,d})$ and let C_- be the distribution function of the last $d-1$ components of a random vector with distribution function C. Let $(V_{1,1}, V_{1,2}, \ldots, V_{1,d}), \ldots, (V_{n,1}, V_{n,2}, \ldots, V_{n,d})$ be a random sample of size n from C and let $(V_{n+1,2}, \ldots, V_{n+1,d}), \ldots, (V_{n+m,2}, \ldots, V_{n+m,d})$ be a random sample of size m from C_-, independent of the random sample from C. Clearly all the $V_{i,j}$ have a uniform-(0,1) distribution. Write $X_i = F_1^{-1}(1-V_{i,1})$, $i = 1, \ldots, n$, and $Y_{l,j} = F_j^{-1}(1-V_{l,j})$, $l = 1, \ldots, n+m$, $j = 2, \ldots, d$. Then $(X_1, Y_{1,2}, \ldots, Y_{1,d}), \ldots, (X_n, Y_{n,2}, \ldots, Y_{n,d})$, and $(Y_{n+1,2}, \ldots, Y_{n+1,d}), \ldots, (Y_{n+m,2}, \ldots, Y_{n+m,d})$ have the distributions as specified in the beginning of Section 2.4.

Consider the univariate empirical distribution functions $\Gamma_{n,j}(s) = \frac{1}{n}\sum_{i=1}^{n} 1_{[0,s]}(V_{i,j})$, $0 \le s \le 1$, $j = 1, 2, \ldots, d$, and $\Gamma_{n+m,j}(t) = \frac{1}{n+m}\sum_{l=1}^{n+m} 1_{[0,t]}(V_{l,j})$, $0 \le t \le 1$, $j = 2, \ldots, d$, and the corresponding uniform tail empirical processes

$$w_{n,j}(s) = \frac{n}{\sqrt{k}}\left[\Gamma_{n,j}\left(\frac{k}{n}s\right) - \frac{k}{n}s\right], \quad 0 \le s \le 1,$$
$$w_{n+m,j}(t) = \frac{n+m}{\sqrt{k_+}}\left[\Gamma_{n+m,j}\left(\frac{k_+}{n+m}t\right) - \frac{k+}{n+m}t\right], \quad 0 \le t \le 1.$$

Now define the Gaussian vector of processes $(W_1, \ldots, W_{2d-1})$, where W_j, $j = 1, \ldots, 2d-1$, is a standard Wiener process on $[0,1]$, and the covariances are as follows:

$$\begin{aligned}
&Cov(W_i(s), W_j(t)) = R_{ij}(s,t), 0 \le s,t \le 1, 1 \le i < j \le d,\\
&Cov(W_i(s), W_j(t)) = \nu R_{i,j-d+1}(s, \beta t), 0 \le s,t \le 1, 1 \le i \le d,\ d+1 \le j \le 2d-1, j \ne i+d-1,\\
&Cov(W_i(s), W_{i+d-1}(t)) = \nu(s \wedge \beta t), 0 \le s,t \le 1,\ 2 \le i \le d.\\
&Cov(W_i(s), W_j(t)) = R_{i-d+1,j-d+1}(s,t), 0 \le s,t \le 1, d+1 \le i < j \le 2d-1. \qquad (2.16)
\end{aligned}$$

Let I denote the identity function on $[0,1]$. Then we have on $(D[0,1])^{2d-1}$, for $0 \le \delta < \frac{1}{2}$, as $n \to \infty$,

$$\left(\frac{w_{n,1}}{I^\delta}, \ldots, \frac{w_{n,d}}{I^\delta}, \frac{w_{n+m,2}}{I^\delta}, \ldots, \frac{w_{n+m,d}}{I^\delta}\right) \xrightarrow{d} \left(\frac{W_1}{I^\delta}, \ldots, \frac{W_d}{I^\delta}, \frac{W_{d+1}}{I^\delta}, \ldots, \frac{W_{2d-1}}{I^\delta}\right). \qquad (2.17)$$

For the proof of this statement, note that the convergence and tightness of every component is well-known, see Corollary 4.2.1 in Csörgő et al. (1986) or Theorem 3 in Einmahl (1992). This also yields the tightness of the entire vector on the left-hand side. It remains to prove the convergence of the finite-dimensional distributions (without the I^δ), which follows from the (general) multivariate central limit theorem. It suffices to compute the limits of the covariances: we perform this computation for the second formula in (2.16); the other three formulas there are essentially special cases of that one. We have

$$\begin{aligned}
&Cov(w_{n,i}(s), w_{n+m,j-d+1}(t)) = Cov\left(\frac{1}{\sqrt{k}}\sum_{l=1}^{n} 1_{[0,\frac{k}{n}s]}(V_{l,i}), \frac{1}{\sqrt{k_+}}\sum_{l=1}^{n+m} 1_{[0,\frac{k_+}{n+m}t]}(V_{l,j-d+1})\right)\\
&= Cov\left(\frac{1}{\sqrt{k}}\sum_{l=1}^{n} 1_{[0,\frac{k}{n}s]}(V_{l,i}), \frac{1}{\sqrt{k_+}}\sum_{l=1}^{n} 1_{[0,\frac{k_+}{n+m}t]}(V_{l,j-d+1})\right)\\
&= \frac{n}{\sqrt{kk_+}} Cov(1_{[0,\frac{k}{n}s]}(V_{1,i}), 1_{[0,\frac{k_+}{n+m}t]}(V_{1,j-d+1}))\\
&= \frac{n}{\sqrt{kk_+}}\left[\mathbb{P}\left(V_{1,i} \le \frac{k}{n}s, V_{1,j-d+1} \le \frac{k_+}{n+m}t\right) - \frac{kk_+}{n(n+m)}st\right]\\
&= \sqrt{\frac{k}{k_+}}\left[\frac{n}{k}\mathbb{P}\left(V_{1,i} \le \frac{k}{n}s, V_{1,j-d+1} \le \frac{k}{n}\frac{n}{k}\frac{k_+}{n+m}t\right) - \frac{k_+}{n+m}st\right]\\
&\quad \to \nu R_{i,j-d+1}(s, \beta t) = Cov(W_i(s), W_j(t)).
\end{aligned}$$

Hence (2.17) is established.

According to de Haan and Ferreira (2006), Chapter 5 and Theorem 2.3.9, we have, as $n \to \infty$,

$$\sqrt{k}(\hat{\gamma}_j - \gamma_j) = -\gamma_j\left(w_{n,j}(1) - \int_0^1 \frac{w_{n,j}(u)}{u}du\right) + \frac{\lambda_j}{1-\rho_j} + o_p(1), \quad j = 1, \ldots, d.$$

Using that $|A_j|$ is regularly varying at ∞ with index ρ_j, we get similarly

$$\sqrt{k_+}(\hat{\gamma}_{j+} - \gamma_j) = -\gamma_j\left(w_{n+m,j}(1) - \int_0^1 \frac{w_{n+m,j}(u)}{u}du\right) + \frac{\lambda_j \beta^{-\rho_j}}{\nu(1-\rho_j)} + o_p(1), \quad j = 2, \ldots, d.$$

Combining all these with (2.17) we obtain

$$\Big(\sqrt{k}(\hat{\gamma}_1-\gamma_1),\ldots,\sqrt{k}(\hat{\gamma}_d-\gamma_d),\sqrt{k_+}(\hat{\gamma}_{2+}-\gamma_2),\ldots,\sqrt{k_+}(\hat{\gamma}_{d+}-\gamma_d)\Big)$$

$$\xrightarrow{d}\left(-\gamma_1\left(W_1(1)-\int_0^1\frac{W_1(u)}{u}du\right)+\frac{\lambda_1}{1-\rho_1},\ldots,-\gamma_d\left(W_d(1)-\int_0^1\frac{W_d(u)}{u}du\right)+\frac{\lambda_d}{1-\rho_d},\right.$$

$$\left.-\gamma_2\left(W_{d+1}(1)-\int_0^1\frac{W_{d+1}(u)}{u}du\right)+\frac{\lambda_2\beta^{-\rho_2}}{\nu(1-\rho_2)},\ldots,-\gamma_d\left(W_{2d-1}(1)-\int_0^1\frac{W_{2d-1}(u)}{u}du\right)+\frac{\lambda_d\beta^{-\rho_d}}{\nu(1-\rho_d)}\right).$$

It is immediate and well-known that this yields the mean vector and the variances as in the proposition. (Note that the components of the left-hand side there are listed in a different order.) It remains to derive the covariances. Again we only consider the case where $1\le i\le d$, $d+1\le j\le 2d-1, j\neq i+d-1$. The other cases are easier and essentially special cases of this one. We have

$$Cov\left(-\gamma_i\left(W_i(1)-\int_0^1\frac{W_i(u)}{u}du\right)+\frac{\lambda_i}{1-\rho_i},-\gamma_{j-d+1}\left(W_j(1)-\int_0^1\frac{W_j(v)}{v}dv\right)\right. \tag{2.18}$$

$$\left.+\frac{\lambda_{j-d+1}\beta^{-\rho_{j-d+1}}}{\nu(1-\rho_{j-d+1})}\right)=\gamma_i\gamma_{j-d+1}\big[E(W_i(1)W_j(1))+\int_0^1\int_0^1\frac{E\big(W_i(u)W_j(v)\big)}{uv}dudv$$

$$-\int_0^1\frac{E\big(W_i(u)W_j(1)\big)}{u}du-\int_0^1\frac{E\big(W_i(1)W_j(v)\big)}{v}dv\big]=\nu\gamma_i\gamma_{j-d+1}\big[R_{i,j-d+1}(1,\beta)$$

$$+\int_0^1\int_0^1\frac{R_{i,j-d+1}(u,\beta v)}{uv}dudv-\int_0^1\frac{R_{i,j-d+1}(u,\beta)}{u}du-\int_0^1\frac{R_{i,j-d+1}(1,\beta v)}{v}dv\big]. \tag{2.19}$$

Observe that by two changes of variables and the homogeneity of order 1 of $R_{i,j-d+1}$:

$$\int_0^1\int_0^1\frac{R_{i,j-d+1}(u,\beta v)}{uv}dudv=\int_0^1\int_0^v\frac{R_{i,j-d+1}(u,\beta v)}{uv}dudv+\int_0^1\int_0^u\frac{R_{i,j-d+1}(u,\beta v)}{uv}dvdu$$

$$=\int_0^1\int_0^1\frac{R_{i,j-d+1}(vu,\beta v)}{uv}dudv+\int_0^1\int_0^1\frac{R_{i,j-d+1}(u,\beta vu)}{uv}dvdu$$

$$=\int_0^1\frac{R_{i,j-d+1}(u,\beta)}{u}du+\int_0^1\frac{R_{i,j-d+1}(1,\beta v)}{v}dv.$$

Hence the covariance is equal to $\nu\gamma_i\gamma_{j-d+1}R_{i,j-d+1}(1,\beta)$. □

Proof of Theorem 2.3.1: From the uniform consistency of the tail copula estimators and the continuity of the tail copulas we have $\hat{H}_{1j}^{-1} \xrightarrow{P} H_{1j}^{-1}$, $j = 1, \dots, d$. This in combination with (2.13) and Corollary 2.3.1 yields

$$\sqrt{k}(\hat{\gamma}_{1,d} - \gamma_1) = \sqrt{k}(\hat{\gamma}_1 - \gamma_1) + \sum_{j=2}^{d} \frac{\gamma_1}{\gamma_j} \frac{H_{1j}^{-1}}{H_{11}^{-1}} \sqrt{k}(\hat{\gamma}_{j+} - \hat{\gamma}_j) + o_p(1). \tag{2.20}$$

Now Corollary 2.3.1 and the continuous mapping theorem yield (2.14). □

Remark 5 In the bivariate case in Section 2.3, the determinant of the matrix H is always positive and hence the additional invertibility *assumption* on H is not needed there.

Chapter 3

Extreme value statistics in semi-supervised models

[Based on joint work with John H.J. Einmahl and Chen Zhou]

Abstract. We consider extreme value analysis in a semi-supervised setting, where we observe, next to the n data on the target variable, $n+m$ data on one or more covariates. This is called the semi-supervised model with n labeled and m unlabeled data. By exploiting the tail dependence between the target variable and the covariates, we derive an estimator for the extreme value index of the target variable in this setting and establish its asymptotic behavior. Our estimator substantially improves the univariate estimator, based on only the n target variable data, in terms of asymptotic variance whereas the asymptotic bias remains unchanged. We present a simulation study in which the asymptotic results are confirmed and also an extreme quantile estimator is derived and its improved performance is shown. Finally the estimation method is applied to rainfall data in France.

Key words. Asymptotic normality, extreme value index, semi-supervised inference, tail

dependence, variance reduction.

3.1 Introduction

The semi-supervised model, initially introduced in machine learning, deals with unbalanced datasets, when the labeled data are harder (more expensive or more time consuming) to obtain than the unlabeled data. For an example data structure in the bivariate case, see Table 1.1. Consider a dataset with one variable of interest, sometimes referred to as the target variable or outcome variable, and one or more covariates. The difficulty for collecting labeled data stems from collecting the target variable, whereas unlabeled data containing only the covariates, i.e. with the target variable missing, can be easily collected. Semi-supervised learning focuses on uncovering the (non-linear) relation between the target variable and the covariates. Estimations and predictions based on such relations and using the additional unlabeled data often show substantially improved performance. For example, for classification analysis see Vapnik (2013) and Zhu and Goldberg (2009); for regression analysis see Wasserman and Lafferty (2008), Azriel et al. (2016) and Chakrabortty and Cai (2018).

Semi-supervised inference aims at estimating parameters or quantities regarding the target variable under the semi-supervised model. Zhang et al. (2019) investigates the general semi-supervised framework and shows how to use the unlabeled data to improve the estimation of the mean of the target variable; for inference on heavy tailed distributions in this framework, see Ahmed and Einmahl (2019).

Extreme value statistics deals with estimation of parameters or quantities related to the tail of a distribution, only making semi-parametric assumptions on this tail. Consequently, most of extreme value methods start with a relatively large number of observations n, but select only $k \ll n$ extreme observations from the full sample for statistical inference. Two techniques are often used in selecting the extreme observations: the peaks-over-threshold (POT) approach which selects the highest k observations, and the block maxima (BM) approach which splits the full sample into k blocks and selects the maxima of each block. Since only k observations are used in estimation, typically consistent estimators have a speed of convergence of $1/\sqrt{k}$. In practice, to obtain accurate estimators for tail parameters/quantities, one needs a sample with a relatively large sample size n to guarantee a sufficient number of extreme observations. In contrast, the semi-supervised model is greatly suitable for statistics of extremes in case data on the target variable are hard to obtain.

The main goal of this paper is to derive in this semi-supervised setting a new, improved pseudo-maximum likelihood estimator (MLE) for a general extreme value index γ and to establish its asymptotic behavior. This extreme value index describes the tail heaviness of a probability distribution. If $\gamma > 0$ the distribution is heavy tailed and has an infinite right endpoint, if $\gamma = 0$ the distribution is light tailed and may have an infinite or finite endpoint, and if $\gamma < 0$ the endpoint is finite, see, e.g., Beirlant et al. (2004) or de Haan and Ferreira (2006) for a thorough treatment of extreme value theory and the corresponding statistical inference. For ease of explanation of our novel estimator, let us assume that there is only one covariate. We estimate γ for the variable of interest initially (that is ignoring the covariate) using the pseudo-MLE $\hat{\gamma}$, see Smith (1987) and Drees et al. (2004). Then, we choose a number g and for the covariate we transform the labeled data empirically (using all the labeled and unlabeled data) such that they obtain an artificial extreme value index g. Using the transformed covariates of the labeled data, we estimate the *known* g by the pseudo-MLE $\hat{g}$, say, and use the difference $\hat{g} - g$ to adapt and substantially improve the initial estimator $\hat{\gamma}$ for the extreme value index γ of the variable of interest. For this adaptation the tail dependence between the target variable and the covariate is crucial. Precise estimation of γ is important for describing the tail heaviness, but it is even more important when estimating extreme quantiles or very small tail probabilities. We further demonstrate the improved performance of an extreme quantile estimator under the semi-supervised model by a simulation study.

Compared to Ahmed and Einmahl (2019), this study has at least three improvements. Firstly, we provide a general result in the context of the relevant tail quantile process (see Lemma 3.6.5). Based on the tail quantile process result, one may improve most estimators based on the POT approach in extreme value statistics under the semi-supervised model. Secondly, we impose no assumptions on the tail of the covariates. When analyzing the tail of the target variable, it is crucial to assume regularity in its tail such as the max-domain of attraction condition in extreme value analysis. However, requiring such conditions for the covariates can be restrictive in applications. Thirdly and most important, our main result is valid for a broader class of distributions for the target variable: we deal with a general extreme value index $\gamma \in \mathbb{R}$ whereas Ahmed and Einmahl (2019) only handles the case $\gamma > 0$. Extending the range of the extreme value index is particularly important for applications where the sign of γ is not known beforehand. For example, when analyzing extreme weather, various studies find that the extreme value index is around zero for different meteorologic variables: for hourly surge level on the English east coast (Coles and Tawn (1991)), for hourly maximum wind speed in Sheffield, UK (Coles and Walshaw (1994)), for wave height and still water level on the Dutch coast (de Haan and de Ronde (1998)) and for daily rainfall in North Holland, The Netherlands (Buishand et al. (2008)).

This paper is organized as follows. In Section 3.2, for clearness of the exposition, we first introduce our adapted estimator for the extreme value index in the semi-supervised model with one covariate and we establish its asymptotic normality. In Section 3.3 we consider the general multivariate semi-supervised setting and present and establish asymptotic normality of the adapted estimator. Section 3.4 is devoted to a simulation study for the setting with one covariate. The improved performance, in terms of variance, of the adapted estimator compared with the initial estimator is shown. An application to rainfall in France can be found in Section 3.5 and the detailed proofs are deferred to Section 3.6.

3.2 Main results: one covariate

Let F be a bivariate distribution function with marginals F_1 and F_2. We assume that F_1 is in the max-domain of attraction of an extreme-value distribution G_γ, where γ is the extreme value index, our parameter of interest. Let the pairs $(X_1, Y_1), \dots, (X_n, Y_n)$ be a random sample from F, and let $(Y_{n+1}, \dots, Y_{n+m})$ be a random sample from F_2, independent from the n pairs. This is the semi-supervised model. Assume that the tail copula R of (X_1, Y_1) exists:

$$R(x,y) = \lim_{t \downarrow 0} \frac{1}{t} \mathbb{P}\left(1 - F_1(X_1) \le tx, 1 - F_2(Y_1) \le ty\right), (x,y) \in [0,\infty]^2 \setminus \{(\infty,\infty)\}. \quad (3.1)$$

Denote the order statistics of $X_i, i = 1, \dots, n$, with $X_{1:n} \le \dots \le X_{n:n}$, and similarly for the $Y_i, i = 1, \dots, n$. We estimate $\gamma > -\frac{1}{2}$ with the often used pseudo-MLE $\hat{\gamma}$ based on $X_{n-k:n}, \dots, X_{n:n}$, for $k \in \{1, \dots, n-1\}$; see Section 3.4 in de Haan and Ferreira (2006).

Define for $i = 1, \dots, n$,

$$\tilde{Y}_i = \begin{cases} \frac{\left(1-\left(F_{n+m}(Y_i)-\frac{1}{2(n+m)}\right)\right)^{-g}-1}{g} & , \quad g \neq 0, \\ -\log\left(1-\left(F_{n+m}(Y_i)-\frac{1}{2(n+m)}\right)\right) & , \quad g = 0, \end{cases} \quad (3.2)$$

where F_{n+m} is the empirical distribution function based on $Y_l, l = 1, \dots, n+m$, and $g > -\frac{1}{2}$ is a number we may choose that mimics an extreme value index. Let the order statistics of $\tilde{Y}_i, i = 1, \dots, n$, be denoted by $\tilde{Y}_{1:n} \le \dots \le \tilde{Y}_{n:n}$, and let $\hat{g}$ be the pseudo-MLE of g based on $\tilde{Y}_{n-k:n}, \dots, \tilde{Y}_{n:n}$, using the same k as before. Of course, since we choose and hence know g, there is no direct need to estimate it. We will show below, however, that the dependence of the difference $\hat{g} - g$ and $\hat{\gamma}$, helps to improve the estimator of γ in the semi-supervised setting.

For the asymptotic theory, we assume that $m = m(n)$ and

$$k \to \infty, \frac{k}{n} \to 0, \sqrt{\frac{n}{n+m}} \to \nu \in (0,1), \quad \text{as } n \to \infty. \quad (3.3)$$

We begin with establishing the joint asymptotic normality of $\hat{\gamma}$ and $\hat{g}$, a crucial result for deriving and showing asymptotic normality of our semi-supervised estimator (SSE) of γ.

For that purpose we need the usual second order condition on the marginal distribution F_1. Let $U_1 = F_1^{-1}(1-1/\cdot)$ be the tail quantile corresponding to F_1. We assume that there exist a positive scale function a, a positive or negative function A, with $\lim_{t\to\infty} A(t) = 0$, and $\rho \leq 0$, such that for $x > 0$,

$$\lim_{t\to\infty} \frac{\frac{U_1(tx)-U_1(t)}{a(t)} - \frac{x^\gamma - 1}{\gamma}}{A(t)} = \Psi(x), \quad \gamma \in \mathbb{R}, \tag{3.4}$$

where

$$\Psi(x) = \begin{cases} \frac{x^{\gamma+\rho}-1}{\gamma+\rho}, & \rho < 0, \\ \frac{1}{\gamma} x^\gamma \log x, & \gamma \neq \rho = 0, \\ \frac{1}{2}\log^2 x, & \gamma = \rho = 0, \end{cases}$$

see de Haan and Ferreira (2006), p. 46.

Proposition 3.2.1 *Assume $\gamma > -\frac{1}{2}$ and choose $g > -\frac{1}{2}$.Assume that F_2 is continuous, (3.1), (3.3) and (3.4) hold, and $\sqrt{k}A(\frac{n}{k}) \to \lambda \in \mathbb{R}$, as $n \to \infty$, then with probability tending to 1, there exist unique maximizers of the likelihood functions based on $\{X_i\}_{i=1}^n$ and $\{\tilde{Y}_i\}_{i=1}^n$, denoted as $(\hat{\gamma}, \hat{g})$, such that*

$$\left(\sqrt{k}\left(\hat{\gamma}-\gamma\right), \sqrt{k}\left(\hat{g}-g\right)\right) \xrightarrow{d} N\left(\left[\frac{\lambda(1+\gamma)}{(1-\rho)(1+\gamma-\rho)}, 0\right], \Sigma\right)$$

where

$$\Sigma = \begin{bmatrix} (1+\gamma)^2 & (1-\nu^2)(1+\gamma)(1+g)R_g \\ (1-\nu^2)(1+\gamma)(1+g)R_g & (1-\nu^2)(1+g)^2 \end{bmatrix},$$

with

$$R_g = R(1,1) + \frac{g-\gamma}{\gamma+g+1}\left((2\gamma+1)\int_0^1 \frac{R(s,1)}{s^{1-\gamma}}ds - (2g+1)\int_0^1 \frac{R(1,t)}{t^{1-g}}dt\right). \tag{3.5}$$

Based on Proposition 3.2.1, we derive the SSE of γ. For this derivation only, take $\lambda = 0$. Then the approximate bivariate normal distribution of $(\hat{\gamma}, \hat{g} - g)$, has mean $[\gamma, 0]$ and estimated covariance matrix

$$\frac{1}{k}\hat{\Sigma} = \frac{1}{k}\begin{bmatrix} (1+\hat{\gamma})^2 & (1-\frac{n}{n+m})(1+\hat{\gamma})(1+g)\hat{R}_g \\ (1-\frac{n}{n+m})(1+\hat{\gamma})(1+g)\hat{R}_g & (1-\frac{n}{n+m})(1+g)^2 \end{bmatrix},$$

where $\hat{R}_g$ is the estimator of R_g, obtained by replacing γ with $\hat{\gamma}$ and the tail copula R with its natural estimator

$$\hat{R}(x,y) = \frac{1}{k}\sum_{i=1}^{n} 1_{[X_i \geq X_{n-[kx]+1:n}, Y_i \geq Y_{n-[ky]+1:n}]}, \quad x, y \geq 0, \tag{3.6}$$

see, e.g., Drees and Huang (1998). Maximizing the thus obtained approximate likelihood function of the single "data point" $(\hat{\gamma}, \hat{g} - g)$ with respect to the unknown γ we obtain as SSE for γ:

$$\hat{\gamma}_g = \hat{\gamma} - \frac{1+\hat{\gamma}}{1+g}\hat{R}_g(\hat{g} - g). \tag{3.7}$$

Now we present the main result of this section, the asymptotic normality of the SSE.

Theorem 3.2.1 *Under the conditions of Proposition 3.2.1, as $n \to \infty$,*

$$\sqrt{k}(\hat{\gamma}_g - \gamma) \xrightarrow{d} N\left(\frac{\lambda(1+\gamma)}{(1-\rho)(1+\gamma-\rho)}, (1+\gamma)^2\left[1-(1-\nu^2)R_g^2\right]\right). \tag{3.8}$$

Remark 3.2.1 *Note that the asymptotic bias of the SSE $\hat{\gamma}_g$ is the same as that of the pseudo-MLE $\hat{\gamma}$ (in Proposition 3.2.1). Therefore, when comparing both estimators we can and will focus on the (relative) reduction of the asymptotic variance which is equal to $(1-\nu^2)R_g^2$. The value of the crucial $R_g \in [-1,1]$ depends on the known g and the unknown γ and R. Note that R_g can indeed be positive, zero, or negative and R_g can exceed $R(1,1)$ even when R is symmetric in its arguments. Nevertheless it is appealing to consider $g = \gamma$, reducing R_g to $R(1,1)$. Since γ is unknown, this would lead to the choice $g = \hat{\gamma}$. However the simulation results show that this random g is often not the best option in terms of the reduction of variance compared to a deterministic g not too far away from γ. Also observe that when g is close to γ and R is symmetric, then $R_g - R(1,1)$ is of order $(g-\gamma)^2$, that is, the variance reduction does not change much with the choice of g and is close to $R(1,1)$. We will see in Section 4 through simulations that the variance reduction is substantial in "standard" semi-supervised settings.*

3.3 Main results: multiple covariates

In this section we consider the more general situation with $d-1$ covariates where $d > 2$. Consider a d-variate distribution F, with marginals $F_1, \ldots, F_d$. We assume again that (only) F_1 is in the max-domain of attraction of an extreme-value distribution G_γ. Let F_- be the distribution function of the last $d-1$ components of a random vector with distribution function F. Let $(X_1, Y_{1,2}, \ldots, Y_{1,d}), \ldots, (X_n, Y_{n,2}, \ldots, Y_{n,d})$ be a random sample of size n from F and let $(Y_{n+1,2}, \ldots, Y_{n+1,d}), \ldots, (Y_{n+m,2}, \ldots, Y_{n+m,d})$ be a random sample of size m from F_-, independent of the d-variate random sample of size n. This is the multivariate semi-supervised setting.

Then, for fixed $j = 2, \ldots, d$, we use all data for the covariates $\{Y_{i,j}\}_{i=1}^{n+m}$ to obtain $\left\{\tilde{Y}_{i,j}\right\}_{i=1}^{n}$ as in (3.2), where we may choose a number $g > -\frac{1}{2}$, that mimics an extreme value index, as before. For $k \in \{1, \ldots, n-1\}$, let, similarly as in the previous section, $\hat{\gamma}$ and $\hat{g}_j$, $j = 2, \ldots, d$, be the pseudo-MLEs of γ and ($d-1$ times) of g, respectively. Assume the existence of the tail copula R_{ij} of the i^{th} and the j^{th} component,

$$R_{ij}(x, y) = \lim_{t \downarrow 0} \frac{1}{t} \mathbb{P}\left(1 - F_i(Y_{1,i}) \le tx, 1 - F_j(Y_{1,j}) \le ty\right), \tag{3.9}$$

where $(x, y) \in [0, \infty]^2 \setminus \{(\infty, \infty)\}, 1 \le i, j \le d$. Here $Y_{1,1}$ is understood as X_1. Again, we first consider the joint asymptotic normality of $\hat{\gamma}$, and $\hat{g}_j, j = 2, \ldots, d$.

Proposition 3.3.1 *Assume $\gamma > -\frac{1}{2}$ and choose $g > -\frac{1}{2}$. Assume that $F_j, j = 2, \ldots, d$, is continuous,* (3.3), (3.4), *and* (3.9) *hold, and as $n \to \infty$, $\sqrt{k}A(\frac{n}{k}) \to \lambda \in \mathbb{R}$, then with probability tending to 1, there exist unique maximizers of the likelihood functions based on $\{X_i\}_{i=1}^n$, $\{\tilde{Y}_{i,2}\}_{i=1}^n$, ..., $\{\tilde{Y}_{i,d}\}_{i=1}^n$, denoted as $(\hat{\gamma}, \hat{g}_2, \ldots, \hat{g}_d)$, such that*

$$\left(\sqrt{k}(\hat{\gamma} - \gamma), \sqrt{k}(\hat{g}_2 - g), \ldots, \sqrt{k}(\hat{g}_d - g)\right) \xrightarrow{d} N\left(\left[\frac{\lambda(\gamma+1)}{(1-\rho)(1+\gamma-\rho)}, 0, \ldots, 0\right], \Sigma_d\right),$$

with $\Sigma_d = \Gamma\Gamma^T \circ H$ *("$\circ$" is the Hadamard or entrywise product), where*

$$\Gamma = \begin{bmatrix} 1+\gamma \\ 1+g \\ \cdot \\ \cdot \\ \cdot \\ 1+g \end{bmatrix}, H = \begin{bmatrix} 1 & h_{12} & \cdot\ \cdot\ \cdot & h_{1d} \\ h_{12} & 1-\nu^2 & \cdot\ \cdot\ \cdot & h_{2d} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ h_{1d} & h_{2d} & \cdot\ \cdot\ \cdot & 1-\nu^2 \end{bmatrix},$$

$h_{1i} = (1-\nu^2)\left[R_{1i}(1,1) + \frac{g-\gamma}{\gamma+g+1}\left[(2\gamma+1)\int_0^1 \frac{R_{1i}(s,1)}{s^{1-\gamma}}ds - (2g+1)\int_0^1 \frac{R_{1i}(1,t)}{t^{1-g}}dt\right]\right]$, *and* $h_{ij} = (1-\nu^2)R_{ij}(1,1), \quad i = 2,\dots,d, j = i+1,\dots,d.$

Very similar to the bivariate case, let $\lambda = 0$ and derive the SSE of γ by using the approximate multivariate normal distribution of $(\hat{\gamma}, \hat{g}_2 - g, \dots, \hat{g}_d - g)$, with mean $[\gamma, 0, \dots, 0]$, and variance $\frac{1}{k}\hat{\Sigma}_d = \frac{1}{k}\hat{\Gamma}\hat{\Gamma}^T \circ \hat{H}$, where for the estimation of R_{ij}, $\hat{R}_{ij}$ is defined like in (3.6). By maximizing the approximate likelihood function of $(\hat{\gamma}, (\hat{g}_2 - g), \dots, (\hat{g}_d - g))$ with respect to γ, we obtain the SSE in this multivariate setting:

$$\hat{\gamma}_g = \hat{\gamma} + \frac{1+\hat{\gamma}}{1+g}\sum_{j=2}^{d} \frac{\hat{H}^{-1}_{1j}}{\hat{H}^{-1}_{11}}(\hat{g}_j - g), \tag{3.10}$$

where $\hat{H}^{-1}_{ij}$ is the entry in the i^{th} row and j^{th} column of the inverse of the matrix $\hat{H}$. The following theorem shows the asymptotic behavior of the improved estimator $\hat{\gamma}_g$.

Theorem 3.3.1 *Assume that H is invertible. Then under the conditions of Proposition 3.3.1, as $n \to \infty$,*

$$\sqrt{k}(\hat{\gamma}_g - \gamma) \xrightarrow{d} N\left(\frac{\lambda(1+\gamma)}{(1-\rho)(1+\gamma-\rho)}, \sigma^2\right), \tag{3.11}$$

where

$$\sigma^2 = (1+\gamma)^2\left(1 + \frac{1}{(H^{-1}_{11})^2}\left[2\sum_{i=1}^{d}\sum_{j=i+1}^{d} H^{-1}_{1i}H^{-1}_{1j}h_{ij} + (1-\nu^2)\sum_{j=2}^{d}(H^{-1}_{1j})^2\right]\right).$$

3.4 Simulation study

In this section we perform for the one-covariate setting a simulation study. First we investigate the finite sample performance of our novel SSE of γ and then we compare in detail the variances of the SSE with those of the pseudo-MLE based on $X_{n-k:n}, \ldots, X_{n:n}$ only. In addition, we estimate an extreme quantile by substituting the SSE $\hat{\gamma}_g$ and a similar SSE $\hat{\sigma}_g$ of the scale a, in the generic formula of the extreme quantile estimator. Again, we compare the variances of this SSE and the classical estimator based on the pseudo-MLEs of γ and a.

We begin with simulating data from the bivariate Cauchy distribution restricted to the first quadrant. This Cauchy density is proportional to

$$(1 + xS^{-1}x^T)^{-3/2}$$

where S is a 2×2 scale matrix with 1 on the diagonal and s off-diagonal. For s we take two values: 0 and 0.8. These data are denoted by $(\check{X}_i, Y_i)$. To obtain our data (X_i, Y_i), where the X_i have extreme value index γ, we transform the $\check{X}_i$, as follows

$$X_i = \begin{cases} \frac{(1-F_s(\check{X}_i))^{-\gamma}-1}{\gamma}, & \gamma \neq 0, \\ -\log(1 - F_s(\check{X}_i)), & \gamma = 0, \end{cases}$$

where F_s is the distribution function of $\check{X}_i$. Simulations are performed for values of γ that are negative, positive or 0.

First, we generate 500 samples of sizes $n = 500, m = 1000$, for $s = 0.8$, and estimate γ using the SSE and the pseudo-MLE for $k = 1, \ldots, 499$. We depict the root mean squared error (RMSE) based on these 500 samples as a function of k. We consider $\gamma = -0.25, 0$, and, 0.25, and take $g = 0$. The RMSE of the SSE (indicated by AMLE in Figure 3.1) is indeed substantially lower than that of the pseudo-MLE for the different values of γ.

Next, we focus on the (relative) variance reduction of the SSE in comparison to the pseudo-MLE. We use the following values of n and m (and k):

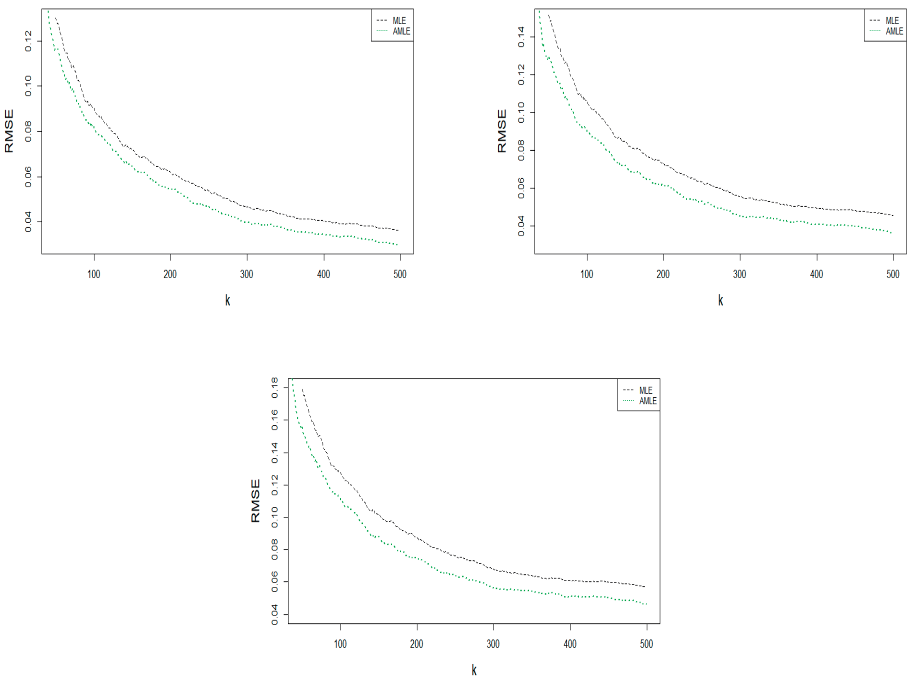


Figure 3.1: RMSE using the pseudo-MLE and the SSE-MLE. From left to right: $\gamma = -0.25, 0, 0.25$.

- $n = 1000$, $m = 500$ (less unlabeled than labeled data) and $k = 250$,
- $n = 1000$, $m = 1000$ (equal number of unlabeled and labeled data) and $k = 250$,
- $n = 500$, $m = 1000$ (more unlabeled than labeled data) and $k = 125$.

Table 3.1 shows the empirical percentages of variance reduction for different values of γ and g. The results are based on $10,000$ replications. We observe that the variance reduction ranges from 10% to more than 30%, hence indeed the SSE has a substantially smaller variance than the pseudo-MLE. By comparing the three panels, we observe that the variance reduction increases substantially with the ratio of the number of unlabeled data m and the number of labeled data n, which is in line with the asymptotic theory. Observe that the actual choice of g does not have a large influence as long as it is somewhat close to γ, a choice that is in practice often feasible.

Table 3.1: Variance reduction for different extreme value indices

(i) $(n, m) = (1000, 500)$

g	γ				g	γ			
	-0.3	-0.2	-0.1	0		0	0.1	0.2	0.3
-0.25	13.8%	14.0%	13.4%	12.0%	0	15.4%	15.6%	15.4%	14.8%
-0.125	12.6%	14.2%	14.8%	14.1%	0.125	15.3%	15.9%	16.1%	15.8%
0	10.4%	13.0%	14.6%	15.4%	0.25	14.7%	15.5%	16.1%	16.4%

(ii) $(n, m) = (1000, 1000)$

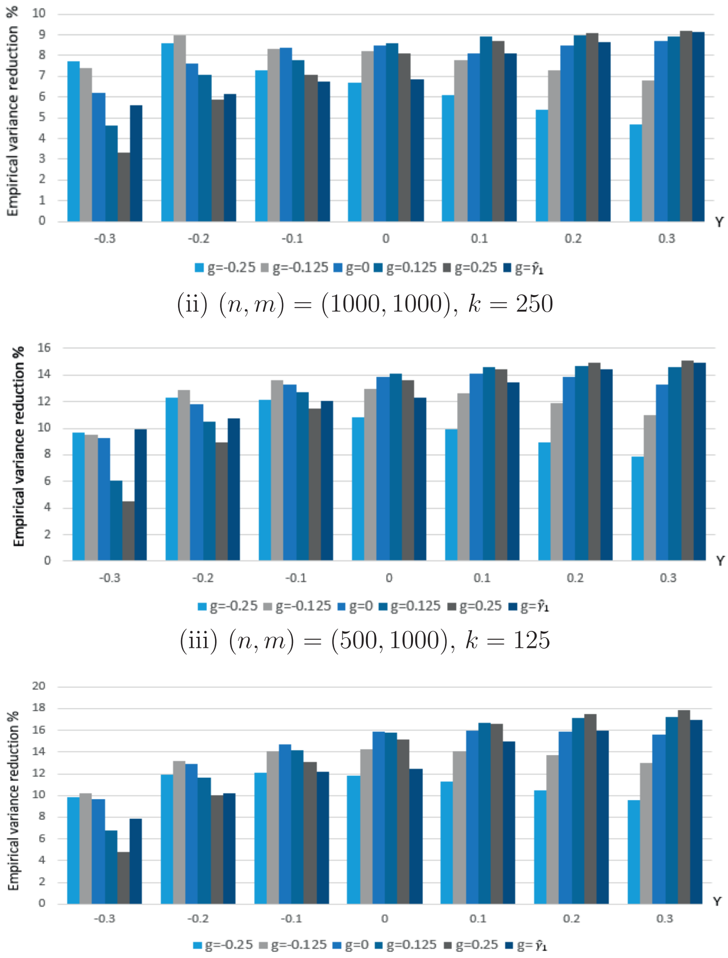
	-0.3	-0.2	-0.1	0		0	0.1	0.2	0.3
-0.25	20.2%	21.1%	20.5%	18.8%	0	23.6%	23.5%	23.1%	22.2%
-0.125	18.9%	21.5%	22.4%	22.0%	0.125	23.4%	24.1%	24.3%	25.0%
0	16.3%	19.9%	22.3%	23.6%	0.25	22.4%	23.7%	24.6%	24.6%

(iii) $(n, m) = (500, 1000)$

	-0.3	-0.2	-0.1	0		0	0.1	0.2	0.3
-0.25	22.4%	25.8%	25.7%	24.5%	0	30.8%	30.2%	29.9%	29.4%
-0.125	21.6%	26.2%	28.0%	28.0%	0.125	31.0%	30.4%	31.0%	30.8%
0	18.6%	24.6%	28.1%	30.8%	0.25	30.0%	30.1%	31.4%	31.5%

Next we investigate in more detail the sensitivity of the variance reduction to the choice of g using a wider range of values of g, including cases where γ and g have opposite sign and the case where $g = \hat{\gamma}$. In these simulations, we take $s = 0$ for the bivariate Cauchy distribution. The results, based on $10,000$ replications, for the aforementioned values of n, m and k are presented in Figure 3.2. Generally, there is always variance reduction, but for $|g - \gamma|$ relatively large the reduction is lower than when g is closer to γ. Additionally the choice of $g = \hat{\gamma}$ often does not result in a better reduction than using fixed g. Observe that for the present range of γ (-0.3 to 0.3) the choice $g = 0$ yields an almost maximal variance reduction.

Figure 3.2: Variance reduction for various combinations of γ and g



(i) $(n, m) = (1000, 500)$, $k = 250$

(ii) $(n, m) = (1000, 1000)$, $k = 250$

(iii) $(n, m) = (500, 1000)$, $k = 125$

Finally we study in more detail the effect of the size of m, the number of unlabeled data, on the variance reduction; again we take $s = 0$. We consider the case where $n = 500$ and let m vary; we choose $g = 0$. The results are based on 500 replications. Table 3.2 shows that the variance reduction approximately doubles when m ranges from 500 to $10,000$.

Table 3.2: Variance reduction for different numbers of unlabeled data m

m				γ			
	-0.3	-0.2	-0.1	0	0.1	0.2	0.3
500	6.0%	8.3%	9.6%	11.4%	11.7%	11.9%	11.8%
1000	8.7%	12.3%	14.4%	15.6%	12.9%	13.0%	15.5%
5000	11.2%	16.4%	19.5%	22.5%	21.8%	22.0%	21.7%
10000	11.9%	17.5%	20.8%	24.6%	23.0%	23.2%	22.9%

The last part of this section is devoted to extending the semi-supervised estimation approach to scale estimation and in particular to extreme quantile estimation. Here we confine ourselves to the choice $g = 0$. The scale $a = a(n/k)$ can be estimated with $\hat{\sigma}$, based on $X_{n-k:n}, \ldots, X_{n:n}$; see Section 3.4 in de Haan and Ferreira (2006) for definitions and results. Somewhat similar to the derivation of the SSE $\hat{\gamma}_g$, the SSE of the scale can be derived as:

$$\hat{\sigma}_0 = \hat{\sigma}\left(1 - \hat{S}_0(\tilde{\sigma}_0 - 1)\right),$$

where $\tilde{\sigma}_0$ is the pseudo-MLE of the scale parameter (which is equal to 1) based on $\tilde{Y}_{n-k:n}, \ldots, \tilde{Y}_{n:n}$, and

$$\hat{S}_0 = \frac{1}{2}\Big((3\hat{\gamma}_0 - 1)\int_0^1 \frac{\hat{R}(1,t)}{t}dt + \hat{\gamma}_0\int_0^1 \frac{\hat{R}(1,t)}{t}\log t dt \\ -(2\hat{\gamma}_0+1)^2\int_0^1 \frac{\hat{R}(s,1)}{s^{1-\hat{\gamma}_0}}ds + 2(\hat{\gamma}_0+2)\hat{R}(1,1)\Big).$$

For very small $p \in (0,1)$, an extreme quantile is defined to be $x_p = F_1^{-1}(1-p)$. It is usually estimated by

$$\hat{x}_p = X_{n-k:n} + \hat{\sigma}\frac{(\frac{k}{np})^{\hat{\gamma}} - 1}{\hat{\gamma}}, \tag{3.12}$$

see Section 4.3 in de Haan and Ferreira (2006). Using our SSEs of γ and a, we define the SSE of x_p as

$$\hat{x}_{p_0} = X_{n-k:n} + \hat{\sigma}_0\frac{(\frac{k}{np})^{\hat{\gamma}_0} - 1}{\hat{\gamma}_0}. \tag{3.13}$$

Note that the estimators $\hat{\sigma}_0$ and $\hat{x}_{p_0}$ are obtained for $g = 0$. In Chapter 4, the two estimators are introduced for all values of g, in both bivariate and multivariate settings. We further derive the asymptotic normality of each estimator under general $g \in \mathbb{R}$.

For n and m as in Table 1, we again simulate $10,000$ times from the bivariate Cauchy distribution with $s = 0$. For γ ranging from -0.3 to 0.3, we aim to estimate the extreme quantile x_p for $p = \frac{1}{n}$. Table 3.3 shows the reduction in the variance of $\hat{\gamma}_0, \hat{\sigma}_0$, and $\hat{x}_{p,0}$, respectively, when compared with the pseudo-MLE estimators based on only $X_{n-k:n}, \ldots X_{n:n}$. Remarkably, the variance reduction when estimating the extreme quantile x_p ranges from 15% to more than 30%. It is much higher than the reductions when estimating γ or a and in almost all cases it even exceeds the sum of both reductions. To conclude this simulation section we briefly compare the here obtained sizeable variance reductions with the reductions $(1-\nu^2)R_g^2$ obtained from the asymptotic theory. It turns out that for the present sample sizes, when estimating γ, the asymptotic theory yields even higher reductions than the ones here obtained, partly due to the variability in the necessary estimation of the tail copulas, which is not reflected in the asymptotic variance. Based on limited comparisons it seems that for positive γ, asymptotic theory and simulations match somewhat better than in case γ is negative. Interestingly, we can show that for $\gamma \geq 0$ the asymptotic variance reduction of the extreme quantile estimator $\hat{x}_{p,0}$ is the same as that of $\hat{\gamma}_0$, cf. Theorem 4.3.1 in de Haan and Ferreira (2006). Now Table 3.3 indicates that for the important extreme quantile estimation, the variance reductions obtained in practice can be even higher than those inferred from the asymptotic theory.

Table 3.3: Variance reduction

	γ	-0.3	-0.2	-0.1	0	0.1	0.2	0.3
$\hat{\gamma}_0$ vs. $\hat{\gamma}$	$n > m$	6.6%	7.6%	9.3%	8.9%	9.9%	9.8%	9.5%
	$n = m$	9.4%	12.0%	13.6%	13.1%	14.5%	13.1%	13.9%
	$n < m$	8.7%	12.9%	14.5%	16.4%	16.5%	16.8%	16.6%
$\hat{\sigma}_0$ vs. $\hat{\sigma}$	$n > m$	4.5%	4.7%	5.4%	4.8%	5.5%	5.4%	5.3%
	$n = m$	6.1%	7.0%	7.5%	6.7%	7.8%	7.7%	7.6%
	$n < m$	6.1%	8.1%	8.4%	8.1%	8.9%	8.9%	8.7%
$\hat{x}_{p,0}$ vs. $\hat{x}_p$	$n > m$	15.7%	15.8%	17.5%	16.0%	16.3%	15.6%	14.9%
	$n = m$	22.6%	25.3%	25.2%	23.5%	23.4%	22.3%	21.2%
	$n < m$	27.6%	30.6%	32.5%	31.2%	31.6%	31.0%	30.5%

3.5 Application

In this section, we demonstrate an application using the SSE for analyzing forecasted precipitation data.

The national French weather service, Météo France, produces daily forecasted precipitation (in mm) at very high resolution (0.1° × 0.1°) covering the mainland of France, between 2012 and 2017. To improve the forecasting model, meteorologists want to check if the forecasted precipitation shares the same distribution as the observed precipitation at the same location, particularly in the right tail. Consequently, the goal of this study is to estimate quantities such as the extreme value index and extreme quantiles of the forecasted precipitation distribution. We focus on forecasting grid points that are close to an actual weather station.

Besides the forecasted precipitation, Météo France records at 123 weather stations the actual daily precipitation, between 1980 and 2017. We pair each weather station with a forecasting grid point that is closest to the station, and regard the two as the same location. When focusing on the fall seasons (91 days per year), at the 123 locations, we have 38 years actual precipitation data (3458 observations) with the last 6 years paired with forecasting data (546 observations). For the last 6 years, the paired data are dependent since the forecasting data are made to forecast the precipitation at the same location on the same day. Part of this dataset has been employed in a study comparing the spatial dependence structure of extreme forecasted precipitation and extreme observed precipitation in southern France; see Oesting and Naveau (2020).[1]

Since it is challenging to conduct extreme value analysis, such as extreme quantile estimation, for the forecasted precipitation based on only 546 observations, we make use of the available information in the actual precipitation to improve the estimation accuracy, exploiting the semi-supervised setting. To validate that our proposed methodology can be applied to the dataset, we perform two pre-tests on the actual precipitation data (3458 observations) at each station. Firstly, we test whether the actual precipitation at each station possesses the same distribution across time, using the test statistic T_2 proposed in Einmahl et al. (2016) (with $k = 200$). Secondly, we test whether the extreme precipitation at each station can be regarded as independent over time, based on testing whether the extremal index is significantly different from 1, using the sliding block estimator proposed in Berghaus and Bücher (2018) (with $b = 80$). We exclude all stations for which any of the two tests rejects the null at the 5% significance level. Eventually, for 91 stations both null hypotheses are not rejected, and we apply our proposed method to these 91 stations.

We use the SSE $\hat{\gamma}_g$ with $g = 0$ to estimate the extreme value index, and compare it with the pseudo-MLE $\hat{\gamma}$. For both estimators, we take $k = 136$. In particular, we estimate the variance reduction factor $(1-\nu^2)R_g^2$ to evaluate the improvement when using the SSE. In addition, we estimate the "once per 10 year" extreme rainfall, i.e. the quantile at the probability level $1 - 1/910$, by (3.12) and (3.13), to compare the impact of using the SSE on practically relevant quantities.

[1] We thank Marco Oesting and Philippe Naveau for providing this dataset.

Table 3.4 shows the results for three selected stations. We select the three stations from very distant areas: one from the south, one from the northwest, and one from the southwest.[2] For the station from the south, Nîmes, the estimated extreme value index is positive indicating a heavy-tailed distribution. The reduction in variance is estimated at 16%. The difference of the two estimates of the extreme value index leads to a substantial difference in the quantile estimates: the quantile estimated using the SSE exceeds the usual quantile estimate with roughly 50%. In contrast, for the station in the northwest, Boulogne sur Mer, the estimated extreme value index is about zero. The difference between the two point estimates is small, with the SSE having 17.5% variance reduction. The two quantile estimates are about the same. Finally, for the station in the southwest, Ciboure, both estimators lead to negative estimates, although not significantly different from zero. The variance reduction is at a pronounced level: 23.3%. The quantile estimate using the SSE is somewhat lower than the usual one.

Table 3.4: Estimation results for three stations

$\hat{\gamma}$ (MLE)	Quantile	$\hat{\gamma}_0$ (SSE)	Quantile SSE	Reduction	Station	LAT	LON
0.358	114.85	0.517	167.82	16.0%	Nîmes	43.86	4.41
-0.002	38.29	-0.018	37.07	17.5%	Boulogne s.M.	50.73	1.60
-0.056	59.32	-0.088	56.36	23.3%	Ciboure	43.39	-1.69

To further analyze the variance reduction factor, we plot the histogram of the estimated variance reduction factors across all 91 locations in Figure 3.3. The variance reduction using the SSE compared to the pseudo-MLE ranges from 9% to 25%, and is on average 16.3%. This confirms the improved performance of the SSE for the forecasted precipitation data.

[2] The last two columns show the latitude and longitude of each station.

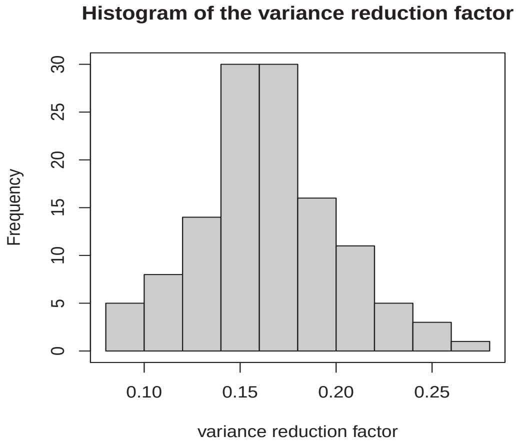


Figure 3.3: Histogram of the variance reduction factor across 91 stations

3.6 Proofs

We first present proofs for the one-covariate (bivariate) case and then extend the proofs to the multivariate case. The asymptotic normality for $\sqrt{k}(\hat{\gamma} - \gamma)$, the first component of the pair in Proposition 3.2.1, is established in Drees et al. (2004). However, we cannot directly use that proof, since we have to keep track of the joint behavior of $\hat{\gamma}$ and $\hat{g}$. Nevertheless, we mimic that proof for both $\hat{\gamma}$ and $\hat{g}$, with observing that $\hat{g}$ is based on dependent observations. In this respect the proof has to be adapted substantially. We begin with various lemmas which are needed for the main proofs.

Let C be a copula corresponding to the distribution function of $(-X_1, -Y_1)$. Let $(V_{1,1}, V_{1,2}), \ldots, (V_{n,1}, V_{n,2})$ be a random sample of size n from C and $V_{n+1,2}, \ldots, V_{n+m,2}$ be a random sample of size m from the uniform-$(0, 1)$ distribution, independent of the random sample from C. Clearly all the $V_{i,j}, i = 1, \ldots, n, j = 1, 2$, have also a uniform-$(0, 1)$ distribution. Write $X_i = F_1^{-1}(1 - V_{i,1}), i = 1, \ldots, n$, and $Y_l = F_2^{-1}(1 - V_{l,2}), l = 1, \ldots, n + m$. Then (X_i, Y_i), $i = 1, \ldots, n$, and $Y_{n+1}, \ldots, Y_{n+m}$ have the distributions as specified in the beginning of Section 3.2.

Consider the following uniform empirical distribution functions:

$$\Gamma_{n,j}(s) = \frac{1}{n}\sum_{i=1}^{n} 1_{[0,s]}(V_{i,j}), \quad 0 \leq s \leq 1, j = 1, 2,$$

$$\Gamma_{n+m}(t) = \frac{1}{n+m}\sum_{l=1}^{n+m} 1_{[0,t]}(V_{l,2}), \quad 0 \leq t \leq 1.$$

The corresponding uniform tail empirical processes are

$$w_{n,j}(s) = \sqrt{k}\left[\frac{n}{k}\Gamma_{n,j}\left(\frac{k}{n}s\right) - s\right], \quad 0 \leq s \leq 1, j = 1, 2,$$

$$w_{n+m}(t) = \sqrt{\frac{(n+m)k}{n}}\left[\frac{n}{k}\Gamma_{n+m}\left(\frac{k}{n}t\right) - t\right], \quad 0 \leq t \leq 1.$$

Define the Gaussian vector of processes (W_1, W_2, W_3), where W_j, $j = 1, 2, 3$, is a standard Wiener process on $[0, T], T > 0$, with covariances:

$$\begin{aligned} Cov(W_1(s), W_2(t)) &= R(s,t), \quad 0 \leq s, t \leq T, \\ Cov(W_1(s), W_3(t)) &= \nu R(s,t), \quad 0 \leq s, t \leq T, \\ Cov(W_2(s), W_3(t)) &= \nu(s \wedge t), \quad 0 \leq s, t \leq T. \end{aligned} \tag{3.14}$$

Let I denote the identity function. Then we have on $(D[0, T])^3$, for all $0 \leq \delta < \frac{1}{2}$, as $n \to \infty$,

$$\left(\frac{w_{n,1}}{I^\delta}, \frac{w_{n,2}}{I^\delta}, \frac{w_{n+m}}{I^\delta}\right) \xrightarrow{d} \left(\frac{W_1}{I^\delta}, \frac{W_2}{I^\delta}, \frac{W_3}{I^\delta}\right). \tag{3.15}$$

The proof of (3.15) is given in Ahmed and Einmahl (2019); note that in there $T = 1$, but the proof for arbitrary $T > 0$ follows similarly. Now using a Skorohod construction we obtain from (3.15) that

$$\sup_{0<s\leq T} \frac{|w_{n,j}(s) - W_j(s)|}{s^\delta} \xrightarrow{a.s.} 0, j = 1, 2, \quad \text{and} \quad \sup_{0<s\leq T} \frac{|w_{n+m}(s) - W_3(s)|}{s^\delta} \xrightarrow{a.s.} 0. \tag{3.16}$$

The processes in (3.16) are different from those in (3.15) but we keep the same notation, since the new vector $(w_{n,1}, w_{n,2}, w_{n+m})$ has the same distribution as the old vector and also the new vector (W_1, W_2, W_3) has the same distribution as the old vector. In the sequel the X_i and Y_i are transformations as above of the uniform-(0,1) random variables on which the $w_{n,j}$ are based. We continue with the processes satisfying (3.16).

For convenience we introduce the following notation. Let f_n, h_n be positive functions on $[l_n, u_n]$. Then we write, as $n \to \infty$,

$$f_n \overset{\mathbb{P}}{\asymp} h_n|_{l_n}^{u_n},$$

if both $f_n(s) = O_{\mathbb{P}}(h_n(s))$ and $h_n(s) = O_{\mathbb{P}}(f_n(s))$ hold uniformly for $s \in [l_n, u_n]$. This notation is useful for the following lemma, which can found in Shorack and Wellner (2009), p. 419.

Lemma 3.6.1 *Let $\Gamma_{n,j}^{-1}, j = 1, 2,$ be the empirical quantile functions corresponding to $\Gamma_{n,j}, j = 1, 2,$ respectively. Then, as $n \to \infty$,*

$$\Gamma_{n,j} \overset{\mathbb{P}}{\asymp} I|_{\Gamma_{n,j}^{-1}(1/(2n))}^{1} \quad \textit{and} \quad \Gamma_{n,j}^{-1} \overset{\mathbb{P}}{\asymp} I|_{1/(2n)}^{1}, \quad j = 1, 2.$$

The following lemma states the weighted convergence of the tail quantile processes corresponding to $\Gamma_{n,j}^{-1}, j = 1, 2$, to the processes $-W_1$ and $-W_2$ in (3.16).

Lemma 3.6.2 *Let $\Gamma_{n,j}^{-1}$, be the empirical quantile functions corresponding to $w_{n,j}$, $j = 1, 2$, in (3.16) and let W_j be as in (3.16), $j = 1, 2$. Then for any $\delta < \frac{1}{2}$, as $n \to \infty$,*

$$\sup_{\frac{1}{2k} \le s \le 1} \frac{|\sqrt{k}(\frac{n}{k}\Gamma_{n,j}^{-1}(\frac{k}{n}s) - s) + W_j(s)|}{s^{\delta}} \xrightarrow{\mathbb{P}} 0. \tag{3.17}$$

Proof of Lemma 3.6.2: Write $v_{n,j}(s) = \sqrt{k}(\frac{n}{k}\Gamma_{n,j}^{-1}(\frac{k}{n}s) - s), j = 1, 2$. Theorem 2.3 in Einmahl (1992) yields, as $n \to \infty$,

$$\sup_{\frac{1}{2k} \le s \le 1} \frac{|v_{n,j}(s) - W_{n,j}(s)|}{s^{\delta}} \xrightarrow{\mathbb{P}} 0, \tag{3.18}$$

where $W_{n,j}$ is an appropriate sequence of standard Wiener processes. Let W be a standard Wiener process and let $\varepsilon > 0$. It is well-known that there exist an $\eta > 0$, such that

$$\mathbb{P}\left(\sup_{0<s\leq\eta} \frac{|W(s)|}{s^\delta} \geq \frac{\varepsilon}{2}\right) \leq \frac{\varepsilon}{2}. \tag{3.19}$$

Combining (3.18) and (3.19) yields, for large n,

$$\mathbb{P}\left(\sup_{\frac{1}{2k}\leq s\leq\eta} \frac{|v_{n,j}(s)|}{s^\delta} \geq \varepsilon\right) \leq \varepsilon. \tag{3.20}$$

Combining (3.19), (3.20), (3.16), and Lemma 1 in Vervaat (1972), yields (3.17). □

The next lemma is very similar to Lemma 3.1 in Drees et al. (2004), but the lemma therein cannot be used here because we need specifically the approximation with the present W_1 in order to obtain the joint behavior of $\hat{\gamma}$ and $\hat{g}$.

Lemma 3.6.3 *Let $\varepsilon > 0$. Assume that* (3.3) *and* (3.4) *hold and $\sqrt{k}A(\frac{n}{k}) = O(1)$, as $n \to \infty$. Then for suitably chosen functions a and A in* (3.4)*, as $n \to \infty$,*

$$\sup_{\frac{1}{2k}\leq s\leq 1} s^{\gamma+1/2+\varepsilon} \left|\sqrt{k}\left(\frac{X_{n-[ks]:n} - U_1(\frac{n}{k})}{a(\frac{n}{k})} - \frac{s^{-\gamma}-1}{\gamma}\right) - s^{-\gamma-1}W_1(s) - \sqrt{k}A\left(\frac{n}{k}\right)\Psi(s^{-1})\right| \xrightarrow{\mathbb{P}} 0.$$

Proof of Lemma 3.6.3: From (3.4) we obtain inequality (2.3.17) in de Haan and Ferreira (2006): for any $\theta, \delta > 0$ to be specified later, there exists $t_0 = t_0(\theta, \delta)$ such that for all $t, tx \geq t_0$,

$$\left|\frac{\frac{U_1(tx)-U_1(t)}{a(t)} - \frac{x^\gamma-1}{\gamma}}{A(t)} - \Psi(x)\right| \leq \theta x^{\gamma+\rho}\max(x^\delta, x^{-\delta}).$$

We replace tx by $1/\Gamma_{n,1}^{-1}(\frac{k}{n}s)$ and t by $\frac{n}{k}$. Then we have, writing $\check{s} = \frac{n}{k}\Gamma_{n,1}^{-1}\left(\frac{k}{n}s\right)$, with probability tending to 1, as $n \to \infty$,

$$\left|\frac{X_{n-[ks]:n} - U_1(\frac{n}{k})}{a(\frac{n}{k})} - \frac{\check{s}^{-\gamma}-1}{\gamma} - A\left(\frac{n}{k}\right)\Psi\left(\check{s}^{-1}\right)\right| \leq \left|A\left(\frac{n}{k}\right)\right|\theta\check{s}^{-\gamma-\rho}\cdot\max(\check{s}^{-\delta}, \check{s}^{\delta}). \tag{3.21}$$

Define $f(s) = \frac{s^{-\gamma}-1}{\gamma}$. Then by a Taylor expansion for some $\check{\Theta}_n(s)$ between $\check{s}$ and s we have

$$f(\check{s}) - f(s) = f'(s)(\check{s} - s) + \frac{f''(\check{\Theta}_n(s))}{2}(\check{s} - s)^2.$$

Lemma 3.6.1 implies $\check{\Theta}_n \overset{\mathbb{P}}{\asymp} I|_{\frac{1}{2k}}^{1}$ and thus $f''\left(\check{\Theta}_n\right) \overset{\mathbb{P}}{\asymp} I^{-\gamma-2}|_{\frac{1}{2k}}^{1}$. Next, by Lemma 3.6.2 and the fact that for all $\delta_1 < \frac{1}{2}$, $\sup_{0\le s\le 1} |W_1(s)|/s^{\delta_1} = O_{\mathbb{P}}(1)$, we have that, as $n \to \infty$,

$$\sup_{\frac{1}{2k}\le s\le 1} \left(\check{s}-s\right)^2/s^{2\delta_1} = O_{\mathbb{P}}\left(\frac{1}{k}\right).$$

This and again Lemma 3.6.2 with $\delta = \delta_1$ yield, as $n \to \infty$, uniformly for all $\frac{1}{2k} \le s \le 1$,

$$f\left(\check{s}\right) - f(s) = -s^{-\gamma-1}\frac{1}{\sqrt{k}}\left(-W_1(s) + s^{\delta_1} o_{\mathbb{P}}(1)\right) + s^{-\gamma-2+2\delta_1} O_{\mathbb{P}}\left(\frac{1}{k}\right).$$

Choose δ_1 such that $\frac{1-\varepsilon}{2} < \delta_1 < \frac{1}{2}$. Then $\delta_1 > \frac{1}{2} - \varepsilon$ and $2\delta_1 + \varepsilon > 1$. Hence, as $n \to \infty$,

$$\sup_{\frac{1}{2k}\le s\le 1} s^{-\frac{3}{2}+\varepsilon+2\delta_1} \le \max\left(1, (2k)^{\frac{3}{2}-\varepsilon-2\delta_1}\right) = o(\sqrt{k}).$$

Therefore, as $n \to \infty$, uniformly for all $\frac{1}{2k} \le s \le 1$,

$$\begin{aligned} f(\check{s}) &= f(s) + \frac{1}{\sqrt{k}} s^{-\gamma-1}\left(W_1(s) + s^{\delta_1} o_{\mathbb{P}}(1) + s^{-1+2\delta_1} O_{\mathbb{P}}\left(\frac{1}{\sqrt{k}}\right)\right) \qquad (3.22) \\ &= f(s) + \frac{1}{\sqrt{k}} s^{-\gamma-1}\left(W_1(s) + s^{1/2-\varepsilon}\left(s^{\delta_1-1/2+\varepsilon} o_{\mathbb{P}}(1) + s^{-3/2+\varepsilon+2\delta_1} O_{\mathbb{P}}\left(\frac{1}{\sqrt{k}}\right)\right)\right) \\ &= \frac{s^{-\gamma}-1}{\gamma} + \frac{1}{\sqrt{k}} s^{-\gamma-1}\left(W_1(s) + s^{1/2-\varepsilon} o_{\mathbb{P}}(1)\right). \end{aligned}$$

From the mean value theorem, for some $\Theta_n(s)$ between $\check{s}$ and s

$$\Psi\left(\check{s}^{-1}\right) = \Psi(s^{-1}) - \Psi'(1/\Theta_n(s))(\Theta_n(s))^{-2}\left(\check{s}-s\right).$$

As above, $\Theta_n \overset{\mathbb{P}}{\asymp} I|_{\frac{1}{2k}}^{1}$, which implies that as $n \to \infty$, uniformly for $\frac{1}{2k} \le s \le 1$,

$$\left|\Psi'(1/\Theta_n(s))(\Theta_n(s))^{-2}\right| = s^{-\gamma-\rho-1}(1 + |\log s|)O_{\mathbb{P}}(1).$$

Hence, using Lemma 3.6.2 with $\delta = \delta_1$ (as above), we have uniformly for $\frac{1}{2k} \le s \le 1$,

$$A\left(\frac{n}{k}\right)\left(\Psi(\check{s}^{-1}) - \Psi(s^{-1})\right) = \frac{1}{\sqrt{k}} A\left(\frac{n}{k}\right) s^{-\gamma-\rho-1+\delta_1}(1 + |\log s|)O_{\mathbb{P}}(1).$$

With δ_1 chosen as above, we have that as $n \to \infty$, uniformly for $\frac{1}{2k} \le s \le 1$,

$$A\left(\frac{n}{k}\right)\Psi(\check{s}^{-1}) = A\left(\frac{n}{k}\right)\Psi(s^{-1}) + \frac{1}{\sqrt{k}} s^{-\gamma-\varepsilon-\frac{1}{2}} o_{\mathbb{P}}(1). \qquad (3.23)$$

Next consider the right-hand side of (3.21), where we take $\delta < 1/2$. Using Lemma 3.6.1, it can be bounded, uniformly for $\frac{1}{2k} \leq s \leq 1$, by

$$\begin{aligned}
&\theta \left| A\left(\frac{n}{k}\right) \right| s^{-\gamma-\rho-\delta} O_{\mathbb{P}}(1) = \theta\sqrt{k} \left| A\left(\frac{n}{k}\right) \right| \frac{1}{\sqrt{k}} s^{-\gamma-\delta} O_{\mathbb{P}}(1) \\
&= \theta \frac{1}{\sqrt{k}} s^{-\gamma-\varepsilon-1/2} s^{\varepsilon+1/2-\delta} O_{\mathbb{P}}(1) = \theta \frac{1}{\sqrt{k}} s^{-\gamma-\varepsilon-1/2} O_{\mathbb{P}}(1).
\end{aligned} \tag{3.24}$$

Now, plugging (3.22), (3.23), and (3.24) into inequality (3.21) and noting that $\theta > 0$ can be chosen arbitrarily small, we obtain the statement in the lemma. □

Define

$$Z_n(s) = \sqrt{k}\left(\frac{X_{n-[ks]:n} - X_{n-k:n}}{a(\frac{n}{k})} - \frac{s^{-\gamma}-1}{\gamma} \right).$$

Then for functions a and A as in Lemma 3.6.3, for any $\varepsilon > 0$, uniformly for $\frac{1}{2k} \leq s \leq 1$,

$$Z_n(s) = s^{-\gamma-1} W_1(s) - W_1(1) + \sqrt{k} A\left(\frac{n}{k}\right) \Psi(s^{-1}) + o_{\mathbb{P}}(1) s^{-\gamma-1/2-\varepsilon}.$$

Hence for $\gamma > -\frac{1}{2}$,

$$\sup_{\frac{1}{2k} \leq s \leq 1} s^{\gamma+1/2+\varepsilon} |Z_n(s)| = O_{\mathbb{P}}(1). \tag{3.25}$$

Proposition 3.6.1 *Under the conditions of Lemma 3.6.3, for $\gamma > -\frac{1}{2}$ and and $\gamma \neq 0$, with probability tending to 1, there exists a unique maximizer of the likelihood function based on $\{X_i\}_{i=1}^n$ denoted as $\hat{\gamma}$, such that as $n \to \infty$,*

$$\sqrt{k}(\hat{\gamma} - \gamma) - \frac{(\gamma+1)^2}{\gamma} \int_0^1 (s^{\gamma} - (2\gamma+1)s^{2\gamma}) Z_n(s) ds = o_{\mathbb{P}}(1),$$

and, for $\gamma = 0$,

$$\sqrt{k}\hat{\gamma} + \int_0^1 (2 + \log s) Z_n(s) ds = o_{\mathbb{P}}(1).$$

Proof of Proposition 3.6.1: The existence of $\hat{\gamma}$ follows from Theorem 4.1 in Zhou (2009). Then, using Lemma 3.6.3 and (3.25) above in conjunction with Lemma 3.2 in Drees et al. (2004). the result is obtained following the same steps as in the proof of Proposition 3.1 in Drees et al. (2004). □

To study the asymptotic behavior of $\hat{g}$ we need the following result. Define

$$\tilde{w}_n(s) = \frac{n}{\sqrt{k}}\left(\Gamma_{n+m}\left(\Gamma_{n,2}^{-1}\left(\frac{k}{n}s\right)\right) - \frac{k}{n}s\right) \quad \text{and} \quad \tilde{W}(s) = \nu W_3(s) - W_2(s).$$

Lemma 3.6.4 *Assume that F_2 is continuous and k satisfies* (3.3)*, then for any $0 \le \delta < \frac{1}{2}$, as $n \to \infty$,*

$$\sup_{\frac{1}{2k} \le s \le 1} \frac{|\tilde{w}_n(s) - \tilde{W}(s)|}{s^\delta} \xrightarrow{\mathbb{P}} 0.$$

Proof of Lemma 3.6.4: We have

$$\tilde{w}_n(s) = \sqrt{\frac{n}{n+m}} w_{n+m}\left(\frac{n}{k}\Gamma_{n,2}^{-1}\left(\frac{k}{n}s\right)\right) + \frac{n}{\sqrt{k}}\left(\Gamma_{n,2}^{-1}\left(\frac{k}{n}s\right) - \frac{k}{n}s\right).$$

Define $\hat{s} = \frac{n}{k}\Gamma_{n,2}^{-1}(\frac{k}{n}s)$. From Lemma 3.6.2 with $j = 2$, (3.3) and (3.19), we see that it suffices to show that, as $n \to \infty$,

$$\sup_{\frac{1}{2k} \le s \le 1} \frac{|w_{n+m}(\hat{s}) - W_3(s)|}{s^\delta} \xrightarrow{\mathbb{P}} 0. \tag{3.26}$$

Let $s_0 \in (0,1)$. We first handle the region $s \ge s_0$. Obviously we have $1/s^\delta \le 1/s_0^\delta$. By Lemma 3.6.2, as $n \to \infty$,

$$\sup_{\frac{1}{2k} \le s \le 1} |\hat{s} - s| \xrightarrow{\mathbb{P}} 0. \tag{3.27}$$

Using this, (3.16), and the uniform continuity of W_3 we obtain, as $n \to \infty$,

$$\sup_{s_0 \le s \le 1} \frac{|w_{n+m}(\hat{s}) - W_3(s)|}{s^\delta} \xrightarrow{\mathbb{P}} 0.$$

It remains to show that for $\varepsilon > 0$ there exists $s_0 \in (0,1)$ such that for large n

$$\mathbb{P}\left(\sup_{\frac{1}{2k} \le s \le s_0} \frac{|w_{n+m}(\hat{s}) - W_3(s)|}{s^\delta} \ge 3\varepsilon\right) \le 3\varepsilon.$$

Using again (3.19), for this it suffices to show that

$$\mathbb{P}\left(\sup_{\frac{1}{2k} \le s \le s_0} \frac{|w_{n+m}(\hat{s})|}{s^\delta} \ge 2\varepsilon\right) \le 2\varepsilon.$$

Using Lemma 3.6.1, the proof is complete if we show that for all $\varepsilon > 0, \kappa > 0$ there exists $s_0 \in (0,1)$ such that for large n

$$\mathbb{P}\left(\sup_{\frac{1}{2k}\le s\le s_0} \frac{|w_{n+m}(\hat{s})|}{\hat{s}^{\delta}} \ge 2\kappa\right) \le \varepsilon.$$

We have

$$\mathbb{P}\left(\sup_{\frac{1}{2k}\le s\le s_0} \frac{|w_{n+m}(\hat{s})|}{\hat{s}^{\delta}} \ge 2\kappa\right) \le \mathbb{P}\left(\sup_{0<t\le 2s_0} \left|\frac{w_{n+m}(t)}{t^{\delta}}\right| \ge \kappa\right) + \mathbb{P}(\hat{s} > 2s_0).$$

From (3.16) and (3.19), we have that for small enough $s_0 \in (0,1)$ the first term on the right is bounded by $\varepsilon/2$ for large n, and using (3.27) we obtain that the second term on the right also does not exceed $\varepsilon/2$ for large n. □

In the following we prove a result for the tail quantile process based on $\{\tilde{Y}_i\}_{i=1}^n$ instead of $\{X_i\}_{i=1}^n$. The proof of the next lemma uses Lemma 3.6.4, which is very similar to but easier than that of Lemma 3.6.3, and hence will be omitted.

Lemma 3.6.5 *Let $\varepsilon > 0$. Assume that F_2 is continuous and that* (3.3) *holds, then, as $n \to \infty$,*

$$\sup_{\frac{1}{2k}\le s\le 1} s^{g+1/2+\varepsilon}\left|\sqrt{k}\left(\frac{\left(\tilde{Y}_{n-[ks]:n} - \frac{(\frac{n}{k})^g-1}{g}\right)}{\left(\frac{n}{k}\right)^g} - \frac{s^{-g}-1}{g}\right) + s^{-g-1}\tilde{W}(s)\right| \xrightarrow{\mathbb{P}} 0.$$

Define

$$H_n(s) := \sqrt{k}\left(\frac{\tilde{Y}_{n-[ks]:n} - \tilde{Y}_{n-k:n}}{(\frac{n}{k})^g} - \frac{s^{-g}-1}{g}\right).$$

Then for any $\varepsilon > 0$, uniformly for $s \in [\frac{1}{2k}, 1]$,

$$H_n(s) = \tilde{W}(1) - s^{-g-1}\tilde{W}(s) + o_{\mathbb{P}}(1)s^{-g-1/2-\varepsilon}.$$

Hence for $g > -\frac{1}{2}$,

$$\sup_{\frac{1}{2k}\le s\le 1} s^{g+1/2+\varepsilon}|H_n(s)| = O_{\mathbb{P}}(1). \tag{3.28}$$

Next we show a version of Lemma 3.2 in Drees et al. (2004) based on $\{\tilde{Y}_i\}_{i=1}^n$.

Lemma 3.6.6 *Assume that F_2 is continuous and k satisfies* (3.3). *Let g_n be a sequence of random variables such that*

$$g_n = g + O_{\mathbb{P}}(k^{-1/2}). \tag{3.29}$$

Then, if $-1/2 < g < 0$ or $g > 0$, as $n \to \infty$,

$$\mathbb{P}\left(1 + g_n \frac{\tilde{Y}_{n-[ks]:n} - \tilde{Y}_{n-k:n}}{(\frac{n}{k})^g} \geq C_n s^{-g}, \text{ for all } s \in \left[\frac{1}{2k}, 1\right]\right) \to 1, \tag{3.30}$$

for some random variables $C_n > 0$ such that $1/C_n = O_{\mathbb{P}}(1)$.
If $g = 0$, as $n \to \infty$,

$$\mathbb{P}\left(1 + g_n\left(\tilde{Y}_{n-[ks]:n} - \tilde{Y}_{n-k:n}\right) \geq \frac{1}{2}, \text{ for all } s \in \left[\frac{1}{2k}, 1\right]\right) \to 1, \tag{3.31}$$

and

$$\sup_{s\in[0,1]} \tilde{Y}_{n-[ks]:n} - \tilde{Y}_{n-k:n} = O_{\mathbb{P}}(\log k). \tag{3.32}$$

Proof of Lemma 3.6.5: Consider first $-1/2 < g < 0$ or $g > 0$. Applying Lemma 3.6.1 to Γ_{n+m} and $\Gamma_{n,2}^{-1}$ yields, as $n \to \infty$,

$$\Gamma_{n+m}\left(\Gamma_{n,2}^{-1}\left(\frac{k}{n}I\right)\right) \overset{\mathbb{P}}{\asymp} \frac{k}{n} I|_{\frac{1}{2k}}^{1}.$$

Define $G_n(s) = \Gamma_{n+m}(\Gamma_{n,2}^{-1}(\frac{k}{n}s)) + \frac{1}{2(n+m)}$, $s \in (0,1]$. Hence, as $n \to \infty$,

$$G_n \overset{\mathbb{P}}{\asymp} \frac{k}{n} I|_{\frac{1}{2k}}^{1}. \tag{3.33}$$

Observe that for $g \neq 0$

$$\begin{aligned}
&s^g\left(1 + g_n \frac{\tilde{Y}_{n-[ks]:n} - \tilde{Y}_{n-k:n}}{(\frac{n}{k})^g}\right) = s^g\left(1 + \frac{g_n}{g}\frac{[(G_n(s))^{-g} - (G_n(1))^{-g}]}{(\frac{n}{k})^g}\right) \\
&= \frac{g_n}{g}\left(\frac{G_n(s)}{(ks)/n}\right)^{-g} + s^g\left[\left(1 - \left(\frac{G_n(1)}{k/n}\right)^{-g}\right) - \left(\frac{g_n}{g} - 1\right)\left(\frac{G_n(1)}{k/n}\right)^{-g}\right] \\
&=: T_1(s) + s^g[T_2 - T_3].
\end{aligned}$$

From (3.33) and $g_n/g \xrightarrow{\mathbb{P}} 1$, we have that $1/\inf_{s\in[1/(2k),1]} T_1(s) = O_{\mathbb{P}}(1)$, as $n \to \infty$. Lemma 3.6.4 for $s = 1$ yields that $T_2 = O_{\mathbb{P}}(1/\sqrt{k})$ and hence, since $g > -1/2$, $\sup_{s\in[1/(2k),1]} s^g \cdot T_2 \xrightarrow{\mathbb{P}} 0$. By the assumption on g_n and again (3.33) we obtain similarly $\sup_{s\in[1/(2k),1]} s^g \cdot T_3 \xrightarrow{\mathbb{P}} 0$. This yields (3.30).

In case $g = 0$, for $1/(2k) \le s \le 1$,

$$\tilde{Y}_{n-[ks]:n} - \tilde{Y}_{n-k:n} = -\log G_n(s) + \log G_n(1) \le 2\log A_n - \log s, \tag{3.34}$$

with

$$A_n = \max\left(\sup_{s\in[\frac{1}{2k},1]} \frac{G_n(s)}{\frac{k}{n}s}, \sup_{s\in[\frac{1}{2k},1]} \frac{\frac{k}{n}s}{G_n(s)} \right).$$

If $g_n \ge 0$, then $1 + g_n\left(\tilde{Y}_{n-[ks]:n} - \tilde{Y}_{n-k:n}\right) \ge 1$. If $g_n < 0$, then for $1/(2k) \le s \le 1$,

$$1 + g_n\left(\tilde{Y}_{n-[ks]:n} - \tilde{Y}_{n-k:n}\right) \ge 1 + g_n(2\log A_n + \log 2 + \log k).$$

Since, as $n \to \infty$, $A_n = O_{\mathbb{P}}(1)$ and $g_n = O_{\mathbb{P}}(k^{-1/2})$, we obtain (3.31). Finally, the sup in (3.32) is attained at $s = 1/(2k)$. Hence, (3.34) yields (3.32). □

Finally, the following proposition provides the asymptotic behavior of the pseudo-MLE based on $\{\tilde{Y}_i\}_{i=1}^n$.

Proposition 3.6.2 *Assume that F_2 is continuous and k satisfies* (3.3). *For $g > -\frac{1}{2}$ and $g \ne 0$, with probability tending to 1, there exists a unique maximizer of the likelihood function based on $\{\tilde{Y}_i\}_{i=1}^n$, denoted as $\hat{g}$, such that, as $n \to \infty$,*

$$\sqrt{k}(\hat{g} - g) - \frac{(g+1)^2}{g} \int_0^1 (s^g - (2g+1)s^{2g}) H_n(s) ds = o_{\mathbb{P}}(1).$$

and, for $g = 0$,

$$\sqrt{k}\hat{g} + \int_0^1 (2 + \log s) H_n(s) ds = o_{\mathbb{P}}(1).$$

Proof of Proposition 3.6.2: The existence of $\hat{g}$ follows the same steps as in the proof of Theorem 4.1 in Zhou (2009). Notice that although $\{\tilde{Y}_i\}_{i=1}^n$ are not i.i.d. observations, Lemma 3.6.5 guarantees that statistics based on the tail quantile process of $\{\tilde{Y}_i\}_{i=1}^n$, e.g. the Hill estimator for $g > 0$, possess similar asymptotic behavior as in the i.i.d. case, with the only difference that the random limit is driven by a proper functional of $\tilde{W}$ instead. Such asymptotic expansions are sufficient to ensure that the steps in the proof of Theorem 4.1 in Zhou (2009) can be realized. □

Then, by using Lemma 3.6.5, (3.28), and Lemma 3.6.6 and following the proof of Proposition 3.1 in Drees et al. (2004), we get the analogous result as in Proposition 3.6.1.

Proof of Proposition 3.2.1: Combining (3.16), Propositions 3.6.1 and 3.6.2 we obtain, as $n \to \infty$,

$$\left(\sqrt{k}(\hat{\gamma}-\gamma), \sqrt{k}(\hat{g}-g)\right) \xrightarrow{d} \left(\Omega, \tilde{\Omega}\right),$$

where

$$\Omega = \frac{(\gamma+1)^2}{\gamma}\int_0^1 (s^\gamma - (2\gamma+1)s^{2\gamma})(s^{-\gamma-1}W_1(s) - W_1(1))ds + \frac{\lambda(\gamma+1)}{(1-\rho)(1+\gamma-\rho)}$$

and

$$\tilde{\Omega} = \frac{(g+1)^2}{g}\int_0^1 (t^g - (2g+1)t^{2g})(\tilde{W}(1) - t^{-g-1}\tilde{W}(t))dt.$$

Since the Wiener processes involved have mean zero, we obtain immediately the mean of the limiting pair. Also the individual variances of the limiting pair follow readily, see Drees et al. (2004). It remains to determine the covariance. Note that $Cov(W_1(s), -\tilde{W}(t)) = (1-\nu^2)R(s,t)$. We have that

$$\begin{aligned}
Cov(\Omega, \tilde{\Omega}) &= (1-\nu^2)\frac{(\gamma+1)^2(g+1)^2}{\gamma g} \\
&\cdot \int_0^1\int_0^1 (s^\gamma - (2\gamma+1)s^{2\gamma})(t^g - (2g+1)t^{2g})(\frac{R(s,t)}{s^{\gamma+1}t^{g+1}} - \frac{R(s,1)}{s^{\gamma+1}} - \frac{R(1,t)}{t^{g+1}} + R(1,1))dsdt \\
&= (1-\nu^2)\frac{(\gamma+1)^2(g+1)^2}{\gamma g}\int_0^1\int_0^1 \left(\frac{1}{st} - \frac{(2g+1)}{st^{1-g}} - \frac{(2\gamma+1)}{s^{1-\gamma}t} + \frac{(2\gamma+1)(2g+1)}{s^{1-\gamma}t^{1-g}}\right) R(s,t) \\
&\quad - \left(\frac{t^g}{s} - \frac{(2g+1)t^{2g}}{s} - \frac{(2\gamma+1)t^g}{s^{1-\gamma}} + \frac{(2\gamma+1)(2g+1)t^{2g}}{s^{1-\gamma}}\right) R(s,1) \\
&\quad - \left(\frac{s^\gamma}{t} - \frac{(2g+1)s^\gamma}{t^{1-g}} - \frac{(2\gamma+1)s^{2\gamma}}{t} + \frac{(2\gamma+1)(2g+1)s^{2\gamma}}{t^{1-g}}\right) R(1,t) \\
&\quad + \left(s^\gamma t^g - (2g+1)s^\gamma t^{2g} - (2\gamma+1)s^{2\gamma}t^g + (2\gamma+1)(2g+1)s^{2\gamma}t^{2g}\right) R(1,1)dsdt.
\end{aligned}$$

Using a change of variables and the first order homogeneity of R, we obtain:

$$\begin{aligned}\int_0^1\int_0^1 \frac{R(s,t)}{st}dsdt &= \int_0^1\int_0^t \frac{R(s,t)}{st}dsdt + \int_0^1\int_0^s \frac{R(s,t)}{st}dtds \\ &= \int_0^1\int_0^1 \frac{R(ts,t)}{st}dsdt + \int_0^1\int_0^1 \frac{R(s,ts)}{st}dtds = \int_0^1 \frac{R(s,1)}{s}ds + \int_0^1 \frac{R(1,t)}{t}dt,\end{aligned}$$

$$\begin{aligned}\int_0^1\int_0^1 \frac{R(s,t)}{s^{1-\gamma}t}dsdt &= \int_0^1\int_0^t \frac{R(s,t)}{s^{1-\gamma}t}dsdt + \int_0^1\int_0^s \frac{R(s,t)}{s^{1-\gamma}t}dtds \\ &= \int_0^1\int_0^1 \frac{R(ts,t)}{(st)^{1-\gamma}}dsdt + \int_0^1\int_0^1 \frac{R(s,ts)}{s^{1-\gamma}t}dtds = \frac{1}{1+\gamma}\left[\int_0^1 \frac{R(s,1)}{s^{1-\gamma}}ds + \int_0^1 \frac{R(1,t)}{t}dt\right],\end{aligned}$$

and similarly

$$\int_0^1\int_0^1 \frac{R(s,t)}{st^{1-g}}dsdt = \frac{1}{1+g}\left[\int_0^1 \frac{R(s,1)}{s}ds + \int_0^1 \frac{R(1,t)}{t^{1-g}}dt\right].$$

Also,

$$\begin{aligned}&\int_0^1\int_0^1 \frac{R(s,t)}{s^{1-\gamma}t^{1-g}}dsdt = \int_0^1\int_0^t \frac{R(s,t)}{s^{1-\gamma}t^{1-g}}dsdt + \int_0^1\int_0^s \frac{R(s,t)}{s^{1-\gamma}t^{1-g}}dtds \\ &= \int_0^1\int_0^1 \frac{R(ts,t)}{s^{1-\gamma}t^{1-\gamma-g}}dsdt + \int_0^1\int_0^1 \frac{R(s,ts)}{s^{1-\gamma-g}t^{1-g}}dtds \\ &= \frac{1}{\gamma+g+1}\left[\int_0^1 \frac{R(s,1)}{s^{1-\gamma}}ds + \int_0^1 \frac{R(1,t)}{t^{1-g}}dt\right].\end{aligned}$$

Substituting the expressions for these four integrals involving $R(s,t)$ in the formula for $Cov(\Omega,\tilde{\Omega})$ above, we obtain that this covariance is equal to $(1-\nu^2)(\gamma+1)(g+1)R_g$. □

Proof of Theorem 3.2.1: From the uniform consistency of $\hat{R}$ on $[0,1]^2$, it can be shown that $\hat{R}_g \xrightarrow{\mathbb{P}} R_g$. Using the latter convergence in combination with Proposition 3.2.1 we obtain that, as $n\to\infty$,

$$\sqrt{k}(\hat{\gamma}_g-\gamma) = \sqrt{k}(\hat{\gamma}-\gamma) - \frac{1+\gamma}{1+g}R_g\sqrt{k}(\hat{g}-g) + o_{\mathbb{P}}(1). \tag{3.35}$$

Now Proposition 3.2.1 in conjunction with the continuous mapping theorem yields (3.8). □

The proof of Proposition 3.3.1 can be given along the same lines as that of Proposition 3.2.1 and will be omitted. Note that the lemmas and propositions needed for the proof of

Proposition 3.2.1 are of univariate nature and that hence immediately very similar lemmas can be stated (and proved) in the more-covariates case. Once these results are given, only a straightforward covariance calculation remains; cf. Ahmed and Einmahl (2019) for the joint weak convergence of all the tail empirical processes involved.

Proof of Theorem 3.3.1: From the uniform consistency of the tail copula estimators we obtain $\hat{H}_{1j}^{-1} \xrightarrow{\mathbb{P}} H_{1j}^{-1}, j = 1, \ldots, d$, which in combination with Proposition 3.3.1 yields that, as $n \to \infty$,

$$\sqrt{k}(\hat{\gamma}_g - \gamma) = \sqrt{k}(\hat{\gamma} - \gamma) + \frac{1+\gamma}{1+g} \sum_{j=2}^{d} \frac{H_{1j}^{-1}}{H_{11}^{-1}} \sqrt{k}(\hat{g}_j - g) + o_{\mathbb{P}}(1).$$

Now Proposition 3.3.1 and the continuous mapping theorem yield (3.11). □

Chapter 4

Extreme quantile estimation in semi-supervised models

4.1 Introduction

Extreme value theory is widely used in analysing various applications in finance, meteorology and environmental studies with rare events. Most applications regarding rare events involve estimation of high quantiles with low tail probabilities and other tail related measures. For instance, estimating risk measures (e.g. Value-at-Risk), finding the ultimate record in a specific athletic event (Einmahl and Magnus, 2008), and estimating the limit of human life span (Aarssen and de Haan (1994) and Einmahl et al. (2019)).

Following the peaks over threshold (POT) method, assume that X has a distribution function F_1. Let x^* be the right endpoint of F_1, $x^* = \sup\{x : F_1(x) < 1\}$, and define the excess distribution function

$$F_{1t}(x) = \mathbb{P}(X \leq x + t | X > t) = \frac{F_1(t+x) - F_1(t)}{1 - F_1(t)}, x > 0.$$

The distribution function F_1 belongs to the max-domain of attraction ($F_1 \in D(G_\gamma)$) if

$$\lim_{t \to x^*} F_{1t}(x\sigma(t)) = H_\gamma(x), \tag{4.1}$$

where $\sigma(t)$ is a positive function and

$$H_\gamma(x) = \begin{cases} 1 - (1 + \gamma x)^{-1/\gamma}, & \gamma \neq 0, \\ 1 - \exp(-x), & \gamma = 0, \end{cases}$$

for $x > 0$ if $\gamma \geq 0$ and $0 < x < -1/\gamma$ if $\gamma < 0$. H_γ is the well-known generalized Pareto distribution (GPD) with γ as the shape parameter, often referred to as the extreme value index (see Balkema and de Haan (1974)).

Let $X_1, \ldots, X_n$ be a random sample from distribution function F_1, where $F_1 \in D(G_\gamma)$. Smith (1987) introduced a pseudo-maximum likelihood estimator (pseudo-MLE) for the extreme value index γ and the scale function $\sigma(t)$, denoted as $\hat{\gamma}$ and $\hat{\sigma}$ respectively. These estimators are obtained by maximizing the likelihood function based on $\{X_i\}_{i=1}^n$, see equation (4) in Drees et al. (2004). The consistency of the pseudo-MLEs is obtained under the first order condition (see Zhou (2009)), that is

$$\lim_{t\to\infty} \frac{U_1(tx) - U_1(t)}{a(t)} = \frac{x^\gamma - 1}{\gamma}, \tag{4.2}$$

where $U_1 = F_1^{-1}(1 - 1/.)$ is the tail quantile of the distribution function F_1, $a(t)$ is a positive function and $a(t) = \sigma(U_1(t))$. Note that (4.2) is an equivalent representation of the max-domain of attraction condition in (4.1). The asymptotic normality of the pseudo-MLEs is proved by Drees et al. (2004) (for $\gamma > -1/2$) under the second order condition in (3.4). For a small probability $p = p(n)$, such that $\lim_{n\to\infty} p(n) = 0$, the extreme quantile x_p is defined as $x_p := \sup\{x | F(x) < 1 - p\}$, or equivalently, $x_p = U_1(1/p)$. Based on the first order condition in (4.2), the extreme quantile estimator (see Dekkers et al. (1989) and de Haan and Rootzén (1993))

$$\hat{x}_p = \hat{U}_1\left(\frac{n}{k}\right) + \hat{a}\left(\frac{n}{k}\right) \frac{(\frac{k}{np})^{\hat{\gamma}} - 1}{\hat{\gamma}}, \tag{4.3}$$

where $\hat{U}_1(\frac{n}{k}) = X_{n-k:n}$, $X_{n-k:n}$ is the $k - th$ order statistic of $\{X_i\}_{i=1}^n$ and $\hat{a}\left(\frac{n}{k}\right) = \hat{\sigma}$.

The pseudo-MLEs depend on the top k observations of the variable of interest $\{X_i\}_{i=1}^n$, which is a small fraction of the available data. The fact that extreme data are hard to get can clearly affect the accuracy of these parameters estimation and more importantly the extreme quantile estimator as a consequence. In Chapter 3, we use the semi-supervised model (SSM) as a suitable option to handle the difficulties related to the availability of extreme data. A semi-supervised estimator (SSE) for the extreme value index is obtained, which shows an improved performance over the pseudo-MLE of the extreme value index. Although it is essential to improve the estimation of the extreme value index, especially since it describes the tail heaviness of the distribution, it is even more crucial to improve the extreme quantile estimation, as it affects the reliability and accuracy of decision making in different applications.

Our main goal in this chapter is to obtain an improved estimator for the extreme quantile. We employ the SSM to first obtain the SSE for the scale $\sigma(U_1(\frac{n}{k})), k = \{1, \ldots, n-1\}$, and prove its asymptotic properties. Then we use both SSEs of the extreme value index and the scale to get a new adapted extreme quantile estimator, and establish its asymptotic behaviour. We demonstrate by simulation the substantial improvement gained in the performance of our adapted extreme quantile estimator compared to the extreme quantile estimator in (4.3), which is based on the pseudo-MLEs for the extreme value index and the scale. In addition, we show the prominent effect of the SSE of the scale on the improvement of the extreme quantile estimator.

This chapter is organized as follows. Section 4.2 previews the main results where we show the asymptotic normality related to the SSE of the scale and the adapted extreme quantile estimator based on one covariate. Section 4.3 introduces the asymptotic results of SSE of the scale and the adapted extreme quantile estimator based on multiple covariates. Using a simulation study, Section 4.4 demonstrates the superior performance of the SSE of the scale and the adapted extreme quantile estimator compared to the pseudo-MLE of the scale and the extreme quantile estimator in (4.3), in terms of variance reduction. Section 4.5 provides proofs of the presented results.

4.2 Main results: one covariate

4.2.1 Scale parameter

We use the SSM in the bivariate setting from Section 3.2. For $\gamma > -\frac{1}{2}$, we first estimate γ and $\sigma\left(U_1\left(\frac{n}{k}\right)\right)$ as $\hat{\gamma}$ and $\hat{\sigma}$, using the pseudo-MLE, based on $X_{n-k:n}, \ldots, X_{n:n}$, which are the top $k+1$ order statistics of $\{X_i\}_{i=1}^n$, for some $k \in \{1, \ldots, n-1\}$. Based on the random sample $Y_1, \ldots, Y_{n+m}$, from the distribution function F_2, we generate $\tilde{Y}_i, i = 1, \ldots, n$, with g as in (3.2). Denote the order statistics of $\{\tilde{Y}_i\}_{i=1}^n$, as $\tilde{Y}_{1:n} \leq \ldots \leq \tilde{Y}_{n:n}$. Using (3.2) $g > -\frac{1}{2}$ and $\left(\frac{n}{k}\right)^g$ are numbers that mimic the the extreme value index and the scale, which are estimated as $\hat{g}$ and $\tilde{\sigma}_g$ using the pseudo-MLE based on $\tilde{Y}_{n-k:n}, \ldots, \tilde{Y}_{n:n}$ using the same k as above.

Proposition 4.2.1 *For $\gamma > -\frac{1}{2}$, choose $g > -\frac{1}{2}$, assume F_2 is continuous and the second order condition* (3.4) *holds. Assume that k satisfies* (3.3), (3.1) *holds, and as $n \to \infty$, $\sqrt{k}A\left(\frac{n}{k}\right) \to \lambda \in \mathbb{R}$, then with probability tending to 1, there exist unique maximizers of the likelihood functions based on $\{X_i\}_{i=1}^n$ and $\{\tilde{Y}_i\}_{i=1}^n$, denoted as $(\hat{\sigma}, \tilde{\sigma}_g)$, such that*

$$\left(\sqrt{k}\left(\frac{\hat{\sigma}}{a\left(\frac{n}{k}\right)} - 1\right), \sqrt{k}\left(\frac{\tilde{\sigma}_g}{\left(\frac{n}{k}\right)^g} - 1\right)\right) \xrightarrow{d} N\left(\left[\frac{-\lambda\rho}{(1-\rho)(1+\gamma-\rho)}, 0\right], Z\right),$$

where

$$Z = \begin{bmatrix} 1 + (1+\gamma)^2 & (1-\nu^2)S_g \\ (1-\nu^2)S_g & (1-\nu^2)(1+(1+g)^2) \end{bmatrix},$$

with

$$\begin{aligned} S_g = {} & (\gamma+2)(g+2)R(1,1) - \frac{g(\gamma+1)}{\gamma}\int_0^1 \frac{R(s,1)}{s}ds - \frac{\gamma(g+1)}{g}\int_0^1 \frac{R(1,t)}{t}dt \\ & + \frac{(\gamma+1)(g+1)(2\gamma+1)}{\gamma g}\Big(\frac{\gamma+1}{g+1} - (g+1)(\gamma+1) + \frac{(\gamma+1)(g+1)(2g+1)}{\gamma+g+1} - 1\Big) \\ & \int_0^1 \frac{R(s,1)}{s^{1-\gamma}}ds + \frac{(\gamma+1)(g+1)(2g+1)}{\gamma g}\Big(\frac{g+1}{\gamma+1} - (g+1)(\gamma+1) \\ & + \frac{(\gamma+1)(g+1)(2\gamma+1)}{\gamma+g+1} - 1\Big)\int_0^1 \frac{R(1,t)}{t^{1-g}}dt. \end{aligned}$$

The SSE of the scale parameter is derived based on Proposition 4.2.1. By Slutsky's theorem, and for convenience of the derivations, we use the approximation of the bivariate distribution for $\left(\frac{\hat{\sigma}-a(\frac{n}{k})}{\hat{\sigma}}, \frac{\tilde{\sigma}_g-(\frac{n}{k})^g}{(\frac{n}{k})^g}\right)$. Assuming $\lambda = 0$, then it is approximated by a normal distribution with mean $[0, 0]$ and estimated covariance

$$\frac{1}{k}\hat{Z} = \frac{1}{k}\left[\begin{array}{cc} 1+(1+\hat{\gamma})^2 & (1-\frac{n}{n+m})\hat{S}_g \\ (1-\frac{n}{n+m})\hat{S}_g & (1-\frac{n}{n+m})(1+(1+g)^2) \end{array}\right],$$

where $\hat{S}_g$ is obtained by substituting γ with $\hat{\gamma}$ and the tail copula is estimated as in (3.6). By maximizing the bivariate likelihood function of $\left(\frac{\hat{\sigma}-a(\frac{n}{k})}{\hat{\sigma}}, \frac{\tilde{\sigma}_g-(\frac{n}{k})^g}{(\frac{n}{k})^g}\right)$ with respect to $a(\frac{n}{k})$, the SSE for the scale

$$\hat{\sigma}_g = \hat{\sigma}\left(1 - \frac{\hat{S}_g}{1+(1+g)^2}\left(\frac{\tilde{\sigma}_g}{(\frac{n}{k})^g} - 1\right)\right).$$

The following theorem shows the main result about the asymptotic distribution for a SSE of the scale.

Theorem 4.2.1 *For $\gamma > -\frac{1}{2}$, choose $g > -\frac{1}{2}$, assume F_2 is continuous,* (3.1)*, and* (3.4) *hold, k satisfies* (3.3)*, and $\sqrt{k}A\left(\frac{n}{k}\right) \to \lambda \in \mathbb{R}$, as $n \to \infty$. Then as $n \to \infty$,*

$$\sqrt{k}\left(\frac{\hat{\sigma}_g}{a\left(\frac{n}{k}\right)} - 1\right) \xrightarrow{d} N\left(\frac{-\lambda\rho}{(1-\rho)(1+\gamma-\rho)}, 1+(1+\gamma)^2 - (1-\nu^2)\frac{S_g^2}{1+(1+g)^2}\right). \quad (4.4)$$

Remark 4.2.1 *Theorem 4.2.1 shows that the asymptotic bias of the SSE is the same as that of the pseudo-MLE, while the asymptotic variance is reduced by $(1-\nu^2)\frac{S_g^2}{1+(1+g)^2}$.*

4.2.2 Adapted extreme quantile

We consider an adapted version of the extreme quantile estimator in (4.3), by plugging in the SSEs for the extreme value index and the scale instead of the MLEs, to get

$$\hat{x}_{p_g} = \hat{U}_1\left(\frac{n}{k}\right) + \hat{\sigma}_g\frac{(\frac{k}{np})^{\hat{\gamma}_g} - 1}{\hat{\gamma}_g}. \quad (4.5)$$

Theorem 4.2.2 *Assume* $\gamma > -\frac{1}{2}$ *and choose* $g > -\frac{1}{2}$*. Let* F_2 *be a continuous distribution function,* (3.1), (3.3) *and* (3.4) *hold, and* $\sqrt{k}A(\frac{n}{k}) \to \lambda \in \mathbb{R}$*, as* $n \to \infty$*. Then as* $n \to \infty$*,*

$$\begin{aligned}
&\sqrt{k}\left(\hat{\gamma}_g - \gamma, \frac{\hat{\sigma}_g}{a\left(\frac{n}{k}\right)} - 1, \frac{X_{n-k:n} - U_1\left(\frac{n}{k}\right)}{a\left(\frac{n}{k}\right)}\right) \qquad (4.6)\\
&\xrightarrow{d} N\left(\left[\frac{\lambda(\gamma+1)}{(1-\rho)(1+\gamma-\rho)}, \frac{-\rho\lambda}{(1-\rho)(1+\gamma-\rho)}, 0\right], K_{\hat{\gamma}_g,\hat{\sigma}_g,X_{n-k:n}}\right),
\end{aligned}$$

$$K_{\hat{\gamma}_g,\hat{\sigma}_g,X_{n-k:n}} = \begin{bmatrix} (1+\gamma)^2\left[1-(1-\nu^2)R_g^2\right] & -(1+\gamma)[1+(1-\nu^2)Q] & (1-\nu^2)M_1 \\ -(1+\gamma)[1+(1-\nu^2)Q] & 1+(1+\gamma)^2-(1-\nu^2)\frac{S_g^2}{1+(1+g)^2} & \gamma+(1-\nu^2)M_2 \\ (1-\nu^2)M_1 & \gamma+(1-\nu^2)M_2 & 1 \end{bmatrix},$$

where $Q = \frac{R_g S_g}{1+(1+g)^2} + \frac{1+\gamma}{1+g}R_g Q_{\hat{g},\hat{\sigma}} + \frac{S_g Q_{\hat{\gamma},\tilde{\sigma}_g}}{1+(1+g)^2}$*,* R_g *is as in* (3.5)*,*

$$\begin{aligned}
Q_{\hat{g},\hat{\sigma}} = {}& \frac{(1+g)^2}{\gamma g}\Big[\left(\frac{(\gamma+1)^2+g-(\gamma+1)^2(g+1)}{(g+1)(\gamma+1)}\right)R(1,1) + \frac{\gamma^2}{\gamma+1}\int_0^1 \frac{R(1,t)}{t}dt + \Big(2\gamma+1\\
&- \frac{(2g+1)(\gamma+1)(2\gamma+1)}{\gamma+g+1} + \frac{g(\gamma+1)(2\gamma+1)}{g+1}\Big)\int_0^1 \frac{R(s,1)}{s^{1-\gamma}}ds + \Big(\frac{2g+1}{g+1} - \frac{(2g+1)(\gamma+1)(2\gamma+1)}{\gamma+g+1}\\
&+ (2g+1)(\gamma+1) - \frac{2g+1}{\gamma+1}\Big)\int_0^1 \frac{R(1,t)}{t^{1-g}}dt\Big],
\end{aligned}$$

$$\begin{aligned}
Q_{\hat{\gamma},\tilde{\sigma}_g} = {}& \frac{(1+\gamma)(1+g)}{\gamma g}\Big[\left(\frac{(g+1)^2+\gamma-(g+1)^2(\gamma+1)}{(g+1)(\gamma+1)}\right)R(1,1) + \frac{g^2}{g+1}\int_0^1 \frac{R(s,1)}{s}ds\\
&+ \Big(\frac{2\gamma+1}{\gamma+1} - \frac{(2\gamma+1)(g+1)(2g+1)}{\gamma+g+1} + (2\gamma+1)(g+1) - \frac{2\gamma+1}{g+1}\Big)\int_0^1 \frac{R(s,1)}{s^{1-\gamma}}ds + \Big(2g+1\\
&- \frac{(2\gamma+1)(g+1)(2g+1)}{\gamma+g+1} + \frac{\gamma(g+1)(2g+1)}{\gamma+1}\Big)\int_0^1 \frac{R(1,t)}{t^{1-g}}dt\Big],
\end{aligned}$$

$$M_1 = \frac{(1+\gamma)(1+g)}{g}R_g\Big[(2g+1)\int_0^1 \frac{R(1,s)}{s^{1-g}}ds - \int_0^1 \frac{R(1,s)}{s}ds - \frac{g}{g+1}R(1,1)\Big],$$

and

$$M_2 = -\frac{1+g}{g(1+(1+g)^2)}S_g\Big[(2g+1)(g+1)\int_0^1 \frac{R(1,s)}{s^{1-g}}ds - \int_0^1 \frac{R(1,s)}{s}ds - \frac{g(g+2)}{g+1}R(1,1)\Big].$$

Based on Theorem 4.2.2, we present the main result in this section about the asymptotic normality of the adapted extreme quantile estimator.

Theorem 4.2.3 *For $\gamma > -\frac{1}{2}$ and choose $g > -\frac{1}{2}$. Assume that F_2 is continuous,* (3.1), (3.3), *and* (3.4) *hold, the second-order parameter ρ is negative or zero with γ negative, and $\sqrt{k}A(\frac{n}{k}) \to \lambda \in \mathbb{R}$, as $n \to \infty$. The probability level p satisfies $np = o(k)$ and $\log(np) = o(\sqrt{k})$, as $n \to \infty$. Then as $n \to \infty$,*

$$\sqrt{k}\frac{\hat{x}_{p_g} - x_p}{a\left(\frac{n}{k}\right) q_\gamma\left(\frac{k}{np}\right)} \xrightarrow{d} N(\lambda b_{x_{p_g}}, \sigma_{x_{p_g}}) \tag{4.7}$$

where, $q_\gamma(t) := \int_1^t s^{\gamma-1}\log(s)ds$, for $t > 1$,

$$b_{\hat{x}_{p_g}} = \begin{cases} \frac{(\gamma+1)}{(1-\rho)(\gamma-\rho+1),} & \gamma \geq 0 \neq \rho, \\ \frac{\rho(1+3\gamma+2\gamma^2)}{(1-\rho)(\gamma-\rho+1)(\gamma+\rho),} & \gamma < 0 \neq \rho, \\ 0, & \gamma < 0 = \rho, \end{cases}$$

and $\sigma_{\hat{x}_{p_g}} = (1+\gamma)^2\left[1-(1-\nu^2)R_g^2\right]$, for $\gamma \geq 0$, otherwise $\sigma_{\hat{x}_{p_g}} = 1 + 4\gamma + 5\gamma^2 + 2\gamma^3 - (1-\nu^2)\Big[(\gamma+1)^2R_g^2 + \gamma^2\frac{S_g^2}{1+(1+g)^2} - 2(\gamma+\gamma^2)Q - 2\gamma^2M_1 + 2\gamma^3M_2\Big]$.

Remark 4.2.2 *Note that the asymptotic bias is the same as that of the standard extreme quantile estimator based on the MLEs, while the asymptotic variance is different. The amount of change for the asymptotic variance of the adapted estimator than the standard estimator is $(1-\nu^2)R_g^2$ if $\gamma \geq 0$ and $(1-\nu^2)\Big[(\gamma+1)^2R_g^2 + \gamma^2\frac{S_g^2}{1+(1+g)^2} - 2(\gamma+\gamma^2)Q - 2\gamma^2M_1 + 2\gamma^3M_2\Big]$ if $\gamma < 0$. In the simulation section, we show how this change is in fact a reduction in the asymptotic variance of the standard extreme quantile estimator.*

4.3 Main results: multiple covariates

4.3.1 Scale parameter

Based on the SSM in the multivariate setting in Section 3.3, we extend the SSE of the scale to depend on a $(d-1)-$dimensional covariate. Let $\hat{\gamma}, \hat{\sigma}, \hat{g}_j$, and $\tilde{\sigma}_{g_j}, j = 2, \ldots, d$, be the pseudo-MLEs for $\gamma, \sigma(U_1\left(\frac{n}{k}\right))$, and $(d-1)$ times of g and $\left(\frac{n}{k}\right)^g$.

Proposition 4.3.1 *Assume $\gamma > -\frac{1}{2}$ and choose $g > -\frac{1}{2}$. Let $F_j, j = 2, \ldots, d$, be a continuous distribution function,* (3.3), (3.4), *and* (3.9) *hold, and as $n \to \infty$, $\sqrt{k}A(\frac{n}{k}) \to \lambda \in \mathbb{R}$, then with probability tending to 1, there exist a unique maximizers of the likelihood functions based on $\{X_i\}_{i=1}^n, \{\tilde{Y}_{i,2}\}_{i=1}^n, \ldots, \{\tilde{Y}_{i,d}\}_{i=1}^n$, denoted as $(\hat{\sigma}, \tilde{\sigma}_{g_2}, \ldots, \tilde{\sigma}_{g_d})$, such that*

$$\begin{aligned}&\left(\sqrt{k}\left(\frac{\hat{\sigma}}{a\left(\frac{n}{k}\right)}-1\right), \sqrt{k}\left(\frac{\tilde{\sigma}_{g_2}}{\left(\frac{n}{k}\right)^g}-1\right), \ldots, \sqrt{k}\left(\frac{\tilde{\sigma}_{g_d}}{\left(\frac{n}{k}\right)^g}-1\right)\right)\\ &\xrightarrow{d} N\left(\left[\frac{-\lambda\rho}{(1-\rho)(1+\gamma-\rho)}, 0, \ldots, 0\right], Z_d\right),\end{aligned}$$

$$Z_d = \begin{bmatrix} 1+(1+\gamma)^2 & z_{12} & \cdot\ \cdot\ \cdot & z_{1d} \\ z_{12} & (1-\nu^2)(1+(1+g)^2) & \cdot\ \cdot\ \cdot & z_{2d} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ z_{1d} & z_{2d} & \cdot\ \cdot\ \cdot & (1-\nu^2)(1+(1+g)^2) \end{bmatrix},$$

$$\begin{aligned}z_{1j} &= (1-\nu^2)[(\gamma+2)(g+2))R_{1j}(1,1) - \frac{g(\gamma+1)}{\gamma}\int_0^1 \frac{R_{1j}(s,1)}{s}ds - \frac{\gamma(g+1)}{g}\int_0^1 \frac{R_{1j}(1,t)}{t}dt\\ &+ \frac{(\gamma+1)(g+1)(2\gamma+1)}{\gamma g}\left(\frac{\gamma+1}{g+1} - (g+1)(\gamma+1) + \frac{(\gamma+1)(g+1)(2g+1)}{\gamma+g+1} - 1\right)\\ &\int_0^1 \frac{R_{1j}(s,1)}{s^{1-\gamma}}ds + \frac{(\gamma+1)(g+1)(2g+1)}{\gamma g}\left(\frac{g+1}{\gamma+1} - (g+1)(\gamma+1) + \frac{(\gamma+1)(g+1)(2\gamma+1)}{\gamma+g+1} - 1\right)\\ &\cdot \int_0^1 \frac{R_{1j}(1,t)}{t^{1-g}}dt] = (1-\nu^2)S_{g_j}, j = 2, \ldots, d,\end{aligned}$$

and

$$z_{ij} = (1-\nu^2)\left[(g+2)^2 R_{ij}(1,1) - (g+1)\left(\int_0^1 \frac{R_{ij}(s,1)}{s}ds + \int_0^1 \frac{R_{ij}(1,t)}{t}dt\right)\right],$$

$$i = 2, \ldots, d, j = i+1, \ldots, d.$$

Similar to the bivariate case, let $\lambda = 0$, the approximate multivariate distribution of $\left(\frac{\hat{\sigma}-a(\frac{n}{k})}{\hat{\sigma}}, \frac{\tilde{\sigma}_{g_2}-(\frac{n}{k})^g}{(\frac{n}{k})^g}, \ldots, \frac{\tilde{\sigma}_{g_d}-(\frac{n}{k})^g}{(\frac{n}{k})^g}\right)$ is normal with mean $[0, 0, \ldots, 0]$ and variance $\frac{1}{k}\hat{Z}_d$, where R_{ij} is estimated as in (3.6). By maximizing the approximate multivariate distribution with respect to $a\left(\frac{n}{k}\right)$, the SSE based on multiple covariates

$$\hat{\sigma}_g = \hat{\sigma}\left(1 + \sum_{j=2}^{d} \frac{\hat{Z}_{1j}^{-1}}{\hat{Z}_{11}^{-1}}\left(\frac{\tilde{\sigma}_{g_j}}{\left(\frac{n}{k}\right)^g} - 1\right)\right), \tag{4.8}$$

where $\hat{Z}_{ij}^{-1}$ is the entry in the i^{th} row and j^{th} column of the inverse of the matrix $\hat{Z}_d$.

Theorem 4.3.1 *Assume Z_d is invertible. Then under the conditions of Proposition 4.3.1, as $n \to \infty$,*

$$\sqrt{k}\left(\frac{\hat{\sigma}_g}{a\left(\frac{n}{k}\right)} - 1\right) \xrightarrow{d} N\left(\frac{-\lambda\rho}{(1-\rho)(1+\gamma-\rho)}, V_{\hat{\sigma}_g}\right), \tag{4.9}$$

where

$$V_{\hat{\sigma}_g} = 1 + (1+\gamma)^2 + \frac{1}{(Z_{11}^{-1})^2}\left((1-\nu^2)(1+(1+g)^2)\sum_{j=2}^{d}(Z_{1j}^{-1})^2 + 2\sum_{i=2}^{d}\sum_{j=i+1}^{d} Z_{1i}^{-1}Z_{1j}^{-1}z_{ij}\right)$$
$$+ \frac{2}{Z_{11}^{-1}}\sum_{j=2}^{d} Z_{1j}^{-1}z_{1j}.$$

4.3.2 Adapted extreme quantile

The adapted extreme quantile estimator is obtained by plugging the multivariate SSE of the extreme value index and scale in (4.5). In the following we show the asymptotic results related to the adapted extreme quantile estimator.

Theorem 4.3.2 *Assume $\gamma > -\frac{1}{2}$ and choose $g > -\frac{1}{2}$. Let $F_j, j = 2, \ldots, d$, be a continuous distribution function,* (3.1), (3.3), *and* (3.4) *hold, and $\sqrt{k}A(\frac{n}{k}) \to \lambda \in \mathbb{R}$, as $n \to \infty$.*

Then as $n \to \infty$,

$$\sqrt{k}\left(\hat{\gamma}_g - \gamma, \frac{\hat{\sigma}_g}{a\left(\frac{n}{k}\right)} - 1, \frac{X_{n-k:n} - U_1\left(\frac{n}{k}\right)}{a\left(\frac{n}{k}\right)}\right) \tag{4.10}$$

$$\xrightarrow{d} N\left(\left[\frac{\lambda(\gamma+1)}{(1-\rho)(1+\gamma-\rho)}, \frac{-\rho\lambda}{(1-\rho)(1+\gamma-\rho)}, 0\right], K_{\hat{\gamma}_g,\hat{\sigma}_g,X_{n-k:n}}\right),$$

$$K_{\hat{\gamma}_g,\hat{\sigma}_g,X_{n-k:n}} = \begin{bmatrix} V_{\hat{\gamma}_g} & -(1+\gamma)[1+(1-\nu^2)Q] & (1-\nu^2)M_1 \\ -(1+\gamma)[1+(1-\nu^2)Q] & V_{\hat{\sigma}_g} & \gamma+(1-\nu^2)M_2 \\ (1-\nu^2)M_1 & \gamma+(1-\nu^2)M_2 & 1 \end{bmatrix},$$

where $V_{\hat{\gamma}_g} = (1+\gamma)^2\left(1+\frac{1}{(H_{11}^{-1})^2}\left[2\sum_{i=1}^{d}\sum_{j=i+1}^{d} H_{1i}^{-1}H_{1j}^{-1}h_{ij} + (1-\nu^2)\sum_{j=2}^{d}(H_{1j}^{-1})^2\right]\right)$, H_{ij}^{-1} *is the entry in the i^{th} row and j^{th} column of the inverse of matrix H, which is defined in Proposition 3.3.1,*

$$Q = \frac{1}{H_{11}^{-1}Z_{11}^{-1}}\Big(\sum_{i=2}^{d} H_{1i}^{-1}Z_{1i}^{-1} - \frac{(1+g)^2}{g}\sum_{i=2}^{d}\sum_{\substack{j=2\\ i\neq j}}^{d} H_{1i}^{-1}Z_{1j}^{-1}Q_{\hat{g}_i,\tilde{\sigma}_{g_i}}\Big) - \frac{1+\gamma}{1+g}\sum_{i=2}^{d}\frac{H_{1i}^{-1}}{H_{11}^{-1}}Q_{\hat{g}_i,\hat{\sigma}}$$
$$- \sum_{i=2}^{d}\frac{Z_{1i}^{-1}}{Z_{11}^{-1}}Q_{\hat{\gamma},\tilde{\sigma}_{g_i}},$$

$$Q_{\hat{g}_i,\tilde{\sigma}_{g_i}} = \frac{(g+1)^2-(g+1)^3+g}{(g+1)^2}R_{ij}(1,1) + \frac{g^2}{g+1}\int\limits_0^1 \frac{R_{ij}(s,1)}{s}ds, i,j = 2,\ldots,d, i\neq j,$$

$$Q_{\hat{g}_i,\hat{\sigma}} = \frac{(1+g)^2}{\gamma g}\Big[\left(\frac{(\gamma+1)^2+g-(\gamma+1)^2(g+1)}{(g+1)(\gamma+1)}\right)R_{i1}(1,1) + \frac{\gamma^2}{\gamma+1}\int_0^1\frac{R_{i1}(1,t)}{t}dt$$
$$+\Big(2\gamma+1-\frac{(2g+1)(\gamma+1)(2\gamma+1)}{\gamma+g+1}+\frac{g(\gamma+1)(2\gamma+1)}{g+1}\Big)\int_0^1\frac{R_{i1}(s,1)}{s^{1-\gamma}}ds + \Big(\frac{2g+1}{g+1}$$
$$-\frac{(2g+1)(\gamma+1)(2\gamma+1)}{\gamma+g+1}+(2g+1)(\gamma+1)-\frac{2g+1}{\gamma+1}\Big)\int_0^1\frac{R_{i1}(1,t)}{t^{1-g}}dt\Big], i = 2,\ldots,d,$$

$$Q_{\hat{\gamma},\tilde{\sigma}_{g_i}} = \frac{(1+\gamma)(1+g)}{\gamma g}\Big[\Big(\frac{(g+1)^2+\gamma-(g+1)^2(\gamma+1)}{(g+1)(\gamma+1)}\Big)R_{1i}(1,1)$$
$$+\frac{g^2}{g+1}\int_0^1 \frac{R_{1i}(s,1)}{s}ds+\Big(\frac{2\gamma+1}{\gamma+1}-\frac{(2\gamma+1)(g+1)(2g+1)}{\gamma+g+1}+(2\gamma+1)(g+1)-\frac{2\gamma+1}{g+1}\Big)$$
$$\int_0^1 \frac{R_{1i}(s,1)}{s^{1-\gamma}}ds+\Big(2g+1-\frac{(2\gamma+1)(g+1)(2g+1)}{\gamma+g+1}+\frac{\gamma(g+1)(2g+1)}{\gamma+1}\Big)\int_0^1\frac{R_{1i}(1,t)}{t^{1-g}}dt\Big],$$
$i=2,\ldots,d,$

$$M_1 = \frac{(1+\gamma)(1+g)}{gH^{-1}_{11}}\sum_{i=2}^{d} H^{-1}_{1i}L_{1i},$$
$$L_{1i} = \Big[\int_0^1 \frac{R_{1i}(1,s)}{s}ds-(2g+1)\int_0^1\frac{R_{1i}(1,s)}{s^{1-g}}ds+\frac{g}{g+1}R_{1i}(1,1)\Big],$$

$$M_2 = \frac{(1+g)}{gZ^{-1}_{11}}\sum_{i=2}^{d} Z^{-1}_{1i}L_{2i},$$
$$L_{2i} = \Big[(2g+1)(g+1)\int_0^1\frac{R_{1i}(1,s)}{s^{1-g}}ds-\int_0^1\frac{R_{1i}(1,s)}{s}ds-\frac{g(g+2)}{g+1}R_{1i}(1,1)\Big], i=2,\ldots,d.$$

Theorem 4.3.3 *For $\gamma > -\frac{1}{2}$ and choose $g > -\frac{1}{2}$. Assume that F_2 is continuous, (3.1), (3.3), and (3.4) hold, the second-order parameter ρ is negative or zero with γ negative, $\sqrt{k}A(\frac{n}{k}) \to \lambda \in \mathbb{R}$, as $n\to\infty$. The probability level p satisfies $np = o(k)$ and $\log(np) = o(\sqrt{k})$, as $n\to\infty$. Then as $n\to\infty$,*

$$\sqrt{k}\frac{\hat{x}_{p_g}-x_p}{a\left(\frac{n}{k}\right)q_\gamma\left(\frac{k}{np}\right)} \xrightarrow{d} N(\lambda b_{x_{p_g}},\sigma_{x_{p_g}}) \tag{4.11}$$

where q_γ and $b_{x_{p_g}}$ as in theorem 4.2.3 and
$\sigma_{\hat{x}_{p_g}} = (1+\gamma)^2\left(1+\frac{1}{(H^{-1}_{11})^2}\left[2\sum_{i=1}^{d}\sum_{j=i+1}^{d}H^{-1}_{1i}H^{-1}_{1j}h_{ij}+(1-\nu^2)\sum_{j=2}^{d}(H^{-1}_{1j})^2\right]\right)$, *for* $\gamma \geq 0$,

otherwise

$$\sigma_{\hat{x}_{pg}} = 1 + 4\gamma + 5\gamma^2 + 2\gamma^3 + \Big(\frac{(1+\gamma)^2}{(H_{11}^{-1})^2}\Big[2\sum_{i=1}^{d}\sum_{j=i+1}^{d} H_{1i}^{-1}H_{1j}^{-1}h_{ij} + (1-\nu^2)\sum_{j=2}^{d}(H_{1j}^{-1})^2\Big]\Big)$$

$$+ \frac{\gamma^2}{(Z_{11}^{-1})^2}\Big[(1-\nu^2)(1+(1+g)^2)\sum_{j=2}^{d}(Z_{1j}^{-1})^2 + 2\sum_{i=2}^{d}\sum_{j=i+1}^{d} Z_{1i}^{-1}Z_{1j}^{-1}z_{ij}\Big]$$

$$+ \frac{2\gamma^2}{Z_{11}^{-1}}\sum_{j=2}^{d} Z_{1j}^{-1}z_{1j} + (1-\nu^2)\Big(2(\gamma+\gamma^2)Q + 2\gamma^2 M_1 - 2\gamma^3 M_2\Big).$$

4.4 Simulation

This section is divided into three parts. First, we investigate the behaviour of the SSE for the scale based on one and two covariates in finite samples simulations. We focus on the performance of the SSE of the scale compared to the pseudo-MLE in terms of variance reduction, using different values of g and m. Second, we study the effect of using SSEs for the extreme value index and the scale in the variance reduction of the extreme quantile estimator using different values of g. Third, we particularly check the effect of using the SSE for the scale on the variance reduction when estimating the extreme quantile.

We simulate $(\tilde{X}_i, Y_i)$ from the Cauchy distribution restricted to the first quadrant in dimensions $d = 2$ and $d = 3$. The Cauchy density is proportional to

$$(1 + xS^{-1}x^T)^{-(1+d)/2}.$$

Here for $d = 2$, S is a 2×2 matrix with 1 on the main diagonal and s off diagonal, and for $d = 3$, S is 3×3 matrix with 1 in the main diagonal, $S_{13} = S_{31} = S_{12} = S_{21} = s$ and $S_{23} = S_{32} = r$. When $d = 2$, we consider $s = 0$ and 0.5. For $d = 3$, we take $r = s = 0$ and 0.5, and $(r, s) = (0, 0.5)$ and $(0.5, 0)$. Then we transform the marginals of $\tilde{X}_i$ to X_i such that X_i has an extreme value index γ,

$$X_i = \begin{cases} \frac{(1-F_s(\tilde{X}_i))^{-\gamma}-1}{\gamma}, & \gamma \neq 0, \\ -\log\Big(1 - F_s(\tilde{X}_i)\Big), & \gamma = 0, \end{cases}$$

where F_s is the distribution function of $\tilde{X}_i$. The obtained $\{(X_i, Y_i)\}_{i=1}^{n}$ and $\{Y_i\}_{i=n+1}^{n+m}$ are the simulated data to be analysed.

First, we focus on the relative variance reduction of the SSE of the scale for the target variable X compared to the pseudo-MLE. We consider $\gamma = -0.3, 0, 0.3$, with different values for g, $n = 500, m = 1000$, and $k = 125$. Note that all the following results are based on $10,000$ replications.

Table 4.1 shows that the reduction ranges from 4% to 25%, for the SSE of the scale with one and two covariates. It is obvious that in case of using two covariates, the reduction substationally increases. Additionally, the increase of the tail dependence between the target variables and the covariate(s) positively affects the variance reduction. By contrast the increase of the tail dependence between the covariates leads to less variance reduction (see Table 2.1 for the approximated values of the tail dependence corresponding to the values of r and s).

Based on samples drawn from the Cauchy distribution with $s = 0$ or $s = r = 0$, in case of the SSE for the scale based on one and two covariates, we study the effect of having varying sizes of unlabelled data on the relative variance reduction. For different values of γ, Figure 4.1 shows a strictly increasing trend of the variance reduction with the increase of the unlabelled data size. The case where there is less or equal size of the unlabelled data than the labelled data, $m \leq n$, does not show a big difference in the variance reduction when using SSE based on one covariate or two covariates. The difference between having one or two covariates clearly increases with the increase of the unlabelled data over the labelled data ($m > n$).

We then show the variance reduction of the SSE of the scale for a wider range of g and γ. Figure 4.2 shows the variance reduction for each γ based on different values of g. The results are obtained using the same sample size setting as mentioned above. According to Figure 4.2, the variance reduction increases by the decrease of absolute difference of γ and g. The case where $g = 0$ has a relatively good performance for different values of γ.

Second, we use the SSEs of the extreme value index and the scale in the extreme

γ	g	$d = 2$		$d = 3$			
		$s = 0$	$s = 0.5$	$s = r = 0$	$s = r = 0.5$	$s = 0$ $r = 0.5$	$s = 0.5$ $r = 0$
	-0.25	4.58%	9.25%	9.21%	16.08%	9.48%	19.58%
-0.3	-0.125	4.92%	9.9%	9.65%	16.5%	9.89%	19.40%
	0	4.36%	11.17%	9.42%	16.06%	9.69%	18.26%
	-0.125	7.48%	12.52%	11.61%	18.42%	10.26%	21.26%
0	0	8.38%	13.61%	12.59%	19.57%	11.16%	21.96%
	0.125	8.97%	14.25%	13.16%	20.12%	11.7%	22%
	0	7.83%	14.77%	12%	19.5%	11.4%	23.56%
0.3	0.125	8.41%	15.99%	12.9%	21.04%	12.46%	24.84%
	0.25	8.79%	16.72%	13.4%	21.96%	13.1%	25.43%

Table 4.1: Variance reduction for the scale parameter for $n = 500, m = 1000$, and $k = 125$

quantile estimator, instead of the pseudo-MLEs, to study how that improves its performance using different values of g. For $n = 500, m = 1000$, and $k = 125$, we estimate the extreme quantile x_p with $p = 0.002$. Table 4.2 shows the results when using SSEs based on one covariate. The extreme quantile variance reduction ranges from 17% to 33%. Table 4.3 shows the case of the SSEs based on two covariates, where the variance reduction reaches 43%. There is only one case where we have the variance of the adapted extreme quantile estimator is larger than the variance of the standard quantile estimator based on the pseudo-MLEs. This happens when considering g far from the true positive γ. Based on these results, it is recommended to consider $g \geq -0.125$.

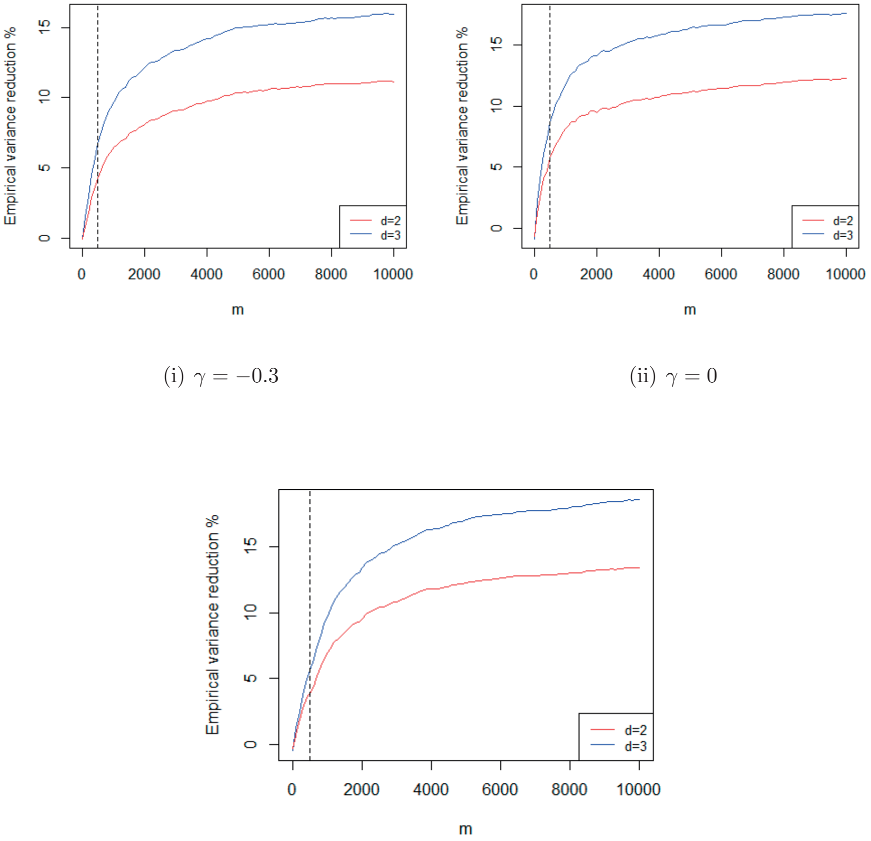
Third, we study the effect of using the SSEs for the extreme value index and the scale, compared to only using the SSE for the extreme value index, on improving the variance of the extreme quantile estimator. We estimate the extreme quantile with $p = 0.002$, and 0.001, with $g = 0$. Figure 4.3 shows the remarkable effect in the variance reduction when

Table 4.2: Variance reduction for the extreme quantile based on one covariate

g	γ						
	-0.3	-0.2	-0.1	0	0.1	0.2	0.3
-0.25	24.07%	25.12%	23.6%	22.52%	20.57%	19.1%	16.56%
-0.125	27.92%	30.07%	29.26%	28.03%	26.74%	25.54%	22.98%
0	27.57%	31.34%	31.89%	31.28%	30.78%	30.02%	27.81%
0.125	23.99%	29.85%	31.95%	32.44%	32.6%	32.42%	30.75%
0.25	20.57%	27.74%	30.95%	32.34%	33.09%	33.46%	32.42%

Table 4.3: Variance reduction for the extreme quantile based on two covariates

g	γ						
	-0.3	-0.2	-0.1	0	0.1	0.2	0.3
-0.25	34.03%	33.91%	27.93%	21.35%	10.89%	0.68%	-6.04%
-0.125	35.83%	39.9%	38.86%	35.46%	33.62%	29.95%	28.02%
0	32.3%	40%	40.87%	38.99%	39.86%	38.69%	38.65%
0.125	25.42%	37.78%	40.33%	39.68%	41.42%	41.38%	42.21%
0.25	16.14%	33.94%	38.52%	39.17%	41.36%	41.98%	43.31%



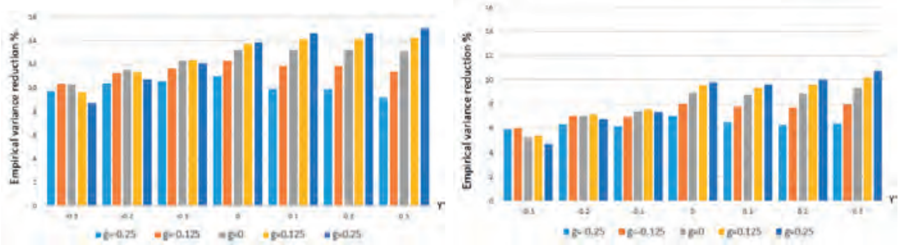
(i) $\gamma = -0.3$

(ii) $\gamma = 0$

(iii) $\gamma = 0.3$

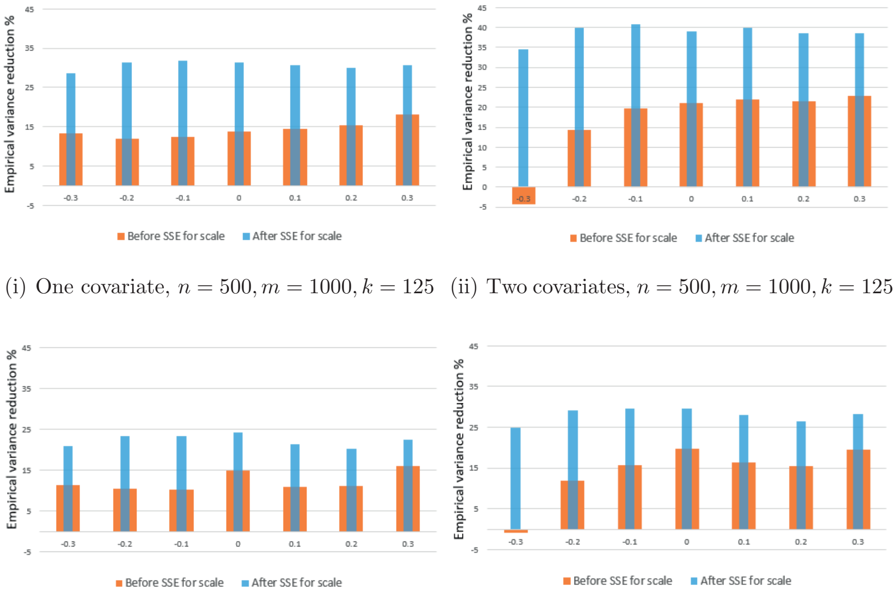
Figure 4.1: Relative variance reduction for $n = 500, k = 125$ and $g = 0$. The vertical line represents $n = m$

using the SSE for the scale, which is doubled in most of the cases. The two cases where there is negative reduction correspond to having 12% and 15.74% as a variance reduction for the SSE of the extreme value index. These reductions are not reflected as a reduction



(i) One covariate (ii) Two covariates

Figure 4.2: Relative variance reduction for $n = 500, k = 125$



(i) One covariate, $n = 500, m = 1000, k = 125$ (ii) Two covariates, $n = 500, m = 1000, k = 125$

(iii) One covariate, $n = m = 1000, k = 250$ (iv) Two covariates, $n = 1000, k = 250$

Figure 4.3: Relative variance reduction for the extreme quantile

in the variance of the extreme quantile estimator before using the SSE of the scale.

4.5 Proofs

In this section we first show the asymptotic normality for $\sqrt{k}\left(\frac{\hat{\sigma}}{a\left(\frac{n}{k}\right)}-1\right)$ and $\sqrt{k}\left(\frac{\tilde{\sigma}_g}{\left(\frac{n}{k}\right)^g}-1\right)$, using the tail quantile processes in Lemma 3.6.3 and Lemma 3.6.5. Second, we incorporate both results to obtain the asymptotic distribution for the SSE of the scale. Third, we derive the joint asymptotic distribution for the SSEs of the extreme value index and the scale and $X_{n-k:n}$. Finally we use the joint asymptotic results to obtain the asymptotic distribution for the adapted extreme quantile estimator. We present the proofs for the results related to the SSE with one covariate, then we extend to the SSE with multiple covariates.

Let C be a copula corresponding to the distribution function of $(-X_1,-Y_1)$. Write $X_i = F_1^{-1}(1-V_{i,1}), i=1,\ldots,n,$ and $Y_l = F_2^{-1}(1-V_{l,2}), l=1,\ldots,n+m,$ where $(V_{1,1},V_{1,2}),\ldots,(V_{n,1},V_{n,2})$ is a random sample of size n from C and $V_{n+1,2},\ldots,V_{n+m,2}$ is a random sample of size m from the uniform-$(0,1)$ distribution, independent of the random sample from C. Consider the uniform empirical distribution functions: $\Gamma_{n,j}(s) = \frac{1}{n}\sum_{i=1}^{n} 1_{[0,s]}(V_{i,j}), \quad 0\le s\le 1, j=1,2,$ $\Gamma_{n+m}(t) = \frac{1}{n+m}\sum_{l=1}^{n+m} 1_{[0,t]}(V_{l,2}), \quad 0\le t\le 1,$ and the corresponding uniform tail empirical processes are $w_{n,j}(s)=\sqrt{k}\left[\frac{n}{k}\Gamma_{n,j}\left(\frac{k}{n}s\right)-s\right], 0\le s\le 1, j=1,2,$ and $w_{n+m}(t)=\sqrt{\frac{(n+m)k}{n}}\left[\frac{n}{k}\Gamma_{n+m}\left(\frac{k}{n}t\right)-t\right], 0\le t\le 1$. Define the Gaussian vector of processes (W_1,W_2,W_3), where W_j, $j=1,2,3$, is a standard Wiener process on $[0,T], T>0$, with covariances as in (3.14). Under proper Skorokhod construction, we get (3.16), that is

$$\sup_{0<s\le T}\frac{|w_{n,j}(s)-W_j(s)|}{s^{\delta}}\xrightarrow{a.s.}0, j=1,2, \quad \text{and} \quad \sup_{0<s\le T}\frac{|w_{n+m}(s)-W_3(s)|}{s^{\delta}}\xrightarrow{a.s.}0,$$

as $n\to\infty$, see Section 3.6 for the proof and definitions related to (3.16).

Define

$$Z_n(s)=\sqrt{k}\left(\frac{X_{n-[ks]:n}-X_{n-k:n}}{a\left(\frac{n}{k}\right)}-\frac{s^{-\gamma}-1}{\gamma}\right).$$

For a suitably chosen functions a and A, based on Lemma 3.6.3, for any $\varepsilon > 0$ uniformly for $\frac{1}{2k} \leq s \leq 1$, as $n \to \infty$,

$$Z_n(s) = s^{-\gamma-1}W_1(s) - W_1(1) + \sqrt{k}A\Big(\frac{n}{k}\Big)\Psi(s^{-1}) + o_{\mathbb{P}}(1)s^{-\gamma-1/2-\varepsilon}.$$

Hence for $\gamma > -\frac{1}{2}$, and $\varepsilon > 0$, $\sup\limits_{\frac{1}{2k}\leq s\leq 1} s^{\gamma+1/2+\varepsilon}|Z_n(s)| = O_{\mathbb{P}}(1)$.

Proposition 4.5.1 *Assume the second order condition* (3.4) *holds, k satisfies* (3.3), *and* $\sqrt{k}A\left(\frac{n}{k}\right) = O(1)$, *as $n \to \infty$. For $\gamma > -1/2$, and $\gamma \neq 0$, with probability tending to 1, there exist a unique maximizer of the likelihood functions based on $\{X_i\}_{i=1}^n$ denoted as $\hat{\sigma}$, such that as $n \to \infty$,*

$$\sqrt{k}\left(\frac{\hat{\sigma}}{a\left(\frac{n}{k}\right)} - 1\right) - \frac{\gamma+1}{\gamma}\int_0^1 ((\gamma+1)(2\gamma+1)s^{2\gamma} - s^{\gamma})Z_n(s)ds = o_{\mathbb{P}}(1).$$

For $\gamma = 0$, as $n \to \infty$,

$$\sqrt{k}\left(\frac{\hat{\sigma}}{a\left(\frac{n}{k}\right)} - 1\right) - \int_0^1 (3 + \log s)Z_n(s)ds = o_{\mathbb{P}}(1).$$

Proof of Proposition 4.5.1: The existence of $\hat{\sigma}$ follows from Theorem 4.1 in Zhou (2009). Using Lemma 3.6.3, (3.25) and Lemma 3.2 from Drees et al. (2004), then following the same steps of Proposition 3.1 proof in Drees et al. (2004) the proposition is proved. □

From Proposition 3.6.1 and Proposition 4.5.1, it follows that for $\gamma > -\frac{1}{2}$ and $\gamma \neq 0$,

$$\begin{aligned}\sqrt{k}(\hat{\gamma} - \gamma) - \frac{(\gamma+1)^2}{\gamma}\sqrt{k}A\left(\frac{n}{k}\right)\int_0^1 (s^{\gamma} - (2\gamma+1)s^{2\gamma})\Psi(s^{-1})ds \\ \xrightarrow{\mathbb{P}} \frac{(\gamma+1)^2}{\gamma}\int_0^1 (s^{\gamma} - (2\gamma+1)s^{2\gamma})(s^{-\gamma-1}W_1(s) - W_1(1))ds,\end{aligned}$$

and

$$\begin{aligned}\sqrt{k}\left(\frac{\hat{\sigma}}{a\left(\frac{n}{k}\right)} - 1\right) - \frac{\gamma+1}{\gamma}\sqrt{k}A\left(\frac{n}{k}\right)\int_0^1 ((\gamma+1)(2\gamma+1)s^{2\gamma} - s^{\gamma})\Psi(s^{-1})ds \\ \xrightarrow{\mathbb{P}} \frac{\gamma+1}{\gamma}\int_0^1 ((\gamma+1)(2\gamma+1)s^{2\gamma} - s^{\gamma})(s^{-\gamma-1}W_1(s) - W_1(1))ds,\end{aligned}$$

as $n \to \infty$. The convergence follows jointly with the same limiting process W_1.

Corollary 4.5.1 *Under the conditions of Proposition 4.5.1, and $\sqrt{k}A\left(\frac{n}{k}\right) \to \lambda$, as $n \to \infty$, as $n \to \infty$,*

$$\sqrt{k}\left(\hat{\gamma} - \gamma, \frac{\hat{\sigma}}{a\left(\frac{n}{k}\right)} - 1\right)$$
$$\xrightarrow{d} N\left(\left[\frac{\lambda(\gamma+1)}{(1-\rho)(1+\gamma-\rho)}, \frac{-\rho\lambda}{(1-\rho)(1+\gamma-\rho)}\right], \left[\begin{array}{cc} (1+\gamma)^2 & -(1+\gamma) \\ -(1+\gamma) & (1+(1+\gamma)^2) \end{array}\right]\right).$$

Define $\tilde{w}_n(s) = \frac{n}{\sqrt{k}}\left(\Gamma_{n+m}\left(\Gamma_{n,2}^{-1}\left(\frac{k}{n}s\right)\right) - \frac{k}{n}s\right)$ and $\tilde{W}(s) = \nu W_3(s) - W_2(s)$, where W_2 and W_3 are defined as in (3.16). Define

$$H_n(s) = \sqrt{k}\left(\frac{\tilde{Y}_{n-[ks]:n} - \tilde{Y}_{n-k:n}}{\left(\frac{n}{k}\right)^g} - \frac{s^{-g}-1}{g}\right),$$

then for any $\varepsilon > 0$ uniformly for $\frac{1}{2k} \leq s \leq 1$, based on Lemma 3.6.5, as $n \to \infty$,

$$H_n(s) = \tilde{W}(1) - s^{-g-1}\tilde{W}(s) + o_{\mathbb{P}}(1)s^{-g-1/2-\varepsilon},$$

and for $g > -\frac{1}{2}$, $\sup\limits_{\frac{1}{2k}\leq s\leq 1} s^{g+1/2+\varepsilon}|H_n(s)| = O_{\mathbb{P}}(1)$.

Proposition 4.5.2 *Assume that F_2 is continuous and k satisfies (3.3). For $g > -1/2$ and $g \neq 0$, with probability tending to 1, there exists a unique maximizer of the likelihood function based on $\{\tilde{Y}_i\}_{i=1}^n$, denoted as $\tilde{\sigma}_g$, such that, as $n \to \infty$,*

$$\sqrt{k}\left(\frac{\tilde{\sigma}_g}{\left(\frac{n}{k}\right)^g} - 1\right) - \frac{g+1}{g}\int_0^1 ((g+1)(2g+1)s^{2g} - s^g)H_n(s)ds = o_{\mathbb{P}}(1),$$

For $g = 0$,

$$\sqrt{k}\left(\tilde{\sigma}_g - 1\right) - \int_0^1 (3 + \log s)H_n(s)ds = o_{\mathbb{P}}(1).$$

Proof of Proposition 4.5.2: The existence of $\tilde{\sigma}_g$ follows the same steps as the proof of Theorem 4.1 in Zhou (2009), similar to the proof of Proposition 3.6.2. Then the proposition is proved following the proof of Proposition 3.1 from Drees et al. (2004) using Lemma 3.6.5, (3.28) and Lemma 3.6.6. □

From Proposition 3.6.2 and Proposition 4.5.2, for $g > -\frac{1}{2}$ and $g \neq 0$, as $n \to \infty$,

$$\sqrt{k}(\hat{g} - g) \xrightarrow{\mathbb{P}} \frac{(g+1)^2}{g} \int_0^1 \left(s^g - (2g+1)s^{2g}\right) \left(\tilde{W}(1) - s^{-g-1}\tilde{W}(s)\right) ds,$$

and

$$\sqrt{k}\left(\frac{\tilde{\sigma}_g}{\left(\frac{n}{k}\right)^g} - 1\right) \xrightarrow{\mathbb{P}} \frac{g+1}{g} \int_0^1 ((g+1)(2g+1)s^{2g} - s^g) \left(\tilde{W}(1) - s^{-g-1}\tilde{W}(s)\right) ds.$$

The joint convergence follows by (3.16) with the same limiting process $\tilde{W}$.

Corollary 4.5.2 *Under the same conditions of Proposition 4.5.2, as $n \to \infty$,*

$$\sqrt{k}\left(\hat{g} - g, \frac{\tilde{\sigma}_g}{\left(\frac{n}{k}\right)^g} - 1\right) \xrightarrow{d} N\left([0,0], (1-\nu^2)\left[\begin{array}{cc} (1+g)^2 & -(1+g) \\ -(1+g) & (1+(1+g)^2) \end{array}\right]\right).$$

Proof of Proposition 4.2.1: Combining (3.16), Proposition 4.5.1, and Proposition 4.5.2. We have that as $n \to \infty$,

$$\left(\sqrt{k}\left(\frac{\hat{\sigma}}{a\left(\frac{n}{k}\right)} - 1\right), \sqrt{k}\left(\frac{\tilde{\sigma}_g}{\left(\frac{n}{k}\right)^g} - 1\right)\right) \xrightarrow{d} (\Sigma_\gamma, \tilde{\Sigma}_g),$$

where $\Sigma_\gamma = \frac{\gamma+1}{\gamma}\int_0^1((\gamma+1)(2\gamma+1)s^{2\gamma} - s^\gamma)(s^{-\gamma-1}W_1(s) - W_1(1))ds - \frac{\lambda\rho}{(1-\rho)(1+\gamma-\rho)}$ and $\tilde{\Sigma}_g = \frac{g+1}{g}\int_0^1((g+1)(2g+1)s^{2g} - s^g)(\tilde{W}(1) - s^{-g-1}\tilde{W}(s))ds$.
Based on the Wiener processes which are involved in the limiting distributions, we can obtain immediately the means and variances of the limiting pair as in the previous corollaries.

It remains to determine the covariance,

$$Cov(\Sigma_\gamma, \tilde{\Sigma}_g) = (1-\nu^2)\frac{(\gamma+1)(g+1)}{\gamma g}\int_0^1\int_0^1 ((\gamma+1)(2\gamma+1)s^{2\gamma} - s^\gamma)((g+1)(2g+1)t^{2g} - t^g)$$
$$\left(\frac{R(s,t)}{s^{\gamma+1}t^{g+1}} - \frac{R(s,1)}{s^{\gamma+1}} - \frac{R(1,t)}{t^{g+1}} + R(1,1)\right) dsdt$$
$$= (1-\nu^2)\frac{(\gamma+1)(g+1)}{\gamma g}\int_0^1\int_0^1 \Big(\frac{(\gamma+1)(2\gamma+1)(g+1)(2g+1)}{s^{1-\gamma}t^{1-g}} - \frac{(\gamma+1)(2\gamma+1)}{s^{1-\gamma}t}$$
$$-\frac{(g+1)(2g+1)}{t^{1-g}s} + \frac{1}{st}\Big)R(s,t) - \Big(\frac{(\gamma+1)(g+1)(2g+1)}{t^{1-g}} - \frac{(\gamma+1)}{t} - \frac{(g+1)(2g+1)}{(\gamma+1)t^{1-g}}$$
$$+\frac{1}{(\gamma+1)t}\Big)R(1,t) - \Big(\frac{(g+1)(\gamma+1)(2\gamma+1)}{s^{1-\gamma}} - \frac{(g+1)}{s} - \frac{(\gamma+1)(2\gamma+1)}{(g+1)s^{1-\gamma}} + \frac{1}{(g+1)s}\Big)R(s,1)dsdt$$
$$+(1-\nu^2)\frac{(\gamma+1)(g+1)}{\gamma g}\left((\gamma+1)(g+1) - \frac{(\gamma+1)}{(g+1)} - \frac{(g+1)}{(\gamma+1)} + \frac{1}{(\gamma+1)(g+1)}\right)R(1,1)$$

$$= (1-\nu^2)\Big[(\gamma g + 2\gamma + 2g + 4)R(1,1) - \frac{g(\gamma+1)}{\gamma}\int_0^1 \frac{R(s,1)}{s}ds - \frac{\gamma(g+1)}{g}\int_0^1 \frac{R(1,t)}{t}dt$$
$$+\frac{(\gamma+1)(g+1)}{\gamma g}\Big(\frac{(2\gamma+1)(\gamma+1)}{g+1} - (g+1)(\gamma+1)(2\gamma+1) + \frac{(\gamma+1)(g+1)(2\gamma+1)(2g+1)}{\gamma+g+1}$$
$$-(2\gamma+1)\Big)\int_0^1 \frac{R(s,1)}{s^{1-\gamma}}ds + \frac{(\gamma+1)(g+1)}{\gamma g}\Big(\frac{(2g+1)(g+1)}{\gamma+1} - (g+1)(\gamma+1)(2g+1)$$
$$+\frac{(\gamma+1)(g+1)(2\gamma+1)(2g+1)}{\gamma+g+1} - (2g+1)\Big)\int_0^1 \frac{R(1,t)}{t^{1-g}}dt\Big] = (1-\nu^2)S_g. \qquad \square$$

Proof of Theorem 4.2.1: It can be shown that $\hat{S}_g \xrightarrow{\mathbb{P}} S_g$, from the uniform consistency of $\hat{R}$ on $[0,1]^2$. Using the bivariate convergence in Proposition 4.2.1 and that $\frac{\hat{\sigma}}{a(\frac{n}{k})} \xrightarrow{\mathbb{P}} 1$, as $n \to \infty$. We obtain, as $n \to \infty$,

$$\sqrt{k}\left(\frac{\hat{\sigma}_g}{a\left(\frac{n}{k}\right)} - 1\right) = \sqrt{k}\left(\frac{\hat{\sigma}}{a\left(\frac{n}{k}\right)} - 1\right) - \frac{S_g}{1+(1+g)^2}\sqrt{k}\left(\frac{\tilde{\sigma}_g}{(\frac{n}{k})^g} - 1\right) + o_{\mathbb{P}}(1). \qquad (4.12)$$

Using Proposition 4.2.1 and the continuous mapping theorem imply (4.4). □

Proof of Theorem 4.2.2: From (4.12), as $n \to \infty$,

$$\sqrt{k}\left(\frac{\hat{\sigma}_g}{a\left(\frac{n}{k}\right)} - 1\right) \xrightarrow{\mathbb{P}} \Sigma_\gamma - \frac{S_g}{1+(1+g)^2}\tilde{\Sigma}_g, \qquad (4.13)$$

where Σ_γ and $\tilde{\Sigma}_g$ are defined as in the proof of Proposition 4.2.1. From (3.35), as $n \to \infty$,

$$\sqrt{k}\left(\hat{\gamma}_g - \gamma\right) \xrightarrow{\mathbb{P}} \Omega - \frac{1+\gamma}{1+g} R_g \tilde{\Omega}, \tag{4.14}$$

see the proof of Proposition 3.2.1 for the definition of Ω and $\tilde{\Omega}$, and R_g is defined as in (3.5). By Lemma 3.6.3, take $s = 1$, then as $n \to \infty$,

$$\sqrt{k}\left(\frac{X_{n-k:n} - U_1\left(\frac{n}{k}\right)}{a(\frac{n}{k})}\right) \xrightarrow{\mathbb{P}} W_1(1). \tag{4.15}$$

Combining (3.16), (4.13), (4.14), and (4.15), we have that, as $n \to \infty$

$$\sqrt{k}\left(\hat{\gamma}_g - \gamma, \frac{\hat{\sigma}_g}{a\left(\frac{n}{k}\right)} - 1, \frac{X_{n-k:n} - U_1\left(\frac{n}{k}\right)}{a\left(\frac{n}{k}\right)}\right) \xrightarrow{d} \left(\Omega - \frac{1+\gamma}{1+g} R_g \tilde{\Omega}, \Sigma_\gamma - \frac{S_g}{1+(1+g)^2}\tilde{\Sigma}_g, W_1(1)\right).$$

We obtain the means and variances of the first two limiting terms from Theorem 3.2.1 and Theorem 4.2.1. Here $W_1(1)$ is independent of Ω, and $Cov(\Sigma_\gamma, W_1(1)) = \gamma$, notice that the latter covariance is incorrect in de Haan and Ferreira (2006), page 139, as it is assumed to be zero.

It remains to calculate covariances between $(\Omega, \tilde{\Sigma}_g)$, $(\tilde{\Omega}, \Sigma_\gamma)$, $(\tilde{\Omega}, W_1(1))$, and $(\tilde{\Sigma}_g, W_1(1))$ as the rest are already calculated in Corollary 4.5.1 and Corollary 4.5.2.

$$\begin{aligned}
&Cov(\Omega, \tilde{\Sigma}_g) = (1-\nu^2)\frac{(1+\gamma)^2(1+g)}{\gamma g}\Big[\Big(\frac{(g+1)^2+\gamma-(g+1)^2(\gamma+1)}{(g+1)(\gamma+1)}\Big)R(1,1) \\
&+\frac{g^2}{g+1}\int_0^1 \frac{R(s,1)}{s}ds + \Big(\frac{2\gamma+1}{\gamma+1} - \frac{(2\gamma+1)(g+1)(2g+1)}{\gamma+g+1} + (2\gamma+1)(g+1) - \frac{2\gamma+1}{g+1}\Big) \\
&\int_0^1 \frac{R(s,1)}{s^{1-\gamma}}ds + \Big((2g+1) - \frac{(2\gamma+1)(g+1)(2g+1)}{\gamma+g+1} + \frac{\gamma(g+1)(2g+1)}{\gamma+1}\Big)\int_0^1 \frac{R(1,t)}{t^{1-g}}dt\Big] \\
&= (1-\nu^2)(1+\gamma)Q_{\hat{\gamma},\hat{\sigma}_g}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
&Cov(\tilde{\Omega}, \Sigma_\gamma) = (1-\nu^2)\frac{(1+\gamma)(1+g)^2}{\gamma g}\Big[\Big(\frac{(\gamma+1)^2+g-(\gamma+1)^2(g+1)}{(g+1)(\gamma+1)}\Big)R(1,1) \\
&+\frac{\gamma^2}{\gamma+1}\int_0^1 \frac{R(1,t)}{t}dt + \Big((2\gamma+1) - \frac{(2g+1)(\gamma+1)(2\gamma+1)}{\gamma+g+1} + \frac{g(\gamma+1)(2\gamma+1)}{g+1}\Big) \\
&\int_0^1 \frac{R(s,1)}{s^{1-\gamma}}ds + \Big(\frac{2g+1}{g+1} - \frac{(2g+1)(\gamma+1)(2\gamma+1)}{\gamma+g+1} + (2g+1)(\gamma+1) - \frac{2g+1}{\gamma+1}\Big)\int_0^1 \frac{R(1,t)}{t^{1-g}}dt\Big] \\
&= (1-\nu^2)(1+\gamma)Q_{\hat{g},\hat{\sigma}},
\end{aligned}$$

Hence the covariances between the limiting terms are:

$$Cov(\Omega - \frac{1+\gamma}{1+g}R_g\tilde{\Omega}, \Sigma_\gamma - \frac{S_g}{1+(1+g)^2}\tilde{\Sigma}_g) = -(1+\gamma)[1+(1-\nu^2)(\frac{R_gS_g}{1+(1+g)^2} + \frac{1+\gamma}{1+g}R_gQ_{\hat{g},\hat{\sigma}}$$
$$+\frac{S_gQ_{\hat{\gamma},\tilde{\sigma}_g}}{1+(1+g)^2})] = -(1+\gamma)[1+(1-\nu^2)Q],$$

$$Cov(-\frac{1+\gamma}{1+g}R_g\tilde{\Omega}, W_1(1)) = (1-\nu^2)\frac{(1+\gamma)(1+g)}{g}R_g\Big[(2g+1)\int_0^1\frac{R(1,s)}{s^{1-g}}ds - \int_0^1\frac{R(1,s)}{s}ds$$
$$-\frac{g}{g+1}R(1,1)\Big] = (1-\nu^2)M_1,$$

$$Cov(\Sigma_\gamma - \frac{S_g}{1+(1+g)^2}\tilde{\Sigma}_g, W_1(1)) = \gamma - (1-\nu^2)\frac{(1+g)}{g(1+(1+g)^2)}S_g\Big[(2g+1)(g+1)\int_0^1\frac{R(1,s)}{s^{1-g}}ds$$
$$-\int_0^1\frac{R(1,s)}{s}ds - (g+2)R(1,1)\Big] = \gamma + (1-\nu^2)M_2. \qquad \square$$

Proof of Theorem 4.2.3: Using Theorem 4.2.2 in combination with proof of Theorem 4.3.1 in de Haan and Ferreira (2006) yield, as $n \to \infty$,

$$\sqrt{k}\frac{\hat{x}_{p_g} - x_p}{a\left(\frac{n}{k}\right)q_\gamma\left(\frac{k}{np}\right)} \xrightarrow{d} \Omega - \frac{1+\gamma}{1+g}R_g\tilde{\Omega} - \gamma_-\left(\Sigma_\gamma - \frac{S_g}{1+(1+g)^2}\tilde{\Sigma}_g\right) + (\gamma_-)^2W_1(1) - \lambda\frac{\gamma_-}{\gamma_- + \rho},$$

where $\gamma_- := \min(0,\gamma)$. The distribution of the limiting random variable is easily seen to be that in (4.7). □

Proof of Proposition 4.3.1 Following similar steps as in the proof of Proposition 4.2.1, Proposition 4.5.2 can be generalized to more covariates. It remains to calculate the covariance terms using the joint weak convergence of all the tail empirical processes (see Ahmed and Einmahl (2019)). □

Proof of Theorem 4.3.1: From the uniform consistency of the tail copula estimators, it follows that $\hat{Z}_{1j}^{-1} \xrightarrow{\mathbb{P}} \hat{Z}_{1j}^{-1}, j = 1,\dots,d$, in combination with Proposition 4.3.1, as $n \to \infty$,

$$\sqrt{k}\left(\frac{\hat{\sigma}_g}{a\left(\frac{n}{k}\right)} - 1\right) = \sqrt{k}\left(\frac{\hat{\sigma}}{a\left(\frac{n}{k}\right)} - 1\right) + \sum_{j=2}^{d}\frac{Z_{1j}^{-1}}{Z_{11}^{-1}}\sqrt{k}\left(\frac{\tilde{\sigma}_{g_j}}{\left(\frac{n}{k}\right)^g} - 1\right) + o_{\mathbb{P}}(1).$$

Now Proposition 4.3.1 and the continuous mapping theorem yield (4.9). □

Proof of Theorem 4.3.2: Using the asymptotic results for the SSE for the extreme value index and the scale in the multivariate setting, the proof follows similar steps as the proof of Theorem 4.2.2. □

Proof of Theorem 4.3.3: Using Theorem 4.3.2 in combination with the proof of Theorem 4.3.1 in de Haan and Ferreira (2006) yield, as $n \to \infty$

$$\sqrt{k}\frac{\hat{x}_{p_g} - x_p}{a\left(\frac{n}{k}\right) q_\gamma\left(\frac{k}{np}\right)} \xrightarrow{d} \Omega - \frac{1+\gamma}{1+g}\sum_{j=2}^{d}\frac{H_{1j}^{-1}}{H_{11}^{-1}}\tilde{\Omega}_j - \gamma_- \left(\Sigma_\gamma - \sum_{j=2}^{d}\frac{\hat{Z}_{1j}^{-1}}{\hat{Z}_{11}^{-1}}\tilde{\Sigma}_{g_j}\right) + (\gamma_-)^2 W_1(1) - \lambda\frac{\gamma_-}{\gamma_- + \rho}.$$

The distribution of the limiting random variable is easily seen to be that in (4.11). □

Chapter 5

Insurance risk and machine learning: Estimating conditional Value-at-Risk using random forest

[Based on joint work with Chen Zhou]

5.1 Introduction

Insurance companies are obliged to calculate solvency capital requirement (SCR), to ensure that they hold sufficient capital to protect policy holders. Under Solvency II, the Value-at-Risk (VaR) of the insurance claims is used to calculate the SCR. As a widely used risk measure, VaR describes the maximum loss within a certain confidence level α which is essentially the quantile of the distribution of the total claims. There are a broad class of parametric and non-parametric estimators for VaR (e.g. Azzalini (1981), Harrell and Davis (1982) and Philippe (2001)). In our context we handle data with heavy tailed distribution and apply extreme value statistics to estimate VaR for the insurance claims.

Extreme value statistics consider fitting a model on the tail distribution of data. There are two main approaches to sample data of extreme events, namely, the Block Maxima (BM) and Peak Over Threshold (POT). The BM approach divides the data into several blocks and considers the maxima of each block. The POT approach selects a certain threshold and considers exceedances over the threshold. In the following, we use the POT approach as it obtains extremes more efficiently.

The existing extreme value statistics often assume that observations are independent and identically distributed (i.i.d.), while in practice this assumption is violated. Embrechts et al. (2003) address data violating the stationarity assumption: they emphasize on the non-stationary pattern which accounts for the structural changes in the observed data such as evolution over time. Other sources of non-stationary pattern can be the survival bias, changes in the economy cycle, business volume, management interactions and regulations.

Davison and Smith (1990) consider estimating the parameters in the POT approach based on covariates, to adjust for different variation such as seasonality, using a parametric linear regression model. Coles (2001) discusses the applicability of the extreme value models in case of non-stationary processes that have systematic changes through time by introducing different method to deal with variation caused by time. Coles (2001) considers time as a covariate and estimates the parameters in the POT approach dynamically using parametric models, which can then be used in estimating VaR. Chavez-Demoulin et al. (2016) extend the idea by Coles (2001) by considering another covariate, the business line, in addition to time for modelling operational risk losses. Chavez-Demoulin et al. (2016) introduce a semi-parametric model using penalized likelihood method for dynamic parameters estimation based on these two covariates. Then they evaluate risk measures such as VaR using the estimated model.

In this chapter we extend the POT approach to incorporate a large number of covariates in the parameters, when modelling heavy tailed response variable. In particular, we characterize the extreme value index and the probability of exceedences as conditional expectations given a set of covariates, then we use the random forest algorithm to estimate them. In addition, we obtain the conditional VaR given the covariates using the conditional estimates of the extreme value index and probability of exceedences. We focus on dealing with categorical covariates which is different from the continuous covariates. The categorical covariates provide less number of possible splits. By contrast the continuous covariates can be more informative when using the random forest algorithm. Our methodology is demonstrated using a loss dataset from an anonymized insurance company which contains a large number of categorical covariates. The performance of VaR is examined via backtesting procedures.

Random forest algorithm is a machine learning scheme proposed by Breiman (2001). It consists of ensembles of trees, where each tree in the ensemble grows based on suitable tuning parameters. Eventually these trees are aggregated to produce estimates for random forest classification or regression model. In many empirical studies, random forest classification and regression models are emerged as a serious competitor to other machine learning techniques (see Svetnik et al. (2003), Díaz-Uriarte and De Andres (2006), and Genuer et al. (2008)). Random forest algorithm can handle a large number of covariates with measuring the predictive power for each covariate. It is flexible when dealing with both linear and non-linear relationships.

Random forest algorithm has been applied in analysing different insurance applications. Lin et al. (2017) use machine learning techniques on life insurance data to obtain classification model for predicting users recommendation of insurance products. They find that random forest algorithm shows a superior performance, especially when dealing with unbalanced classes. Alshamsi (2014) uses random forest and other machine learning techniques to predict the customers choices for different services based on car insurance data, where random forest algorithm outperforms the other used techniques. Staudt and Wagner (2021) examine predication models for the claim severity in collision car insurance data. The random forest is used to model the claim severity and log-normal transformation of the severity. They show that the log-normal transformation is preferable to apply with the right skewed claims in the specific used application.

Our study is related to other stream of literature that combines tail estimation with machine learning. Fissler et al. (2021) introduce a deep neural network regression model for estimating actuarial claim size, where the threshold for the large claims is given in terms of a quantile of the conditional claim size distribution. Farkas et al. (2020) introduce the generalized Pareto regression trees, that combines the extreme value theory with the regression trees algorithm to estimate a conditional generalized Pareto distribution based on covariates. Velthoen et al. (2021) estimate a conditional generalized Pareto distribution and the intermediate threshold based on covariates using gradient boosting procedure. The conditional parameters estimated by gradient boosting are then used to estimate conditional extreme quantiles.

In this chapter, we combine the extreme value statistics with the random forest algorithm by providing a coherent framework to relate the extreme value index and the probability of exceedence with covariates. The use of random forest classification and regression models for estimating these quantities turns to be a natural choice. Moreover, we focus on insurance relevant quantities by further estimating the conditional VaR_α. Our approach differs from existing studies such as Velthoen et al. (2021) and Staudt and Wagner (2021). Velthoen et al. (2021) consider a fixed probability of exceedence, but allowing for varying thresholds based on the covariates. By contrast we choose to consider a fixed threshold but allowing for varying probability of exceedence based on covariates. The two studies are different but have analogous choices. Staudt and Wagner (2021) directly estimate the claim severity based on covariates using random forest model. By contrast we consider the conditional distribution based on covariates, estimate the conditional parameters using random forest models, and obtain the conditional extreme quantile such as the conditional VaR_α.

The structure of the chapter is as follows. Section 5.2 previews the classical extreme value theory focusing on the estimation of VaR. Section 5.3 introduces our proposed model setup, derives the conditional parameters, discusses the random forest algorithms to estimate the conditional parameters and the backtesting methods to evaluate the performance of the conditional VaR. Section 5.4 is devoted to a simulation study which assesses the performance of our proposed methodology. In Section 5.5, we apply our methodology to a real insurance dataset consisting of claim loss.

5.2 Classical Extreme Value Theory

Consider a random variable Y following a distribution function F, which belongs to the max-domain of attraction of an extreme value distribution H_γ. That is, there exists a sequence of constants $a_n > 0$ and $b_n \in \mathbb{R}$, such that

$$\lim_{n\to\infty} F^n\left(a_n y + b_n\right) = H_\gamma(y),$$

where

$$H_\gamma(y) = \begin{cases} \exp(-(1+\gamma y)^{-1/\gamma}) & \text{if} \quad 1+\gamma y > 0 \text{ and } \gamma \neq 0, \\ \exp(-e^{-y}) & \text{if} \quad \gamma = 0. \end{cases}$$

Here γ is the extreme value index, which describes the heaviness of the tail of the distribution.

We focus on the case with a positive extreme value index ($\gamma > 0$), then F is called a heavy tailed distribution. Theorem 1.2.1 in de Haan and Ferreira (2006), shows that F belongs to the max-domain of attraction with $\gamma > 0$ if and only if

$$\lim_{t\to\infty} \mathbb{P}(Y > ty | Y > t) = y^{-1/\gamma}. \tag{5.1}$$

We intend to estimate VaR_α of Y which is defined as: $VaR_\alpha = \inf\{y|F(y) \geq \alpha\}$, for a confidence level $0 < \alpha < 1$. If F is a continuous distribution, then $\mathbb{P}(Y > VaR_\alpha) = 1 - \alpha$. Motivated by (5.1) with a properly chosen high threshold u, an estimator of the VaR_α can be given as

$$\widehat{VaR_\alpha} = u\left(\frac{\hat{g}}{1-\alpha}\right)^{\hat{\gamma}}, \tag{5.2}$$

where $\hat{g} = \widehat{\mathbb{P}(Y > u)}$ is a proper estimate for the exceedence probability (e.g. the empirical counter parts) and $\hat{\gamma}$ is the estimated extreme value index (see Weissman (1978)).

To obtain an estimate for γ, we use the Hill estimator. The intuition of the Hill estimator is as follows: From Remark 1.2.3 in de Haan and Ferreira (2006), the domain of attraction condition in (5.1) is equivalent to

$$\lim_{t\to\infty} E\left(\log\left(\frac{Y}{t}\right)\middle| Y > t\right) = \gamma. \tag{5.3}$$

Based on (5.3), Hill (1975) introduces the Hill estimator of the extreme value index

$$\hat{\gamma}_H = \frac{1}{k}\sum_{i=0}^{k-1} \log Y_{n-i:n} - \log Y_{n-k:n}, \tag{5.4}$$

where $Y_{1:n} \leq Y_{2:n} \leq \ldots \leq Y_{n:n}$ are the order statistics of $\{Y_i\}_{i=1}^n$, and k is an intermediate sequence such that $k \to \infty$, and $\frac{k}{n} \to 0$ as $n \to \infty$.

5.3 Proposed Methodology

We extend the classical approach to a conditional model. Consider a d-dimensional covariate X, such that $(Y, X) \in \mathbb{R}^{d+1}$ is dependent random vector. For $\gamma > 0$, assume that

$$\lim_{t\to\infty} \mathbb{P}(Y > ty | Y > t, X = x) = y^{-1/\gamma(x)}.$$

$\gamma(x)$ is a continuous function on $\mathbb{R}^d \to \mathbb{R}^+$. The function $\gamma(x)$ can be considered as a conditional extreme value index. Similar to (5.3), we can derive that

$$\gamma(x) = \lim_{t\to\infty} E\left(\log\left(\frac{Y}{t}\right)\middle| Y > t, X = x\right). \tag{5.5}$$

Define the conditional VaR_α of Y given a set of covariates $X = x$, as $VaR_\alpha(x) := VaR_\alpha(Y|X = x)$, which satisfies $\mathbb{P}(Y \geq VaR_\alpha(x)|X = x) = 1 - \alpha$. For a high threshold

u,

$$\frac{\mathbb{P}(Y \geq VaR_\alpha(x)|X = x)}{\mathbb{P}(Y \geq u|X = x)} \approx \left(\frac{VaR_\alpha(x)}{u}\right)^{-1/\gamma(x)}.$$

Define

$$g(x) = \mathbb{P}(Y \geq u|X = x) = E(I_{Y>u}|X = x),$$

where $I[.]$ is an indicator function that equals to 1 when $Y > u$ and 0 otherwise. Then we get

$$\frac{1-\alpha}{g(x)} \approx \left(\frac{VaR_\alpha(x)}{u}\right)^{-1/\gamma(x)}.$$

Thus an estimator of $VaR_\alpha(x)$ can be

$$\widehat{VaR_\alpha(x)} = u\left(\frac{\hat{g}(x)}{1-\alpha}\right)^{\hat{\gamma}(x)}, \tag{5.6}$$

where $\hat{\gamma}(x)$ and $\hat{g}(x)$ are proper estimators of $\gamma(x)$ and $g(x)$.

Since $g(x)$ is the conditional expectation of the indicator function $I_{Y>u}$ given the covariates $X = x$, we can estimate it using a classification model. In addition, the limit in (5.5) implies that

$$\gamma(x) \approx E\left(\log\left(\frac{Y}{u}\right)\middle| Y > u, X = x\right) = E\left(Z\middle|X = x\right),$$

where $Z = \log\left(\frac{Y}{t}\middle|Y > t\right)$. Therefore $\gamma(x)$ is approximately the conditional expectation of Z given the covariates $X = x$ and it can be estimated using a regression model. We use non-parametric ensemble random forest algorithm to estimate the classification and regression models.

5.3.1 Estimation of $\gamma(x)$

Suppose our observations $\{Y_i, X_i\}_{i=1}^n$, where $X_i = (X_{i1}, \ldots, X_{id})$ are d-dimensional covariates. Define $\{i|Y_i > u\} = \{i_1, \ldots, i_m\}$ which are the indices corresponding to high losses above the threshold u. For each $1 \leq l \leq m$, we consider $Z_l = \log\left(\frac{Y_{i_l}}{u}\right)$ and the corresponding covariates X_{i_l} in a regression model to obtain estimate for $\gamma(x)$.

Random forest regression consists of combination of regression trees. Each regression tree is based on a bootstrapped sample drawn with replacement from the original sample upto the same sample size. On average, that leads to use only 63% genuine observations in the original sample. The rest are denoted as out-of-bag sample and used to get unbiased estimation of the model error. For each categorical variable, we use the one-hot-encoding method which transform each category into dummy variable then we drop one of the dummy variables to avoid the dummy variable trap that may lead to multicollinearity. In the following, X_{ij} refers to the dummy variables and the number of dummies produced from the all categorical covariates is D.

The regression tree starts with the root node C_0, which contains all bootstrapped observations for this tree, then it splits into two child nodes C_l and C_r as follows: Randomly select $[\sqrt{D}]$ dummies from all D dummy variables, denoted as $X_{.j}, j \in U \subset \{1, \ldots, D\}$. The algorithm tests all possible splits among all dummies $X_{.j}, j \in U$. The set of splits is defined as S where each split s depends on the value of one covariate. The splitting rule for categorical variables is based on the elements belonging to a particular class. Each dummy variable can be used only once for splitting. The best split s^* is the one that maximizes the decrease of the least squared error

$$s^* = \arg\max_{s \in S} I(C_0) - I(X_{.j}),$$

where $I(C_0) = \frac{1}{|C_0|} \sum\limits_{i \in C_0} (Z_i - \bar{Z}(C_0))^2$ is the mean squared error in the root node C_0 with $\bar{Z}(C_0)$ is the average of observations in C_0, $I(X_{.j}) = \pi_l I(C_l) + \pi_r I(C_r)$, π_l and π_r are the fractions of observations in each child node C_l and C_r respectively, and $I(C_l)$ and $I(C_r)$ are the mean squared error in the nodes C_l and C_r defined in an analogous way. The next step is to split the nodes C_l and C_r, which are regarded as parent nodes, following the same procedure as above. The splitting process is repeated till a minimum node size is reached. The final node where there is no more splits allowed, is denoted as the terminal node. For a given covariate $X = x$, following the regression tree structure leads to a terminal node. Then, the predicted value of the response variable from this regression tree becomes the average of observations in the terminal node.

Based on different bootstrapping samples, we grow B regression trees. For tree b, given a covariate $X = x$, the predicted value following regression tree b is denoted as $T_r(x; \Theta_b)$, where $b = 1, \dots, B$, and Θ_b describes the structure of the regression tree in terms of splits and nodes. By aggregating the results of the B regression trees, the random forest estimator for the conditional extreme value index becomes

$$\hat{\gamma}(x) = \frac{1}{B} \sum_{b=1}^{B} T_r(x; \Theta_b). \tag{5.7}$$

See Figure 5.1 for the random forest scheme.

5.3.2 Estimation of $g(x)$

We estimate $g(x)$ in two steps. First we use the random forest algorithm to build a classification model, then we calibrate the results of the classification model to obtain the conditional probability estimates.

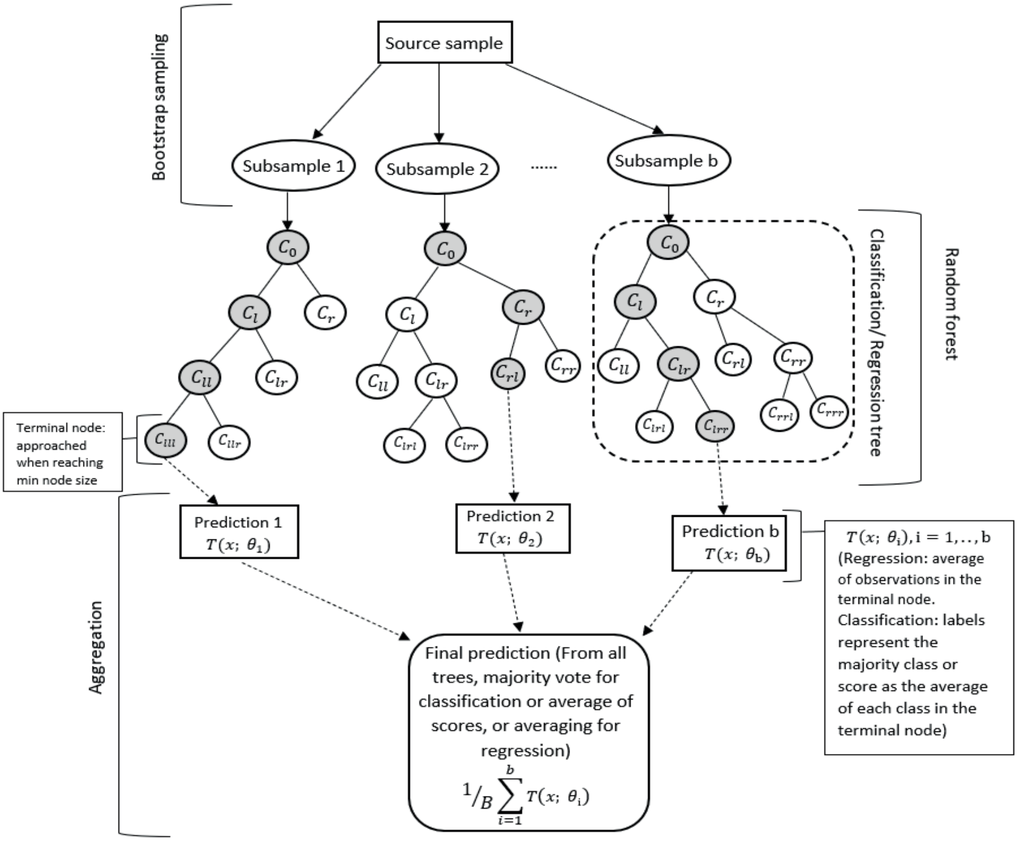
Random Forest for Classification

We estimate $g(x)$ by conditional expectation $E(V_i|X = x)$ using random forest classification model, where

$$V_i = \begin{cases} 1 & \text{if} \quad Y_i > u, \\ 0 & \text{if} \quad Y_i \leq u, \end{cases},$$

$i = 1, \dots, n$ and X is a d-dimensional covariates, $X_i = (X_{i1}, \dots, X_{id})$. Random forest classification model has the same structure as the random forest regression model as in Figure 5.1. In the classification tree the splitting at each node is in a similar way as the regression tree but with different splitting criteria. Classification tree starts by splitting the root node C_0 which contains all observations, where the best split s^* is the one that maximize the Gini gain

$$s^* = \arg\max_{s \in S} I_G(C_0) - I_G(X_{.j}), \quad j \in U,$$



Bootstrap sampling
Source sample
Subsample 1
Subsample 2
......
Subsample b
C_0
C_l
C_r
C_{ll}
C_{lr}
C_{rl}
C_{rr}
C_{lll}
C_{llr}
C_{lrl}
C_{lrr}
C_{rrl}
C_{rrr}
Terminal node: approached when reaching min node size
Classification/ Regression tree
Random forest
Prediction 1 $T(x;\ \theta_1)$
Prediction 2 $T(x;\ \theta_2)$
Prediction b $T(x;\ \theta_b)$
$T(x;\ \theta_i), i = 1,..,b$ (Regression: average of observations in the terminal node. Classification: labels represent the majority class or score as the average of each class in the terminal node)
Aggregation
Final prediction (From all trees, majority vote for classification or average of scores, or averaging for regression)
$1/B \sum_{i=1}^{b} T(x;\ \theta_i)$

Figure 5.1: Random forest scheme

where $I_G(C_0) = 1 - \pi_{m,0}^2 - \pi_{m,1}^2$ is the Gini impurity for C_0 and $\pi_{m,0}$ and $\pi_{m,1}$ are fraction of observations belongs to each class, $I_G(X_{.j}) = \pi_l I_G(C_l) + \pi_r I_G(C_r)$, π_l and π_r are the fraction of observations at each child node C_l and C_r respectively, and $I_G(C_l)$ and $I_G(C_r)$ are the Gini impurity of C_l and C_r. Note that the maximum Gini impurity is 0.5. Starting from the root node, we run the splitting process in a similar way till reaching terminal node where a minimum size of a node is reached. For tree b, Θ_b describes the structure of the tree. Given a covariate $X = x$, the classification score is denoted as $T_c(x; \Theta_b)$ which represents the average of votes for each class in the terminal node.

Based on different bootstrapping samples, we grow B classification trees. The classification score is the average of scores from the B classification trees in the random forest as follow

$$\hat{f}(x) = \frac{1}{B}\sum_{b=1}^{B} T_c(x; \Theta_b).$$

Note that this average cannot be interpreted as a classification probability. We further calibrate the probability of $\mathbb{P}(V = 0|X = x)$ based on the classification score $f(x)$.

Probability Calibration and Evaluation

Following Niculescu-Mizil and Caruana (2005), we use calibration to produce robust probability estimates that match the expected distribution of probabilities for each class. We use a parametric calibration method, which is known as sigmoid method. That is, we assume that

$$P(V = 0|X = x) = \frac{1}{1 + \exp(\beta_0 f(x) + \beta_1)}$$

where β_0 and β_1 are the parameters. We estimate them by minimizing the negative log likelihood as follows

$$(\hat{\beta}_0, \hat{\beta}_1) = \min_{(\beta_0,\beta_1)} -\sum_{i=1}^{n} V_i \log(p_i) + (1 - V_i)\log(1 - p_i),$$

where $p_i = \frac{1}{1+\exp(\beta_0 \hat{f}(X_i)+\beta_1)}$, and V_i is the observed class. After estimating β_0 and β_1 the calibrated probability is estimated as

$$\frac{1}{1 + \exp(\hat{\beta}_0 \hat{f}(X_i) + \hat{\beta}_0)}.$$

The calibrated probabilities are evaluated using Brier score (Brier et al., 1950). Brier score is computed as the mean squared error of the calibrated probability and the observed class label. It measures the accuracy of predicted probabilities; see Ovadia et al. (2019) and Savage (1971). The Brier score ranges between 0 and 1, where a lower values of the score indicates a better prediction.

5.3.3 Variable Importance

We evaluate the importance of each covariate based on the relative importance method, as in Hastie et al. (2009) Section 10.13.1. The relative importance helps to understand which covariates are more crucial in the prediction of the dependent variable.

Recall the construction of the regression (or classification) tree as in Section 5.3.1 (or 5.3.2). The splits are chosen to minimize the residual sum of squares (or to increase the gini gain). More informative splits are those resulting in large decrease in the residual sum of squares. The relative importance of a given covariate is obtain by considering all the splits related to this covariate, and summing up the decrease in the residual sum of squares due to those splits.

5.3.4 Backtesting

In this section we discuss the evaluation methods to asses the quality of our proposed conditional of VaR_α. For that purpose we use two backtesting methods.

Proportion of Failures Test (PoF)

The PoF test is one of the widely used standard tests for VaR_α and known as the unconditional test. It tests if the probability of exceedances p based on the estimated VaR_α is significantly different from $1-\alpha$ ($H_0 : p = 1-\alpha$). The PoF test is conducted using the likelihood-ratio (LR) statistic (Kupiec, 1995)

$$LR = -2\ln\left(\frac{(1-\hat{p})^{n-r}\hat{p}^r}{\alpha^{n-r}(1-\alpha)^r}\right),$$

where $\hat{p} = \frac{r}{n}$ and r is the number of exceedances and n is the total number of observations used for the backtesting. Under the null hypothesis, the LR statistic asymptotically follows χ^2 distribution with one degree of freedom.

Comparative Test

We use the comparative test by Nolde and Ziegel (2017) to compares two methods: the conditional VaR_α given a set of covariates $(VaR_\alpha(x))$ as in (5.6) and the unconditional VaR_α as in (5.2). The comparative test depends on using a score function which is strictly consistent to VaR_α.

For the backtesting one can use the following null hypothesis:

H_0^- : The conditional VaR_α predicts at least as well as the unconditional VaR_α

We use the following test statistic

$$T = \frac{\Delta\bar{S}}{\hat{\sigma}_n/\sqrt{n}},$$

where $\Delta\bar{S} = \frac{1}{n}\sum_{i=1}^{n}(S(VaR_\alpha(x_i), Y_i) - S(VaR_\alpha, Y_i))$, and $S(r, Y_i) = (1 - \alpha - 1_{\{Y_i>r\}})G(r) + 1_{\{Y_i>r\}}G(Y_i), i = 1, \ldots, n$, here n is the number of observations used for the backtesting. We consider a zero and first degree homogeneous scoring functions, i.e, $G(r) = \log(r)$ and $G(r) = r$ respectively. $\hat{\sigma}_n$ is the standard deviation of $\Delta\bar{S}$. Under H_0^-, the test statistic T has expected value less than or equal to zero. Under certain mixing assumptions detailed in Giacomini and White (2006), T is asymptotically normally distributed with variance 1. The backtest is passed if at a fixed confidence level η, H_0^- is not rejected when $1-\Phi(T) \geq \eta$.

In case of comparative backtesting, a more conservative approach is adopted using the null hypothesis:

H_0^+ : The conditional VaR_α predicts at most as well as the unconditional VaR_α.

This null hypothesis can also be tested using T, which has expected value greater than or equal to zero under H_0^+. At a fixed confidence level η, H_0^+ is rejected when $\Phi(T) \leq \eta$.

The decision is taken in comparative backtesting based on the three regions procedure proposed by Fissler et al. (2015). We consider that the conditional VaR_α fails the comparative backtesting if H_0^- is rejected at level η, then the conditional VaR_α is in the red region. The conditional VaR_α passes the comparative backtesting if H_0^+ is rejected, and it falls in the green region. The conditional VaR_α needs further investigation if neither H_0^- or H_0^+ can be rejected, and the conditional VaR_α is in the yellow region.

5.4 Simulation

We conduct a simulation study to evaluate the performance of our proposed methodology. We generate a random sample of size $n = 10000, 20000$ and 30000, for 5-covariates $X_i = (X_{i1}, \ldots, X_{i5}), 1 \le i \le n$, such that X_{ij} are independent across $j = 1, \ldots, 5$, and each from a multinomial distribution. That is

- $\{X_{i1}\}_{i=1}^n \sim multinom([0.7, 0.2, 0.1], n), \{X_{i2}\}_{i=1}^n \sim multinom([0.3, 0.5, 0.1, 0.1], n),$
- $\{X_{i3}\}_{i=1}^n \sim multinom([0.4, 0.3, 0.2, 0.01, 0.09], n), \{X_{i4}\}_{i=1}^n \sim multinom([0.7, 0.2, 0.06, 0.04], n),$
- $\{X_{i5}\}_{i=1}^n \sim multinom([0.8, 0.1, 0.1], n).$

We construct $\gamma(x)$ and $g(x)$ based on the generated covariates, here we consider

Case I: $\gamma_1(x_i) = 0.15 + 0.7I(x_{i3} = 1) + 0.93I(x_{i2} = 2), g_1(x_i) = 0.1 + 0.05I(x_{i2} = 1) + 0.1I(x_{i1} = 2),$

Case II: $\gamma_2(x_i) = 0.1 + 0.6I(x_{i3} = 1) + 0.8I(x_{i2} = 2), g_2(x_i) = 0.1 + 0.11I(x_{i2} = 1) + 0.08I(x_{i1} = 2).$

For each i, we generate a Bernoulli random variable p_i with probability $g(x_i)$. Then we simulate Y_i as

$$Y_i = \begin{cases} u\tilde{Y_i} & , \quad p_i = 1, \\ (\frac{F_i - F_i u^{0.1} + u^{0.1}}{u^{0.1}})^{-10} & , \quad p_i = 0, \end{cases}$$

where $\tilde{Y_i} \sim PD(\gamma(x_i))$, $F_i \sim Uniform(0, 1)$, and u is a prespecified threshold.

Our main goal is to check the ability of our methodology to precisely estimate the conditional VaR_α of Y. We start with checking the estimation of the conditional extreme value index $\gamma(x)$ and conditional probability of exceedence $g(x)$. We use all the generated covariates as potential covariates for the random forest algorithm to check the ability of the algorithm in selecting the important covariates.

First we examine the estimation of $\gamma(x)$ using random forest regression model. To obtain estimates for $\gamma(x)$, we only consider observations above threshold u, to get Z_l as defined in Section 5.3.1 then we regress it on the set of all covariates. Based on the g functions, the number of exceedences for $n = 10000, 20000, 30000$ are on average $1350, 2700, 4050$ for case I and $1490, 2980, 4470$ for case II. To evaluate the performance of the proposed estimator, we calculate the rooted mean squared error as

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \left(\frac{\hat{\gamma}(x_i)}{\gamma(x_i)} - 1 \right)^2}{n}}.$$

For the random forest regression model, we need to decide the stopping criteria which is the minimum size of node. Note that the maximum depth is reached when all leaves contain less than the minimum node size. Therefore we test the performance of the random forest regression for estimating $\gamma(x)$ for a range of values of the minimum node size. The remaining main parameters for the random forest are chosen as follows: the number of trees is 100 and the minimum sample size required to be at a leaf node is 1. The random forest algorithm for the regression and classification are implemented in Python using scikit-learn package. Figure 5.2 and Figure 5.3 show that the RMSE decreases by the increase of minimum size of the node, except the case $n = 10000$. Based on Figure 5.2 and Figure 5.3, we choose the optimal minimum size of node as 320 where the RMSE ranges between 0.056 to 0.076.

Second, we consider the conditional probability of exceedence $g(x)$. Similarly we inspect the optimal minimum size of node by considering a range of values for the minimum node size. We take a higher range than in the regression case as we are dealing with all observations not only the exceedences. Then we calculate the corresponding RMSE of $g(x)$ to each minimum node size based on 100 times simulation. Figures 5.4-5.5 show a different behaviour to the regression figures. Based on Figures 5.4-5.5, it is not recommended to

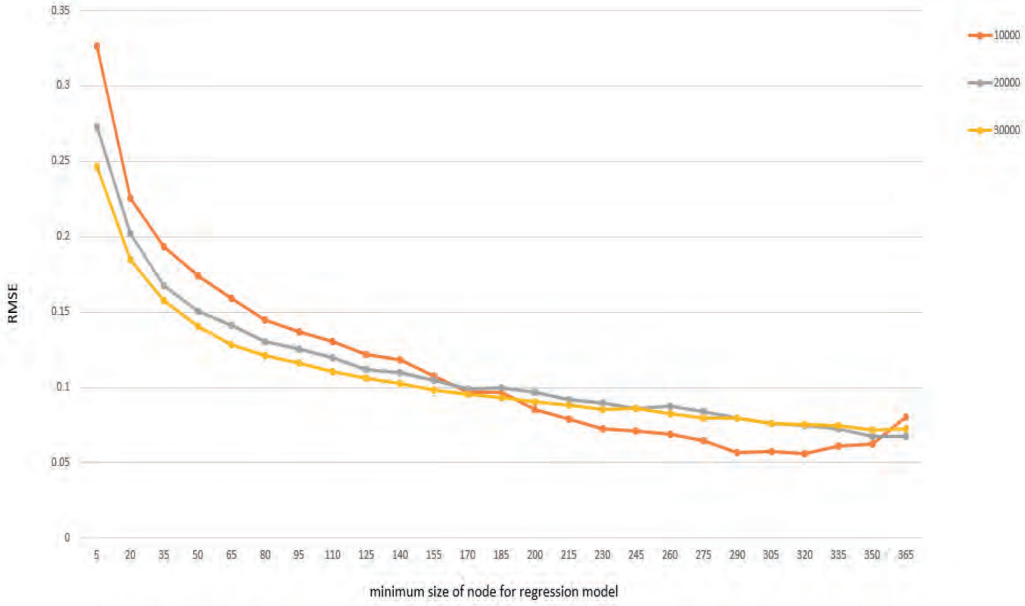


Figure 5.2: RMSE of $\gamma_1(x)$ for different values of minimum node size in random forest regression model based on 100 simulation samples.

set very high minimum size of the node. We decide that for $g_1(x)$ and $g_2(x)$ the optimal minimum node size are 1400 and 1100 respectively where the RMSE ranges between 0.077 to 0.011 and 0.094 to 0.12.

Another important point is to check the sensitivity of using different threshold in the random forest regression and classification model. To study the models sensitivity, we choose different thresholds that produce different percentage of exceedences and calculate the RMSE for $\gamma(x)$ and $g(x)$ for each threshold. Take $n = 20000$, the thresholds are selected such that the percent of exceedences are $[8\%, 10\%, 12\%, 14\%, 16\%, 18\%, 20\%]$, where the case of having 14% exceedences is the closest to the threshold u which is used in generating the data.

For the early chosen optimal minimum node size of the regression and classification

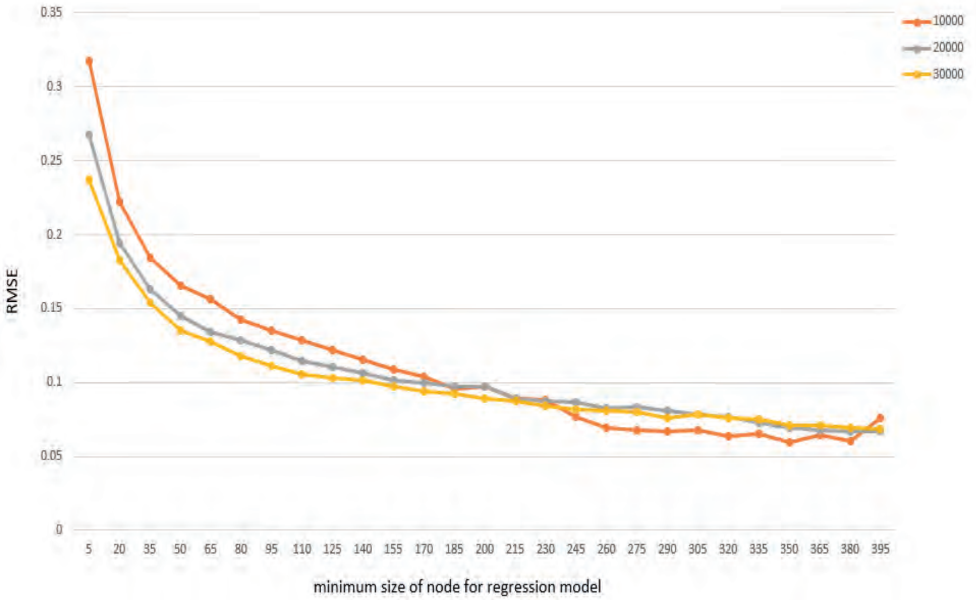


Figure 5.3: RMSE of $\gamma_2(x)$ for different values of minimum node size in random forest regression model based on 100 simulation samples.

models, Figures 5.6-5.7 show that the regression model is less sensitive to the change of the threshold. The RMSE for $\gamma(x)$ increases with the increase on the exceedence percentage. Hence if we include more observation that does not belong to the tail of the distribution, that will affect the estimation of the extreme value index. However, having less observations does not deteriorate the performance. By contrast, the classification model seems to be more sensitive towards the change of the threshold. This can be explained by the fact that we are dealing with unbalanced classes. The change in the percent of the majority and the minority classes may require different tuning for the parameters.

Finally we estimates the conditional VaR_α, for $\alpha = 0.90, 0.95$ and 0.99. Using minimum node size for the regression and classification models as 320 and 1100 respectively. The

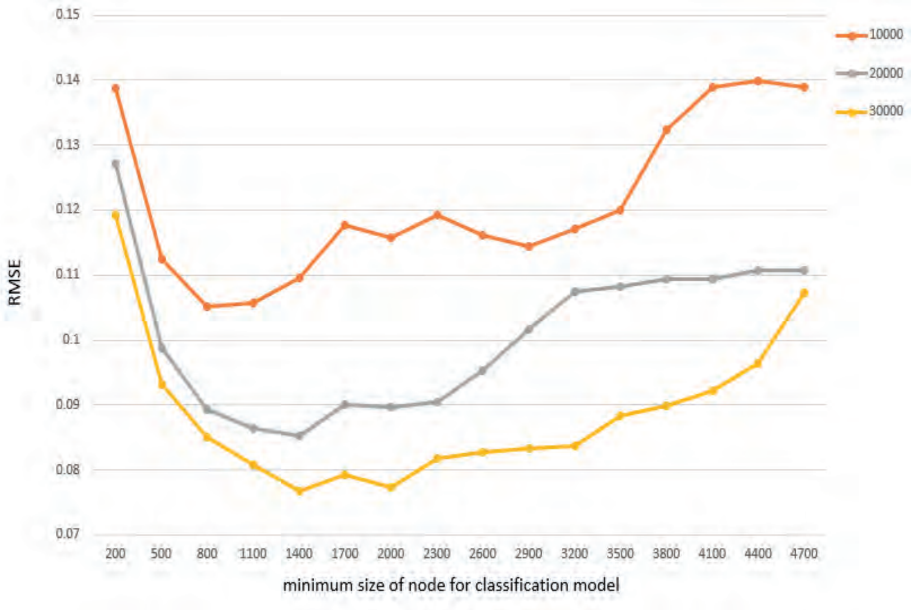


Figure 5.4: RMSE of $g_1(x)$ for different values of minimum node size in random forest classification model

following results are based on 500 times simulations and using the threshold u. Table 5.1 shows that the RMSE for the $VaR_{0.90}$ decreases by the increase of the data size, but it increases in the more extreme quantile specially $VaR_{0.99}$ which is mainly affected by the estimation of $\gamma(x)$.

One important advantage for the random forest models is their ability to rank the importance of different input covariates in the fitting process. Here we test the accuracy of this feature by calculating the importance in different simulations, then check the percent of getting the covariates in $\gamma(x)$ and $g(x)$ as the top ranked covariates. Table 5.2 shows a very high ability for the regression model to detect the covariates importance correctly. Table 5.3 shows good results for the top ranked covariates, although we observe that the

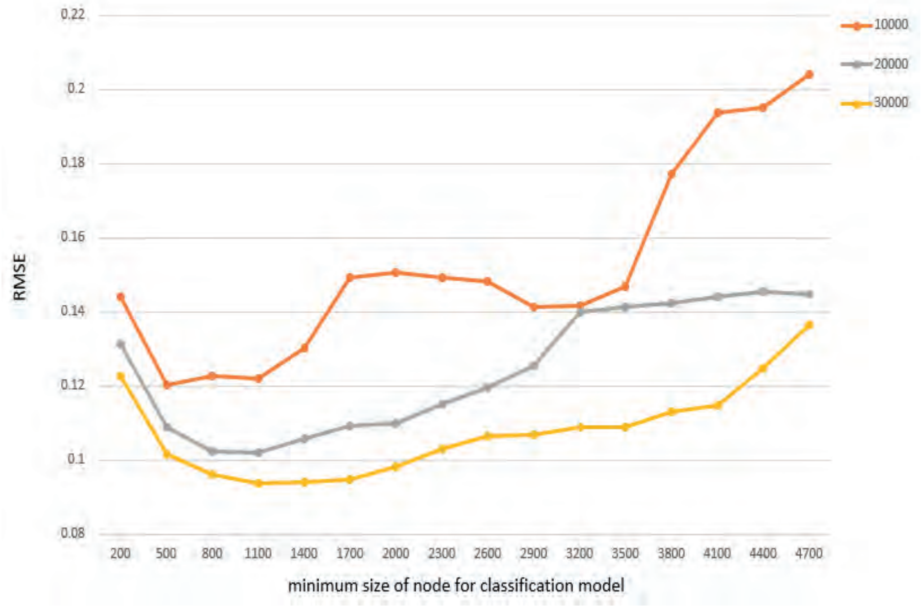


Figure 5.5: RMSE of $g_2(x)$ for different values of minimum node size in random forest classification model

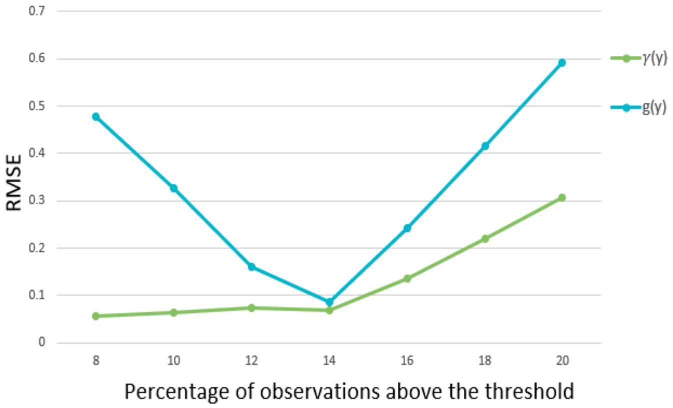


Figure 5.6: RMSE of $g_1(x)$ for different percentage of observations above the threshold

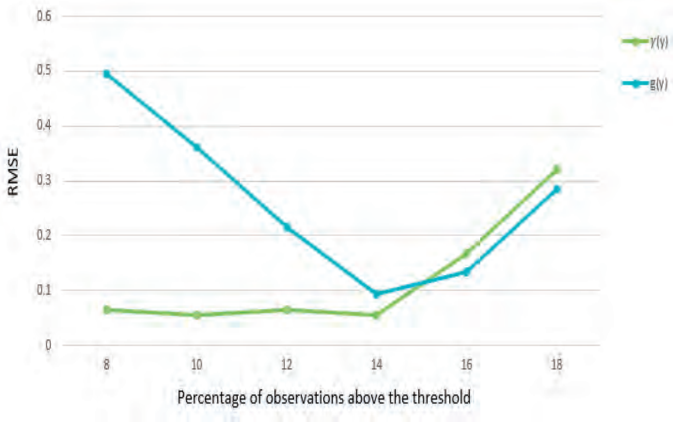


Figure 5.7: RMSE of $g_2(x)$ for different percentage of observations above the threshold

second covariate appears more as third important but it still shows in the top three ranked covariates.

5.5 Application

In this section we apply our proposed methodology to loss claims data from an anonymized insurance company. The dataset consists of 71297 claims due to Hail in the time period between 1/2007 and 12/2016. Table 5.4 presents the summary statistics for the claims grouped by the year of occurrence. Note that the average kurtosis is around 885 and it ranges from 205 to 2706 which generally indicates to a heavy tailed distribution. Figure 5.8 shows the histogram of the data, which illustrates the long tail of the distribution.

There are two main covariates related to each claim, which are the type of insurance and the segment. The two covariates contain four different categories. The type of insurance includes engineering (construction projects, mostly building and infrastructure), greenhouse (a special type of property, namely glasshouses and other buildings related to horticulture), motor(mainly passenger cars, delivery vans and trucks), and property (build-

n	10000	20000	30000
$\hat{\gamma}_1(x)$	0.0574	0.0741	0.0766
$\hat{g}_1(x)$	0.1028	0.0880	0.0831
$\widehat{VaR_{0.90}}(x)$	0.1147	0.0971	0.0923
$\widehat{VaR_{0.95}}(x)$	0.1310	0.1164	0.1204
$\widehat{VaR_{0.99}}(x)$	0.2179	0.2380	0.2876
$\hat{\gamma}_2(x)$	0.0663	0.0755	0.0768
$\hat{g}_2(x)$	0.099	0.0906	0.0837
$\widehat{VaR_{0.90}}(x)$	0.0837	0.0775	0.0728
$\widehat{VaR_{0.95}}(x)$	0.0964	0.0933	0.0943
$\widehat{VaR_{0.99}}(x)$	0.172	0.1847	0.2126

Table 5.1: RMSE for the conditional estimators

		$\gamma_1(x)$		$\gamma_2(x)$	
n	Rank	$I(X_{.2} = 2)$	$I(X_{.3} = 1)$	$I(X_{.2} = 2)$	$I(X_{.3} = 1)$
10000	First	99.8%	0.2%	99.8%	0.2%
	Second	0.2%	99.8%	0.2%	99.8%
20000	First	100%	0%	100%	0%
	Second	0%	100%	0%	100%
30000	First	100%	0%	100%	0%
	Second	0%	100%	0%	100%

Table 5.2: Importance ranking for the covariates used in the true $\gamma(x)$

		$g_1(x)$		$g_2(x)$	
n	Rank	$I(X_{\cdot 1} = 2)$	$I(X_{\cdot 2} = 1)$	$I(X_{\cdot 1} = 2)$	$I(X_{\cdot 2} = 1)$
	First	99.8%	0.4%	0%	100%
10000	Second	0.2%	23.6%	55.6%	0%
	Third	0%	76%	44.4%	0%
	First	100%	0%	0%	100%
20000	Second	0%	17.6%	62.2%	0%
	Third	0%	82.4%	37.8%	0%
	First	100%	0%	0%	100%
30000	Second	0%	19.8%	76.4%	0%
	Third	0%	80.2%	23.6%	0%

Table 5.3: Importance ranking for the covariates used in the true $g(x)$

Year	Mean	Std	Skewness	Kurtosis
2007	4176.70	14236.68	17.22	403.29
2008	3175.52	8091.11	23.99	776.21
2009	4289.81	13814.71	11.93	205.06
2010	4490.87	22891.20	18.28	395.39
2011	2555.05	6518.95	27.36	1130.40
2012	3093.77	10627.10	27.12	969.87
2013	3215.10	7719.03	14.34	274.35
2014	4733.62	24335.14	18.26	429.85
2015	3792.38	29622.65	42.85	2705.67
2016	11175.00	151442.32	31.95	1157.30

Table 5.4: Summary statistics for claims by year the claim occurred

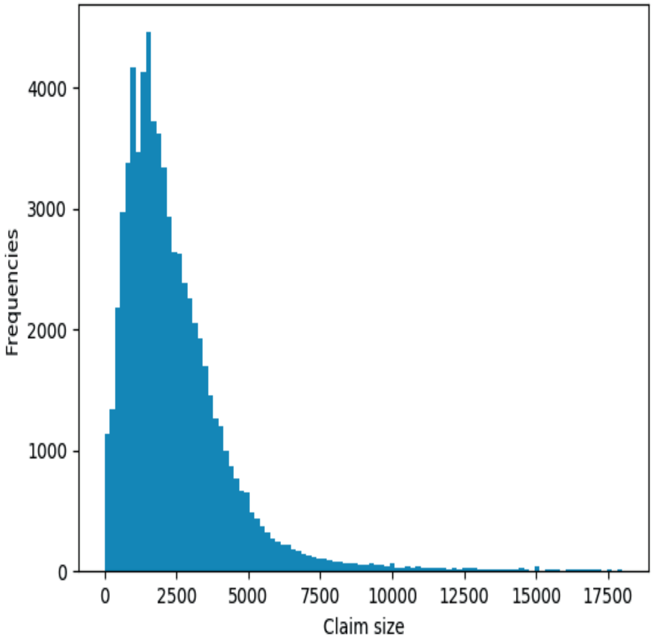


Figure 5.8: Histogram of claims distribution

ings and their contents), while the segment includes agricultural (risks related to farmers), commercial (risks belong to companies), horticultural (risks related to glasshouses), and residential (risks related to clients not being companies (e.g. houses)). The rest of the covariates are date and postal code. Regarding the date, we use the months to generate a new categorical covariate that represents the raining months (the covariate takes the value 1 in months between May to October and 0 otherwise). Finally we use the postal code to generate other categorical covariate to account for the location effect. Figure 5.9 shows the association between the main four covariates calculated by Cramer's V measure (Cramér, 1946). We observe that there is no strong dependence between them which may affect the models used for estimation of the conditional parameters. After applying the one hot encoding and dropping one dummy from each categorical variable, we end up using 94 dummies in the random forest models.

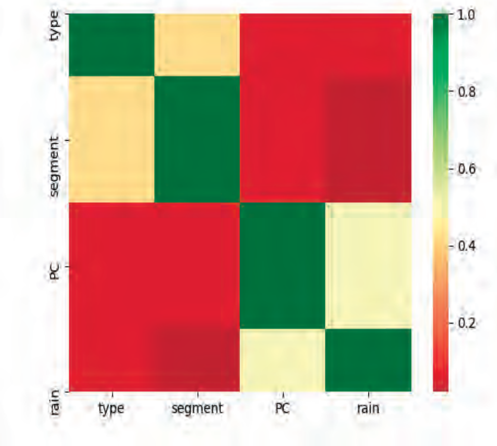


Figure 5.9: Cramer's V matrix between the main covariates

In the following we discuss estimation of the conditional VaR_α. First we address the threshold selection issue and how we divide the data into training and testing samples to apply the random forest algorithm. Second, we use the random forest regression model to estimate the $\gamma(x)$ and investigate whether the exceedences indeed follow approximately Pareto distribution using the Q-Q plot. Third, we estimate $g(x)$ using random forest classification model and use Brier score to evaluate the model performance. Lastly, we use the random forest models produced from the training sample to estimate the conditional VaR_α for the testing sample and use backtesting to evaluate the performance of the estimation.

One critical issue of the analysis is the selection of appropriate high threshold. For that purpose we start by using the Hill plot, where the Hill estimator in (5.4) is plotted for different values of k using all data. Figure 5.10 indicates that the Hill estimator is not stable for a wide range of k, which may indicate the presence of a mixed distribution of the data. To assess the sensitivity and check for robustness, we use three thresholds where we assume the tail starts from the last 30%, 20% and 10% of the data. The three thresholds correspond to three levels of k at $7129, 14260$, and 21390 respectively. To apply the random forest algorithm, We divide the data into training sample which includes data in the time period from 2007 till 2013, the rest belongs to the testing sample.

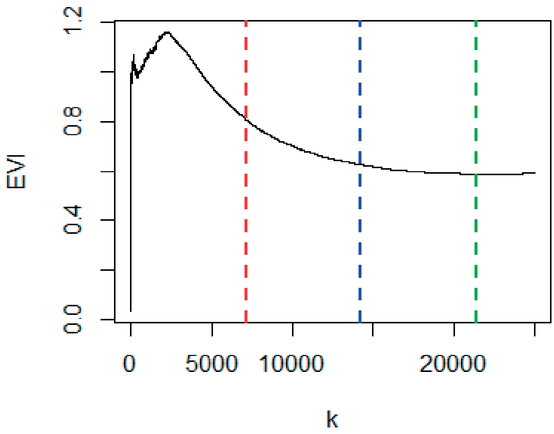
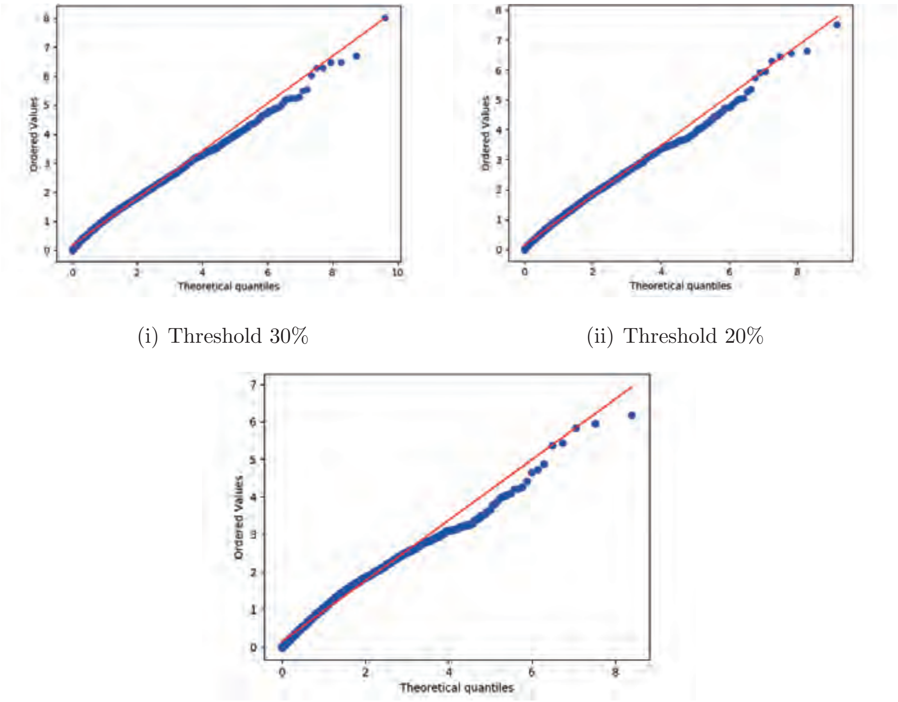


Figure 5.10: Extreme value index based on the Hill estimators for different values of k.

We begin by estimating $\gamma(x)$ using random forest regression model. We use the training sample to regress $\log\left(\frac{Y}{u}|Y>u\right)$ on the covariates, with the minimum node size selected at 320, similar to the simulation. We use the Q-Q plot to check if $\log\left(\frac{Y}{u}|Y>u\right)/\gamma(x)$ has a standard exponential distribution. Figure 5.11 shows a good fit for the exceedences using different thresholds. Between the three selected thresholds, the threshold 20% shows the best fit compared to the rest. In the following we show only the results based on 20% threshold. The results for the other thresholds do not show a remarkable difference, they are available upon request.



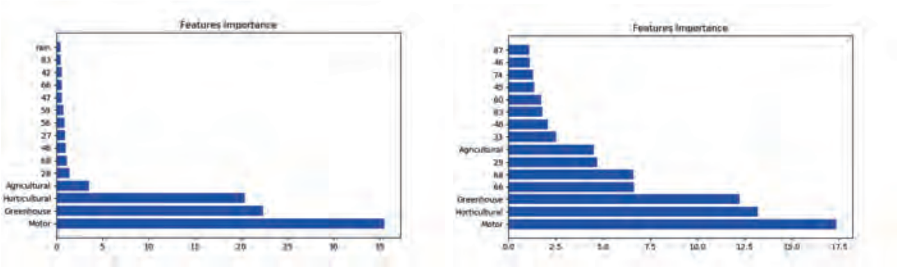
(i) Threshold 30%

(ii) Threshold 20%

(iii) Threshold 10%

Figure 5.11: Q-Q plot to check whether the large claim losses are approximately Pareto distributed. Note that the data are correctly fitted when it gets closer to the red 45-degree line.

We apply the random forest classification model to estimate the conditional probability $\mathbb{P}(Y > u|X = x)$ in the training sample, and use the Brier score to evaluate the model performance in the training and testing samples. The Brier score ranges between 0 and 1, when the score gets close to 0 that indicates to good performance for the model. Using minimum node size of 1100 as in the simulation, the Brier score for the training and testing samples are 0.1390 and 0.1675. Next, Figure 5.12 shows the variable importance for the top 15 covariates that affect the estimation of the conditional extreme value index and the conditional probability model. In the analysis, we left out the dummies indicating building and their contents in the type of insurance and risk related to houses in the segment. Figure 5.12 shows that the main source of large losses are the damaged vehicles, other buildings related to horticulture and the damaged glasshouses. However, the losses from damaged vehicles due to hails occurs more frequently than losses related to glasshouses and other buildings related to horticulture.



(i) Regression model

(ii) Classification model

Figure 5.12: Variable importance for the random forest models

Finally we estimate an out-of-sample conditional VaR_α, using confidence levels of $0.99, 0.95, 0.90$. Tables 5.5 shows the backtesting results based on 5% confidence level for all the thresholds. Regarding PoF test, we do not reject the null hypothesises only in the case of $VaR_{0.99}$. This is potentially due to the superiority of extreme value analysis in estimating extreme quantiles. The comparative test focuses on the magnitude of violation using score function. For $0-$homogeneous score function, the conditional VaR always outperforms compared to the standard VaR. For $1-$homogeneous score function the conditional VaR still outperforms for $VaR_{0.90}$ and $VaR_{0.95}$ otherwise it falls in the yellow region but it never performs worse than the standard VaR.

By contrast, Table 5.6 shows the backtesting results when using only the top important covariates in each model, defined as variables with a variable importance greater than 5%.

α	PoF test			Comparative test	
	# Violations(%)	P-value	descision	0- Homo	1 - Homo
0.90	4788 (13.69)	0	Reject	Green	Green
0.95	2345 (6.71)	0	Reject	Green	Green
0.99	379 (1.08)	0.3062	Do not Reject	Green	Yellow

Table 5.5: Tests result $VaR_{\alpha}(x)$

α	PoF test			Comparative test	
	# Violations(%)	P-value	descision	0- Homo	1 - Homo
0.90	3342 (9.55)	0.0665	Do not Reject	Green	Green
0.95	1870 (5.35)	0.0521	Do not Reject	Green	Green
0.99	379 (1.08)	0.3062	Do not Reject	Green	Green

Table 5.6: Tests result $VaR_{\alpha}(x)$ based on the top important covariates

In Table 5.6 we observe that the PoF test does not reject the null hypothesis for the three considered VaR_{α}. Table 5.7 shows the results of the PoF test for the standard VaR_{α}, where the null hypothesis is rejected for all cases of VaR_{α}. Additionally, the average conditional $VaR_{0.90}(x)$ and $VaR_{0.95}(x)$ are lower compared to the standard VaR without having higher percentage of violations than expected. These results are more desirable for companies and banks, our method leads to VaR estimates that would comply with the regulations while corresponding to a lower capital requirements. Based on the comparative test (Table 5.5 and 5.6) the conditional VaR_{α} outperforms the standard VaR_{α} for both $0-$homogeneous and $1-$homogeneous score functions. In particular, we conclude that estimation of VaR is substantially improved when considering only the top important variables.

α	PoF test		
	# of Violations (%)	P-value	decision
0.90	2654 (7.89)	0	Reject
0.95	1444 (4.13)	0	Reject
0.99	607 (1.74)	0	Reject

Table 5.7: Tests result for VaR_α

5.6 Conclusion

In this chapter, we use the random forest regression and classification models to estimate the conditional extreme value index $\gamma(x)$ and the conditional probability of exceedance $g(x)$, respectively. We consider a large number of covariates, where random forest models manage to rank the most important covariates that affect the estimation of the two parameters. We then use the conditional estimators to obtain the conditional VaR_α.

We show that our methodology is able to produce more accurate estimate for the conditional VaR_α using a simulation study. With applying our methodology to insurance data, we estimate the conditional VaR_α using a training sample while backtest its performance in an out-of-sample set up. We compare the conditional VaR_α to the standard VaR_α. In the application of the insurance data, our proposed methodology outperforms the standard approach, especially when using the top important covariates selected by the random forest model.

Our current approach still bears some limitations. One extension can be based on the fact that using the top important covariates significantly affects the performance. One may consider to use other machine learning models to calculate the variable importance while using the random forest for estimation or the other way around.

Algorithm 1 Random Forest Algorithm

```
Input: Training set $D_n$
Output: Random forest classifier $T(x, \Theta_b), b = 1, \ldots, B$.
for $b = 1, \ldots, B$ do
    Generate $d_n$ bootstrapped sample with replacement to the same sample size from $D_n$.
    Randomly select $p = \sqrt{d}$ covariate out of all $d$ covariates.
    Generate a decision tree and select the best split $s^*$, from the set of all splits $S = \{1, \ldots, s\}$ based on the set of randomly selected covariates $p$, that optimize the splitting criterion.
    The splitting rule for categorical variables is based on the elements belonging to particular category.
    Splitting is repeated till reaching minimum node size.
    The terminal node, where the tree stops splitting, has the output $T(x, \Theta_b)$.
end for
To make a prediction for a new point $x$ :
Regression: $\hat{f}_{rf}(x) = \frac{1}{B}\sum_{b=1}^{B} T(x; \Theta_b)$.
Classification: the random forest label is $T_{rf}(x) =$ majority vote $\{T(x; \Theta_b)\}_1^B$, and the random forest score is $\hat{f}_{rf}(x) = \frac{1}{B}\sum_{b=1}^{B} T(x; \Theta_b)$.
```

Bibliography

Aarssen, K. and L. de Haan (1994). On the maximal life span of humans. *Mathematical Population Studies 4*(4), 259–281.

Ahmed, H. and J. H. J. Einmahl (2019). Improved estimation of the extreme value index using related variables. *Extremes 22*, 553–569.

Alshamsi, A. S. (2014). Predicting car insurance policies using random forest. In *2014 10th International Conference on Innovations in Information Technology (IIT)*, pp. 128–132. IEEE.

Azriel, D., L. D. Brown, M. Sklar, R. Berk, A. Buja, and L. Zhao (2016). Semi-supervised linear regression. *arXiv preprint arXiv:1612.02391*.

Azzalini, A. (1981). A note on the estimation of a distribution function and quantiles by a kernel method. *Biometrika 68*(1), 326–328.

Balkema, A. A. and L. de Haan (1974). Residual life time at great age. *The Annals of Probability*, 792–804.

Beirlant, J., Y. Goegebeur, J. Segers, and J. Teugels (2004). *Statistics of Extremes: Theory and Applications.* Wiley.

Berghaus, B. and A. Bücher (2018). Weak convergence of a pseudo maximum likelihood estimator for the extremal index. *The Annals of Statistics 46*(5), 2307–2335.

Breiman, L. (2001). Random forests. *Machine learning 45*(1), 5–32.

Brier, G. W. et al. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review 78*(1), 1–3.

Buishand, T., L. de Haan, and C. Zhou (2008). On spatial extremes: with application to a rainfall problem. *The Annals of Applied Statistics 2*(2), 624–642.

Caeiro, F. and M. I. Gomes (2015). Threshold selection in extreme value analysis. In D. Dey and J. Yan (Eds.), *Extreme Value Modeling and Risk Analysis: Methods and Applications*, pp. 69–87. Chapman and Hall/CRC.

Chakrabortty, A. and T. Cai (2018). Efficient and adaptive linear regression in semi-supervised settings. *The Annals of Statistics 46*(4), 1541–1572.

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

Coles, S. (2001). *An introduction to statistical modeling of extreme values*, Volume 208. Springer.

Coles, S. G. and J. A. Tawn (1991). Modelling extreme multivariate events. *Journal of the Royal Statistical Society: Series B (Methodological) 53*(2), 377–392.

Coles, S. G. and D. Walshaw (1994). Directional modelling of extreme wind speeds. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 43*(1), 139–157.

Cramér, H. (1946). Mathematical methods of statistics, 1946. *Department of Mathematical SU*.

Csörgő, M., S. Csörgő, L. Horváth, and D. M. Mason (1986). Weighted empirical and quantile processes. *The Annals of Probability 14*, 31–85.

Davison, A. C. and R. L. Smith (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society: Series B (Methodological) 52*(3), 393–425.

de Haan, L. (1971). A form of regular variation and its application to the domain of attraction of the double exponential distribution. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 17*(3), 241–258.

de Haan, L. (1990). Fighting the arch–enemy with mathematics '. *Statistica neerlandica 44*(2), 45–68.

de Haan, L. and J. de Ronde (1998). Sea and wind: multivariate extremes at work. *Extremes 1*(1), 7.

de Haan, L., C. G. de Vries, and C. Zhou (2009). The expected payoff to internet auctions. *Extremes 12*(3), 219–238.

de Haan, L. and A. Ferreira (2006). *Extreme Value Theory: an Introduction*. Springer.

de Haan, L. and H. Rootzén (1993). On the estimation of high quantiles. *Journal of Statistical Planning and Inference 35*(1), 1–13.

Dekkers, A. L., J. H. Einmahl, and L. de Haan (1989). A moment estimator for the index of an extreme-value distribution. *The Annals of Statistics*, 1833–1855.

Díaz-Uriarte, R. and S. A. De Andres (2006). Gene selection and classification of microarray data using random forest. *BMC bioinformatics 7*(1), 1–13.

Donnelly, C. and P. Embrechts (2010). The devil is in the tails: actuarial mathematics and the subprime mortgage crisis. *Astin Bulletin 40*(1), 1–33.

Drees, H., A. Ferreira, and L. de Haan (2004). On maximum likelihood estimation of the extreme value index. *The Annals of Applied Probability*, 1179–1201.

Drees, H. and X. Huang (1998). Best attainable rates of convergence for estimators of the stable tail dependence function. *Journal of Multivariate Analysis 64*, 25–47.

Einmahl, J. H. (1992). Limit theorems for tail processes with application to intermediate quantile estimation. *Journal of Statistical Planning and Inference 32*, 137–145.

Einmahl, J. H., L. de Haan, and C. Zhou (2016). Statistics of heteroscedastic extremes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 78*(1), 31–51.

Einmahl, J. H. and J. R. Magnus (2008). Records in athletics through extreme-value theory. *Journal of the American Statistical Association 103*(484), 1382–1391.

Einmahl, J. J., J. H. Einmahl, and L. de Haan (2019). Limits to human life span through extreme value theory. *Journal of the American Statistical Association 114*(527), 1075–1080.

Embrechts, P., H. Furrer, and R. Kaufmann (2003). Quantifying regulatory capital for operational risk. *Derivatives Use, Trading and Regulation 9*(3), 217–233.

Embrechts, P., C. Klüppelberg, and T. Mikosch (2013). *Modelling extremal events: for insurance and finance*, Volume 33. Springer Science & Business Media.

Farkas, S., O. Lopez, and M. Thomas (2020). Cyber claim analysis through generalized pareto regression trees with applications to insurance.

Fisher, R. A. and L. H. C. Tippett (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical proceedings of the Cambridge philosophical society*, Volume 24, pp. 180–190. Cambridge University Press.

Fissler, T., M. Merz, and M. V. Wüthrich (2021). Deep quantile and deep composite model regression. *arXiv preprint arXiv:2112.03075*.

Fissler, T., J. F. Ziegel, and T. Gneiting (2015). Expected shortfall is jointly elicitable with value at risk-implications for backtesting. *arXiv preprint arXiv:1507.00244*.

Genuer, R., J.-M. Poggi, and C. Tuleau (2008). Random forests: some methodological insights. *arXiv preprint arXiv:0811.3619*.

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

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

Gomes, M. I. and A. Guillou (2015). Extreme value theory and statistics of univariate extremes: a review. *International Statistical Review 83*, 263–292.

Harrell, F. E. and C. Davis (1982). A new distribution-free quantile estimator. *Biometrika 69*(3), 635–640.

Hastie, T., R. Tibshirani, and J. Friedman (2009). The elements of statistical learning. Springer Series in Statistics. Springer, New York, second edition. Data mining, inference and prediction.

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

Hosking, J. R. and J. R. Wallis (1987). Parameter and quantile estimation for the generalized pareto distribution. *Technometrics 29*(3), 339–349.

Huang, X. (1992). *Statistics of bivariate extreme values.* Thesis Publishers Amsterdam.

Klüppelberg, C. and T. Mikosch (1997). Large deviations of heavy-tailed random sums with applications in insurance and finance. *Journal of Applied Probability*, 293–308.

Kupiec, P. (1995). Techniques for verifying the accuracy of risk measurement models. *The J. of Derivatives 3*(2).

Lay, T. and T. C. Wallace (1995). *Modern Global Seismology.* Academic Press.

Lin, W., Z. Wu, L. Lin, A. Wen, and J. Li (2017). An ensemble random forest algorithm for insurance big data analysis. *Ieee access 5*, 16568–16575.

Niculescu-Mizil, A. and R. Caruana (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pp. 625–632.

Nolde, N. and J. F. Ziegel (2017). Elicitability and backtesting: Perspectives for banking regulation. *The annals of applied statistics 11*(4), 1833–1874.

Oesting, M. and P. Naveau (2020). Spatial modeling of heavy precipitation by coupling weather station recordings and ensemble forecasts with max-stable processes. *arXiv preprint arXiv:2003.05854*.

Ovadia, Y., E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. V. Dillon, B. Lakshminarayanan, and J. Snoek (2019). Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *arXiv preprint arXiv:1906.02530*.

Philippe, J. (2001). Value at risk: the new benchmark for managing financial risk. *NY: McGraw-Hill Professional*.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association 66*(336), 783–801.

Shorack, G. R. and J. A. Wellner (2009). *Empirical Processes with Applications to Statistics*. SIAM.

Smith, R. L. (1987). Estimating tails of probability distributions. *The Annals of Statistics 15*, 1174–1207.

Staudt, Y. and J. Wagner (2021). Assessing the performance of random forests for modeling claim severity in collision car insurance. *Risks 9*(3), 53.

Svetnik, V., A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston (2003). Random forest: a classification and regression tool for compound classification and qsar modeling. *Journal of chemical information and computer sciences 43*(6), 1947–1958.

Thomas, M., M. Lemaitre, M. L. Wilson, C. Viboud, Y. Yordanov, H. Wackernagel, and F. Carrat (2016). Applications of extreme value theory in public health. *Plos one 11*(7).

Vapnik, V. (2013). *The Nature of Statistical Learning Theory.* Springer.

Velthoen, J., C. Dombry, J.-J. Cai, and S. Engelke (2021). Gradient boosting for extreme quantile regression. *arXiv preprint arXiv:2103.00808*.

Vervaat, W. (1972). Functional central limit theorems for processes with positive drift and their inverses. *Probability Theory and Related Fields 23*(4), 245–253.

Von Mises, R. (1936). La distribution de la plus grande de n valuers. *Rev. math. Union interbalcanique 1*, 141–160.

Wasserman, L. and J. D. Lafferty (2008). Statistical analysis of semi-supervised regression. In *Advances in Neural Information Processing Systems*, pp. 801–808.

Weissman, I. (1978). Estimation of parameters and large quantiles based on the k largest observations. *Journal of the American Statistical Association 73*(364), 812–815.

Zhang, A., L. D. Brown, and T. T. Cai (2019). Semi-supervised inference: General theory and estimation of means. *The Annals of Statistics 47*(5), 2538–2566.

Zhou, C. (2009). Existence and consistency of the maximum likelihood estimator for the extreme value index. *Journal of Multivariate Analysis 100*(4), 794–815.

Zhu, X. and A. B. Goldberg (2009). Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning 3*(1), 1–130.

CENTER DISSERTATION SERIES

CentER for Economic Research, Tilburg University, the Netherlands

No.	Author	Title	ISBN	Published
638	Pranav Desai	Essays in Corporate Finance and Innovation	978 90 5668 639 0	January 2021
639	Kristy Jansen	Essays on Institutional Investors, Asset Allocation Decisions, and Asset Prices	978 90 5668 640 6	January 2021
640	Riley Badenbroek	Interior Point Methods and Simulated Annealing for Nonsymmetric Conic Optimization	978 90 5668 641 3	February 2021
641	Stephanie Koornneef	It's about time: Essays on temporal anchoring devices	978 90 5668 642 0	February 2021
642	Vilma Chila	Knowledge Dynamics in Employee Entrepreneurship: Implications for parents and offspring	978 90 5668 643 7	March 2021
643	Minke Remmerswaal	Essays on Financial Incentives in the Dutch Healthcare System	978 90 5668 644 4	July 2021
644	Tse-Min Wang	Voluntary Contributions to Public Goods: A multi-disciplinary examination of prosocial behavior and its antecedents	978 90 5668 645 1	March 2021
645	Manwei Liu	Interdependent individuals: how aggregation, observation, and persuasion affect economic behavior and judgment	978 90 5668 646 8	March 2021
646	Nick Bombaij	Effectiveness of Loyalty Programs	978 90 5668 647 5	April 2021
647	Xiaoyu Wang	Essays in Microeconomics Theory	978 90 5668 648 2	April 2021
648	Thijs Brouwer	Essays on Behavioral Responses to Dishonest and Anti-Social Decision-Making	978 90 5668 649 9	May 2021
649	Yadi Yang	Experiments on hold-up problem and delegation	978 90 5668 650 5	May 2021
650	Tao Han	Imperfect information in firm growth strategy: Three essays on M&A and FDI activities	978 90 5668 651 2	June 2021

No.	Author	Title	ISBN	Published
651	Johan Bonekamp	Studies on labour supply, spending and saving before and after retirement	978 90 5668 652 9	June 2021
652	Hugo van Buggenum	Banks and Financial Markets in Microfounded Models of Money	978 90 5668 653 6	August 2021
653	Arthur Beddock	Asset Pricing with Heterogeneous Agents and Non-normal Return Distributions	978 90 5668 654 3	September 2021
654	Mirron Adriana Boomsma	On the transition to a sustainable economy: Field experimental evidence on behavioral interventions	978 90 5668 655 0	September 2021
655	Roweno Heijmans	On Environmental Externalities and Global Games	978 90 5668 656 7	August 2021
656	Lenka Fiala	Essays in the economics of education	978 90 5668 657 4	September 2021
657	Yuexin Li	Pricing Art: Returns, Trust, and Crises	978 90 5668 658 1	September 2021
658	Ernst Roos	Robust Approaches for Optimization Problems with Convex Uncertainty	978 90 5668 659 8	September 2021
659	Joren Koëter	Essays on asset pricing, investor preferences and derivative markets	978 90 5668 660 4	September 2021
660	Ricardo Barahona	Investor Behavior and Financial Markets	978 90 5668 661 1	October 2021
660	Stefan ten Eikelder	Biologically-based radiation therapy planning and adjustable robust optimization	978 90 5668 662 8	October 2021
661	Maciej Husiatyński	Three essays on Individual Behavior and New Technologies	978 90 5668 663 5	October 2021
662	Hasan Apakan	Essays on Two-Dimensional Signaling Games	978 90 5668 664 2	October 2021
663	Ana Moura	Essays in Health Economics	978 90 5668 665 9	November 2021
664	Frederik Verplancke	Essays on Corporate Finance: Insights on Aspects of the General Business Environment	978 90 5668 666 6	October 2021

No.	Author	Title	ISBN	Published
665	Zhaneta Tancheva	Essays on Macro-Finance and Market Anomalies	978 90 5668 667 3	November 2021
666	Claudio Baccianti	Essays in Economic Growth and Climate Policy	978 90 5668 668 0	November 2021
667	Hongwei Zhang	Empirical Asset Pricing and Ensemble Machine Learning	978 90 5668 669 7	November 2021
668	Bart van der Burgt	Splitsing in de Wet op de vennootschapsbelasting 1969 Een evaluatie van de Nederlandse winstbelastingregels voor splitsingen ten aanzien van lichamen	978 90 5668 670 3	December 2021
669	Martin Kapons	Essays on Capital Markets Research in Accounting	978 90 5668 671 0	December 2021
670	Xolani Nghona	From one dominant growth mode to another: Switching between strategic expansion modes	978 90 5668 672 7	December 2021
671	Yang Ding	Antecedents and Implications of Legacy Divestitures	978 90 5668 673 4	December 2021
672	Joobin Ordoobody	The Interplay of Structural and Individual Characteristics	978 90 5668 674 1	February 2022
673	Lucas Avezum	Essays on Bank Regulation and Supervision	978 90 5668 675 8	March 2022
674	Oliver Wichert	Unit-Root Tests in High-Dimensional Panels	978 90 5668 676 5	April 2022
675	Martijn de Vries	Theoretical Asset Pricing under Behavioral Decision Making	978 90 5668 677 2	June 2022
676	Hanan Ahmed	Extreme Value Statistics using Related Variables	978 90 5668 678 9	June 2022

Hanan Emad Galal Ahmed (Cairo, Egypt, 1993) obtained a bachelor's degree in statistics (2014) and a Master's in statistics (2016) from Cairo University (Egypt). In 2017, she started her Ph. D. at Tilburg University as a part of the Department of Econometrics and Operations Research.

Extreme value theory mainly handles data related to rare events, that are originated as a consequence of huge changes, such as earthquakes or large changes in asset pricing. Various fields have such applications where they use extreme value theory, for instance, meteorology, health, internet auctions, sports, risk management in finance, and insurance.

This dissertation contains five chapters that involve the use of the extreme value theory. Chapter 2 provides a novel methodology for improving the extreme value index estimation based on covariates in the case of heavy-tailed distributions. An application of the earthquakes and the related financial losses is used to show the improvement obtained when applying the proposed methodology. Chapters 3 and 4 introduce further generalizations for the proposed methodology in Chapter 2, by considering less assumptions, all cases for the extreme value index, and extending to the scale parameter and the estimation of the extreme quantile. In Chapter 3, An application of rainfall in France is used to demonstrate the use of the generalized methodology introduced in these two chapters. Chapter 5 combines the use of the extreme value theory with machine learning techniques. In Chapter 5, the machine learning technique is used to obtain an estimator of the Value-at-Risk (VaR) based on related covariates. An insurance application is used to show the effectiveness of the combined methodology of extreme value theory with machine learning.

ISBN: 978 90 5668 678 9
DOI: 10.26116/xwzn-p778

TILBURG SCHOOL OF ECONOMICS
AND MANAGEMENT

CentER Dissertation Series