Estimation of the Probability of a Failsure Set

de Haan and Ferreira(2006)

November 25, 2020

Chapter 8 of "Extreme value theory: An Introduction"

Presented by Liujun Chen.

Outline

Introduction

- In this Chapter, we we want to estimate $P((X, Y) \in C_n)$. Clearly, there is no observation (or very few) in the failure set.
- One Example. The wave height (HmO) and still water level (SWL) have been recorded during 828 storm events that are relevant for the Pettemer Zeewering. The failure set:

$$C = \{(HmO, SWL) : 0.3HmO + SWL > 7.6\}.$$

Basic Assumptions

- There exist normalizing functions $a_1 > 0, a_2 > 0$ and b_1, b_2 real , and a distribution function G with nondegenerate marginals, such that for all continuity points (x, y) of G,

$$\lim_{t\to\infty} F^t(a_1(t)x + b_1(t), a_2(t)y + b_2(t)) = G(x, y).$$

- Moreover, we choose the functions a_1, a_2, b_1, b_2 such that

$$G(x, \infty) = exp(-(1 + \gamma_1 x)^{-1/\gamma_1}), \quad 1 + \gamma_1 x > 0,$$

and

$$G(\infty, y) = exp(-(1 + \gamma_2 y)^{-1/\gamma_2}), \quad 1 + \gamma_2 y > 0.$$

Exponential Measure

With the exponential measure defined in Section 6.1.3,

$$\lim_{t\to\infty} P\{\left((1+\gamma_1\frac{X-b_1(t)}{a_1(t)})^{1/\gamma_1},(1+\gamma_2\frac{Y-b_2(t)}{a_2(t)})^{1/\gamma_2}\right)\in Q\}=v(Q),$$

for all Borel sets $Q\subset\mathbb{R}_+^2$ with $\inf_{(x,y)\in Q}\max(x,y)>0$ and $v(\partial Q)=0$.
Then for any $a>0$, we know that

$$v(aQ)=a^{-1}v(Q).$$

The Probability of a Failure Set

Now, we write the probability we want to estimate in terms of the transformed variables:

$$p_n := P((X,Y) \in C_n)$$
$$= P\{((1+\gamma_1 \frac{X-b_1(t)}{a_1(t)})^{1/\gamma_1}, (1+\gamma_2 \frac{Y-b_2(t)}{a_2(t)})^{1/\gamma_2}) \in Q_n\}, \tag{8.1.6}$$

with

$$Q_n := \{((1+\gamma_1 \frac{x-b_1(t)}{a_1(t)})^{1/\gamma_1}, (1+\gamma_2 \frac{y-b_2(t)}{a_2(t)})^{1/\gamma_2}) : (x,y) \in C_n\}.$$

We divide the set Q_n by a large positive constant c_n such that Q_n/c_n contains a small portion of the observations. This way we can estimate $v(Q_n/c_n)$ and hence $v(Q_n) := v(Q_n/c_n)/c_n$.

The Probability of a Failure Set

Summing up, the procedure involves the following steps:

- Marginal transformations

$$X_i \to \left(1 + \gamma_1 \frac{X_i - b_1(t)}{a_1(t)}\right)^{1/\gamma_1}$$
$$Y_i \to \left(1 + \gamma_2 \frac{Y_i - b_2(t)}{a_2(t)}\right)^{1/\gamma_2}$$

  in order to transform the marginal distribution approximately to a standard Pareto distribution.

- Use the homogeneity property of the measure v, in order to pull the transformed failure set to the observations.

The Probability of a Failure Set

- Let k be an intermediate sequence, *i.e.*, $k = k(n) \to \infty, k/n \to 0, n \to \infty$.
- Suppose the failure set C_n can be written as

$$C_n = \{\big(a_1(\frac{n}{k})\frac{(c_n x)^{\gamma_1} - 1}{\gamma_1} + b_1(\frac{n}{k}), \\ a_2(\frac{n}{k})\frac{(c_n y)^{\gamma_2} - 1}{\gamma_2} + b_2(\frac{n}{k})\big) : (x, y) \in S\}, \tag{8.1.9}$$

where c_n is a positive sequence and S is a fixed open set of $\mathbb{R}^2$, and the marginal transformations applied to C_n give the set $c_n S$ (called Q_n before).

The Probability of a Failure Set

Then, for some fixed Borel set $S \subset \mathbb{R}_+^2$ with $\inf_{(x,y)\in Q} \max(x,y) > 0$ and $v(\partial Q) = 0$, we can write (8.1.6) as

$$P\{((1+\gamma_1 \frac{X - b_1(\frac{n}{k})}{a_1(\frac{n}{k})})^{1/\gamma_1}, (1+\gamma_2 \frac{Y - b_2(\frac{n}{k})}{a_2(\frac{n}{k})})^{1/\gamma_2}) \in c_n S\}.$$

This is approximately equal to

$$\frac{k}{n} v(c_n S) = \frac{k}{n c_n} v(S)$$

And, this leads to the estimator

$$\hat{p}_n := \frac{k}{n c_n} \hat{v}(\hat{S}).$$

Note that S is not known since $\gamma_1, \gamma_2, a_1, a_2, b_1, b_2$ are not known.

Two treatment for c_n

- Up to this point we have dealt with c_n as if it were known. This way it is to be chosen (under certain bounds) by the statistician.
- An alternative way to deal with c_n is to incorporate it in the problem itself, and consequently to estimate it along with the other unknown quantities.

Some comments about $v(S)$????

- In the above discussion we assumed v(S) positive, and this will be the case considered in the next section.
- In fact this is the case if the random variables X and Y are not asymptotically independent or S contains (at least part of) the axis

$$\{(x, y) : x > 0 \,\text{and}\, y = 0\} \vee \{(x, y) : x = 0 \,\text{and}\, y > 0\}. \qquad (8.1.12)$$

- The case $v(S) = 0$ is discussed in Section 8.3. Clearly $v(S) = 0$ under asymptotic independence and if S is contained in a set of the form $(x, \infty) \times (y, \infty)$, for some $x, y > 0$.

Outline

Assumption

Define

$$q_n := (1 + \gamma_1 \frac{v_n - b_1(\frac{n}{k})}{a_1(\frac{n}{k})})^{1/\gamma_1}, \quad r_n := (1 + \gamma_2 \frac{w_n - b_2(\frac{n}{k})}{a_2(\frac{n}{k})})^{1/\gamma_2}.$$

and assume that $0 < \lim_{n\to\infty} q_n/r_n < \infty$; this avoids the predominance of one marginal over the other so that the problem does not become a univariate one in the limit.

First Approach: c_n known

Further Assumptions:

- $$\sqrt{k}(\hat{\gamma}_i - \gamma_i, \frac{\hat{a}_i(\frac{n}{k})}{a_i(\frac{n}{k})} - 1, \frac{\hat{b}_i(\frac{n}{k}) - b_i(\frac{n}{k})}{a_i(\frac{n}{k})}) = (O_p(1), O_p(1), O_p(1)).$$
- $v(\partial S) = 0$ and $v(S) > 0$, and c_n a sequence of positive numbers with $c_n \to \infty$.
- Suppose $0 < q_n/r_n < \infty$ (this condition imply that q_n/r_n does not depend on n),
  $$\lim_{t \to \infty} \frac{w_{\gamma_1 \wedge \gamma_2}(c_n)}{\sqrt{k}} = 0, \tag{8.2.5}$$
  where
  $$w_\gamma(t) = t^{-\gamma} \int_1^t s^{\gamma-1} \log s ds, t > 1.$$

Some Remarks about the condition

- The estimation of $\gamma_i, a_i(n/k), b_i(n/k)$, is known from the univariate extreme value statistics.
- Note that the relation between k = k(n) and c n may restrict the range of possible values of the marginal extreme value indices. For $\gamma_1 \wedge \gamma_2 < 0$, condition (8.2.5) implies

$$\lim_{t\to\infty} \frac{c_n^{-(\gamma_1\wedge\gamma_2)}}{\sqrt{k}} = \lim_{t\to\infty} k^{-1/2-(\gamma_1\wedge\gamma_2)}(\frac{k}{c_n})^{(\gamma_1\wedge\gamma_2)} = 0.$$

For instance, if we want to allow $k/c_n = O(1)$, we must have $k^{-1/2-(\gamma_1\wedge\gamma_2)} \to 0$, which is true only if $\gamma_1 \wedge \gamma_2 > -\frac{1}{2}$.

First Approach: c_n known

Then, with

$$\hat{p}_n := \frac{1}{nc_n}\sum_{i=1}^{n} 1_{\left\{\left((1+\hat{\gamma}_1\frac{X_i-\hat{b}_1(\frac{n}{k})}{\hat{a}_1(\frac{n}{k})})^{1/\hat{\gamma}_1},(1+\hat{\gamma}_2\frac{Y_i-\hat{b}_2((\frac{n}{k})}{\hat{a}_2(\frac{n}{k})})^{1/\hat{\gamma}_2}\right)\in\hat{S}\right\}},$$

where

$$\hat{S} := \Big\{\Big(\frac{1}{c_n}(1+\hat{\gamma}_1\frac{x-\hat{b}_1(\frac{n}{k})}{\hat{a}_1(\frac{n}{k})})^{1/\hat{\gamma}_1}, \frac{1}{c_n}(1+\hat{\gamma}_2\frac{y-\hat{b}_2(\frac{n}{k})}{\hat{a}_2(\frac{n}{k})})^{1/\hat{\gamma}_2}\Big) : (x,y)\in C_n\Big\},$$

we have

$$\frac{\hat{p}_n}{p_n} \xrightarrow{p} 1.$$

Alternative Approach: Estimate c_n.

Define for some $r > 0$,

$$c_n := \frac{\sqrt{q_n^2 + r_n^2}}{r}, \tag{8.2.9}$$

where q_n and r_n are as in (8.2.2). Under the same condition as Theorem 8.2.1, and with $\gamma_1 \wedge \gamma_2 > -1/2$, define

$$\hat{q}_n := (1 + \hat{\gamma}_1 \frac{v_n - \hat{b}_1(\frac{n}{k})}{\hat{a}_1(\frac{n}{k})})^{-1/\hat{\gamma}_1},$$

$$\hat{r}_n := (1 + \hat{\gamma}_2 \frac{w_n - \hat{b}_2(\frac{n}{k})}{\hat{a}_2(\frac{n}{k})})^{-1/\hat{\gamma}_2}$$

$$\hat{c}_n := \frac{\sqrt{\hat{q}_n^2 + \hat{r}_n^2}}{r},$$

for some $r > 0$ (to be chosen by the statistician). We can prove the consistency. And under much stronger condition, we can prove the asymptotical normality.

Outline

Failure Set Contained in an Upper Quadrant

Let us start with the failure set as an upper quadrant. From (8.1.1) one gets

$$\lim_{t\to\infty} P\Big(\frac{X-b_1(t)}{a_1(t)} > x \text{ or } \frac{Y-b_2(t)}{a_2(t)} > y\Big) = -\log G(x,y)$$

and hence

$$\begin{aligned}\lim_{t\to\infty} P\Big(\frac{X-b_1(t)}{a_1(t)} > x \text{ and } \frac{Y-b_2(t)}{a_2(t)} > y\Big)\\ = \log G(x,y) - \log G(x,\infty) - \log G(\infty,y),\end{aligned}$$

and in case of asymptotic independence the right-hand side is identically zero.

Failure Set Contained in an Upper Quadrant

More generally if Q is any Borel set contained in $[u,\infty)\times[v,\infty)$, with $u,v>0$ and $v(\partial Q)=0$, under asymptotic independence of (X,Y),

$$\lim_{t\to\infty} P\{((1+\gamma_1\frac{X-b_1(t)}{a_1(t)})^{1/\gamma_1},(1+\gamma_2\frac{Y-b_2(t)}{a_2(t)})^{1/\gamma_2})\in Q\}=0,$$

This gives too little information on the probability of the set Q.we propose the following refinement of (8.1.4), which will lead to a new limit measure v: for $x,y>0$,

$$\lim_{t\to\infty} r(t)P\{(1+\gamma_1\frac{X-b_1(t)}{a_1(t)})^{1/\gamma_1}>x,\text{ and }(1+\gamma_2\frac{Y-b_2(t)}{a_2(t)})^{1/\gamma_2}>y\}$$

exists, and it is positive and finite.

Failure Set Contained in an Upper Quadrant

Then, we can redifine the exponential measure v as follows: for any Borel set Q in $\mathbb{R}_+^2$ with $\inf_{(x,y)\in Q} \max(x, y) > 0$ and $v(\partial Q) = 0$, let

$$v(Q) := \lim_{t\to\infty} P\{((1+\gamma_1 \frac{X-b_1(t)}{a_1(t)})^{1/\gamma_1}, (1+\gamma_2 \frac{Y-b_2(t)}{a_2(t)})^{1/\gamma_2}) \in Q\}.$$

Moreover, it follows that the function r is regularly varying with index greater than or equal to 1. Also, in the proof of Theorem 6.1.9, it follows that

$$v(aQ) = a^{-1/\eta} v(Q).$$

Failure Set Contained in an Upper Quadrant

We are now ready to proceed with the estimation of p n , which closely follows the reasoning developed in the previous section. Using again (8.1.8),

$$\begin{aligned} p_n &= P((X, Y) \in C_n) \\ &= P\Big\{\Big((1+\gamma_1 \frac{X - b_1(\frac{n}{k})}{a_1(\frac{n}{k})})^{1/\gamma_1}, (1+\gamma_2 \frac{Y - b_2(\frac{n}{k})}{a_2(\frac{n}{k})})^{1/\gamma_2}\Big) \in c_n S\Big\}. \end{aligned}$$

which is approximately equal to

$$\frac{v(c_n S)}{r(\frac{n}{k})} = \frac{v(S)}{c_n^{1/\eta} r(\frac{n}{k})}.$$

The estimation procedure is the same as before. And the consistency can be proved.