Extreme Value Theory for Anomaly Detection

TransferLab, appliedAI Institute for Europe gGmbH

March 5, 2024

1 Taking a look back at the theory

So, what have we really done and why does it make sense to use the GEV for such problems? What kind of guarantees does the Fisher-Tipett-Gnedenko theorem give us about the quality of the fit?
Well, the truth is, not too many. First notice the following exact equality:

$$P(M_n < z) = P(X_1 < z \text{ and } X_2 < z ... \text{ and } X_n < z) = F^n(z) \tag{1}$$

So, if we know the cumulative distribution, there is no need to resort to the GEV. Typically, of course, we do not know it. The above equality implies:

$$\lim P(M_n < z) = \begin{cases} 0 & \text{if } F(z) < 1 \\ 1 & \text{otherwise} \end{cases} \tag{2}$$

We actually always know the exact limit of the distribution of the block-maxima! It is degenerate (either a step function of identical zero). In fact, this degenerate distribution can be seen as a limit of the GEV. It would correspond to normalizing constants $a_n = 1,\ b_n = 0$.
While this observation is very simple and the difference between the cdf of block maxima $P(M_n < z)$ and its degenerate limit does decrease as n increases, this limiting distribution is unexpressive and fitting it to data does not provide probabilistic insight.

Q: How many parameters does the exact limit of F^n have? What would we get if we fit it to data?
A:

Introducing the normalizing constants a_n and b_n *might* allow the distribution of renormalized block maxima to converge to something non-trivial. It also might not.

In applications we usually care about modeling M_n for a *fixed* n_0 (or maybe for a few selected n_i). An arbitrary series of a_n and b_n that at some point helps convergence does not directly address our needs. In fact, this is also not what we do — by fitting the GEV parameters to data for our selected n_0 we automatically find the *best* a_{n_0} and b_{n_0} that minimize the difference between $F^{n_0}(z)$ and $G(z)$.
Clearly $G(z)$ is much more expressive than the degenerate exact limit and could potentially provide a good fit. So, the convergence that we really care about is to answer the question: How well do the best fits of $G(z)$ for fixed n — let us call them $G_n(z)$ — approximate the distributions $F^n(z)$ as n increases? One could e.g. be interested in the infinity norm

$$\Delta_n := \sup_z |F^n(z) - G_n(z)| \tag{3}$$

This is not the same as asking how well $G(z)$ approximates some rescaled variant of $F^n(z)$ with n-dependent normalization constants! That would be

$$\tilde{\Delta}_n(a_n, b_n) := \sup_z |F^n(a_n z + b_n) - G(z)| \tag{4}$$

In the latter question, the choice of normalization constants matters, in the former it does not — they are implicitly determined by the best fit for each n. Since for Δ_n the a_n, b_n have been optimized, one could reasonably expect a relation of the type

$$\Delta_n \approx \min_{a_n, b_n} \tilde{\Delta}_n(a_n, b_n) \tag{5}$$

to hold.
It is easy to see that given some normalizing sequences a_n, b_n, the convergence to a GEV is possible, than with other sequences $\tilde{a}_n, \tilde{b}_n$ with some $a > 0, b$ such that

$$\lim_{n\to\infty} \frac{\tilde{a}_n}{a_n} = a \quad , \quad \lim_{n\to\infty} \frac{b_n - \tilde{b}_n}{a_n} = b \tag{6}$$

the rescaled $\frac{M_n - \tilde{b}_n}{\tilde{a}_n}$ also converges to a GEV of the same type (with the same ξ). This is often formulated that a distribution F has *a fixed domain of attraction*. However, the error rates $\tilde{\Delta}_n(\tilde{a}_n, \tilde{b}_n)$ would be different from those associated to a_n, b_n.
Unfortunately, theoretical bounds for the quantity of interest Δ_n are hard to come by — we are not aware of any. They also highly depend on the fitting procedure, which is non-trivial, as we have seen above. There are some bounds for quantities of the type $\tilde{\Delta}_n(\tilde{a}_n, \tilde{b}_n)$ (see the annotated literature reference) but they are rather loose and not really helpful in practice. Therefore, the EVT theorems are more of a motivation for selecting distribution families for fitting than a rigorous approach with guarantees. In practice the convergence and fit tend to work pretty well, though.

2 Proofing the Fisher-Gnedenko-Tripet theorem

One may wonder how the statement of the Fisher-Gnedenko-Tripet theorem is obtained without providing bounds on convergence. The reason is that the limiting distribution of (renormalized) maxima must have a very special property — it must be max-stable. It is instructive to go through a part of the proof to get a feeling for the EVT theorems. We will do so in this exercise.

Definition 2.1. *A cumulative distribution function $D(z)$ is called max-stable iff for all $n \in \mathbb{N}$ $\exists$ $\alpha_n > 0, \beta_n \in \mathbb{R}$ such that*

$$D^n(z) = D(\alpha_n z + \beta_n). \tag{7}$$

Prove that from $\lim_{n\to\infty} P\left(\frac{M_n - b_n}{a_n} < z\right) = G(z)$ follows that $G(z)$ is max-stable.

This goes a long way towards proving the first EVT theorem. One can easily compute that the GEV distribution is max-stable and with more effort one can also prove that any max-stable distribution belongs to the GEV family. Thus, the proof of the theorem is very implicit and does not involve any convergence rates or bounds.

3 Deriving the second theorem of EVT

Use the approximation $\ln(1+x) \approx 1+x$ for $|x| \ll 1$ and $F(z) \approx 1$ for large enough z to derive.

$$P(X-u<y \mid X>u) \approx 1-\left(1+\frac{\xi\ y}{\tilde{\sigma}}\right)^{-\frac{1}{\xi}} \tag{8}$$

for large enough u (this is a slightly less formal derivation of Pickards' et. al. theorem). One could equivalently write

$$P(X-u>y \mid X>u) \approx \left(1+\frac{\xi\ y}{\tilde{\sigma}}\right)^{-\frac{1}{\xi}}. \tag{9}$$

What is the relation between $\tilde{\sigma}$ and the normalizing coefficients of the first theorem of EVT?

4 Peaks over thresholds

So far we have only used the first theorem of EVT. As you might have noticed above, it can be somehow wasteful when it comes to data efficiency. Since the GEV is fitted on block-maxima, a huge number of data points remain unused for parameter estimation. The second theorem of EVT gives rise to a more efficient approach
The above equation can be used to estimate the entire tail of the cdf F of X from a sample of size N obtained by sampling repeatedly from F. First note that for a single u we can approximate the cdf through the sample statistics as:

$$1 - F(u) = P(X > u) \approx \frac{N_u}{N} \tag{10}$$

where N_u is the number of samples with values above u. Interpreting u as a threshold, we will call those samples *peaks over threshold* (PoT) and N_u is simply their count.

Q: What should u and the data set fulfill in order for the above approximation to be accurate?
A: It should be small enough such that many data points are larger than it. Then the approximation in $P(X > u) \approx \frac{N_u}{N}$ holds (the estimator is not too biased).

Now we can perform a series of approximations for $z > u$ to get to the tail-distribution. First using $P(X > u) \approx \frac{N_u}{N}$ we get

$$\begin{aligned} P(X > z) &= P(X > z \cap X > u) \\ &= P(X > z \mid X > u) P(X > u) \\ &\approx \frac{N_u}{N} P(X > z \mid X > u). \end{aligned} \tag{11}$$

Now we use the GDP theorem to approximate

$$\begin{aligned} P(X > z \mid X > u) &= P(X - u > z - u \mid X > u) \\ &\approx \left(1 + \frac{\xi(z-u)}{\tilde{\sigma}}\right)^{-\frac{1}{\xi}}. \end{aligned} \tag{12}$$

Putting everything together gives

$$P(X > z) \approx \frac{N_u}{N} \left(1 + \frac{\xi(z-u)}{\tilde{\sigma}}\right)^{-\frac{1}{\xi}}. \tag{13}$$

Q: Intuitively, what does u need to fulfill for both approximations to hold?
A: u should be small enough such that the approximation $P(X > u) \approx \frac{N_u}{N}$ holds and sufficiently large such that the generalized pareto distribution is a good estimate of the tail of the distribution for values larger than u. Intuitively, it should be at the *beginning of the tail*, where for values larger than u only the tail behavior plays a role - i.e. no more local extrema or other specifics of the underlying distribution of the data.
