EVA 2023 Software tutorial

Conditional extremes

Thomas Opitz

Biostatistics and Spatial Processes, INRAE, Avignon

Extreme-Value Analysis Conference 2023

Milano, 30 June 2023

Setting of conditional extremes

Consider a random vector $\boldsymbol{X} = (X_1, \ldots, X_d)$.
We denote by $\boldsymbol{X}_{-j}$ the vector $\boldsymbol{X}$ with component X_j removed for $j \in \{1, \ldots, d\}$.

What is possible conditional behavior of

$$\boldsymbol{X}_{-j} \mid (X_j > u)$$

for a high threshold u with $\Pr(X_j > u) > 0$ small?

Asymptotic framework as $u \to u^\star = F_{X_j}^{-1}(1^-)$?

Flexible statistical models?

$\Rightarrow$ Multivariate conditional extremes (Heffernan & Tawn, 2004, JRSSB)

Asymptotic framework

- **Standard exponential tails** $\Pr(X_j > x) \approx c \times \exp(-x)$ as $x \to \infty$ (with $c > 0$)
- Useful choice: **Standard Laplace distribution** F_{Laplace} with density $\frac{1}{2}\exp(-|x|)$
  $\Rightarrow$ **Pretransformation** $X_j \mapsto F^{-1}_{\text{Laplace}}\left(F_{X_j}(X_j)\right)$.

Asymptotic setting for X_j with exponential tails

We assume: there are normalizing function vectors $\boldsymbol{a}_{|j}$ and $\boldsymbol{b}_{|j} > \boldsymbol{0}$ such that

$$\Pr\left(\left\{\frac{X_k - a_{k|j}(X_j)}{b_{k|j}(X_j)}\right\}_{k\neq j} \leq \boldsymbol{z}_{-j} \,\middle|\, X_j \geq u\right) \to \Pr\left(\boldsymbol{Z}_{|j} \leq \boldsymbol{z}_{-j}\right), \quad u \to \infty,$$

with nondegenerate **residual random vector** $\boldsymbol{Z}_{|j}$ with components $Z_{k|j}$ for $k \neq j$.
(Note: $\boldsymbol{z}_{-j}$ is $\boldsymbol{z} = (z_1, \ldots, z_d)$ with component z_j removed.)

- We have $(X_j - u) \mid (X_j > u) \to E \sim \mathrm{Exp}(1)$ and $E \perp \boldsymbol{Z}_{|j}$
- Location-scale transformation ensures non-degenerate limit
- Under mild conditions, we can replace the condition $X_j \geq u$ by $X_j = u$
- By considering $j \in \{1, \ldots, d\}$, we get d limit relations.

Statistical modeling of conditional extremes

In practice, we fix a threshold u and consider the location/scale normalizations as parameters to be estimated. We focus on the approach as implemented in `texmex`.

General model

Given $\boldsymbol{X}$ with exponential tails, we fix a high threshold u and assume

$$\boldsymbol{X}_{-j} \mid (X_j > u) \quad = \quad \boldsymbol{a}_{|j}(X_j) + \boldsymbol{b}_{|j}(X_j)\boldsymbol{Z}_{|j}$$

with deterministic parametric functions $\boldsymbol{a}_{|j}$ and $\boldsymbol{b}_{|j} > \mathbf{0}$, and residual vector $\boldsymbol{Z}_{|j}$.

Semiparametric nonlinear regression model

Given $\boldsymbol{X}$ with exponential tails, we fix a high threshold u and assume

$$\boldsymbol{X}_{-j} \mid (X_j > u) \quad = \quad \boldsymbol{\alpha}_{|j} X_j + X_j^{\boldsymbol{\beta}_{|j}} \boldsymbol{Z}_{|j}$$

with $\boldsymbol{\alpha}_{|j} \in [-1, 1]^{d-1}$ and $\boldsymbol{\beta}_{|j} \in (-\infty, 1]^{d-1}$, and with unspecified residuals $\boldsymbol{Z}_{|j}$.

Pseudo-likelihood estimation: Estimate $\boldsymbol{\alpha}_{|j}, \boldsymbol{\beta}_{|j}$ by assuming $\boldsymbol{Z}_{|j} \sim \mathcal{N}_{d-1}(\boldsymbol{\mu}_{|j}, \mathrm{diag}(\boldsymbol{\sigma}^2_{|j}))$ with nuisance parameters $\boldsymbol{\mu}_{|j}$ and $\boldsymbol{\sigma}^2_{|j}$.

Bivariate dependence properties

Assume without loss of generality $\boldsymbol{X} = (X_1, X_2)$.

Tail correlation / χ-measure

$\chi = \lim_{u\to 1} \chi_u \in [0,1]$ where

$$\chi(u) = \Pr(F_{X_2}(X_2) > u \mid F_{X_1}(X_1) > u) = \frac{\Pr(F_{X_1}(X_1) > u, F_{X_2}(X_2) > u)}{\Pr(F_{X_1}(X_1) > u)}, \quad u \in (0,1).$$

Interpretation:
Asymptotic dependence for $\chi > 0$, otherwise asymptotic independence

Behavior of the conditional extremes model for $X_2 \mid (X_1 > u)$

- Asymptotic dependence only if $\alpha_{2|1} = 1$ (and $\beta_{2|1} = 0$)
- Positive extremal dependence if $\alpha_{2|1} > 0$
- Negative extremal dependence if $\alpha_{2|1} < 0$
- Near independence if $\alpha_{2|1} = 0$ and $\beta_{2|1} = 0$

Usually we assume $0 < \beta_{2|1} \le 1$ to avoid strange limiting behavior.

Strategies for marginal normalization

- Real data usually do not come with exponential tails.
- We therefore have to transform marginal distributions using a Probality Integral Transform, by default to the Laplace distribution.
- For the components X_k different from the conditioning component X_j, we also model the values below the threshold u. Therefore, we need a transformation of the whole distribution and not only for threshold exceedances.
- The default approach is to use a Generalized Pareto distribution $\mathrm{GPD}(\xi_j, \sigma_{u,j})$ above u, with threshold exceedance probability p_j, and to use the empirical distribution (that is, the rank transform) below the threshold.
- In certain modeling contexts, other choices may be relevant:
  - Pretransform data using only the empirical distribution (for bulk and tail).
  - Use a parametric distribution that jointly models bulk and tail, such as a variant of Extended Generalized Pareto Distributions (EGPD).
  - Use a different way of combining a bulk model and tail model, such as linear interpolation of regression quantiles for a grid of non-extreme quantile levels (such as $0.01, 0.02, \ldots, 0.95$).
  - If the marginal distributions depend on covariates, then using the rank transform is difficult.

Marginal normalisation (default approach)

We have to estimate marginal distributions $\hat{F}_j(x)$ for $j = 1, \ldots, d$.

Let us write $X_{j,1}, X_{j,2}, \ldots, X_{j,n}$ for the sample of component j.

Two options for obtaining a threshold u_j:

- Fix the quantile u_j, and estimate the exceedance probability $p_j = \Pr(X_j > u)$
- Fix the exceedance probability p_j, and estimate the quantile $u_j = F_j^{-1}(p_j)$

Estimated marginal distribution

We fit a $\mathrm{GPD}(\hat{\xi}_j, \hat{\sigma}_{u,j})$ for threshold exceedances above u and use the empirical distribution below:

$$\hat{F}_j(x) = \begin{cases} F_{j,n}(x), & x \leq u, \\ 1 - p_j + p_j \times \mathrm{GPD}(x - u; \hat{\xi}_j, \hat{\sigma}_{u,j}), & x > u. \end{cases}$$

Uncertainty quantification

How can we obtain standard errors and confidence intervals for the parameters?

The pretransformation with estimated marginal distribution must be considered.

`texmex` implements a **non-standard bootstrap approach** to obtain standard errors and other uncertainty measures for estimates of marginal GPD parameters and $\hat{\alpha}_{|j}$, $\hat{\beta}_{|j}$; see Heffernan & Tawn (2004).

Simulation with **extrapolation** towards very-low-probability events is possible by using a higher threshold $v > u$ for X_j.

Generation of a new extreme event conditional to $X_j > v$

- Simulate $x_j \sim v + \mathrm{Exp}(1)$
- Draw $\boldsymbol{z}_{|j}$ at random from the empirical distribution of residuals
- Set $\boldsymbol{x}_{-j} = \hat{\boldsymbol{\alpha}}_{|j} + x_j^{\hat{\boldsymbol{\beta}}_{|j}} \boldsymbol{z}_{|j}$
- Backtransform $\boldsymbol{x}_{-j}$ to original marginal scale of data:

$$x_k \mapsto \hat{F}_j^{-1}\left(F_{\mathrm{Laplace}}(x_k)\right), \qquad \text{for } k \neq j$$

- Return the simulated vector $\boldsymbol{x}$.

To predict $\Pr(E)$ for various types of risk regions, we can repeatedly simulate from the conditional models and then construct a Monte-Carlo estimate.

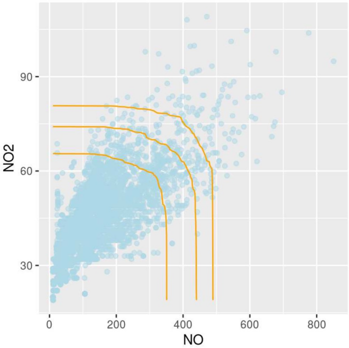
Joint exceedance curves

Among the numerous inferences that can be drawn based on conditional extremes models, joint exceedance curves and their graphical display are implemented in `texmex`.

In a bivariate setting for $1 \leq j_1 < j_2 \leq d$ and for a given probability p_0,
the **joint exceedance curve** is defined as the set

$$\{(x_1, x_2) : \Pr\left(X_{j_1} > x_1, X_{j_2} > x_2\right) = p\}$$

Simulation-based joint exceedance curves for three extreme probability levels:



Issues of self-consistency

Consider the bivariate case and the two conditional extremes models for

$$X_2^L \mid X_1^L > u, \quad X_1^L \mid X_2^L > u$$

- ⚠ Models where $\Pr(X_2 > x \mid X_1 > u) > c \times \exp(-x)$ are not consistent with the marginally normalized exponential tails
- Both conditional models fully characterize the behavior of (X_1, X_2) in the region $\{(x_1, x_2) \in \mathbb{R}^2 \mid x_1 > u, x_2 > u\}$.
  However, it is challenging to impose conditions on the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ and on the residual distributions $\boldsymbol{Z}$ that ensure consistency of the two models.
- This does not prevent the model from providing very useful estimations and predictions!

Some necessary consistency conditions were stated by Keef et al. (2013) and are imposed in `texmex` when using Laplace marginal distributions.

Pros and cons of models for conditional extremes

Pros

- Modeling flexibility beyond asymptotic stability and asymptotic dependence
- Semiparametric formulation and uncertainty assessment is strongly data-driven
- Complete software implementation with state-of-the-art visualization

Cons

- Lack of self-consistency of the conditional distributions
  (usually only some necessary conditions are imposed)
- Semi-parametric approach does not scale well to higher dimensions
- Asymptotic dependence summaries (such as $\overline{\chi}$) can vary with threshold u

There is various ongoing research on theoretical properties and modeling extensions!

Wrap-up: Approaches for conditional extremes

- Main implementation in the `texmex` package, which further provides implementations for marginal Peaks-Over-Threshold modeling (including declustering strategies and estimation of the Extremal Index).

- Theoretical self-consistency issues are no impedement to robust statistical modeling in practice.

- Many recent works on spatial and spatiotemporal conditional extremes where $\boldsymbol{Z}$ is a Gaussian process (sometimes marginally transformed).