EVA 2023 Software tutorial

Times series extremes

Léo Belzile, **Thomas Opitz**

Biostatistics and Spatial Processes, INRAE, Avignon

Extreme-Value Analysis Conference 2023

Milano, 30 June 2023

Plan for this part

Modeling aspects, implementations and worked-out code examples for two settings:

- **Extremes in time series**
- **Conditional extremes**

Note: In the following, **Belzile et al. (2023+)** refers to our software review,

Belzile, L. R., Dutang, C., Northrop, P. J., & Opitz, T. (2022). A modeler's guide to extreme value software. arXiv preprint arXiv:2205.07714.

A near-exhaustive list of available implementations on the CRAN is given in the Task View of Extreme-Value Analysis:

https://cran.r-project.org/web/views/ExtremeValue.html

Modeling aspects for time series extremes

- Exploration and summaries for temporal clustering: extremogram, extremal index...
- Estimation of marginal tails under serial dependence
- Models for serially dependent extremes, such as Markov chains
- Nonstationarity (e.g. seasonality)

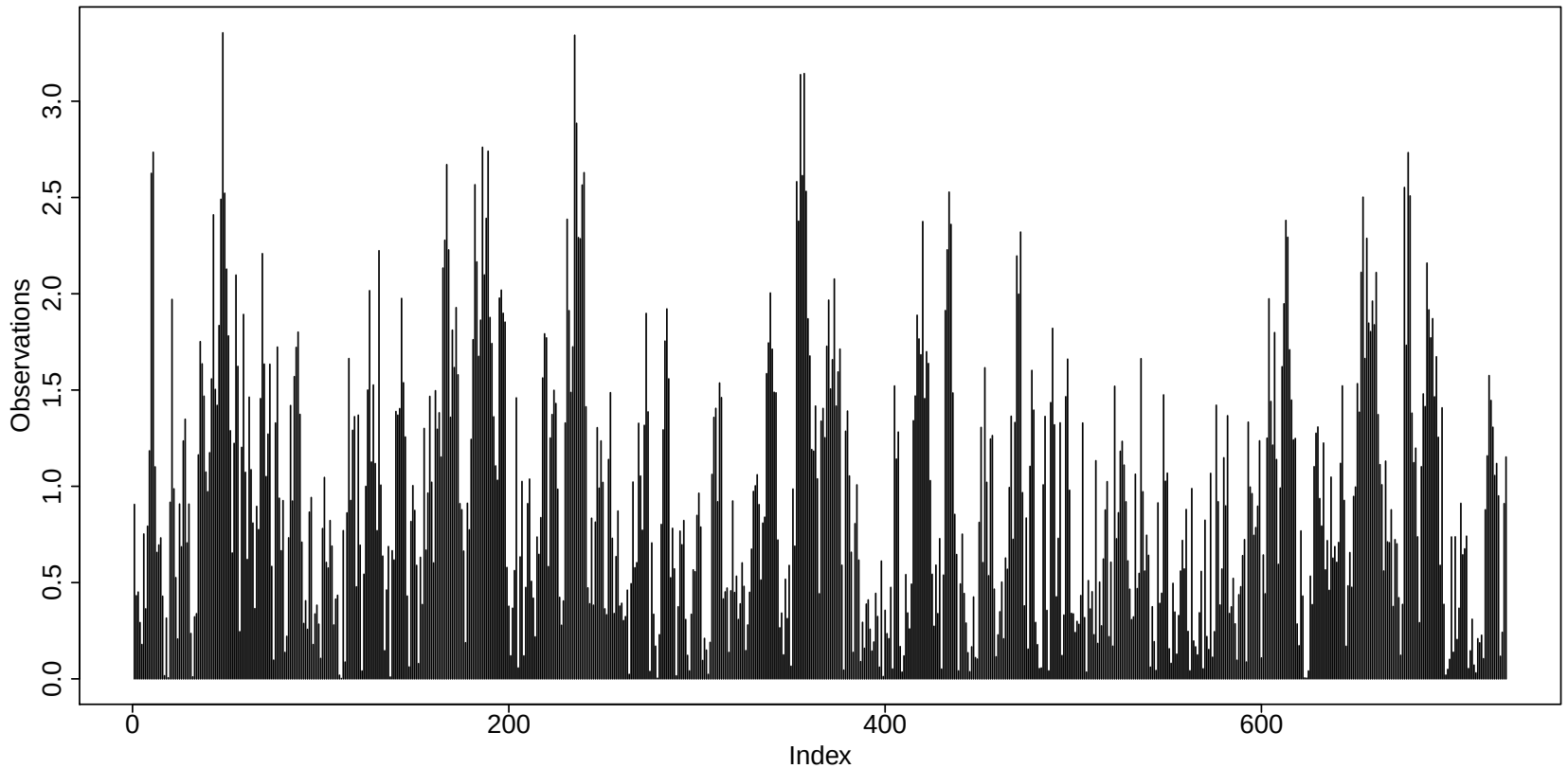
We have to decide if temporal extremal dependence is simply a nuisance for estimation of univariate tails, or if we seek to fully characterize it using a generative model.

Illustration: A stationary dependent sequence

The default assumption for many theoretical results and estimation techniques is stationarity of the time series

$$X_1, X_2, \ldots$$

If the observed series is not stationary, this may require steps of data pre-processing (choice of a nonstationary threshold; marginal pre-transformation of data...)



Stationarity

If not stated otherwise, we assume that the variables $X_1, X_2, \ldots$ define a **stationary time series**, such that blocks of the same length possess the same joint distribution:

$$(X_{k_1+1}, \ldots, X_{k_1+m})^T \stackrel{d}{=} (X_{k_2+1}, \ldots, X_{k_2+m})^T \quad \text{for any } k_1, k_2, m \geq 1.$$

Possible consequences of dependent observations:

- A nondegenerate asymptotic distribution of maxima $M_n = \max_{i=1}^n X_i$ may not exist, or may be different from the GEV distribution.
- Consecutive excesses over a high threshold may be dependent, for instance by arising in **clusters**.

$D(u_n)$-condition

For stationary processes, this standard mixing condition allows defining the **extremal index** and preserves the GEV and GPD limits for maxima and threshold exceedances, respectively.

Extremogram (Tail autocorrelation function)

Autocorrelation functions are important tools in classical time series analysis.
The **extremogram** is an extreme-value analogue.
Suppose that $X_i \sim F$, $i = 1, 2, \ldots$

Given a time lag $h \in \{0, 1, 2 \ldots\}$, we consider the conditional exceedance probability

$$\chi(h; u) = \Pr(F(X_{i+h}) > u \mid F(X_i) > u) = \frac{\Pr(F(X_{i+h}) > u, F(X_i) > u)}{\Pr(F(X_i) > u)}, \quad u \in (0, 1).$$

Here, we define the **extremogram** as the limit (if it exists)

$$\chi(h) = \lim_{u \to 1} \chi(h; u) \in [0, 1], \quad h = 0, 1, 2 \ldots$$

- The general definition of the extremogram in the literature allows for extreme event sets that are more general than $[u, 1]$ but the above formulation is commonly used in practice.
- The function $\chi(h)$ is sometimes also called **tail autocorrelation function** or **auto-tail dependence function**.

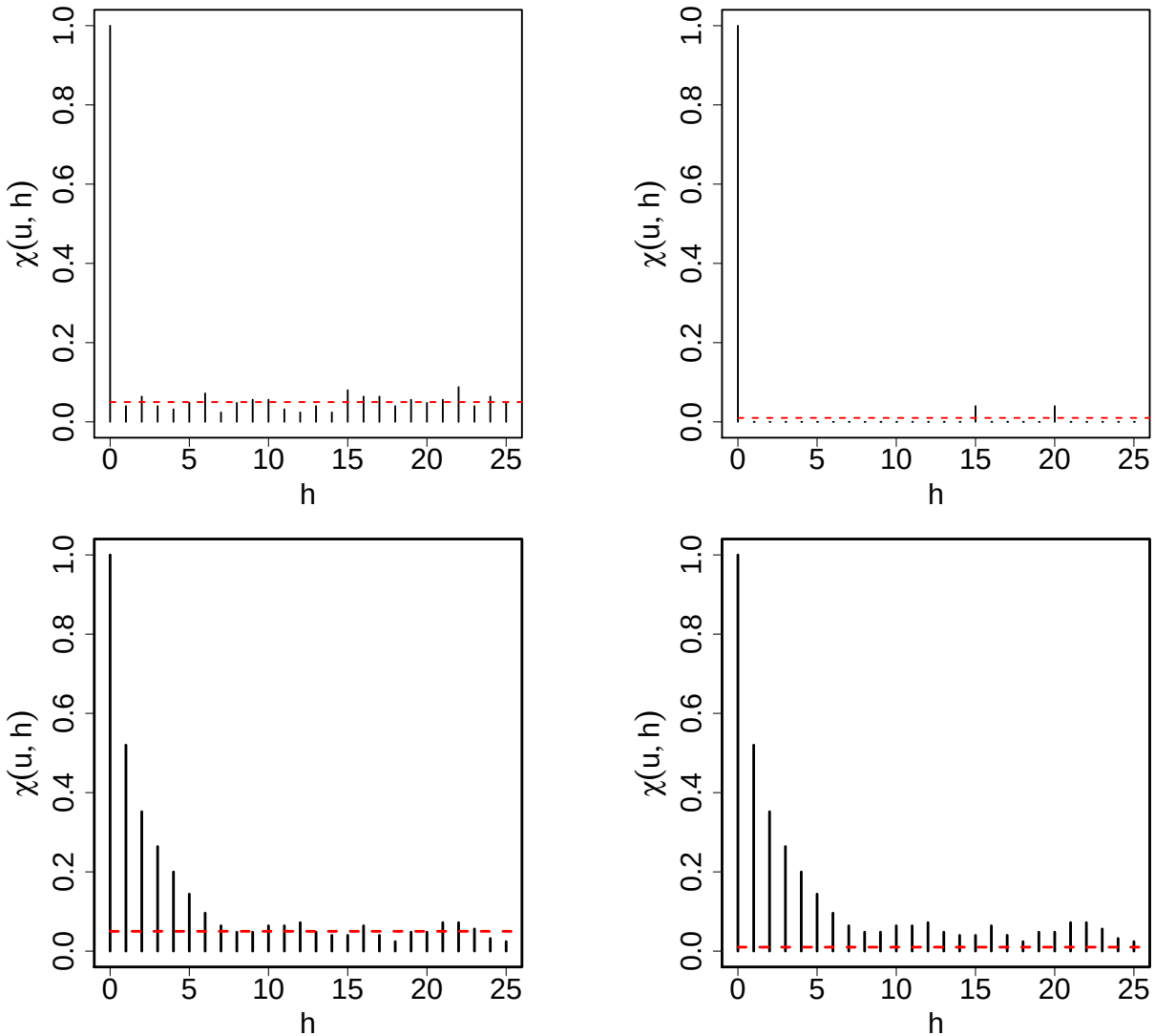
Properties of the extremogram

- $\chi(h)$ characterizes co-occurrence probabilities of high values at temporal lag h. There are also variants for spatial processes where h is spatial distance.
- By definition, $\chi(0) = 1$.
- With independent observations, $\chi(h) = 0$ for $h > 0$ (but $\chi(h; u) = 1 - u > 0$ for "subasymptotic" $u < 1$).
- The series $\{X_i\}$ is called **asymptotically independent at lag h if $\chi(h) = 0$**.
- For data, an **empirical version** can be estimated using $\chi(h; u)$ with empirical probabilities, that is, with F replaced by the empirical distribution function F_n.
- **Cross-extremogram:** With two time series $X_i^{(1)} \sim F_1$ and $X_i^{(2)} \sim F_2$, $i = 1, 2, \ldots$, we can consider

$$\chi_{12}(h) = \lim_{u \to 1} \Pr\left(F_2(X_{i+h}^{(1)}) > u \mid F_1(X_i^{(1)}) > u\right)$$

Illustration: Empirical tail autocorrelation function

Top row: independence; bottom row: asymptotic dependence
Left column: $u = 0.95$; right column: $u = 0.99$
Dashed red line corresponds to theoretical $\chi(h; u)$ for independence.



R implementations of the extremogram

- Dedicated package `extremogram`, but not (yet) very stable; provides also cross-extremograms and confidence bounds

- `atdf` function in `extRemes` package provides a time-series version of χ (extremogram) and also of $\overline{\chi}$
  (⚠ notation used in the package is ρ instead of χ)

The extremal index

The **extremal index** $\theta \in (0, 1]$ can be defined as the **reciprocal of the limit of the expected cluster size** in exceedances above an increasingly high threshold:

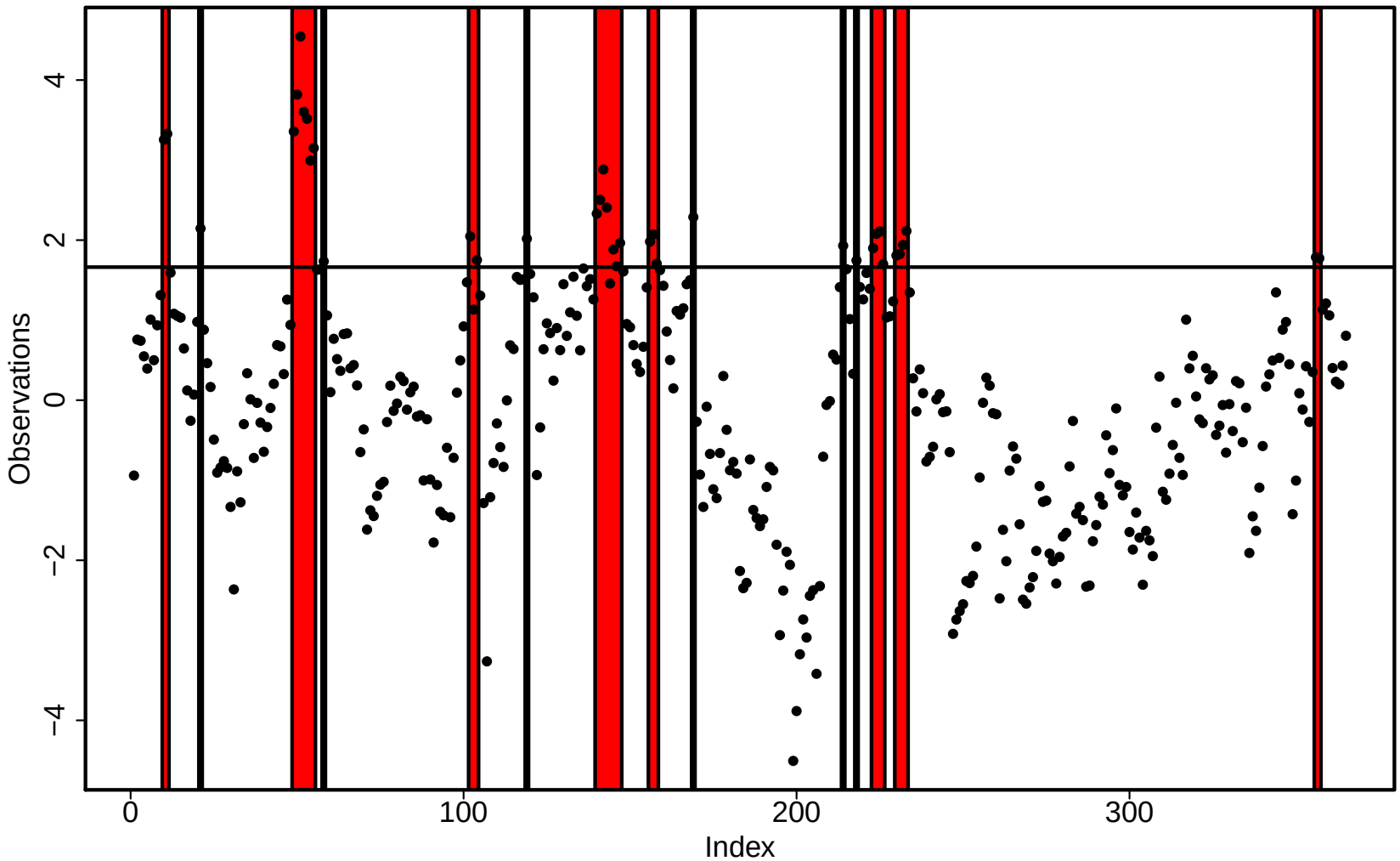
$$\frac{1}{\theta} = \lim_{n \to \infty} \mathbb{E}\left(\sum_{i=1}^{\rho_n} \mathbb{I}(X_i > u_n) \mid M_{\rho_n} > u_n \right)$$

where

- $M_{\rho_n} = \max_{i=1}^{\rho_n} X_i$,
- $\rho_n \to \infty$ and $\rho_n / n \to 0$ as $n \to \infty$,
- $\mathbb{I}(A) = 1$ if the event A occurs, and 0 otherwise, is the indicator function,
- threshold u_n with $n(1 - F(u_n)) \to \lambda$ with some $0 < \lambda < \infty$.

Interpretation: If at least one X_i exceeds a very high threshold in a block of consecutive observations, then on average there are $1/\theta$ "nearby" observations that exceed this threshold.

Illustration: Clusters of exceedances



The average length of the intervals in red tends to $1/\theta$ when we increase the threshold (black horizontal line).

Example: Estimation based on the runs-method

Estimation of the extremal index through threshold exceedances:

$$\hat{\theta} = \frac{1}{\text{average cluster length}}$$

How can we define and identify clusters in practice?

The runs method

Principle: two exceedances are part of the same cluster if there are at most $k-1$ consecutive non-exceeding observations in-between, with a tuning parameter $k \geq 1$.

1. Fix a high threshold u, such as the empirical 95%-quantile of data.
2. Fix k. Often $k = 1$ or $k = 2$.
3. Look for the index i_0 with first threshold exceedance $X_{i_0} > u$ in X_i, $i = 1, 2 \ldots$ $\Rightarrow$ the first cluster begins at i_0.
4. If at least one of $X_{i_0+1}, \ldots, X_{i_0+k}$ exceeds u, then $i_0 + 1$ is still part of the first cluster.
5. Iterate until k **consecutive non exceedances** above u are found.
6. The first cluster goes from i_0 until the last found exceedance.
7. Continue and iterate through Steps 3 to 6 to detect the second cluster, third cluster, and so on.

Extremal index: available estimators

Taken from Belzile et al. (2023+)

estimator	reference	tuning parameter(s)
runs	Smith and Weissman (1994)	run length, threshold
blocks (blocks 1)	Smith and Weissman (1994)	block size, threshold
modified blocks (blocks 2)	Smith and Weissman (1994)	block size, threshold
intervals (FS)	Ferro and Segers (2003)	threshold
iterative least squares (ILS)	Süveges (2007)	threshold
K-gaps	Süveges and Davison (2010)	run length K, threshold
semiparametric maxima (SPM)	Northrop (2015)	block size

Table 6: Overview of some direct estimators of the extremal index with associated references and tuning parameters.

Extremal index: available R implementations

Taken from Belzile et al. (2023+)

package	estimator(s)	estimation	UQ	diagnostics
evd	runs, FS	`exi`	no	`exiplot`
evir	blocks 2	`exindex`	no	`exindex`
extRemes	runs, FS	`extremalindex`	yes	—
exdex	ILS	`iwls`	no	—
	K-gaps	`kgaps`	yes	`choose_uk`
	SPM	`spm`	yes	`choose_b`
fExtremes	runs	`runTheta`	no	`exindexPlot`
	blocks 1	`clusterTheta`	no	`exindexesPlot`
	blocks 2	`blocktheta`	no	
	intervals	`ferrosegersTheta`	no	
mev	ILS, FS	`ext.index`	no	`ext.index`
	K-gaps	`ext.index`	no	`infomat.test`, `ext.index`
POT	runs	`fitexi`	no	`exiplot`
revdbayes	K-gaps	`kgaps_post`	yes	—
texmex	FS	`extremalIndex`	yes	`extremalIndexRangeFit`
tsxtreme	runs	`thetaruns`	yes	—

Table 7: Comparison of R packages for the direct estimation of the extremal index. Estimator(s): name(s) of the estimators available; estimation: function name(s) for estimation; uncertainty quantification (UQ): are methods for estimating uncertainty provided?; diagnostics: function names(s) for choosing tuning parameters.

Declustering for marginal POT modeling

If exceedances in threshold-exceedance models may be dependent, we can obtain an (approximately) independent sample by considering only the most extreme observation among each "cluster" of extremes.

Steps for peaks-over-threshold with declustering

1. Use an **empirical rule to define clusters of exceedances**, for instance the runs method.
2. Identify the **maximum excess of each cluster**.
3. Assume cluster maxima to be independent, with their excesses $X_j - u > 0$ following the GPD
4. Estimate the **GPD** for the sample of threshold excesses of cluster maxima.

Interestingly, under the usual mixing conditions, the GPD limit distribution for the maximum excess is exactly the same as the one for excesses without declustering!

⚠ We could apply standard GPD estimators without declustering but then uncertainty estimates (e.g. confidence intervals) would be biased.

⚠ Results of the GPD model have to be interpreted with respect to cluster maxima.

R implementations of declustering and GPD estimation

- `evd` package, function `clusters`
  - argument `r` for k of runs estimator
  - `cmax` argument to extract cluster maxima
  - threshold can be time-varying
  - plotting is available
- `POT` package, function `clust`
  - argument `tim.cond` for k of runs estimator
  - `clust.max` argument to extract cluster maxima
  - plotting is available
- `extRemes` package, function `decluster`
  - argument `r` for k of runs estimator
  - application of various cluster functions, such as maximum
  - result can be used directly for declustered GPD estimation with `fevd` function
- `texmex` package, function `declust`
  - argument `r` for k of runs estimator
  - cluster maxima are returned
  - result can be used directly for declustered GPD estimation with `evm` function

Estimation without explicit declustering: package `lite`, function `flite`

- Based on an appropriately "post-processed" likelihood assuming independent exceedances
- Estimates and estimation uncertainty for (θ, ξ, σ_u) and exceedance probability p_u

Models for extremal dependence in time series

As before, we assume that $X_1, X_2 \ldots$ is a stationary time series.

Some options for modeling the serial dependence:

- Apply a model for **multivariate extremes** of d-variate random vectors to $(X_t, X_{t+1}, \ldots, X_{t+d})$ by splitting the time series into d-vectors
- Apply a model for **conditional extremes** to $(X_{t+1}, \ldots, X_{t+d}) \mid (X_t > u)$
  $\rightsquigarrow$ `texmex` and `tsxtreme` packages
- Apply a **first-order Markov chain model** to exceedances above u:
  - Marginal model $(X_t - u) \mid (X_t > u) \sim \mathrm{GPD}(\xi, \sigma_u)$
  - Dependence model: assume the first-order Markov chain property

$$X_{t+1} \mid (X_t, X_{t-1}, \ldots, X_1) \quad \stackrel{d}{=} \quad X_{t+1} \mid X_t$$

  - Use the conditional distributions of a bivariate Multivariate Generalized Pareto Distribution to model $X_{t+1} \mid X_t$ if $\max(X_t, X_{t+1}) > u$.

  $\rightsquigarrow$ `fitmcgpd` function in `POT` package
  (joint estimation of marginal and dependence parameters is possible)

Time series extremes: Overview of R implementations

Taken from Belzile et al. (2023+)

reference	package	function(s)	area
Fawcett and Walshaw (2012)	**texmex**	`declust, evm`	**m**
Fawcett and Walshaw (2012)	**lite**	`flite`	**m**
Durrieu et al. (2018)	**extremefit**	`hill.ts`	**m**
Davis and Mikosch (2009)	**extremogram**	`extremogram1, bootconf1, ...`	**e**
Lugrin et al. (2016)	**tsxtreme**	`depfit, dep2fit`	**d**
Smith et al. (1997)	**evd**	`evmc`	**d**
Smith et al. (1997)	**POT**	`fitmcgpd, simmc`	**d**
Noven et al. (2018)	**ev.trawl**	`FullPL, rtrawl`	**d**
Hees et al. (2021)	**CTRE**	`MLestimates`	**d**

Table 8: Overview of packages and main functions for modeling time series extremes by area: marginal modelling (**m**); exploratory analysis (**e**); dependence modelling (**d**).

Strategies for handling marginal nonstationarity

Often, data are marginally nonstationary (intra-day variability, seasonalities...):

- Estimation and interpretation of clusters and of the extremal index could still make sense in a nonstationary setting but with possibly very large clusters due to the nonstationarity.
- Sometimes, we can simply subset the data (for instance, use only a specific season) to get approximate stationarity.
- To remove marginal nonstationarities, it makes sense to use regression models where time is a covariate, for instance by defining a nonstationary threshold as a high quantile estimated through quantile regression.
- **Example:** If estimation is based only on exceedance indicators $\mathbb{I}(X_i > u)$, we can estimate a high nonstationary quantile $\tilde{u}_i$ and then use the threshold $u = 0$ for $X_i - \tilde{u}_i$.

Wrap-up: Approaches for time series extremes

- **Exploration and dependence summaries:**
  - **Extremal index**: a key summary parameter that can be estimated and interpreted
  - **Extremogram:**
    - An analogue of the classical autocorrelation function based on the χ-measure to explore the strength of temporal dependence among extremes at fixed lags $h = 1, 2, \ldots$
    - Variants using $\overline{\chi}$ to assess asymptotic independence have also been proposed (called $\overline{\rho}$ in the `extRemes` package and Λ_τ in the `POT` package).
- **Marginal estimation:**
  - With the block-maximum approach, we can proceed as in the i.i.d. case, if the dependence between blocks is negligible.
  - With the POT-approach using the GPD, we can also proceed similar to the i.i.d. case if we first "decluster" the threshold exceedances.
- **Joint marginal and dependence modeling:**
  - More complex in general, only few available models.
  - Markov chain assumption provides useful models.
  - Any multivariate extremes models (such as conditional extremes) could be used to model extremes arising for random vectors composed of d consecutive time steps.
- **Assumption of asymptotic stability:** In case of serial asymptotic independence, clusters could ultimately vanish at very high thresholds levels u, but this is hard to detect in practice if there are too few observed exceedances for such levels.