Two Different Problems
Practical Issues
EVT for Stationary Sequences
Data Analysis
Concluding Remarks

Extremes of Dependent Sequences

Debbie J. Dupuis

Department of Management Sciences

HEC Montréal

Two Different Problems
Practical Issues
EVT for Stationary Sequences
Data Analysis
Concluding Remarks

Outline

- Two Different Problems : Examples
- Practical Issues
- Extreme Value Theory for Stationary Sequences
- Data Analysis
- Concluding Remarks

Two Different Problems
Practical Issues
EVT for Stationary Sequences
Data Analysis
Concluding Remarks

Two problems

Our concern could be

1. the largest value, i.e. estimating the marginal tail ; or
2. the aggregate effect of extreme observations occurring one after the other, i.e. estimating the structure of clusters of extreme values.

Two Different Problems
Practical Issues
EVT for Stationary Sequences
Data Analysis
Concluding Remarks

Examples

Analyzing daily temperature :

1. one very hot day may cause much discomfort (marginal tail) ;
2. a heat wave is unforgiving for societies and infrastructures unable to cope or adapt (cluster of extreme values).

Analyzing daily rainfall :

1. river flooding may be caused by one extreme rainfall event (marginal tail) ;
2. river flooding may be caused by the ground already being saturated with water due to high precipitation during several days (cluster of extreme values).

Two Different Problems
Practical Issues
EVT for Stationary Sequences
Data Analysis
Concluding Remarks

Marginal Analysis

Essentially three different approaches :

(1) identify independent clusters of exceedances and construct a new data set by taking the maximum from each cluster. One obtains an iid sample whose tail behavior can be analyzed using standard techniques from classical EVT for iid data.
(2) apply classical tail estimators (for iid samples) directly to all exceedances observed in the time series. However, to construct CIs, one needs results on their asymptotic behavior that hold true under mild assumptions on the serial dependence structure.
(3) fit a parametric model of the serial dependence to the data and infer the tail behavior of the time series from a suitable analysis of the residuals.

Two Different Problems
Practical Issues
EVT for Stationary Sequences
Data Analysis
Concluding Remarks

Modeling of Clusters of Extreme Values

(1) Markov chain modeling of clusters of extreme values [e.g., Smith, Coles and Tawn (1997), Bortot and Tawn (1998), Sisson and Coles (2003)]

(2) Some results on empirical processes of cluster functionals [Drees and Rootzén (2010)]

Two Different Problems
Practical Issues
EVT for Stationary Sequences
Data Analysis
Concluding Remarks

Marginal Analysis - Some Difficulties

(1) (identify independent clusters of exceedances, take cluster maximum, and use classical EVT on iid sample) :

- (i) cluster maximum is not the quantity of interest. Eg. Laurini and Tawn (2009) and estimation of Value-at-Risk ;
- (ii) clusters may be hard to identify

(2) (apply classical tail estimators directly to all exceedances) : need asymptotic behavior under the assumed serial dependence structure, there are very few results, eg. Drees (2000, 2002, 2003)

(3) (fit parametric model of serial dependence to data and analyze iid residuals) : can give completely misleading estimates even if the deviation from the assumed linear time series model is moderate, eg. Drees (2008)

Two Different Problems
Practical Issues
EVT for Stationary Sequences
Data Analysis
Concluding Remarks

Definition

A sequence of rv $X_1, X_2, \ldots$ is strictly stationary if its finite-dimensional distributions are invariant under shifts of time, i.e.

$$(X_{t_1}, \ldots, X_{t_m}) \stackrel{d}{=} (X_{t_1+h}, \ldots, X_{t_m+h})$$

for any choice of indices $t_1 < \ldots < t_m$ and integers h. $\{$ but X_{t_1} need not be independent of X_{t_1+1} or $X_{t_1+h}, \ldots\}$

$\Rightarrow$ It is impossible to build a general extreme value theory for the class of all stationary sequences. E.g.

1. Assume $X_n = X \ \forall n$. This relation defines a stationary sequence and

$$\mathrm{P}(M_n \leq x) = \mathrm{P}(X \leq x) = F(x) \quad x \in \mathbb{R}$$

   $\Rightarrow$ the distribution of the sample maxima can be any distribution F. (not reasonable basis for a general theory)

2. X_n an iid sequence

$$\mathrm{P}\left(\frac{M_n - b_n}{a_n} \leq z\right) \approx G(z)$$

   where G is GEV if non-degenerate.

We need some condition in between, must be reasonable for applications yet mathematically tractable so as to get the form of the limiting distribution.

Condition $D(u_n)$

[Leadbetter, Lindgren and Rootzén (1983), Leadbetter and Rootzén (1988)]

For any integers p, q, and n

$$1 \leq i_1 < \ldots i_p < j_1 < \ldots < j_q \leq n$$

such that $j_1 - i_p \geq l$ we have

$$\left| \mathrm{P}\left(\max_{i \in A_1 \cup A_2} X_i \leq u_n \right) - \mathrm{P}\left(\max_{i \in A_1} X_i \leq u_n \right) \mathrm{P}\left(\max_{i \in A_2} X_i \leq u_n \right) \right| \leq \alpha(n, l)$$

where $A_1 = \{i_1, \ldots, i_p\}$, $A_2 = \{j_1, \ldots, j_q\}$ and $\alpha(n, l) \to 0$ as $n \to \infty$ for some $l = l_n = o(n)$.

NOTE :

- For sequences of independent rv, the difference $\mathrm{P}() - \mathrm{P}()\mathrm{P}()$ is exactly 0 for any sequence u_n.
- Condition $D(u_n)$ is a distributional mixing condition, weaker than most of the classical forms of dependence restrictions.
- For Gaussian sequences with autocorrelation ρ_n at lag n, the $D(u_n)$ is satisfied as soon as $\rho_n \log n \to 0$ as $n \to \infty$. This is much weaker than the geometric decay assumed by autoregressive models.
- Condition $D(u_n)$ ensures that, for sets of rv far enough apart, $\mathrm{P}() - \mathrm{P}()\mathrm{P}()$ is sufficiently close to zero to have no effect on the limit laws for extremes.

Theorem 1

[Leadbetter (1974)]
Let $X_1, X_2, \ldots$ be a stationary process and $M_n = \max\{X_1, \ldots, X_n\}$. Then, if $\{a_n > 0\}$ and $\{b_n\}$ are sequences of constants such that

$$\lim_{n\to\infty} \mathrm{P}\left(\frac{M_n - b_n}{a_n} \leq z\right) = G(z)$$

where G is a non-degenerate df, and the $D(u_n)$ condition is satisfied with $u_n = a_n z + b_n \; \forall z \in \mathbb{R}$, then G is GEV.

$\Rightarrow$ Provided a series has limited long-range dependence at extreme levels (in the sense that the $D(u_n)$ condition makes sense), maxima of stationary series follow the same distributional limit laws as those of independent series.

[Note that we are not saying that the parameters of the GEV are the same as those of the corresponding independent sequence.]

Two Different Problems
Practical Issues
EVT for Stationary Sequences
Data Analysis
Concluding Remarks

Theorem 2

[Leadbetter (1983)]
Let $X_1, X_2, \ldots$ be a stationary process and $X_1^*, X_2^*, \ldots$ be an associated sequence. If

$$\mathrm{P}\left(\frac{M_n^* - b_n}{a_n} \leq z\right) \to G_1(z)$$

as $n \to \infty$ for normalizing sequences $\{a_n > 0\}$ and $\{b_n\}$, where G_1 is a non-degenerate df, if $D(u_n)$ holds with $u_n = a_n x + b_n$ for each x such that $G(x) > 0$ and if $\mathrm{P}\left(\frac{M_n - b_n}{a_n} \leq x\right)$ converges for some x, then

$$\mathrm{P}\left(\frac{M_n - b_n}{a_n} \leq z\right) \to G_2(z)$$

where $G_2(z) = G_1^{\theta}(z)$ for a constant θ such that $0 < \theta \leq 1$.

BASIC IDEA :

1. If maxima of a stationary series converge, then the limit distribution is related to the limiting distribution of an independent series according to $G_2(z) = G_1^\theta(z)$.
2. Recall that maxima of a stationary series will converge provided an appropriate $D(u_n)$ condition is satisfied.
3. Recall that $G_1(z)$ has to be GEV, so $G_1^\theta(z)$ is also GEV. In fact,

$$\begin{aligned} G_1^\theta(z) &= \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}^\theta \\ &= \exp\left\{-\left[\theta^{-\xi}+\theta^{-\xi}\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} \\ &= \exp\left\{-\left[1+\xi\left(\frac{z-\mu^*}{\sigma^*}\right)\right]^{-1/\xi}\right\} \end{aligned}$$

where $\mu^* = \mu - \frac{\sigma}{\xi}(1-\theta^\xi)$, $\sigma^* = \sigma\theta^\xi$. Note that the shape parameter ξ remains the same.

The quantity θ as defined by $G_2(z) = G_1^\theta(z)$ is the extremal index.

Two Different Problems
Practical Issues
EVT for Stationary Sequences
Data Analysis
Concluding Remarks

Alternatively, without making use of the associated series, we have the following.

Definition

Let $X_1, X_2, \ldots$ be a strictly stationary process with marginal df F and θ a non-negative number. Assume that for every $\tau > 0$ there exists a sequence $\{u_n\}$ such that

$$\begin{aligned} \lim_{n\to\infty} n(1 - F(u_n)) &= \tau, \\ \lim_{n\to\infty} \mathrm{P}(M_n \le u_n) &= \exp(-\theta\tau). \end{aligned}$$

Then θ is called the extremal index of the sequence $\{X_n\}$.

NOTE :

- We essentially have
$\mathrm{P}(\max(X_1, X_2, \ldots, X_n) \leq x) \approx F^{\theta n}(x)$
- θ is the limiting reciprocal mean cluster size :
$\theta =$ (limiting mean cluster size)$^{-1}$ (limiting in the sense of clusters of exceedances of increasingly high thresholds).
- Independent series $\Rightarrow \theta = 1$
Stationary series with $\theta = 1 \not\Rightarrow$ independent
A series for which $\theta = 1$ means that dependence is negligible at asymptotically high levels, but not necessarily so at "extreme" levels that are relevant for any particular application.

Relevance of θ

Consider a simple example from Weissman (1994) :

Assume a dyke has to be built at the seashore to protect against floods with 95% certainty for the next 100 years.

Suppose that it has been established that the 99.9% and 99.95% quantiles of the annual wave-height are

$$z_{0.001} = 10 \ m \quad \text{and} \quad z_{0.0005} = 11 \ m.$$

If the annual maxima are iid, then the dyke should be 11 m high since $0.9995^{100} \approx 0.95$

If the annual maxima are stationary with extremal index $\theta = 0.5$, then a height of 10 m is sufficient since $0.999^{50} \approx 0.95$.

Block Maxima

How are these asymptotic results used to draw inferences on stationary series? Easy.

If data ≈ a realization of a process satisfying $D(u_n)$ condition, it is still appropriate to model block maxima using the GEV family (as in iid case).

The parameters themselves are different than those that would have been obtained had the series been iid, but since they are estimated from the data, it doesn't matter.

There is a price to pay however.

- The accuracy of the GEV family as an approximation to the distribution of block maxima is likely to diminish with increased level of dependence.
- Effective number of observations is reduced from n to $n\theta$.

Two Different Problems
Practical Issues
EVT for Stationary Sequences
Data Analysis
Concluding Remarks

Threshold Models

Modifications are required

For stationary series, asymptotic arguments imply that the marginal distribution of excesses of a high threshold is GP.

But extremes have a tendency to cluster and we do not have a result for the joint distribution of neighboring excesses.

The most widely-adopted approach is **declustering**.

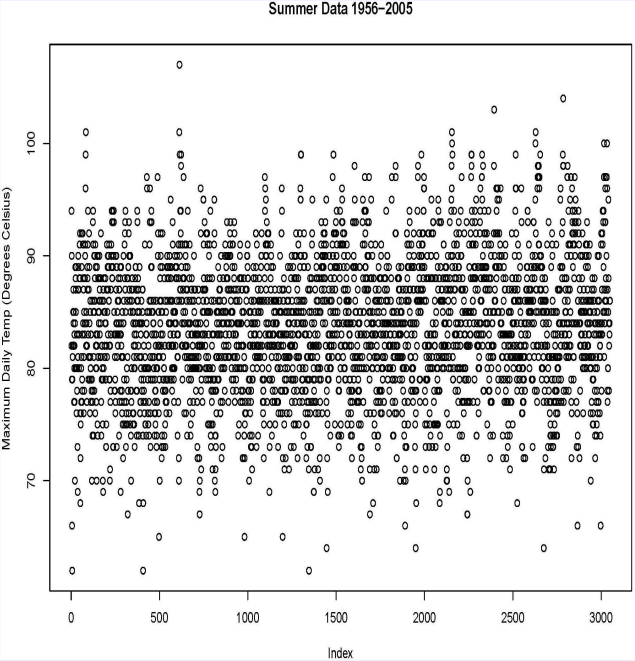
Two Different Problems
Practical Issues
EVT for Stationary Sequences
Data Analysis
Concluding Remarks

Consider the maximum daily temperature in New York City.

Data are from January 1, 1956 to December 31, 2005.

Consider only July and August, hoping for a stationary series.

The following plots show the impact of the threshold selection and the cluster definition.


Summer Data 1956–2005
Maximum Daily Temp (Degrees Celsius)
Index
70
80
90
100
0
500
1000
1500
2000
2500
3000

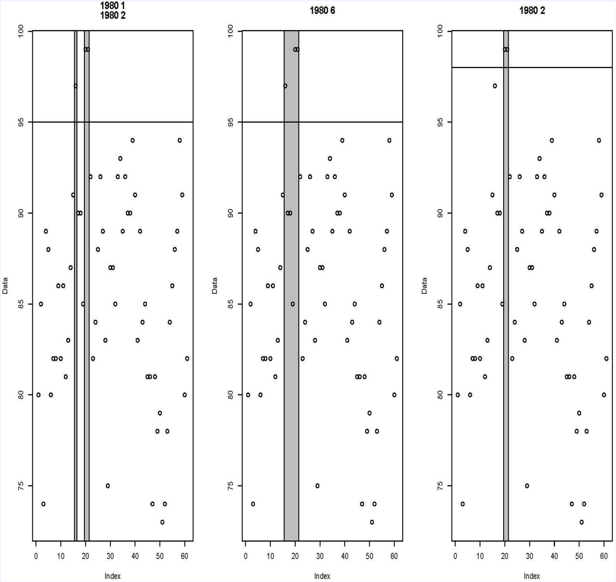


Fig.: Maximum temperatures for New York for summer 1980 ; run length = 3/4/3 ; threshold = 95%/95%/98%.

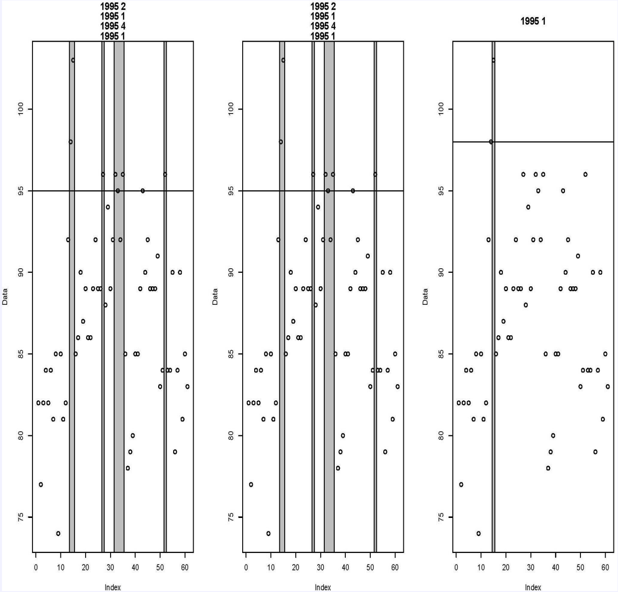


Fig.: Maximum temperatures for New York for summer 1995 ; run length = 3/4/3 ; threshold = 95%/95%/98%.

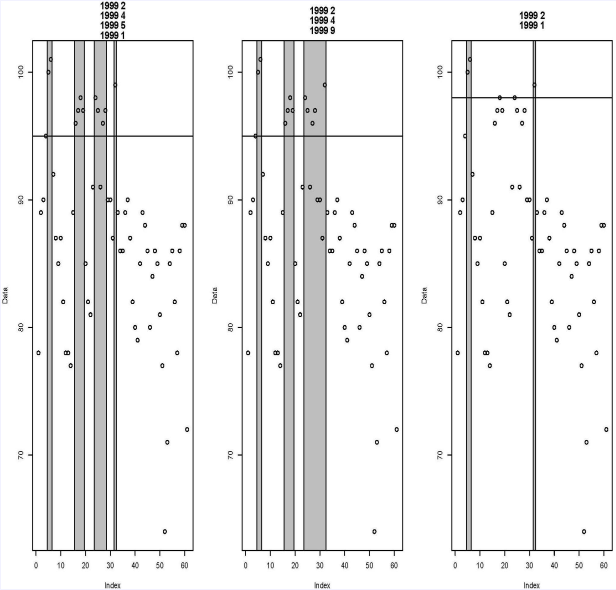


Fig.: Maximum temperatures for New York for summer 1999 ; run length = 3/4/3 ; threshold = 95%/95%/98%.

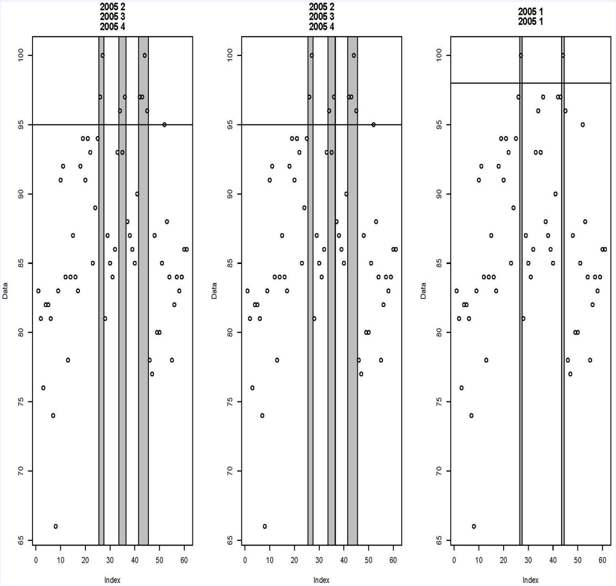


Fig.: Maximum temperatures for New York for summer 2005 ; run length = 3/4/3 ; threshold = 95%/95%/98%.

Estimating θ

1. Blocks Method
2. Extremal Index as Reciprocal of the Mean Cluster Size
3. Runs Method
4. Two-thresholds Method
5. Intervals Estimator

We can find estimates of θ based on the blocks method, the reciprocal of mean cluster size interpretation, the runs method and the intervals estimator using the `fExtremes` package in R.

Title :
Extremal Index from Block Method

Call :
blockTheta(x = as.timeSeries(x), block = 3, quantiles = qthres)

Extremal Index :

	quantiles	thresholds	N	K	theta
1	0.870	90	388	287	0.812
2	0.883	91	294	226	0.826
3	0.897	91	294	226	0.826
4	0.911	92	218	174	0.843
5	0.925	92	218	174	0.843
6	0.939	93	164	136	0.866
7	0.953	94	117	102	0.900
8	0.967	95	91	74	0.831
9	0.981	97	39	39	1.013
10	0.995	99	11	15	1.371

Two Different Problems
Practical Issues
EVT for Stationary Sequences
Data Analysis
Concluding Remarks

Title :
Extremal Index from Reciprocal Cluster Method

Call :
clusterTheta(x = as.timeSeries(x), block = 3, quantiles = qthres)

Extremal Index :

	quantiles	thresholds	N	K	theta
1	0.870	90	388	287	0.739
2	0.883	91	294	226	0.768
3	0.897	91	294	226	0.768
4	0.911	92	218	174	0.798
5	0.925	92	218	174	0.798
6	0.939	93	164	136	0.829
7	0.953	94	117	102	0.871
8	0.967	95	91	74	0.813
9	0.981	97	39	39	1.000
10	0.995	99	11	15	1.363

Title :
Extremal Index from Run Method

Call :
runTheta(x = as.timeSeries(x), block = 3, quantiles = qthres)

Extremal Index :

	quantiles	thresholds	N	theta
1	0.870	90	388	0.569
2	0.883	91	294	0.619
3	0.897	91	294	0.619
4	0.911	92	218	0.674
5	0.925	92	218	0.674
6	0.939	93	164	0.719
7	0.953	94	117	0.752
8	0.967	95	91	0.758
9	0.981	97	39	0.820
10	0.995	99	11	0.909

Title :
Extremal Index from Ferro-Segers Method

Call :
ferrosegersTheta(x = as.timeSeries(x), quantiles = qthres)

Extremal Index :

	Threshold	Quantiles	RunLength	Clusters	theta
1	0.870	90	17	53	0.484
2	0.883	91	17	54	0.465
3	0.897	91	17	54	0.465
4	0.911	92	16	52	0.460
5	0.925	92	16	52	0.460
6	0.939	93	15	49	0.429
7	0.953	94	9	46	0.386
8	0.967	95	10	36	0.397
9	0.981	97	12	18	0.454
10	0.995	99	17	7	0.609

Two Different Problems
Practical Issues
EVT for Stationary Sequences
Data Analysis
Concluding Remarks

and choosing an estimate of θ is not easy...

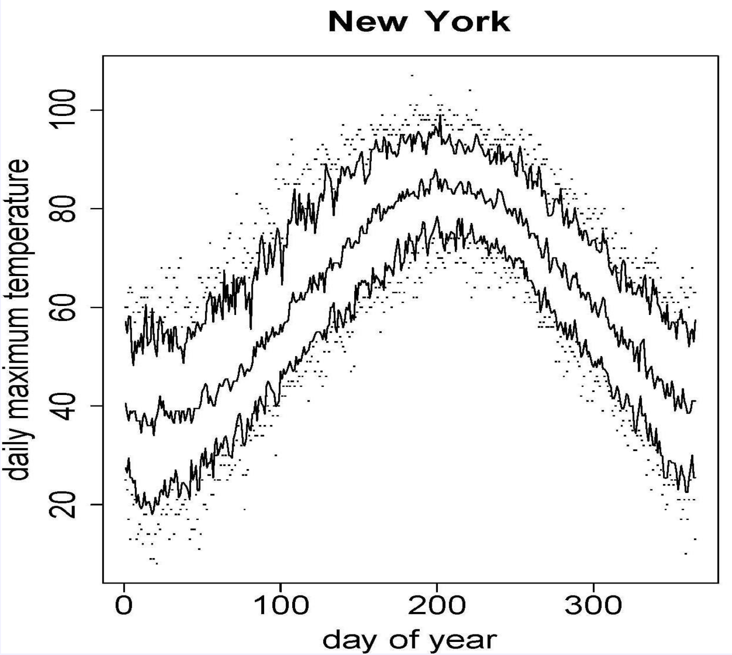
	thres	Blocks	Recip	Runs'	Int's
1	90	0.812	0.739	0.569	0.484
2	91	0.826	0.768	0.619	0.465
3	91	0.826	0.768	0.619	0.465
4	92	0.843	0.798	0.674	0.460
5	92	0.843	0.798	0.674	0.460
6	93	0.866	0.829	0.719	0.429
7	94	0.900	0.871	0.752	0.386
8	95	0.831	0.813	0.758	0.397
9	97	1.013	1.000	0.820	0.454
10	99	1.371	1.363	0.909	0.609

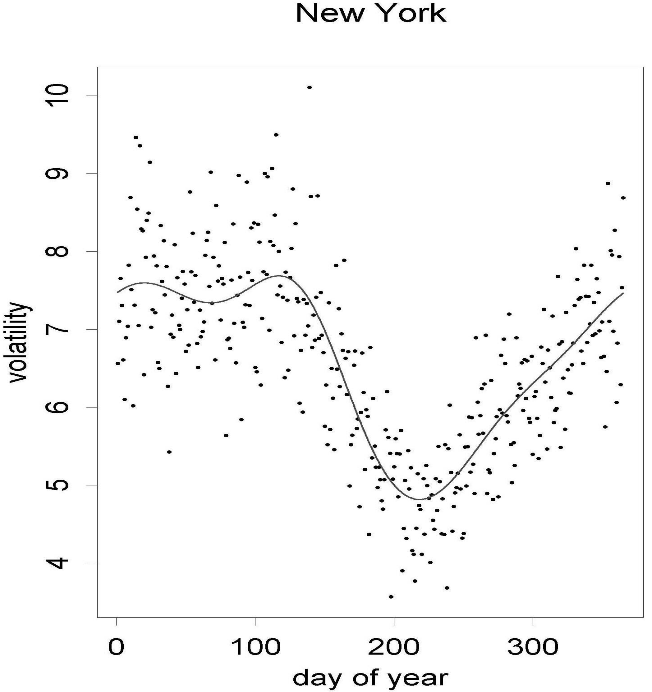
(Can get bootstrap CIs for intervals estimator of θ using the `extRemes` package in R.)

Two Different Problems
Practical Issues
EVT for Stationary Sequences
Data Analysis
Concluding Remarks

Extremal Dependence or Non-stationarity ?

What is the impact of non-stationarity on the estimated extremal index ?


New York
daily maximum temperature
20
40
60
80
100
0
100
200
300
day of year


New York
volatility
day of year
10
9
8
7
6
5
4
0
100
200
300

Tab.: Estimated GPD parameters. ξ_1 (ξ_2) is the estimated shape parameter for the period before (after) the change-year η. Estimated extremal index θ (using intervals estimator) at threshold u_z is in last column.

Season	u_z	β	ξ_1	ξ_2	θ
1	2.03	0.60 (0.10)	-0.38 (0.17)	-0.65 (0.13)	0.80
2	2.35	0.49 (0.08)	-0.49 (0.14)	0.05 (0.17)	0.92
3	1.99	0.37 (0.03)	-0.11 (0.07)	-0.04 (0.10)	0.98

Season 1 : Day 1 to Day 52 (change-year 1975)
Season 2 : Day 53 to Day 125 (change-year 1975)
Season 3 : Day 126 to Day 365 (change-year 1991)

Two Different Problems
Practical Issues
EVT for Stationary Sequences
Data Analysis
Concluding Remarks

Concluding Remarks

- can relax assumption of iid to stationary sequences
- extremal properties of special classes of stationary sequences have been studied :
  - Markov chains - O'Brien (1987), Smith (1992), Perfekt (1994)
  - moving average processes - Rootzén (1986)
  - ARCH process - de Haan et al (1989)
- marginal analysis vs cluster analysis
- stationary vs non-stationary
  - meteorological data typically have a strong seasonal component
  - macro-economic data often show an upward or downward trend

References

Drees, H. (2000). Weighted approximations of tail processes for β-mixing random variables. *Ann. Appl. Probab.*, 10, 1274–1301.

Drees, H. (2002). Tail empirical processes under mixing conditions. In *Empirical Process Techniques or Dependent Data (H. G. Dehling, T. Mikosch and M. Sorensen, eds.)*, Birkhäuser, Boston, 325–342.

Drees, H. (2003). Extreme quantile estimation for dependent data with applications to finance. *Bernoulli*, 9 617–657.

Drees, H. (2008). Some aspects of extreme value statistics under serial dependence. *Extremes*, 11, 35–53.

Drees, H. and H. Rootzén (2010). Limit theorems for empirical processes of cluster functionals. *Ann. Stat.*, 38, 2145–2186.

Dupuis, D.J. (2011). Modeling waves of extreme temperature : The changing tails of four cities. *JASA*, to appear.

Ferro, C.A.T. and J. Segers (2003). Inference for clusters of extreme values. *J.R.S.S. B*, 65, 545–556.

Laurini, F. and J.A. Tawn (2003). newblock New estimators for the extremal index and other cluster characteristics. *Extremes*, 6, 189–211.

Leadbetter, M.R. (1974). On extreme values in stationary sequences. *Z. Wahrsch. Verw. Gebiete*, 28, 289–303.

Leadbetter, M.R. (1983). Extreme and local dependence in stationary sequences. *Z. Wahrsch. Verw. Gebiete*, 65, 291–306.

Leadbetter, M.R., G. Lindgren and H. Rootzén (1983). *Extremes and Related Properties of Random Sequences and Processes.* New York : Springer-Verlag.

Leadbetter, M.R., and H. Rootzén (1988). Extremal theory for stochastic processes. *Ann. Probab.*, 16, 431–478.

O'Brien, G.L. (1987). Extreme values for stationary and Markov sequences. *Ann. Probab.*, 15, 281-291.

Perfekt, R. (1994). Extremal Behaviour of Stationary Markov Chains with Applications. *Ann. Appl. Probab.*, 4, 529–548.

References (continued...)

Rootzén, H. (1986). Extreme Value Theory for Moving Average Processes. *Ann. Probab.*, 14, 612–652.

Smith, R.L. (1992). The Extremal Index for a Markov Chain. *J. Appl. Probab.*, 29, 37–45.

Smith, R.L., J.A. Tawn and S.G. Coles (1997). Markov Chain Models for Threshold Exceedances. *Biometrika*, 84, 249–268.

Weissman, I. (1994). On the extremal index of stationary sequences. Preprint, Technion, Israel.