Application of Extreme Value Theory to Tennis Serve Record Estimation

Blue Yonder Meetup
Ivan Marevic, 26.10.2023

Agenda

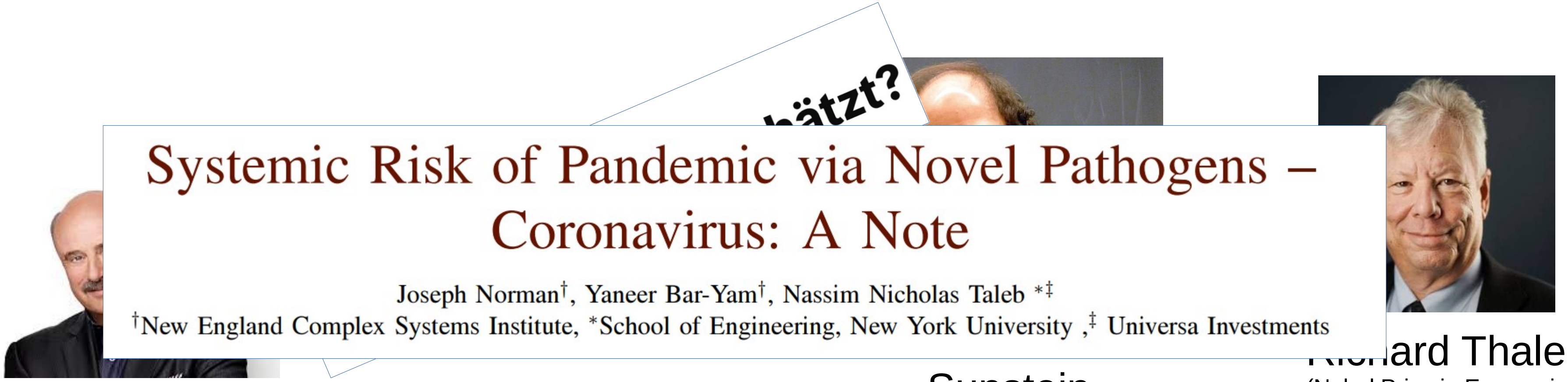
1. Introduction
2. What is Extreme Value Theory (EVT)?
3. EVT Basics
4. Applications of EVT and Tennis Serve Record Estimation
5. Summary & Next Steps

Introduction

Introduction

How did I get interested in Extreme Value Theory (EVT)?

1. During COVID-19 → plenty of time
2. Downplay of COVID-19 by “experts”



Systemic Risk of Pandemic via Novel Pathogens – Coronavirus: A Note

Joseph Norman[†], Yaneer Bar-Yam[†], Nassim Nicholas Taleb [*‡]
[†]New England Complex Systems Institute, [*]School of Engineering, New York University ,[‡] Universa Investments

Richard Thaler
(Nobel Price in Economics)

Sunstein

Dr. Phil

“More people drown in a swimming pool than are killed by Covid”

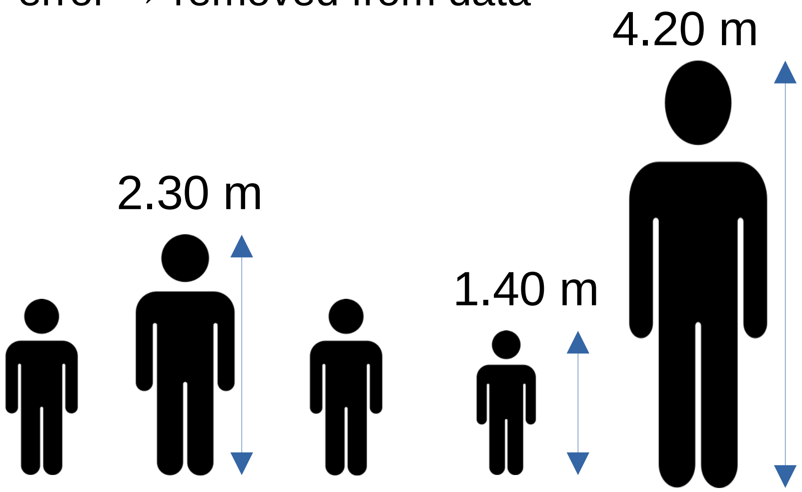
“Most people in North America and Europe do not need to worry much about the risk of contracting the disease. That's true even for people who are traveling to nations such as Italy that have seen outbreaks of the disease.”

What is Extreme Value Theory?

What is Extreme Value Theory?

Outliers

- Data point *out of* the natural range of variation
- Might be caused by measurement error → removed from data



Extremes

- Data point *within* natural range of variation
- Characterized by unusual magnitude or temporal scarcity

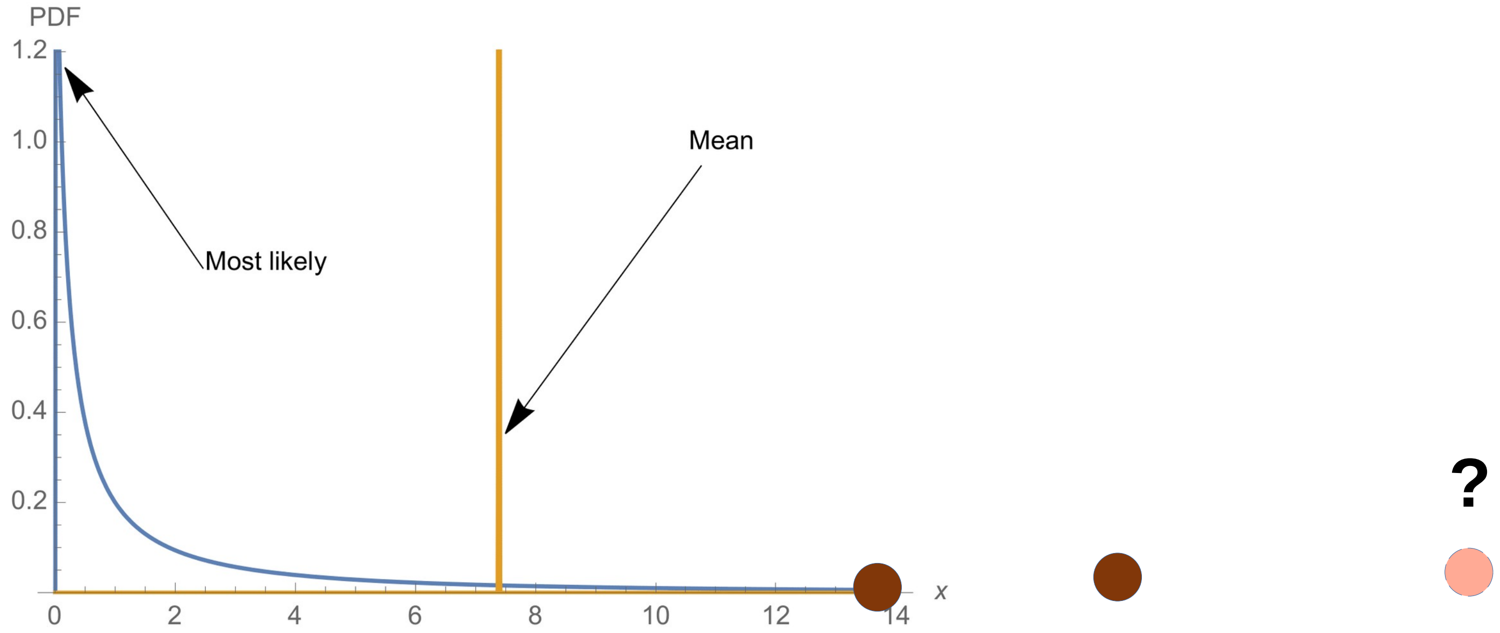
→ gain in stock option of 40%
→ gain in stock option of 800%

What is Extreme Value Theory?

What are we trying to do with EVT?

- Check if the phenomenon we are investigating is fat tailed or not?
- Find the endpoint of a distribution (if it exists)

endpoint **x* =** ●



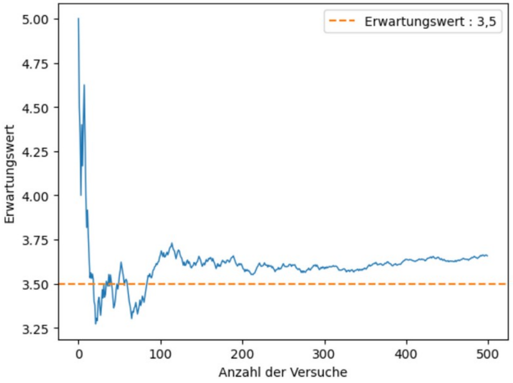
EVT Basics

EVT Basics

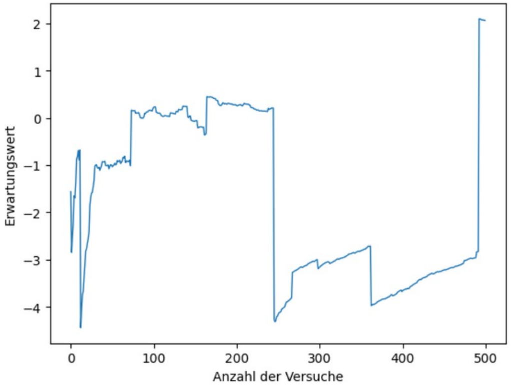
Why do we need EVT?

- Assumption of normality does not hold in many cases
- Future extremes are not in the current data
- Many statistical concepts/operations fail to work on fat-tailed distributions

Die roll E(X):



Cauchy E(X):



EVT Basics

- Given X_1, X_2, …, X_n iid random variables, the central limit theorem is concerned with with the limit behavior of partial sums $X_1 + X_2 + \ldots + X_n$ as $n \to \infty$
- EVT is concerned with $\max(X_1, X_2, \ldots, X_n)$ or $\min(X_1, X_2, \ldots, X_n)$ as $n \to \infty$

Example scenarios:

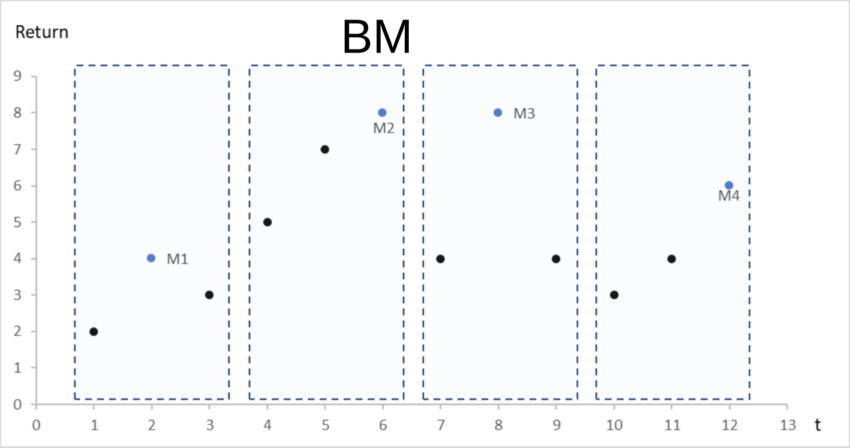
- Tires of a car wear out over time and over a long time the accumulated decay results in failure → partial sum exceeds a threshold (CLT)
- Tire blows up when hitting the sidewalk → partial maxima exceed threshold (EVT)

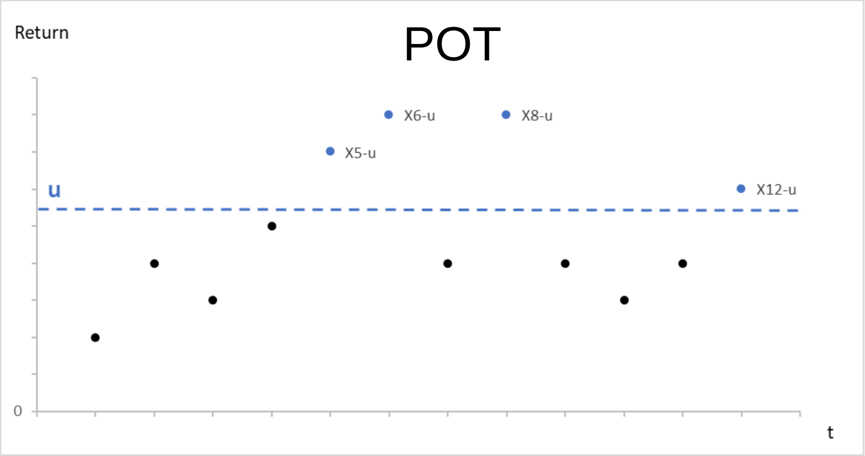
EVT Basics

How do we get the maxima?

- For non time series data define maximum per unit of interest
  - Max. age per person
  - Max. tennis serve speed per player in a dataset

- For time series data we can either use *the block maxima (BM)* or *peaks over threshold (POT)* methods





EVT Basics

Extreme Value Distribution

Let *F* be the underlying distribution function and x* its right endpoint.

$$x^* := sup(x : F(x) < 1)$$

$$P(max(X_1, ..., X_n \leq x) = P(X_1 \leq x, X_2 \leq x, ..., X_n \leq x)) = F^n(x)$$

For $x < x^*$ this converges to 0 and to 1 for $x \geq x^*$

Thus, some normalization is necessary:

$$\boxed{\frac{max(X_1, X_2, ..., X_n) - b_n}{a_n}} \longrightarrow \boxed{\lim_{x \to \infty} F^n(a_n x + b_n) = G(x)}$$

$a_n > 0$ and b_n real $(n = 1, 2, ...)$

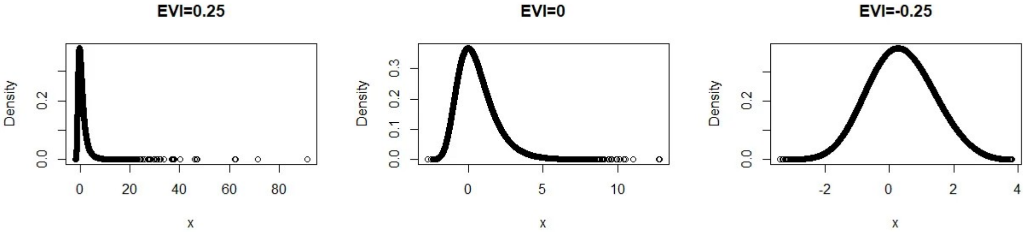
EVT Basics

Subclasses of the Extreme Value Distribution (G)

Parametrized general form by introducing the extreme value index ξ:

$$G_{\xi}(x) = exp(-(1 + \xi x)^{-1/\xi})$$

ξ case	consequence	distribution class
ξ > 0	the right endpoint of the distribution is infinity → heavy right tail (major moments do not exist)	Frechet
ξ = 0	the right endpoint of the distribution equals infinity → light tailed (moments exist)	Gumbel
ξ < 0	The right endpoint of the distribution exists → short tail	Reverse-Weibull



EVT Basics

Estimating the endpoint x*

As a reminder we need to estimate $x^* := sup(x : F(x) < 1)$

It can be shown that $\lim_{x\to\infty} F^n(a_n x + b_n) = G(x)$ ≡ $\frac{x^\xi - 1}{\xi}$ with x > 0

For ξ < 0 this yields the following estimator for j = 1, 2:

$$\boxed{\hat{x^*} = \hat{b} - \frac{\hat{a}_j}{\hat{\xi}_j}}$$

Applications of EVT & Tennis Serve Record Estimation

Applications of EVT

Overview of application areas of EVT

Dike height

Pandemics

Wars

Skyscraper height

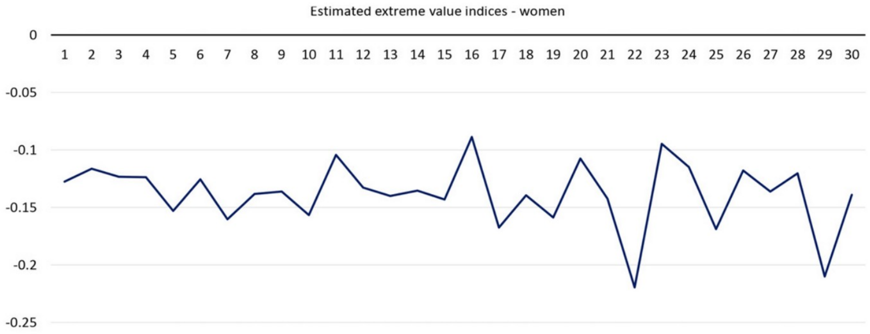
Athletic records

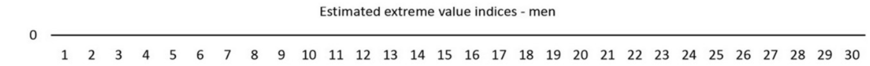
Life span

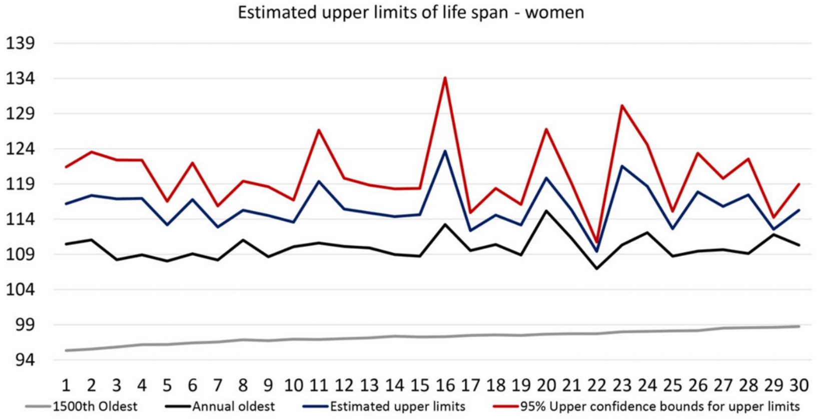
Applications of EVT

Human Lifespan Endpoint Estimation (Einmahl et al. 2019)

- All analyses done for each year in a 30 year span
- Estimate EVI to check if endpoint is finite
- Estimate the endpoint and derive 95% CI for the upper bound







Tennis Serve Record Estimation

Background

- the current officially recorded tennis serve records:
  - Women: 210.8 km/h (Sabine Lisicki)
  - Men: 253 km/h (John Isner)
- These records were recorded at events on the official major circuits (WTA and ATP)
- Unofficial records
  - Women: 220 km/h (Georgina Garcia Perez)
  - Men: 263 km/h (Sam Groth)

?

Tennis Serve Record Estimation

Research Questions

1. What is the fastest tennis serve possible for women and men given today's technology and equipment?

2. What is the quality of the estimated record (e.g. estimate of expected number of exceedances of the estimated record)?

3. How plausible are the unofficial serve records given the endpoint estimates?

Tennis Serve Record Estimation

The Data

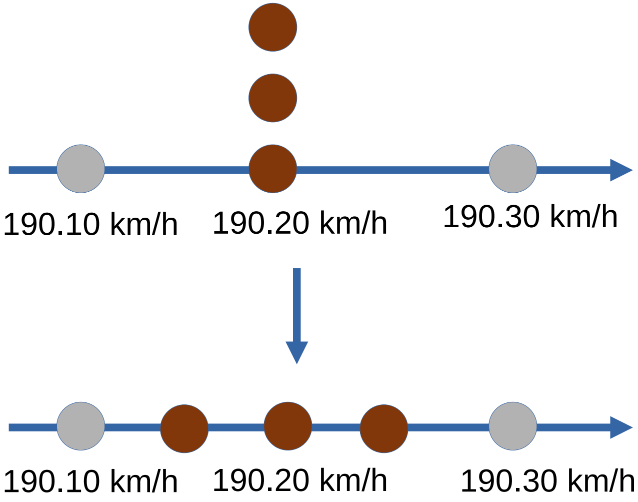
- point-by-point data from the 4 major tournaments since 2011[1].
- inclusion of official recordings of serve speeds that were not in this dataset (incl. world record)
- extraction of serve speed maximum for each player and correction of misspellings
  - every player occurred once in the final dataset
  - Multiple identical entries where smoothed as the estimation procedure is sensitive to clusters:

$$s_j = (d_j - .001) + .01\frac{2j-1}{2m}, \quad j = 1, ..., m.$$

- d=serve speed
- m=number of players in cluster

Examples of misspellings:

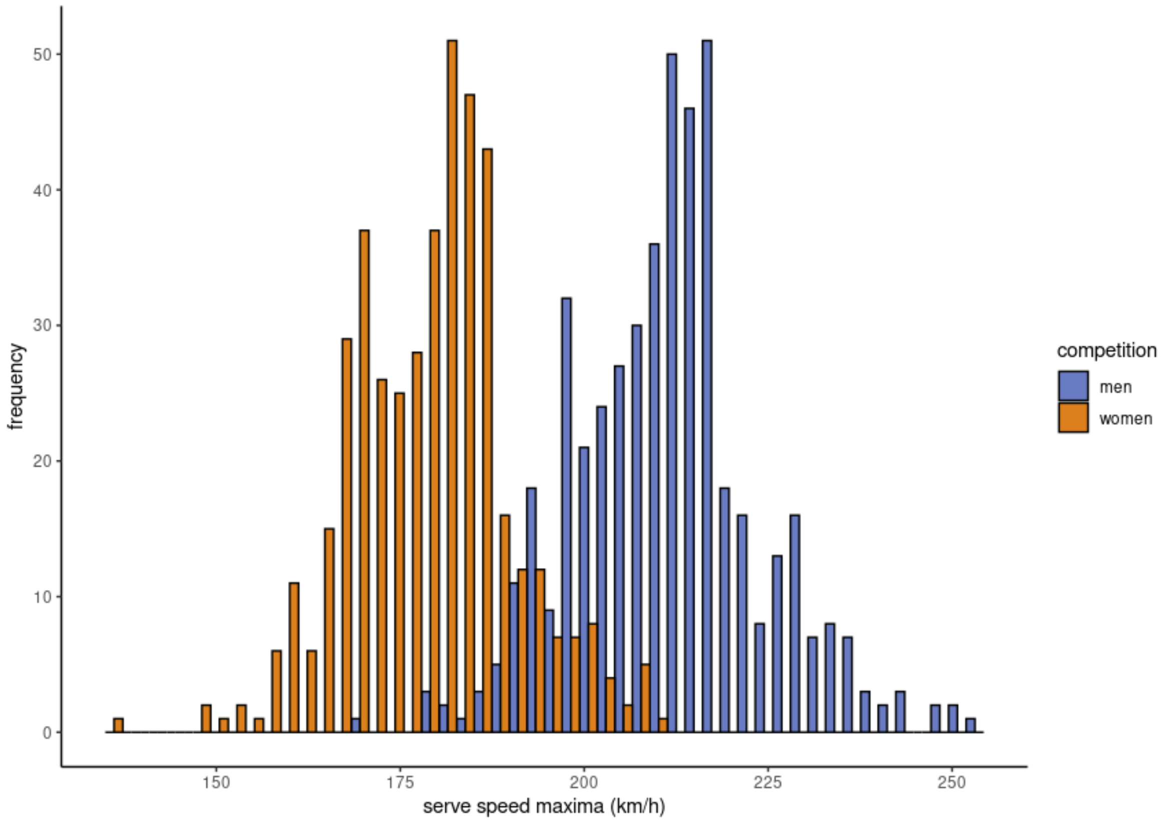
- DiMitrov vs. Dimitrov
- Ramos-Vinolas vs. Ramos Vinolas



[1] https://github.com/JeffSackmann/tennis_slam_pointbypoint

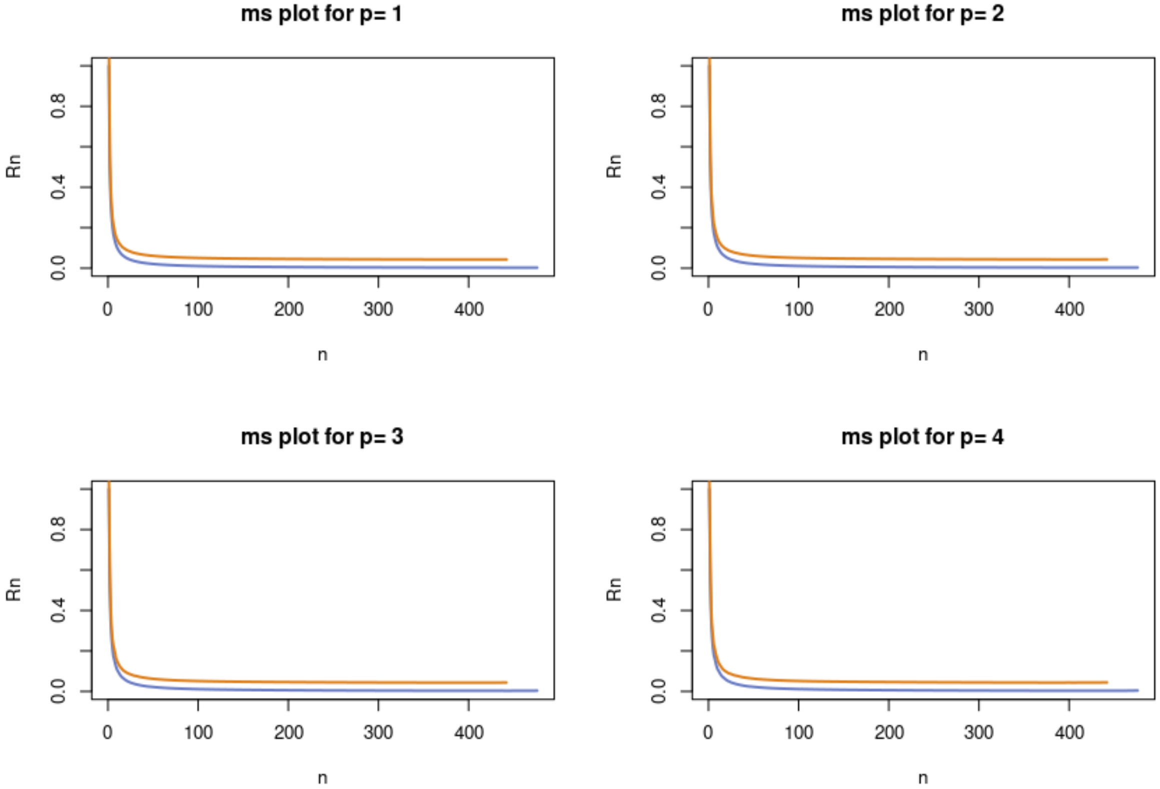
Tennis Serve Record Estimation

Serve speed distributions for men and women



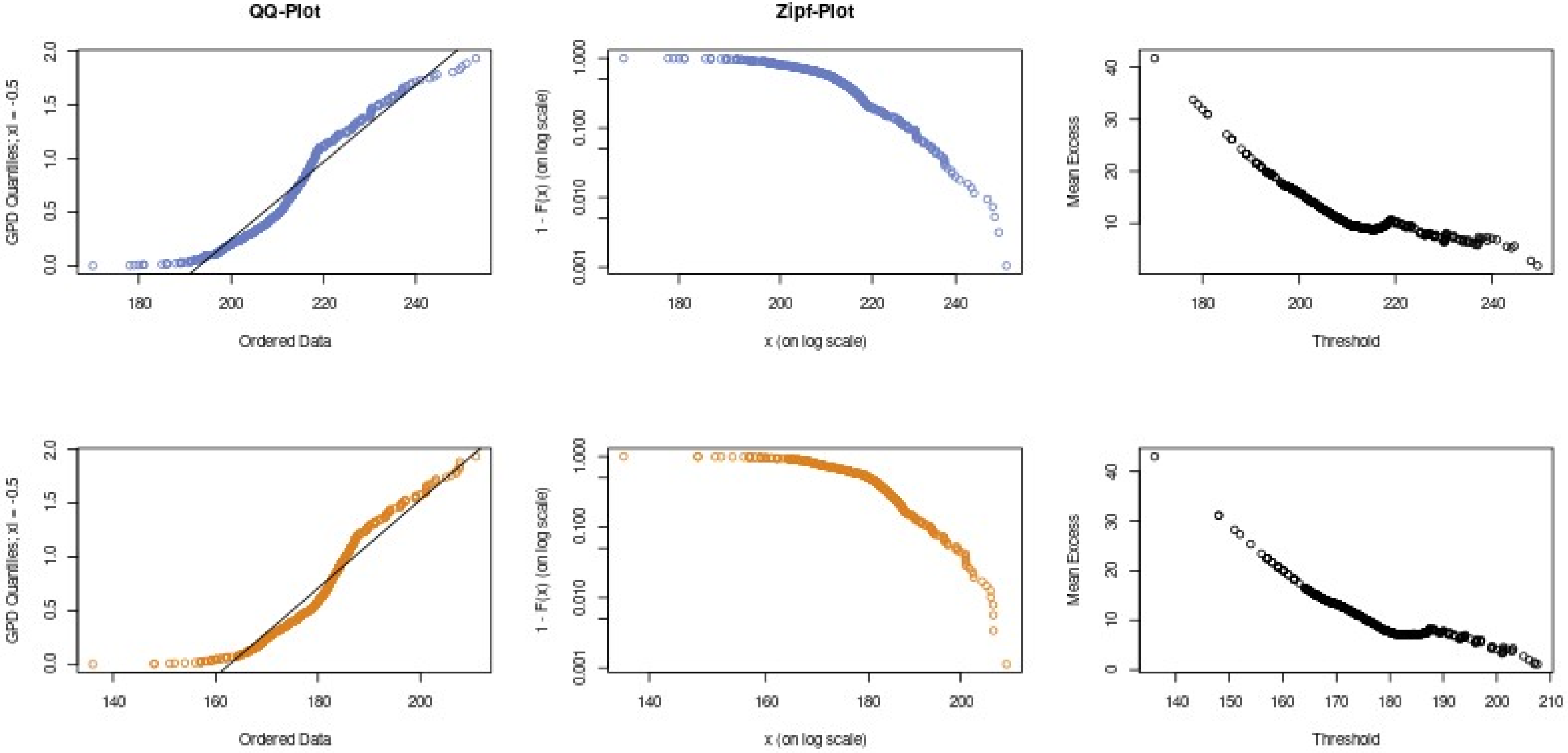
Tennis Serve Record Estimation

Test for normality: MS Plot



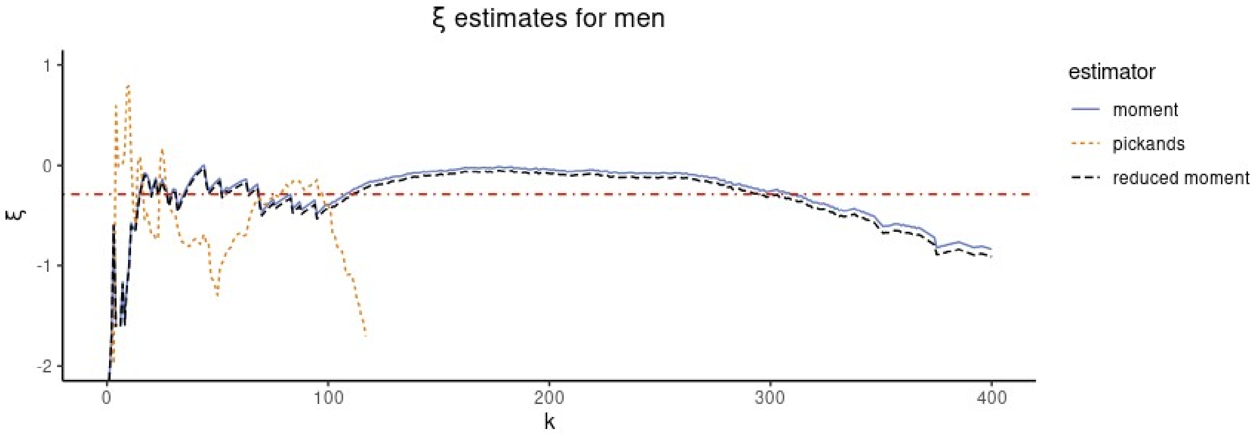
Tennis Serve Record Estimation

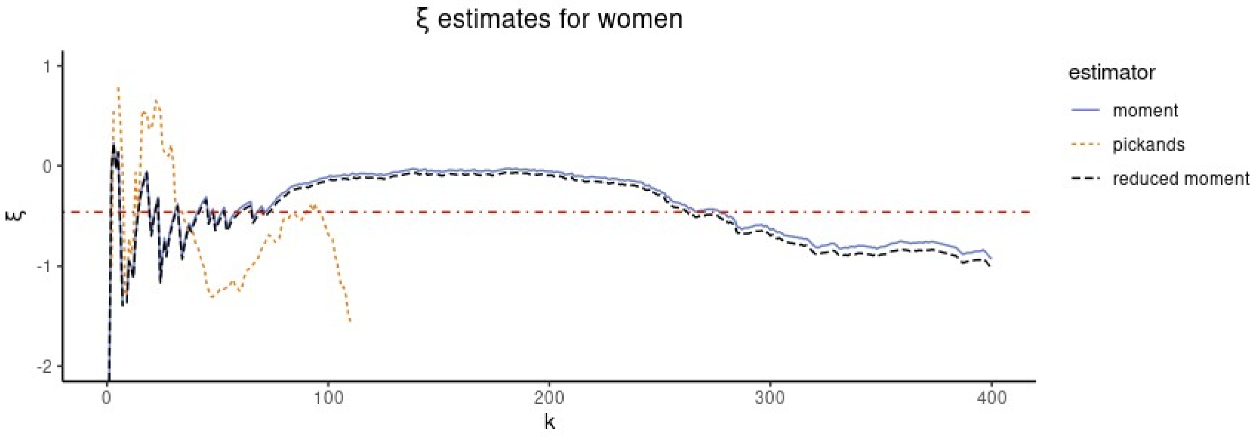
Fat tails inspection: QQ, Zipf & ME Plot



Tennis Serve Record Estimation

Extreme Value Index Estimation (ξ)


ξ estimates for men


ξ estimates for women

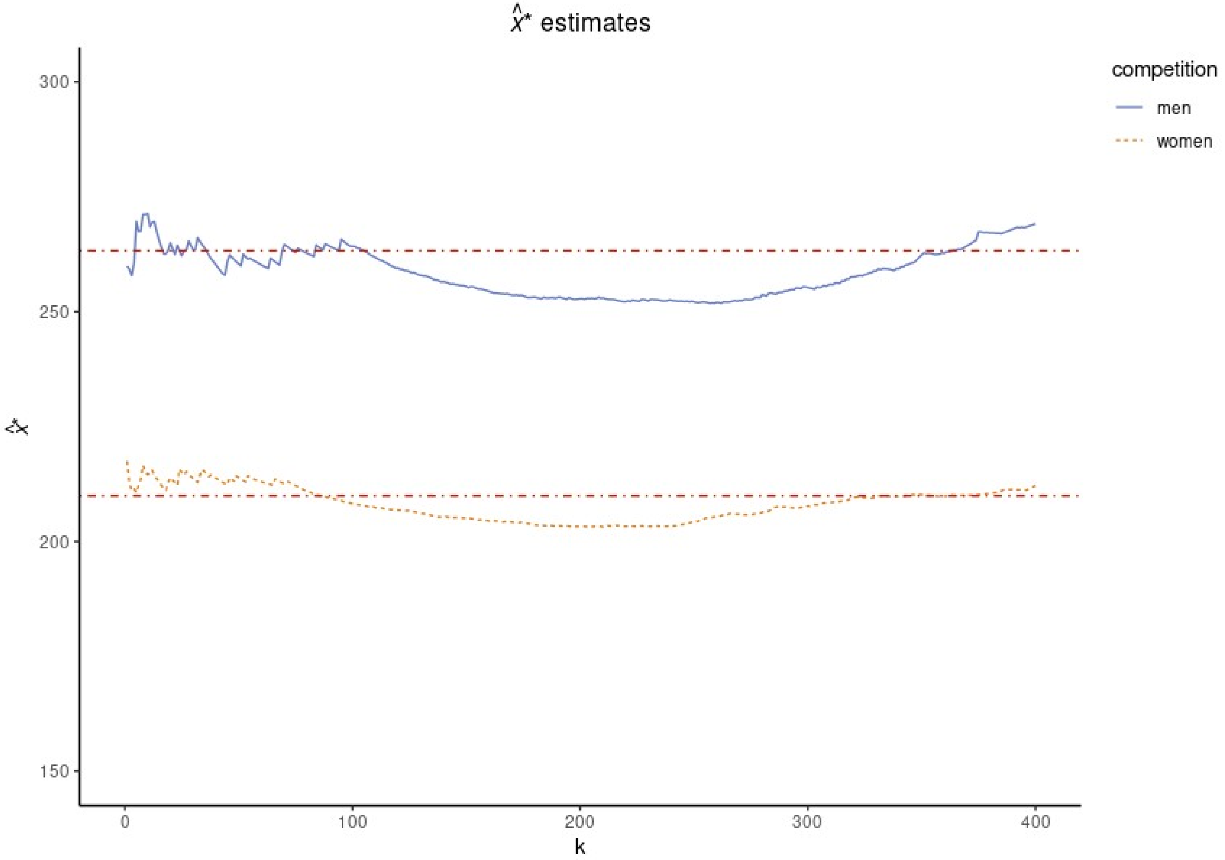
- EVI estimators:
  - Pickand’s
  - moment
  - reduced moment

- first stable region of at least 50 data points
- take mean over stable region
- $\xi < 0$ for both women and men

→ endpoint exists

Tennis Serve Record Estimation

Endpoint Estimation (x^*)



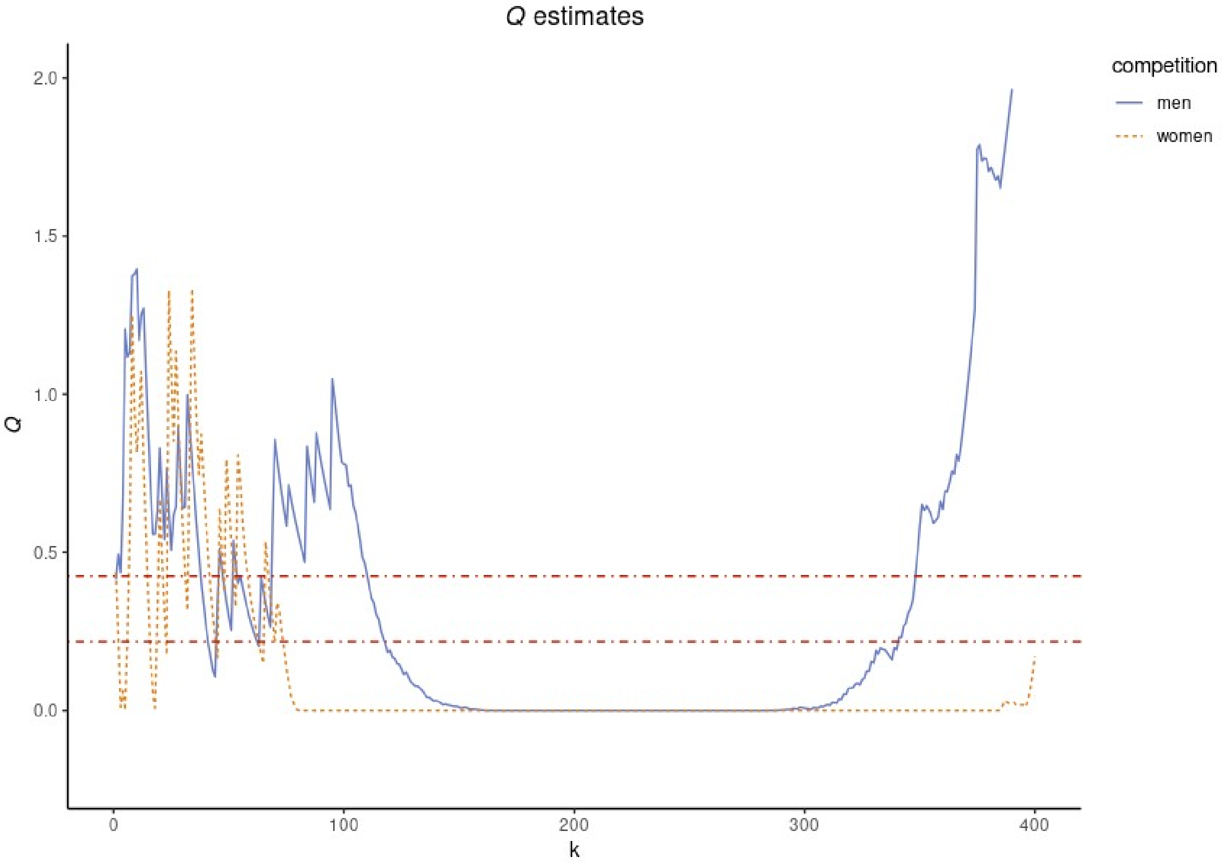
Endpoint estimation:

- first stable region of at least 50 data points
- take mean over stable region

- x^* women: 211.43 km/h
- x^* men: 262.65 km/h

Tennis Serve Record Estimation

Quality of Estimated Endpoint (Q)



- Number of exceedances of estimated endpoint given current record $X_{n,n}$

- Q women: 0.42
- Q men: 0.21

- Q < 0.5 which indicates a good quality → record is unlikely to be exceeded in general

- record is more likely to be broken for women compared to men

Tennis Serve Record Estimation

Plausibility of Unofficial Serve Records

- Construction of confidence interval (CI) around the estimated endpoint given ξ:

$$s^2 = \frac{(1-\xi)^2(1-3\xi+4\xi^2)}{(1-2\xi)(1-3\xi)(1-4\xi)}$$

$$SE = \frac{\sqrt{s^2}}{\sqrt{n}}$$

→ $\text{CI}_{\text{upper}} = \text{X}_{\text{mean}} + (\text{SE} * 1.96)$

- Check if upper bound includes the unoffical records:

  - Women UB: 211.75 km/h
  - Men UB: 263.47 km/h

Unofficial records:

Women:

- *214 km/h (A. Sabalenka)*
- *220 km/h (G. Garcia Perez)*

Men:

- *257.5 km/h (A. Olivetti)*
- *263 km/h (S. Groth)*

Summary & Next Steps

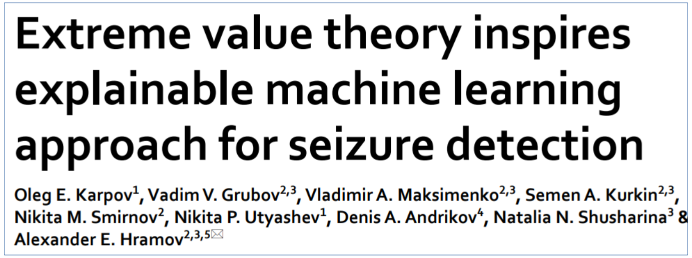
Summary & Next Steps

Summary:

- tennis serve record endpoints can be estimated for women and men competitions
- Quality of estimated endpoints is good
- Unofficial records seem plausible for men but not for women

Next Steps:

- Tennis serve record estimation on year by year basis → technology effects
- Investigate how EVT applicable to ML methods



Extreme value theory inspires explainable machine learning approach for seizure detection

Oleg E. Karpov[1], Vadim V. Grubov[2,3], Vladimir A. Maksimenko[2,3], Semen A. Kurkin[2,3], Nikita M. Smirnov[2], Nikita P. Utyashev[1], Denis A. Andrikov[4], Natalia N. Shusharina[3] & Alexander E. Hramov[2,3,5]✉

Thanks for your attention!

Code:

https://github.com/imarevic/evt_ts

Contact:

Ivan Marevic

imarevic89@gmail.com

https://www.linkedin.com/in/ivan-marevic-9776a8243/