Exploratory data analysis of extreme values using non-parametric kernel methods

Boris Beranger[1,2], Tarn Duong [3], Scott Sisson [2]

[1]Theoretical and Applied Statistics Laboratory, UPMC - Paris 6

[2]School of Mathematics and Statistics, UNSW, Australia

[3]Computer Science Laboratory, Paris-North University - Paris 13

EVA, 15th June 2015

Outline

- Motivation
- Kernel Density Estimators
- Simulation Study
- Real Data Application
- Conclusion

Motivation

- Goal: Projection of extreme events, calculation of return levels
- e.g. Climate (rainfall, wind, temperature, ...)
- Numerous models in the literature
- Problem: Which one is the most appropriate ?

Motivating Example (1)

Perkins et al. (2013): AR4 models (28) to investigate changes in temperature extremes

Model evaluation based on 3 skills:

1. Means
2. PDFs
3. Tails: Observed histogram Z_o is surrogate of the true density. Tail index is

$$T = \sum_{i=1}^{n} W_i |Z_o^i - Z_m^i|$$

where W_i is the weight of bin i, Z_o and Z_m are the observed and modeled frequencies.

Motivating Example (2)

Drawbacks:

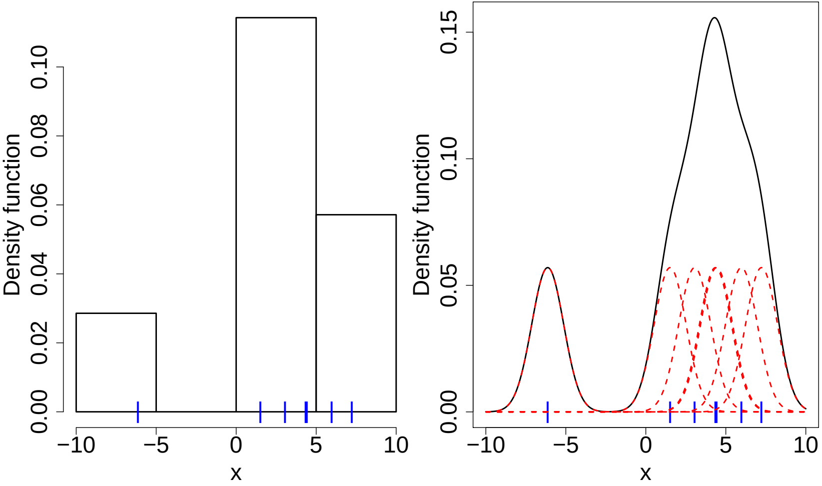
- Comparison of continuous models:
  - Discretization $\Rightarrow$ distortion of the model
- Data driven choices: bin width, bin weights, ...
- Unsuitable for multivariate extremes

Solution: Non-parametric Kernel Density Estimators (KDE)

- Continuous and robust (less arbitrary choices, can be applied to different datasets) $\Rightarrow$ Refinement of existing method
- Works with multi-variables $\Rightarrow$ Multivariate extension

KDE (1)

How do they work?



KDE (2)

What? A KDE is given by

$$\hat{f}_X(x;h) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

where K = kernel and h = bandwidth.

Why?

- Not affected as much by the mass of the data
- Good overall properties (continuity, smoothness, fast cv)

Drawback: noise/bias at the boundary of the support

$\Longrightarrow$ Transformation to focus on the tail and reduce bumps

Framework for Tail Estimation

(Random sample)

$$X \sim f_X$$

$$\Downarrow$$

(Tail sample)

$$X^{[u]} \equiv X|X > u, X^{[u]} \in (u, \infty)$$

$$\Downarrow$$

(Monotonic transformation)

$$Y = t(X^{[u]}),\ Y \sim f_Y$$

$$\Downarrow$$

$$f_{X^{[u]}}(x) = |t'(x)| f_Y(t(x))$$

$$\Downarrow$$

$$\hat{f}_Y(y; h) = n^{-1} \sum_{i=1}^{n} K_h(y - Y_i)$$

$$\Downarrow$$

(Tail density estimator)

$$\hat{f}_{X^{[u]}}(x; h) = |t'(t^{-1}(y)| \hat{f}_Y(y; h)$$

Main Result

Definition 1 (Mean Integrated Square Error - MISE). For the density estimator $\hat{f}_Y$, the MISE is

$$\text{MISE}\,\hat{f}_Y(\cdot;h) = \mathbb{E}\int_{\mathbb{R}}[\hat{f}_Y(y;h) - f(y)]^2 dy.$$

Theorem 1 (Minimal MISE of $\hat{f}_{X^{[u]}}$). Under suitable regularity conditions, as $n \to \infty$,

$$\inf_{h>0} \text{MISE}\,\hat{f}_{X^{[u]}}(\cdot;h) - \left\{\inf_{h>0} \text{MISE}\,\hat{f}_Y(\cdot;h)\right\} = O(n^{-4/5})$$

In other words:

- Bandwidth selection and estimation for transformed data Y retains same asymptotic optimality as original data $X^{[u]}$
- Can use existing results/algorithms

Simulation Study (1)

Targets (3): Fréchet, Gumbel and Generalized Pareto (GPD)

1. Generate 2000 replicates
2. Tail sample: $u = 95\%$ quantile, target tail density $f_{X^{[u]}}$
3. Transformation: $t(x) = \log(x - u)$
4. Fit: parametric models (3), histogram and kernel
5. Iterate 400 times
6. Comparisons:
   - 6.1 L_2 distance between target and fitted densities, e.g. $\int_u^\infty [\hat{f}_{X^{[u]}}(x) - f_{X^{[u]}}(x)]^2 \, dx$
   - 6.2 T_h and T_k: histogram and Kernel based tail indices for $u^* = 99\%$ quantile to avoid boundary bias at $x = u$ affecting model selection, e.g. $T = \int_{u^*}^\infty |\hat{f}_{X^{[u]}}(x) - f_{X^{[u]}}(x)| \, dx$

Simulation Study (2)

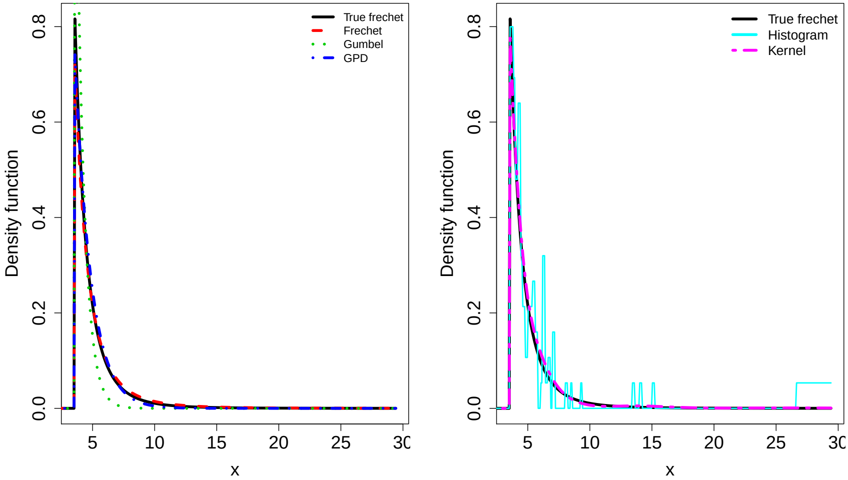


Figure: Parametric (left) and non-parametric (right) estimators of a Fréchet tail density.

Simulation Study (3)

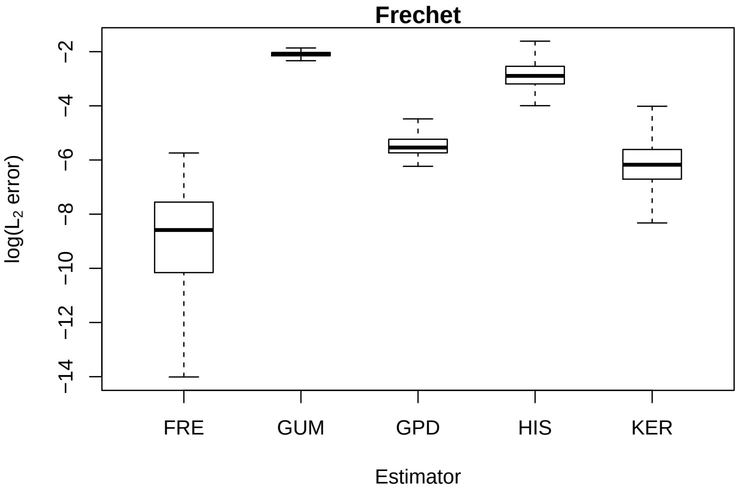


Figure: Boxplot of the L_2 distances between estimated densities and target Fréchet density.

Simulation Study (4)

	T_h			T_k		
Target	Fréchet	Gumbel	GDP	Fréchet	Gumbel	GDP
Fréchet	0.120	0.202	0.678	0.937	0	0.063
Gumbel	0.400	0.592	0.008	0.595	0.400	0.005
GPD	0.012	0.915	0.073	0.067	0.035	0.898

Table: Proportion of accepting a parametric model using histogram and kernel based tail indices.

Remark: True model is Gumbel: $\bar{T}_h = 0.361$ whereas $\bar{T}_k = 0.027$.

Real Data Application (1)

- Data: Daily max temperatures in Sydney for 1911-2005 (36890 obs).
- Comparison with physical models and histogram/KDEs
- Perkins et al. (2007): Histogram as surrogate for model densities
- Model selection:
  - T_h: CCMC.CESM, MPI.ESM.MR, CCMS.CMS
  - T_k: MPI.ESM.MR, MIROC5, HadGEM2.CC
  - Same 5 worst models

Real Data Application (2)

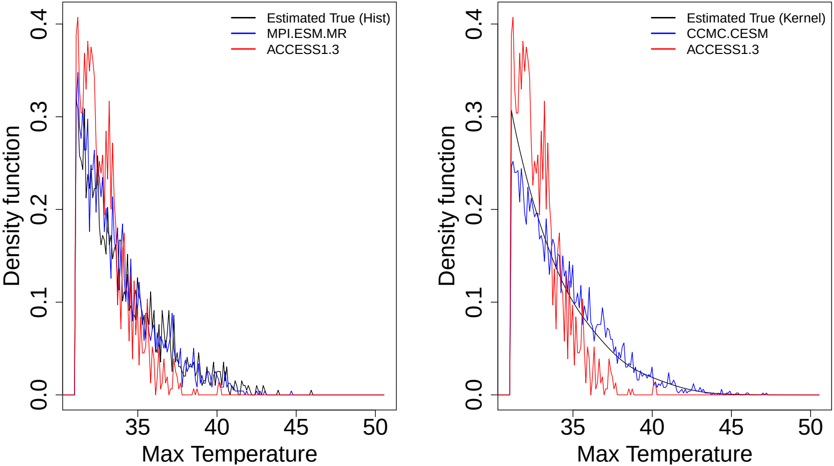


Figure: Best and worst models according to the histogram (left) and kernel (right) based tail indices.

Conclusion

Results:

- Model selection method for extreme values
- More robust and continuous estimator of the tail density
- Efficiency proved for univariate simulated data
- Application to temperature data

Work in progress:

- Extension of the simulations to the bivariate case
- Bivariate real data application (max and min temperatures)

Many thanks for your attention!