Threshold Selection by Distance Minimization

(work in progress)

Holger Drees

University of Hamburg

Workshop on Self-Similarity, Long-Range Dependence and Extremes
June 2018

based on joint work with Anja Janßen (KTH Stockholm), and
Sid Resnick and Tiandong Wang (Cornell)

POT-analysis of heavy tails

X_i, $1 \leq i \leq n$, iid observations with cdf $F \in D(G_{1/\alpha})$, $\alpha > 0$, i.e. as $t \to \infty$

$$\frac{1 - F(tx)}{1 - F(t)} \to x^{-\alpha}, \qquad \forall x > 0.$$

Hill estimator of α:

$$\hat{\alpha}_{n,k} := 1 \Big/ \left[\frac{1}{k-1} \sum_{i=1}^{k-1} \log \frac{X_{n-i+1:n}}{X_{n-k+1:n}} \right]$$

where $X_{j:n}$ denotes the jth smallest order statistic.

Hill estimator is essentially ML estimator if k largest observations behave like Pareto random variables.

Performance strongly depends on choice of k:

- k must be sufficiently small such that Pareto approximation is justified ($\rightsquigarrow$ small bias)
- k must be sufficiently large such that average is taken over many observations ($\rightsquigarrow$ small variance)

POT-analysis of heavy tails

X_i, $1 \le i \le n$, iid observations with cdf $F \in D(G_{1/\alpha})$, $\alpha > 0$, i.e. as $t \to \infty$

$$\frac{1 - F(tx)}{1 - F(t)} \to x^{-\alpha}, \qquad \forall x > 0.$$

Hill estimator of α:

$$\hat{\alpha}_{n,k} := 1 \Bigg/ \left[\frac{1}{k-1} \sum_{i=1}^{k-1} \log \frac{X_{n-i+1:n}}{X_{n-k+1:n}} \right]$$

where $X_{j:n}$ denotes the jth smallest order statistic.

Hill estimator is essentially ML estimator if k largest observations behave like Pareto random variables.

Performance strongly depends on choice of k:

- k must be sufficiently small such that Pareto approximation is justified ($\rightsquigarrow$ small bias)
- k must be sufficiently large such that average is taken over many observations ($\rightsquigarrow$ small variance)

Threshold selection

$$\hat{\alpha}_{n,k} := 1 \Big/ \left[\frac{1}{k-1} \sum_{i=1}^{k-1} \log \frac{X_{n-i+1:n}}{X_{n-k+1:n}} \right]$$

Several procedures for data-dependent selection of k have been suggested, e.g. using

- plug-in methods: Hall & Welsh ('85), ...
- resampling: Hall ('90), Danielsson et al. ('01), Gomes & Oliveira ('01), ...
- Lepskii method: D. & Kaufmann ('98), ...
- using log-spacings: Guillou & Hall ('01), Beirlant et al. ('04), ...
- distance minimization: Clauset, Shalizi & Newman (2009) (over 2700 citations)

Idea: Choose k such that the Kolmogorov-Smirnov distance between empirical cdf of exceedances over $X_{n-k+1:n}$ and fitted Pareto distribution is minimal.

More precisely, minimize

$$D_{n,k} := \sup_{y \geq 1} \left| \frac{1}{k-1} \sum_{i=1}^{k-1} 1_{(y,\infty)} \left(\frac{X_{n-i+1:n}}{X_{n-k+1:n}} \right) - y^{-\hat{\alpha}_{n,k}} \right|$$

Threshold selection

$$\hat{\alpha}_{n,k} := 1 \Big/ \left[\frac{1}{k-1} \sum_{i=1}^{k-1} \log \frac{X_{n-i+1:n}}{X_{n-k+1:n}} \right]$$

Several procedures for data-dependent selection of k have been suggested, e.g. using

- plug-in methods: Hall & Welsh ('85), ...
- resampling: Hall ('90), Danielsson et al. ('01), Gomes & Oliveira ('01), ...
- Lepskii method: D. & Kaufmann ('98), ...
- using log-spacings: Guillou & Hall ('01), Beirlant et al. ('04), ...
- distance minimization: Clauset, Shalizi & Newman (2009) (over 2700 citations)

Idea: Choose k such that the Kolmogorov-Smirnov distance between empirical cdf of exceedances over $X_{n-k+1:n}$ and fitted Pareto distribution is minimal.

More precisely, minimize

$$D_{n,k} := \sup_{y \geq 1} \left| \frac{1}{k-1} \sum_{i=1}^{k-1} 1_{(y,\infty)} \left(\frac{X_{n-i+1:n}}{X_{n-k+1:n}} \right) - y^{-\hat{\alpha}_{n,k}} \right|$$

Threshold selection

$$\hat{\alpha}_{n,k} := 1 \Big/ \left[\frac{1}{k-1} \sum_{i=1}^{k-1} \log \frac{X_{n-i+1:n}}{X_{n-k+1:n}} \right]$$

Several procedures for data-dependent selection of k have been suggested, e.g. using

- plug-in methods: Hall & Welsh ('85), ...
- resampling: Hall ('90), Danielsson et al. ('01), Gomes & Oliveira ('01), ...
- Lepskii method: D. & Kaufmann ('98), ...
- using log-spacings: Guillou & Hall ('01), Beirlant et al. ('04), ...
- distance minimization: Clauset, Shalizi & Newman (2009) (over 2700 citations)

Idea: Choose k such that the Kolmogorov-Smirnov distance between empirical cdf of exceedances over $X_{n-k+1:n}$ and fitted Pareto distribution is minimal.

More precisely, minimize

$$D_{n,k} := \sup_{y \geq 1} \left| \frac{1}{k-1} \sum_{i=1}^{k-1} 1_{(y,\infty)}\left(\frac{X_{n-i+1:n}}{X_{n-k+1:n}}\right) - y^{-\hat{\alpha}_{n,k}} \right|$$

Threshold selection

$$\hat{\alpha}_{n,k} := 1 \Big/ \left[\frac{1}{k-1} \sum_{i=1}^{k-1} \log \frac{X_{n-i+1:n}}{X_{n-k+1:n}} \right]$$

Several procedures for data-dependent selection of k have been suggested, e.g. using

- plug-in methods: Hall & Welsh ('85), ...
- resampling: Hall ('90), Danielsson et al. ('01), Gomes & Oliveira ('01), ...
- Lepskii method: D. & Kaufmann ('98), ...
- using log-spacings: Guillou & Hall ('01), Beirlant et al. ('04), ...
- distance minimization: Clauset, Shalizi & Newman (2009) (over 2700 citations)

Idea: Choose k such that the Kolmogorov-Smirnov distance between empirical cdf of exceedances over $X_{n-k+1:n}$ and fitted Pareto distribution is minimal.

More precisely, minimize

$$D_{n,k} := \sup_{y \geq 1} \left| \frac{1}{k-1} \sum_{i=1}^{k-1} 1_{(y,\infty)} \left(\frac{X_{n-i+1:n}}{X_{n-k+1:n}} \right) - y^{-\hat{\alpha}_{n,k}} \right|$$

Threshold selection

$$\hat{\alpha}_{n,k} := 1 \Big/ \left[\frac{1}{k-1} \sum_{i=1}^{k-1} \log \frac{X_{n-i+1:n}}{X_{n-k+1:n}} \right]$$

Several procedures for data-dependent selection of k have been suggested, e.g. using

- plug-in methods: Hall & Welsh ('85), ...
- resampling: Hall ('90), Danielsson et al. ('01), Gomes & Oliveira ('01), ...
- Lepskii method: D. & Kaufmann ('98), ...
- using log-spacings: Guillou & Hall ('01), Beirlant et al. ('04), ...
- distance minimization: Clauset, Shalizi & Newman (2009) (over 2700 citations)

Idea: Choose k such that the Kolmogorov-Smirnov distance between empirical cdf of exceedances over $X_{n-k+1:n}$ and fitted Pareto distribution is minimal.

More precisely, minimize

$$D_{n,k} := \sup_{y \geq 1} \left| \frac{1}{k-1} \sum_{i=1}^{k-1} 1_{(y,\infty)}\left(\frac{X_{n-i+1:n}}{X_{n-k+1:n}} \right) - y^{-\hat{\alpha}_{n,k}} \right|$$

Threshold selection

$$\hat{\alpha}_{n,k} := 1 \Big/ \left[\frac{1}{k-1} \sum_{i=1}^{k-1} \log \frac{X_{n-i+1:n}}{X_{n-k+1:n}} \right]$$

Several procedures for data-dependent selection of k have been suggested, e.g. using

- plug-in methods: Hall & Welsh ('85), ...
- resampling: Hall ('90), Danielsson et al. ('01), Gomes & Oliveira ('01), ...
- Lepskii method: D. & Kaufmann ('98), ...
- using log-spacings: Guillou & Hall ('01), Beirlant et al. ('04), ...
- distance minimization: Clauset, Shalizi & Newman (2009) (over 2700 citations)

Idea: Choose k such that the Kolmogorov-Smirnov distance between empirical cdf of exceedances over $X_{n-k+1:n}$ and fitted Pareto distribution is minimal.

More precisely, minimize

$$D_{n,k} := \sup_{y \geq 1} \left| \frac{1}{k-1} \sum_{i=1}^{k-1} 1_{(y,\infty)} \left(\frac{X_{n-i+1:n}}{X_{n-k+1:n}} \right) - y^{-\hat{\alpha}_{n,k}} \right|$$

Threshold selection by distance minimization

$$\text{minimize} \quad D_{n,k} := \sup_{y \geq 1} \left| \frac{1}{k-1} \sum_{i=1}^{k-1} 1_{(y,\infty)}\left(\frac{X_{n-i+1:n}}{X_{n-k+1:n}}\right) - y^{-\hat{\alpha}_{n,k}} \right|$$

Rationale:

- If Pareto approximation is accurate for top k order statistics, then $D_{n,k}$ is of stochastic order $k^{-1/2}$, i.e. it shrinks with increasing k
- If below threshold u cdf is poorly approximated by Pareto cdf, $D_{n,k}$ quickly increases as k increases such that $X_{n-k:n}$ shrinks below u.

Indeed, it seems plausible that procedure yields k converging at the "optimal rate".

However, even if all observations are exact Pareto, $D_{n,k}$ will be minimal for k much smaller than n due to random fluctuations.

Threshold selection by distance minimization

$$\text{minimize} \quad D_{n,k} := \sup_{y \geq 1} \left| \frac{1}{k-1} \sum_{i=1}^{k-1} 1_{(y,\infty)}\left(\frac{X_{n-i+1:n}}{X_{n-k+1:n}}\right) - y^{-\hat{\alpha}_{n,k}} \right|$$

Rationale:

- If Pareto approximation is accurate for top k order statistics, then $D_{n,k}$ is of stochastic order $k^{-1/2}$, i.e. it shrinks with increasing k
- If below threshold u cdf is poorly approximated by Pareto cdf, $D_{n,k}$ quickly increases as k increases such that $X_{n-k:n}$ shrinks below u.

Indeed, it seems plausible that procedure yields k converging at the "optimal rate".

However, even if all observations are exact Pareto, $D_{n,k}$ will be minimal for k much smaller than n due to random fluctuations.

Threshold selection by distance minimization

$$\text{minimize} \quad D_{n,k} := \sup_{y \geq 1} \left| \frac{1}{k-1} \sum_{i=1}^{k-1} 1_{(y,\infty)}\left(\frac{X_{n-i+1:n}}{X_{n-k+1:n}} \right) - y^{-\hat{\alpha}_{n,k}} \right|$$

Rationale:

- If Pareto approximation is accurate for top k order statistics, then $D_{n,k}$ is of stochastic order $k^{-1/2}$, i.e. it shrinks with increasing k
- If below threshold u cdf is poorly approximated by Pareto cdf, $D_{n,k}$ quickly increases as k increases such that $X_{n-k:n}$ shrinks below u.

Indeed, it seems plausible that procedure yields k converging at the "optimal rate".

However, even if all observations are exact Pareto, $D_{n,k}$ will be minimal for k much smaller than n due to random fluctuations.

Gaussian approximation: α known

Assume $F(x) = 1 - x^{-\alpha}$ $(x > 1)$ with **known** $\alpha > 0$. Consider KS distance

$$\begin{aligned}
\bar{D}_{n,k} &:= \sup_{y \geq 1} \left| \frac{1}{k-1} \sum_{i=1}^{k-1} 1_{(y,\infty)} \left(\frac{X_{n-i+1:n}}{X_{n-k+1:n}} \right) - y^{-\alpha} \right| \\
&= \max_{1 \leq i < k} \left| \left(\frac{X_{n-i+1:n}}{X_{n-k+1:n}} \right)^{-\alpha} - \frac{i}{k} \right| + O(k^{-1}) \\
&=^d \max_{1 \leq i < k} \left| \frac{U_{i:n}}{U_{k:n}} - \frac{i}{k} \right| + O(k^{-1})
\end{aligned}$$

for iid uniform rv's U_j.

Approximation of uniform order statistics by Brownian motion yields

$$n^{1/2} \bar{D}_{n,\lceil nt \rceil} \to \sup_{0 < z \leq 1} z \left| \frac{W(tz)}{tz} - \frac{W(t)}{t} \right|$$

weakly in $D(0,1]$.

Gaussian approximation: α known

Assume $F(x) = 1 - x^{-\alpha}$ $(x > 1)$ with **known** $\alpha > 0$. Consider KS distance

$$\begin{aligned}
\bar{D}_{n,k} &:= \sup_{y \geq 1} \left| \frac{1}{k-1} \sum_{i=1}^{k-1} 1_{(y,\infty)}\left(\frac{X_{n-i+1:n}}{X_{n-k+1:n}}\right) - y^{-\alpha} \right| \\
&= \max_{1 \leq i < k} \left| \left(\frac{X_{n-i+1:n}}{X_{n-k+1:n}}\right)^{-\alpha} - \frac{i}{k} \right| + O(k^{-1}) \\
&=^d \max_{1 \leq i < k} \left| \frac{U_{i:n}}{U_{k:n}} - \frac{i}{k} \right| + O(k^{-1})
\end{aligned}$$

for iid uniform rv's U_j.

Approximation of uniform order statistics by Brownian motion yields

$$n^{1/2} \bar{D}_{n,\lceil nt \rceil} \to \sup_{0 < z \leq 1} z \left| \frac{W(tz)}{tz} - \frac{W(t)}{t} \right|$$

weakly in $D(0, 1]$.

"Early stopping"

$$n^{1/2}\bar{D}_{n,\lceil nt\rceil} \to \sup_{0<z\le 1} z\left|\frac{W(tz)}{tz} - \frac{W(t)}{t}\right|$$

One might thus expect that the value k for which $\bar{D}_{n,k}$ is minimized behaves like nT^* with

$$T^* := \arg\min_{0<t\le 1} \sup_{0<z\le 1} z\left|\frac{W(tz)}{tz} - \frac{W(t)}{t}\right|.$$

Despite

$$\sup_{0<z\le 1} z\left|\frac{W(tz)}{tz} - \frac{W(t)}{t}\right| =^d t^{-1/2} \sup_{0<z\le 1} z\left|\frac{W(z)}{z} - W(1)\right|,$$

with non-negligible probability, t^* will be substantially smaller than 1, leading to too small a value for k.

"Early stopping"

$$n^{1/2}\bar{D}_{n,\lceil nt\rceil} \to \sup_{0<z\le 1} z\left|\frac{W(tz)}{tz}-\frac{W(t)}{t}\right|$$

One might thus expect that the value k for which $\bar{D}_{n,k}$ is minimized behaves like nT^* with

$$T^* := \arg\min_{0<t\le 1}\sup_{0<z\le 1} z\left|\frac{W(tz)}{tz}-\frac{W(t)}{t}\right|.$$

Despite

$$\sup_{0<z\le 1} z\left|\frac{W(tz)}{tz}-\frac{W(t)}{t}\right| =^d t^{-1/2}\sup_{0<z\le 1} z\left|\frac{W(z)}{z}-W(1)\right|,$$

with non-negligible probability, t^* will be substantially smaller than 1, leading to too small a value for k.

Gaussian approximation: α unknown

If α is unknown and replaced with the Hill estimator, process convergence becomes more involved.

Theorem

Suppose $F(x) = 1 - cx^{-\alpha}$ $(x > c^{1/\alpha})$.

1. *For all* $k = k_n = o(n)$

$$\inf_{2 \le j \le k} n^{1/2} D_{n,j} \xrightarrow{(P)} \infty.$$

2.

$$\begin{aligned} & n^{1/2} D_{n,\lceil nt \rceil} \\ \to \quad & \sup_{0<z\le 1} \left| \left(\int_0^1 \frac{W(tx)}{tx} dx - \frac{W(t)}{t} \right) z \log z + \left(\frac{W(tz)}{tz} - \frac{W(t)}{t} \right) z \right| \\ =: \quad & \sup_{0<z\le 1} |Y(t,z)| \end{aligned}$$

weakly in $D(0,1]$.

Gaussian approximation: α unknown

If α is unknown and replaced with the Hill estimator, process convergence becomes more involved.

Theorem

Suppose $F(x) = 1 - cx^{-\alpha}$ $(x > c^{1/\alpha})$.

1. *For all* $k = k_n = o(n)$

$$\inf_{2 \leq j \leq k} n^{1/2} D_{n,j} \xrightarrow{(P)} \infty.$$

2.

$$\begin{aligned} & n^{1/2} D_{n,\lceil nt \rceil} \\ \to \quad & \sup_{0<z\leq 1} \left| \left(\int_0^1 \frac{W(tx)}{tx} dx - \frac{W(t)}{t} \right) z \log z + \left(\frac{W(tz)}{tz} - \frac{W(t)}{t} \right) z \right| \\ =: \quad & \sup_{0<z\leq 1} |Y(t,z)| \end{aligned}$$

weakly in $D(0,1]$.

Asymptotic behavior of selected threshold

Let $k^* := \arg\min_{2\le k\le n} D_{n,k}$

Corollary

Suppose $F(x) = 1 - cx^{-\alpha}$ $(x > c^{1/\alpha})$. *Then*

$$\frac{k^*}{n} \to \arg\inf_{t\in(0,1]} \sup_{0<z\le 1} |Y(t,z)| =: T^*,$$

provided the process $(\sup_{0<z\le 1} |Y(t,z)|)_{t\in(0,1]}$ *has a unique point of minimum a.s.*

In that case,

$$n^{1/2}(\hat{\alpha}_{n,k^*} - \alpha) \to \alpha\left(\int_0^1 \frac{W(T^* x)}{T^* x}\,dx - \frac{W(T^*)}{T^*}\right) \quad \textit{weakly.}$$

The limit rv is not normally distributed.

Asymptotic behavior of selected threshold

Let $k^* := \arg\min_{2\leq k\leq n} D_{n,k}$

Corollary

Suppose $F(x) = 1 - cx^{-\alpha}$ $(x > c^{1/\alpha})$. *Then*

$$\frac{k^*}{n} \to \arg\inf_{t\in(0,1]} \sup_{0<z\leq 1} |Y(t,z)| =: T^*,$$

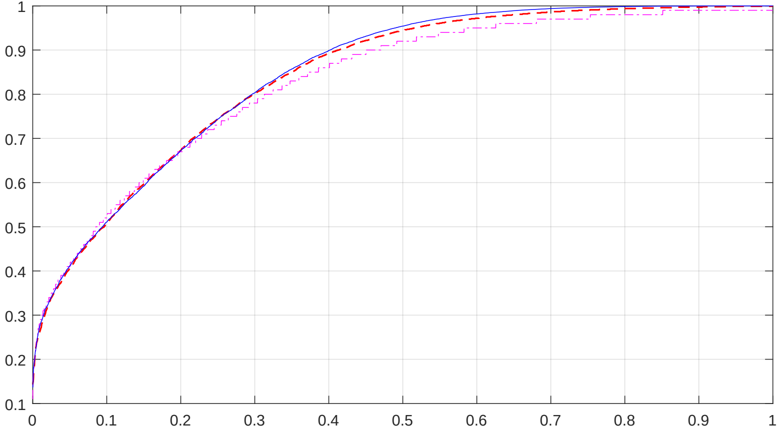
provided the process $(\sup_{0<z\leq 1} |Y(t,z)|)_{t\in(0,1]}$ *has a unique point of minimum a.s.*

In that case,

$$n^{1/2}(\hat{\alpha}_{n,k^*} - \alpha) \to \alpha\left(\int_0^1 \frac{W(T^* x)}{T^* x}\, dx - \frac{W(T^*)}{T^*}\right) \quad \textit{weakly.}$$

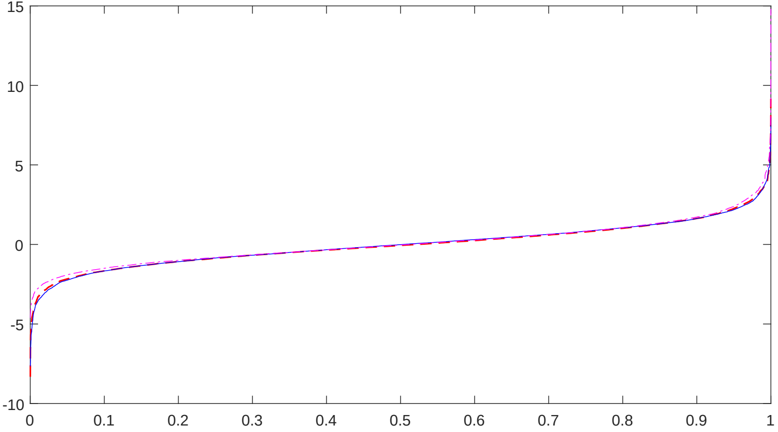
The limit rv is not normally distributed.

Distribution of k^*/n



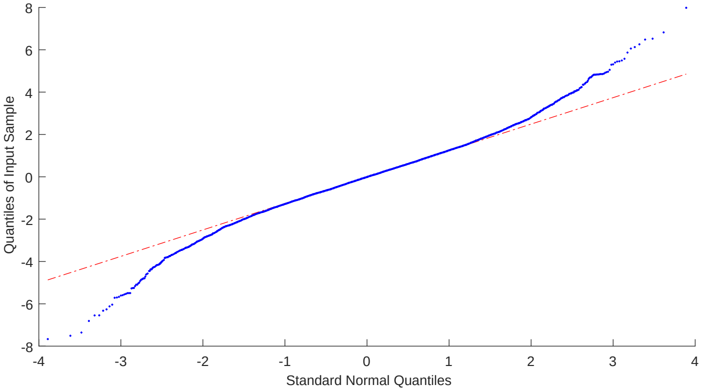
Quantile function of T^*/n for sample sizes $n = 100$ (magenta dash-dotted), $n = 1000$ (red dashed), and limit (blue solid)

Distribution of $\hat{\alpha}_{n,k^*}$



Quantile function of $n^{1/2}(\hat{\alpha}_{n,k^*} - \alpha)$ for sample sizes $n = 100$ (magenta dash-dotted), $n = 1000$ (red dashed), and limit (blue solid)

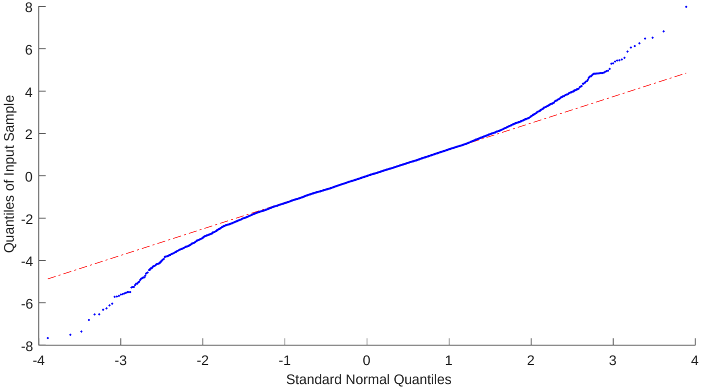
Limit distribution of $\hat{\alpha}_{n,k^*}$



Normal-QQ-plot for limit distribution of $n^{1/2}(\hat{\alpha}_{n,k^*} - \alpha)$

In the limit, the variance is about 1.95 times the variance of $\hat{\alpha}_{n,n}$

Limit distribution of $\hat{\alpha}_{n,k^*}$



Normal-QQ-plot for limit distribution of $n^{1/2}(\hat{\alpha}_{n,k^*} - \alpha)$

In the limit, the variance is about 1.95 times the variance of $\hat{\alpha}_{n,n}$

Structural breaks

In Clauset et al. (2009) (and similar papers) it is assumed that above some threshold u F equals a Pareto cdf, while below it has a different structure.

Selection procedures should yield k such that $X_{n-k+1:n}$ is close to u.

There is no obvious asymptotic setting in which to embed such a situation.

However, simulations suggest that $k^*/(n(1-F(u)))$ roughly behaves like T^* if break is sufficiently clear and n is large.

Hence procedures often selects too small a k, i.e. too high a threshold.

Structural breaks

In Clauset et al. (2009) (and similar papers) it is assumed that above some threshold u F equals a Pareto cdf, while below it has a different structure.

Selection procedures should yield k such that $X_{n-k+1:n}$ is close to u.

There is no obvious asymptotic setting in which to embed such a situation.

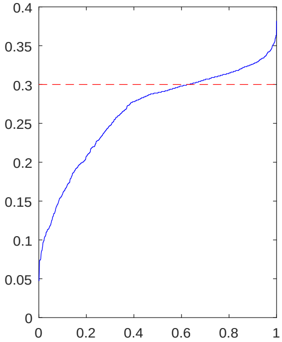
However, simulations suggest that $k^*/(n(1-F(u)))$ roughly behaves like T^* if break is sufficiently clear and n is large.

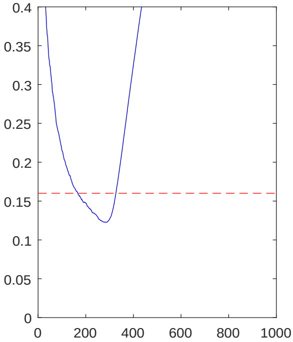
Hence procedures often selects too small a k, i.e. too high a threshold.

Simulation

$$1 - F(x) = \begin{cases} x^{-2}, & x > x_0, \\ cx^{-4}, & x_0 \geq x > c^{1/4} \end{cases}$$

with x_0, c such that $1 - F(x_0) = 0.3$, F continuous.





Left: qf of k^*/n for $n = 1000$; red line indicates break point
Right: RMSE of Hill estimator as function of k; red line indicates RMSE of $\hat{\alpha}_{n,k^*}$

increase of RMSE and of SD $\approx 31\%$

Second order condition

Assume, as $t \downarrow 0$,

$$\frac{\frac{F^{\leftarrow}(1-tx)}{F^{\leftarrow}(1-t)} - x^{-1/\alpha}}{A(t)} \to \psi(x), \qquad \forall\, x > 0,$$

with $A(t) \downarrow 0$, regularly varying at 0 with index $\rho > 0$,
$\psi(x)$ not a multiple of $x^{-1/\alpha}$.

Then there exists sequence $\tilde{k} = \tilde{k}_n \to \infty, \tilde{k} = o(n)$ such that $\tilde{k}^{1/2} A(\tilde{k}/n) \to 1$.

SD, bias balanced iff $k \asymp \tilde{k}$ and then $\hat{\alpha}_{n,k}$ converges with the optimal rate $\tilde{k}^{-1/2}$ (among all deterministic intermediate sequences k). Moreover, AMSE $\hat{\alpha}_{n,k}$ is minimal iff $k \sim c\tilde{k}$ for some constant c depending on α, ρ, ψ.

Most threshold selection methods mentioned in the beginning yield random $\bar{k} \sim c\tilde{k}$ under suitable conditions.

In this setting, minimizer of $D_{n,j}$ can be analyzed only if minimization is restricted to $j \leq k$ for some intermediate sequence k.

Second order condition

Assume, as $t \downarrow 0$,

$$\frac{\frac{F^{\leftarrow}(1-tx)}{F^{\leftarrow}(1-t)} - x^{-1/\alpha}}{A(t)} \to \psi(x), \qquad \forall\, x > 0,$$

with $A(t) \downarrow 0$, regularly varying at 0 with index $\rho > 0$,
$\psi(x)$ not a multiple of $x^{-1/\alpha}$.

Then there exists sequence $\tilde{k} = \tilde{k}_n \to \infty, \tilde{k} = o(n)$ such that $\tilde{k}^{1/2} A(\tilde{k}/n) \to 1$.

SD, bias balanced iff $k \asymp \tilde{k}$ and then $\hat{\alpha}_{n,k}$ converges with the optimal rate $\tilde{k}^{-1/2}$ (among all deterministic intermediate sequences k). Moreover, AMSE $\hat{\alpha}_{n,k}$ is minimal iff $k \sim c\tilde{k}$ for some constant c depending on α, ρ, ψ.

Most threshold selection methods mentioned in the beginning yield random $\bar{k} \sim c\tilde{k}$ under suitable conditions.

In this setting, minimizer of $D_{n,j}$ can be analyzed only if minimization is restricted to $j \leq k$ for some intermediate sequence k.

Second order condition

Assume, as $t \downarrow 0$,

$$\frac{\frac{F^{\leftarrow}(1-tx)}{F^{\leftarrow}(1-t)} - x^{-1/\alpha}}{A(t)} \to \psi(x), \qquad \forall\, x > 0,$$

with $A(t) \downarrow 0$, regularly varying at 0 with index $\rho > 0$,
$\psi(x)$ not a multiple of $x^{-1/\alpha}$.

Then there exists sequence $\tilde{k} = \tilde{k}_n \to \infty, \tilde{k} = o(n)$ such that $\tilde{k}^{1/2} A(\tilde{k}/n) \to 1$.

SD, bias balanced iff $k \asymp \tilde{k}$ and then $\hat{\alpha}_{n,k}$ converges with the optimal rate $\tilde{k}^{-1/2}$ (among all deterministic intermediate sequences k). Moreover, AMSE $\hat{\alpha}_{n,k}$ is minimal iff $k \sim c\tilde{k}$ for some constant c depending on α, ρ, ψ.

Most threshold selection methods mentioned in the beginning yield random $\bar{k} \sim c\tilde{k}$ under suitable conditions.

In this setting, minimizer of $D_{n,j}$ can be analyzed only if minimization is restricted to $j \leq k$ for some intermediate sequence k.

Second order condition

Assume, as $t \downarrow 0$,

$$\frac{\frac{F^{\leftarrow}(1-tx)}{F^{\leftarrow}(1-t)} - x^{-1/\alpha}}{A(t)} \to \psi(x), \qquad \forall\, x > 0,$$

with $A(t) \downarrow 0$, regularly varying at 0 with index $\rho > 0$,
$\psi(x)$ not a multiple of $x^{-1/\alpha}$.

Then there exists sequence $\tilde{k} = \tilde{k}_n \to \infty, \tilde{k} = o(n)$ such that $\tilde{k}^{1/2} A(\tilde{k}/n) \to 1$.

SD, bias balanced iff $k \asymp \tilde{k}$ and then $\hat{\alpha}_{n,k}$ converges with the optimal rate $\tilde{k}^{-1/2}$ (among all deterministic intermediate sequences k). Moreover, AMSE $\hat{\alpha}_{n,k}$ is minimal iff $k \sim c\tilde{k}$ for some constant c depending on α, ρ, ψ.

Most threshold selection methods mentioned in the beginning yield random $\bar{k} \sim c\tilde{k}$ under suitable conditions.

In this setting, minimizer of $D_{n,j}$ can be analyzed only if minimization is restricted to $j \leq k$ for some intermediate sequence k.

Asymptotics under second order condition

Theorem

1. $\inf_{2\le j\le k} \tilde{k}^{1/2} D_{n,j} \to \infty$ *for all intermediate sequences* $k = o(\tilde{k})$
2. $$\tilde{k}^{1/2} D_{n,\lceil \tilde{k} t\rceil} \to \sup_{0<z\le 1} \left| Y(t,z) - \left(\int_0^1 x^{1/\alpha}\psi(x)\,dx \cdot z\log z + \alpha z^{1/\alpha}\psi(z) \right) t^\rho \right|$$
   weakly in $D(0,\infty)$.
3. *If* $\tilde{k} = o(k)$, $k = o(n)$ *then, for all* $0 < t_0 < t_1 < \infty$
   $$\inf_{t\in[t_0,t_1]} \tilde{k}^{1/2} D_{n,\lceil kt\rceil} \to \infty.$$

This suggests (but doesn't prove) that

$$k^*/\tilde{k} \to \arg\inf_{0<t<\infty} \sup_{0<z\le 1} \left| Y(t,z) - \left(\int_0^1 x^{1/\alpha}\psi(x)\,dx \cdot z\log z + \alpha z^{1/\alpha}\psi(z) \right) t^\rho \right|.$$

Asymptotics under second order condition

Theorem

1. $\inf_{2\le j\le k} \tilde{k}^{1/2} D_{n,j} \to \infty$ *for all intermediate sequences* $k = o(\tilde{k})$
2. $$\tilde{k}^{1/2} D_{n,\lceil \tilde{k}t\rceil} \to \sup_{0<z\le 1} \left| Y(t,z) - \left(\int_0^1 x^{1/\alpha}\psi(x)\,dx \cdot z\log z + \alpha z^{1/\alpha}\psi(z)\right) t^{\rho} \right|$$
   weakly in $D(0,\infty)$.
3. *If* $\tilde{k} = o(k)$, $k = o(n)$ *then, for all* $0 < t_0 < t_1 < \infty$
   $$\inf_{t\in[t_0,t_1]} \tilde{k}^{1/2} D_{n,\lceil kt\rceil} \to \infty.$$

This suggests (but doesn't prove) that

$$k^*/\tilde{k} \to \arg\inf_{0<t<\infty} \sup_{0<z\le 1} \left| Y(t,z) - \left(\int_0^1 x^{1/\alpha}\psi(x)\,dx \cdot z\log z + \alpha z^{1/\alpha}\psi(z)\right) t^{\rho} \right|.$$

Asymptotics under second order condition

Theorem

1. $\inf_{2\le j\le k} \tilde{k}^{1/2} D_{n,j} \to \infty$ *for all intermediate sequences* $k = o(\tilde{k})$
2. $$\tilde{k}^{1/2} D_{n,\lceil \tilde{k} t\rceil} \to \sup_{0<z\le 1} \left| Y(t,z) - \left(\int_0^1 x^{1/\alpha}\psi(x)\,dx \cdot z\log z + \alpha z^{1/\alpha}\psi(z)\right) t^{\rho} \right|$$
   weakly in $D(0,\infty)$.
3. *If* $\tilde{k} = o(k)$, $k = o(n)$ *then, for all* $0 < t_0 < t_1 < \infty$
   $$\inf_{t\in[t_0,t_1]} \tilde{k}^{1/2} D_{n,\lceil kt\rceil} \to \infty.$$

This suggests (but doesn't prove) that

$$k^*/\tilde{k} \to \arg\inf_{0<t<\infty} \sup_{0<z\le 1} \left| Y(t,z) - \left(\int_0^1 x^{1/\alpha}\psi(x)\,dx \cdot z\log z + \alpha z^{1/\alpha}\psi(z)\right) t^{\rho} \right|.$$

Asymptotics under second order condition

Theorem

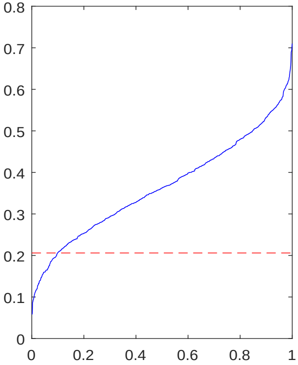
1. $\inf_{2 \le j \le k} \tilde{k}^{1/2} D_{n,j} \to \infty$ *for all intermediate sequences* $k = o(\tilde{k})$
2. $$\tilde{k}^{1/2} D_{n,\lceil \tilde{k} t \rceil} \to \sup_{0<z\le 1} \left| Y(t,z) - \left(\int_0^1 x^{1/\alpha}\psi(x)\,dx \cdot z \log z + \alpha z^{1/\alpha}\psi(z) \right) t^\rho \right|$$
   weakly in $D(0,\infty)$.
3. *If* $\tilde{k} = o(k)$, $k = o(n)$ *then, for all* $0 < t_0 < t_1 < \infty$
   $$\inf_{t \in [t_0,t_1]} \tilde{k}^{1/2} D_{n,\lceil kt \rceil} \to \infty.$$

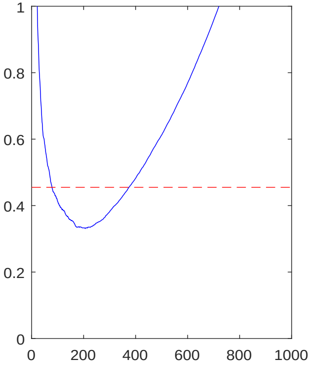
This suggests (but doesn't prove) that

$$k^*/\tilde{k} \to \arg\inf_{0<t<\infty} \sup_{0<z\le 1} \left| Y(t,z) - \left(\int_0^1 x^{1/\alpha}\psi(x)\,dx \cdot z \log z + \alpha z^{1/\alpha}\psi(z) \right) t^\rho \right|.$$

Simulations: Fréchet distribution

$F(x) = \exp(-x^{-4}), \quad x > 0$

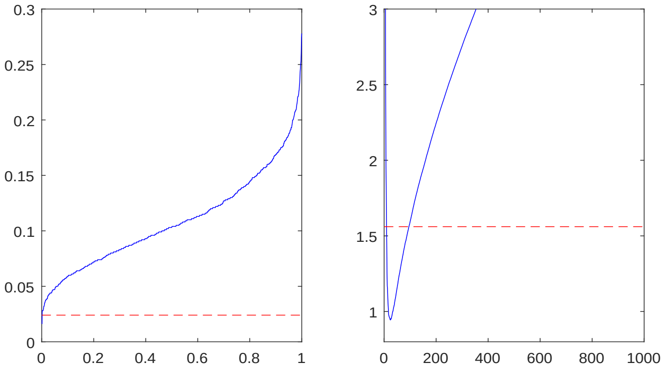




Left: qf of k^*/n for $n = 1000$; red line indicates RMSE minimizing value
Right: RMSE of Hill estimator as function of k; red line indicates RMSE of $\hat{\alpha}_{n,k^*}$

Simulations: Student's t-distribution

F Student's t cdf with 4 degrees of freedom



Left: qf of k^*/n for $n = 1000$; red line indicates RMSE minimizing value
Right: RMSE of Hill estimator as function of k; red line indicates RMSE of $\hat{\alpha}_{n,k^*}$

Loss of efficiency

Increase of RMSE and standard deviation relative to Hill estimator with deterministic k minimizing the RMSE; sample size $n = 1000$

		distance minimization		Lepskii's method
F	α	RMSE	SD	RMSE
Frechet	1	41%	22%	12%
	5	37%	14%	12%
t	1	32%	30%	15%
	4	63%	-28%	14%
	10	49%	-62%	30%
Stable	1/2	37%	13%	30%
log-gamma	3	35%	-32%	9%

Linear preferential attachment networks

LPAN are oriented graphs successively built starting from a core network; in each step one of the following randomly chosen procedures is applied

(a) add new node and edge from this node to an existing node w; latter is chosen with probability proportional to number of existing incoming edges of w plus a constant δ_{in};

(b) add new edge from existing node v to existing node w; pair is chosen with probability proportional to (number of existing outgoing edges of v plus a constant δ_{out}) $\times$ (number of existing incoming edges of w plus a constant δ_{in});

(c) add new node and edge from an existing node v this node; v is chosen with probability proportional to number of existing outgoing edges of v plus a constant δ_{out}

Linear preferential attachment networks

LPAN are oriented graphs successively built starting from a core network; in each step one of the following randomly chosen procedures is applied

(a) add new node and edge from this node to an existing node w; latter is chosen with probability proportional to number of existing incoming edges of w plus a constant δ_{in};

(b) add new edge from existing node v to existing node w; pair is chosen with probability proportional to (number of existing outgoing edges of v plus a constant δ_{out}) $\times$ (number of existing incoming edges of w plus a constant δ_{in});

(c) add new node and edge from an existing node v this node; v is chosen with probability proportional to number of existing outgoing edges of v plus a constant δ_{out}

Linear preferential attachment networks

LPAN are oriented graphs successively built starting from a core network;
in each step one of the following randomly chosen procedures is applied

(a) add new node and edge from this node to an existing node w;
latter is chosen with probability proportional to number of existing incoming edges of w plus a constant δ_{in};

(b) add new edge from existing node v to existing node w;
pair is chosen with probability proportional to (number of existing outgoing edges of v plus a constant δ_{out}) $\times$ (number of existing incoming edges of w plus a constant δ_{in});

(c) add new node and edge from an existing node v this node;
v is chosen with probability proportional to number of existing outgoing edges of v plus a constant δ_{out}

Linear preferential attachment networks

LPAN are oriented graphs successively built starting from a core network; in each step one of the following randomly chosen procedures is applied

(a) add new node and edge from this node to an existing node w; latter is chosen with probability proportional to number of existing incoming edges of w plus a constant δ_{in};

(b) add new edge from existing node v to existing node w; pair is chosen with probability proportional to (number of existing outgoing edges of v plus a constant δ_{out}) $\times$ (number of existing incoming edges of w plus a constant δ_{in});

(c) add new node and edge from an existing node v this node; v is chosen with probability proportional to number of existing outgoing edges of v plus a constant δ_{out}

Asymptotics of linear preferential attachment networks

Let

n: total number of nodes

$n_i^{(in)}$: number of nodes with i incoming edges

$n_i^{(out)}$: number of nodes with i outgoing edges

Ballobás et al. (2003):

$(n_i^{(in)}/n)_{i\in\mathbb{N}_0}$, $(n_i^{(out)}/n)_{i\in\mathbb{N}_0}$ converge to pmf of distribution with Pareto type tail; exponents $\alpha^{(in)}, \alpha^{(out)}$ can be calculated from probabilities of three procedures and $\delta_{in}, \delta_{out}$

(see Samorodnitsky et al. (2016) and Wang & Resnick (2016) for results on joint multivariate regular variation)

In the following simulations, in-degrees are observed;
note that observations are not iid.

Asymptotics of linear preferential attachment networks

Let

n: total number of nodes

$n_i^{(in)}$: number of nodes with i incoming edges

$n_i^{(out)}$: number of nodes with i outgoing edges

Ballobás et al. (2003):

$(n_i^{(in)}/n)_{i\in\mathbb{N}_0}$, $(n_i^{(out)}/n)_{i\in\mathbb{N}_0}$ converge to pmf of distribution with Pareto type tail;
exponents $\alpha^{(in)}, \alpha^{(out)}$ can be calculated from probabilities of three procedures and $\delta_{in}, \delta_{out}$

(see Samorodnitsky et al. (2016) and Wang & Resnick (2016) for results on joint multivariate regular variation)

In the following simulations, in-degrees are observed;
note that observations are not iid.

Asymptotics of linear preferential attachment networks

Let

n: total number of nodes

$n_i^{(in)}$: number of nodes with i incoming edges

$n_i^{(out)}$: number of nodes with i outgoing edges

Ballobás et al. (2003):

$(n_i^{(in)}/n)_{i\in\mathbb{N}_0}$, $(n_i^{(out)}/n)_{i\in\mathbb{N}_0}$ converge to pmf of distribution with Pareto type tail;
exponents $\alpha^{(in)}, \alpha^{(out)}$ can be calculated from probabilities of three procedures and $\delta_{in}, \delta_{out}$

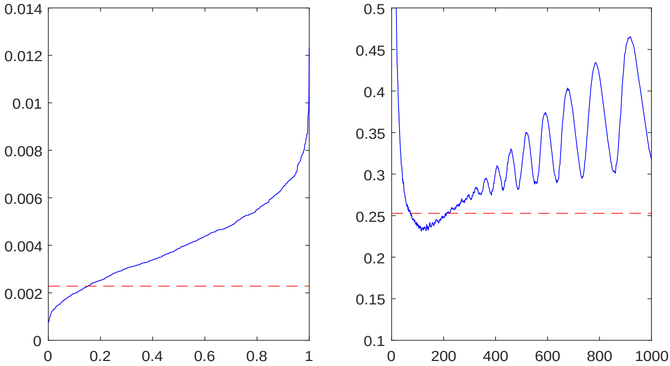
(see Samorodnitsky et al. (2016) and Wang & Resnick (2016) for results on joint multivariate regular variation)

In the following simulations, in-degrees are observed;
note that observations are not iid.

Simulations: LPAN

Model: probability of procedures (a)/(b)/(c): 0.3 / 0.5 / 0.2

$\delta_{in} = 2, \quad \delta_{out} = 1 \qquad (\Rightarrow \alpha = 2.5)$



Left: qf of k^*/n for $n = 50,000$; red line indicates RMSE minimizing value

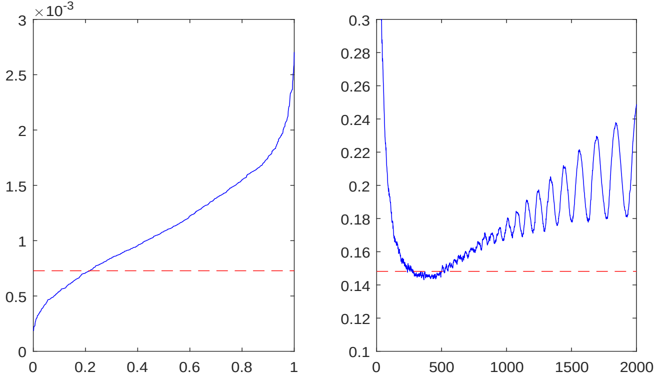
Right: RMSE of Hill estimator as function of k; red line indicates RMSE of $\hat{\alpha}_{n,k^*}$

increase of RMSE $\approx$ 9% (relative to optimal fixed k)

Simulations: LPAN (cont.)

Model: probability of procedures (a)/(b)/(c): 0.3 / 0.5 / 0.2

$\delta_{in} = 2, \quad \delta_{out} = 1 \qquad (\Rightarrow \alpha = 2.5)$



Left: qf of k^*/n for $n = 500,000$; red line indicates RMSE minimizing value
Right: RMSE of Hill estimator as function of k; red line indicates RMSE of $\hat{\alpha}_{n,k^*}$

increase of RMSE $\approx 4\%$ (relative to optimal fixed k)

Simulations: LPAN (cont.)

Q.: Why does minimum distance selection perform so much better for LPAN data than for iid data under second order condition?

Possible answers: Because of

- large sample size,
- discrete data,
- dependence,
- conceptual difference to iid setting?

Simulations: LPAN (cont.)

Q.: Why does minimum distance selection perform so much better for LPAN data than for iid data under second order condition?

Possible answers: Because of

- large sample size,
- discrete data,
- dependence,
- conceptual difference to iid setting:

Simulations: LPAN (cont.)

Q.: Why does minimum distance selection perform so much better for LPAN data than for iid data under second order condition?

Possible answers: Because of

- large sample size, but e.g. for iid Cauchy much worse behavior
- discrete data,
- dependence,
- conceptual difference to iid setting:

Simulations: LPAN (cont.)

Q.: Why does minimum distance selection perform so much better for LPAN data than for iid data under second order condition?

Possible answers: Because of

- large sample size, but e.g. for iid Cauchy much worse behavior
- discrete data, but e.g. for iid discretized Fréchet much worse behavior
- dependence,
- conceptual difference to iid setting:

  in iid setting, α has same clear meaning as exponent of regular variation of $1-F$ for all n

  in LPAN, for any fixed n, distribution of in-degrees does not have a power tail, i.e. α is meaningful only for $n \to \infty$!
  For fixed n, there is no true α. Hence calculated RMSE has a completely different meaning than in an iid setting.
  Thus, here the RMSE may be mainly caused by difference between cdf of in-degrees and limit cdf, not by a feature of the estimators.

Simulations: LPAN (cont.)

Q.: Why does minimum distance selection perform so much better for LPAN data than for iid data under second order condition?

Possible answers: Because of

- large sample size, but e.g. for iid Cauchy much worse behavior
- discrete data, but e.g. for iid discretized Fréchet much worse behavior
- dependence, maybe, but why?
- conceptual difference to iid setting:

  in iid setting, α has same clear meaning as exponent of regular variation of $1 - F$ for all n

  in LPAN, for any fixed n, distribution of in-degrees does not have a power tail, i.e. α is meaningful only for $n \to \infty$!
  For fixed n, there is no true α. Hence calculated RMSE has a completely different meaning than in an iid setting.
  Thus, here the RMSE may be mainly caused by difference between cdf of in-degrees and limit cdf, not by a feature of the estimators.

Simulations: LPAN (cont.)

Q.: Why does minimum distance selection perform so much better for LPAN data than for iid data under second order condition?

Possible answers: Because of

- large sample size, but e.g. for iid Cauchy much worse behavior
- discrete data, but e.g. for iid discretized Fréchet much worse behavior
- dependence, maybe, but why?
- conceptual difference to iid setting:

  in iid setting, α has same clear meaning as exponent of regular variation of $1 - F$ for all n

  in LPAN, for any fixed n, distribution of in-degrees does not have a power tail, i.e. α is meaningful only for $n \to \infty$!
  For fixed n, there is no true α. Hence calculated RMSE has a completely different meaning than in an iid setting.
  Thus, here the RMSE may be mainly caused by difference between cdf of in-degrees and limit cdf, not by a feature of the estimators.

Simulations: LPAN (cont.)

Q.: Why does minimum distance selection perform so much better for LPAN data than for iid data under second order condition?

Possible answers: Because of

- large sample size, but e.g. for iid Cauchy much worse behavior
- discrete data, but e.g. for iid discretized Fréchet much worse behavior
- dependence, maybe, but why?
- conceptual difference to iid setting:

  in iid setting, α has same clear meaning as exponent of regular variation of $1 - F$ for all n

  in LPAN, for any fixed n, distribution of in-degrees does not have a power tail, i.e. α is meaningful only for $n \to \infty$!
  For fixed n, there is no true α. Hence calculated RMSE has a completely different meaning than in an iid setting.
  Thus, here the RMSE may be mainly caused by difference between cdf of in-degrees and limit cdf, not by a feature of the estimators.

Simulations: LPAN (cont.)

Q.: Why does minimum distance selection perform so much better for LPAN data than for iid data under second order condition?

Possible answers: Because of

- large sample size, but e.g. for iid Cauchy much worse behavior
- discrete data, but e.g. for iid discretized Fréchet much worse behavior
- dependence, maybe, but why?
- conceptual difference to iid setting:

  in iid setting, α has same clear meaning as exponent of regular variation of $1 - F$ for all n

  in LPAN, for any fixed n, distribution of in-degrees does not have a power tail, i.e. α is meaningful only for $n \to \infty$!
  For fixed n, there is no true α. Hence calculated RMSE has a completely different meaning than in an iid setting.
  Thus, here the RMSE may be mainly caused by difference between cdf of in-degrees and limit cdf, not by a feature of the estimators.

Thank you for your attention!