New issues in extremes: imperfect extremes, extremal clustering in high dimension, causality and privacy in extreme value analysis

Gennady Samorodnitsky

Importance of extreme value analysis

Importance of extreme value analysis

- Everyone is aware of importance of understanding extremes.

Importance of extreme value analysis

- Everyone is aware of importance of understanding extremes.
- Example: extreme weather conditions:

Importance of extreme value analysis

- Everyone is aware of importance of understanding extremes.
- Example: extreme weather conditions:
  - heat waves,

Importance of extreme value analysis

- Everyone is aware of importance of understanding extremes.
- Example: extreme weather conditions:
  - heat waves,
  - periods of extreme cold,

Importance of extreme value analysis

- Everyone is aware of importance of understanding extremes.
- Example: extreme weather conditions:
  - heat waves,
  - periods of extreme cold,
  - increase in the number and intensity of hurricanes,

Importance of extreme value analysis

- Everyone is aware of importance of understanding extremes.
- Example: extreme weather conditions:
  - heat waves,
  - periods of extreme cold,
  - increase in the number and intensity of hurricanes,
  - record precipitation resulting in unprecedented floods.

- Extreme value analysis has long and distinguished history.

- Extreme value analysis has long and distinguished history.
- New issues are arising in extreme value analysis, related to:

- Extreme value analysis has long and distinguished history.
- New issues are arising in extreme value analysis, related to:
  - **big data**: curse of dimensionality, damaged extremes, ...

- Extreme value analysis has long and distinguished history.
- New issues are arising in extreme value analysis, related to:
  - **big data**: curse of dimensionality, damaged extremes, ...
  - **algorithms and machine learning**: clustering, causality, privacy, ...

- Extreme value analysis has long and distinguished history.
- New issues are arising in extreme value analysis, related to:
  - **big data**: curse of dimensionality, damaged extremes, ...
  - **algorithms and machine learning**: clustering, causality, privacy, ...
- How to deal with these issues in extreme value analysis?

Issues we will not have time to cover: Privacy

Issues we will not have time to cover: Privacy

- We own a data set X_r, $r \in A$.

Issues we will not have time to cover: Privacy

- We own a data set X_r, $r \in A$.
- An agent wants to estimate the mean and/or the median of the data.

Issues we will not have time to cover: Privacy

- We own a data set X_r, $r \in A$.
- An agent wants to estimate the mean and/or the median of the data.
- Can we release the data to the agent in a useful form while satisfying certain privacy requirements?

Issues we will not have time to cover: Privacy

- We own a data set X_r, $r \in A$.
- An agent wants to estimate the mean and/or the median of the data.
- Can we release the data to the agent in a useful form while satisfying certain privacy requirements?
- Example: not reveal clearly if a particular person is in set A.

- Mathematical notions: differential privacy, local differential privacy, ...

- Mathematical notions: differential privacy, local differential privacy, ...
- Typical algorithms: truncate data, add noise before release.

- Mathematical notions: differential privacy, local differential privacy, ...
- Typical algorithms: truncate data, add noise before release.
- If the agent wants to estimate extremal characteristics in the data, such algorithms may be useless.

Issues we will not have time to cover: Potential outcomes

Issues we will not have time to cover: Potential outcomes

- The goal: study the effect of a new treatment.

Issues we will not have time to cover: Potential outcomes

- The goal: study the effect of a new treatment.
- Some individuals are given the treatment, some placebo.

Issues we will not have time to cover: Potential outcomes

- The goal: study the effect of a new treatment.
- Some individuals are given the treatment, some placebo.
- Random assignment mechanism: n individuals, $D_i = 1$ or 0, if ith individual is given the treatment or placebo.

Issues we will not have time to cover: Potential outcomes

- The goal: study the effect of a new treatment.
- Some individuals are given the treatment, some placebo.
- Random assignment mechanism: n individuals, $D_i = 1$ or 0, if ith individual is given the treatment or placebo.
- Covariates $X_1, \ldots, X_n$;

Issues we will not have time to cover: Potential outcomes

- The goal: study the effect of a new treatment.
- Some individuals are given the treatment, some placebo.
- Random assignment mechanism: n individuals, $D_i = 1$ or 0, if ith individual is given the treatment or placebo.
- Covariates $X_1, \ldots, X_n$; $e(x) = P(D_i = 1 | X_i = x)$.

- Y_i response of interest.

- Y_i response of interest.

  Need to estimate: $\mu_{\text{treat}} = E(Y_i|\text{treatment})$.

- Y_i response of interest.

  Need to estimate: $\mu_{\text{treat}} = E(Y_i|\text{treatment})$.

- The IPW estimator:

$$\hat{\mu}_{\text{treat}} = \frac{1}{n}\sum_{i=1}^{n} \frac{D_i}{e(X_i)} Y_i.$$

- Y_i response of interest.

  Need to estimate: $\mu_{\text{treat}} = E(Y_i|\text{treatment})$.

- The IPW estimator:

$$\hat{\mu}_{\text{treat}} = \frac{1}{n} \sum_{i=1}^{n} \frac{D_i}{e(X_i)} Y_i.$$

- Extremes appear if $e(X_i)$ can be close to 0.

- We will work in the context of heavy-tailed extremes.

- We will work in the context of heavy-tailed extremes.
- Much of the discussion can be naturally translated to light-tailed extremes.

- We will work in the context of heavy-tailed extremes.
- Much of the discussion can be naturally translated to light-tailed extremes.
- The context: regular variation, univariate and multivariate.

Regular variation

Random variable X: regularly varying right tail, exponent $\alpha > 0$ if

$$\lim_{x \to \infty} \frac{P(X > tx)}{P(X > x)} = t^{-\alpha}, \text{ any } t > 0.$$

Regular variation

Random variable X: regularly varying right tail, exponent $\alpha > 0$ if

$$\lim_{x \to \infty} \frac{P(X > tx)}{P(X > x)} = t^{-\alpha}, \text{ any } t > 0.$$

X has balanced regularly varying tail, exponent $\alpha > 0$ if

Regular variation

Random variable X: regularly varying right tail, exponent $\alpha > 0$ if

$$\lim_{x \to \infty} \frac{P(X > tx)}{P(X > x)} = t^{-\alpha}, \text{ any } t > 0.$$

X has balanced regularly varying tail, exponent $\alpha > 0$ if

1. $|X|$ has regularly varying right tail, exponent $\alpha > 0$,

Regular variation

Random variable X: regularly varying right tail, exponent $\alpha > 0$ if

$$\lim_{x\to\infty} \frac{P(X > tx)}{P(X > x)} = t^{-\alpha}, \text{ any } t > 0.$$

X has balanced regularly varying tail, exponent $\alpha > 0$ if

1. $|X|$ has regularly varying right tail, exponent $\alpha > 0$,
2. tail balance:

$$\lim_{x\to\infty} \frac{P(X > x)}{P(|X| > x)} \text{ exists.}$$

Random vector $\mathbf{X} = (X_1, \ldots, X_d)$ has regularly varying tails, exponent $\alpha > 0$ if

Random vector $\mathbf{X} = (X_1, \ldots, X_d)$ has regularly varying tails, exponent $\alpha > 0$ if

1. $\|\mathbf{X}\|$ has regularly varying right tail, exponent $\alpha > 0$,

Random vector $\mathbf{X} = (X_1, \ldots, X_d)$ has regularly varying tails, exponent $\alpha > 0$ if

1. $\|\mathbf{X}\|$ has regularly varying right tail, exponent $\alpha > 0$,
2. stabilization of the directional distribution: as $x \to \infty$,

$$P\big(\mathbf{X}/\|\mathbf{X}\| \in \cdot \,\big|\, \|\mathbf{X}\| > x\big) \Rightarrow \Gamma(\cdot) \quad \text{weakly on } S_{d-1}.$$

Random vector $\mathbf{X} = (X_1, \ldots, X_d)$ has regularly varying tails, exponent $\alpha > 0$ if

1. $\|\mathbf{X}\|$ has regularly varying right tail, exponent $\alpha > 0$,
2. stabilization of the directional distribution: as $x \to \infty$,

$$P\big(\mathbf{X}/\|\mathbf{X}\| \in \cdot \,\big|\, \|\mathbf{X}\| > x\big) \Rightarrow \Gamma(\cdot) \quad \text{weakly on } S_{d-1}.$$

- Γ: the spectral measure of $\mathbf{X}$.

- The tail exponent α describes how heavy tails are.

- The tail exponent α describes how heavy tails are.
- The spectral measure Γ describes the likely directions of the extremes.

- The tail exponent α describes how heavy tails are.
- The spectral measure Γ describes the likely directions of the extremes.
- Two of most important tasks of extreme value analysis:

- The tail exponent α describes how heavy tails are.
- The spectral measure Γ describes the likely directions of the extremes.
- Two of most important tasks of extreme value analysis:

  estimation of the tail exponent and the spectral measure from data

MIssing extremes

MIssing extremes

- Suppose we are given 1-dimensional observations; but several of largest values were removed.

MIssing extremes

- Suppose we are given 1-dimensional observations;
  but several of largest values were removed.

- Can we still estimate the right tail of the observations?

MIssing extremes

- Suppose we are given 1-dimensional observations;
  but several of largest values were removed.
- Can we still estimate the right tail of the observations?
- **Examples**: malicious actions, human lifetimes, ...

- $X_1, X_2, \ldots, X_n$: i.i.d., regularly varying right tail.

- $X_1, X_2, \ldots, X_n$: i.i.d., regularly varying right tail.
- A common estimator of the tail exponent α:

- $X_1, X_2, \ldots, X_n$: i.i.d., regularly varying right tail.
- A common estimator of the tail exponent α: Hill estimator.

- $X_1, X_2, \ldots, X_n$: i.i.d., regularly varying right tail.
- A common estimator of the tail exponent α: Hill estimator.
- Order the observations: $X_{(1)} \geq X_{(2)} \geq \cdots \geq X_{(n)}$.

- $X_1, X_2, \ldots, X_n$: i.i.d., regularly varying right tail.
- A common estimator of the tail exponent α: Hill estimator.
- Order the observations: $X_{(1)} \geq X_{(2)} \geq \cdots \geq X_{(n)}$.
- Choose $1 \leq k < n$ and construct an estimator

$$H_n(k) = \frac{1}{k} \sum_{i=1}^{k} \log X_{(i)} - \log X_{(k+1)}.$$

- If $k = k_n \to \infty$, $k_n/n \to 0$, then

$$H_n(k_n) \to 1/\alpha \quad \text{in probability.}$$

- If $k = k_n \to \infty$, $k_n/n \to 0$, then

$$H_n(k_n) \to 1/\alpha \quad \text{in probability.}$$

- Asymptotic normality of the Hill estimator also holds under second order regular variation.

- If $k = k_n \to \infty$, $k_n/n \to 0$, then

$$H_n(k_n) \to 1/\alpha \quad \text{in probability.}$$

- Asymptotic normality of the Hill estimator also holds under second order regular variation.

- F: the cdf of the observations, $F^{\leftarrow}$: the generalized inverse.

- If $k = k_n \to \infty$, $k_n/n \to 0$, then

$$H_n(k_n) \to 1/\alpha \quad \text{in probability.}$$

- Asymptotic normality of the Hill estimator also holds under second order regular variation.

- F: the cdf of the observations, $F^{\leftarrow}$: the generalized inverse.

- Quantile function: $U(t) = F^{\leftarrow}(1 - 1/t)$, $t > 1$;

- If $k = k_n \to \infty$, $k_n/n \to 0$, then

$$H_n(k_n) \to 1/\alpha \quad \text{in probability.}$$

- Asymptotic normality of the Hill estimator also holds under second order regular variation.

- F: the cdf of the observations, $F^{\leftarrow}$: the generalized inverse.

- Quantile function: $U(t) = F^{\leftarrow}(1 - 1/t)$, $t > 1$;
  it is regularly varying with exponent $1/\alpha$.

Second-order regular variation

Second-order regular variation

Assume that:

Second-order regular variation

Assume that:

- There is $\rho \leq 0$ and $A : (0, \infty) \to \mathbb{R}$ such that

$$\lim_{t \to \infty} \frac{\log U(tx) - \log U(t) - \alpha^{-1} \log x}{A(t)} = \frac{x^{\rho} - 1}{\rho}, \quad \text{all } x \geq 1.$$

Second-order regular variation

Assume that:

- There is $\rho \leq 0$ and $A : (0, \infty) \to \mathbb{R}$ such that

$$\lim_{t \to \infty} \frac{\log U(tx) - \log U(t) - \alpha^{-1} \log x}{A(t)} = \frac{x^{\rho} - 1}{\rho}, \quad \text{all } x \geq 1.$$

- In Hill estimator:

$$\sqrt{k_n} A(n/k_n) \to \lambda \in \mathbb{R}.$$

- Under 2nd order regular variation:

$$\sqrt{k_n}\big(H_n(k_n) - 1/\alpha\big) \Rightarrow N(\big(\lambda/(1-\rho), 1/\alpha^2\big).$$

- Under 2nd order regular variation:

$$\sqrt{k_n}\big(H_n(k_n) - 1/\alpha\big) \Rightarrow N(\big(\lambda/(1-\rho), 1/\alpha^2\big).$$

- Hill estimator is asymptotically normal, with asymptotic bias.

- Under 2nd order regular variation:

$$\sqrt{k_n}\big(H_n(k_n) - 1/\alpha\big) \Rightarrow N(\big(\lambda/(1-\rho), 1/\alpha^2\big).$$

- Hill estimator is asymptotically normal, with asymptotic bias.

- Suppose now that several upper order statistics are missing.

- Under 2nd order regular variation:

$$\sqrt{k_n}\big(H_n(k_n) - 1/\alpha\big) \Rightarrow N(\big(\lambda/(1-\rho), 1/\alpha^2\big).$$

- Hill estimator is asymptotically normal, with asymptotic bias.
- Suppose now that several upper order statistics are missing.
- Unaware of that we construct Hill estimator.

- Under 2nd order regular variation:

$$\sqrt{k_n}\big(H_n(k_n) - 1/\alpha\big) \Rightarrow N(\big(\lambda/(1-\rho), 1/\alpha^2\big).$$

- Hill estimator is asymptotically normal, with asymptotic bias.
- Suppose now that several upper order statistics are missing.
- Unaware of that we construct Hill estimator.
- What does Hill estimator show?

- Suppose $[\delta k_n]$ upper order statistics are missing; $\delta = 0$ a possibility.

- Suppose $[\delta k_n]$ upper order statistics are missing; $\delta = 0$ a possibility.

- Can we still estimate α and unknown δ?

- Suppose $[\delta k_n]$ upper order statistics are missing; $\delta = 0$ a possibility.
- Can we still estimate α and unknown δ?
- Assume second-order regular variation condtions hold.

- Suppose $[\delta k_n]$ upper order statistics are missing; $\delta = 0$ a possibility.
- Can we still estimate α and unknown δ?
- Assume second-order regular variation condtions hold.
- We evaluate Hill estimator at θk_n remaining upper order statistics.

- Original observations: $X_1, \ldots, X_n$.

- Original observations: $X_1, \ldots, X_n$.
- Order statistics: $X_{(1)} \geq X_{(2)} \geq \cdots \geq X_{(n)}$.

- Original observations: $X_1, \ldots, X_n$.
- Order statistics: $X_{(1)} \geq X_{(2)} \geq \cdots \geq X_{(n)}$.
- Observed order statistics: $X_{([\delta k_n]+1)} \geq X_{([\delta k_n]+2)} \geq \cdots \geq X_{(n)}$.

- Original observations: $X_1, \ldots, X_n$.
- Order statistics: $X_{(1)} \geq X_{(2)} \geq \cdots \geq X_{(n)}$.
- Observed order statistics: $X_{([\delta k_n]+1)} \geq X_{([\delta k_n]+2)} \geq \cdots \geq X_{(n)}$.
- The Hill Estimator Without Extremes (HEWE) process:

$$H_n(k_n;\theta) = \frac{1}{\lfloor \theta k_n \rfloor} \sum_{i=1}^{\lfloor \theta k_n \rfloor} \log X_{(\lfloor \delta k_n \rfloor + i)} - \log X_{(\lfloor \delta k_n \rfloor + \lfloor \theta k_n \rfloor + 1)}$$

- Original observations: $X_1, \ldots, X_n$.
- Order statistics: $X_{(1)} \geq X_{(2)} \geq \cdots \geq X_{(n)}$.
- Observed order statistics: $X_{([\delta k_n]+1)} \geq X_{([\delta k_n]+2)} \geq \cdots \geq X_{(n)}$.
- The Hill Estimator Without Extremes (HEWE) process:

$$H_n(k_n;\theta) = \frac{1}{\lfloor \theta k_n \rfloor} \sum_{i=1}^{\lfloor \theta k_n \rfloor} \log X_{(\lfloor \delta k_n \rfloor + i)} - \log X_{(\lfloor \delta k_n \rfloor + \lfloor \theta k_n \rfloor + 1)}$$

$= 0$ if $\theta < 1/k_n$.

Theorem

Theorem

Under second-order regular variation,

$$\left(\sqrt{k_n} \left(H_n(k_n; \theta) - \alpha^{-1} g_\delta(\theta) \right) - \lambda b_{\delta,\rho}(\theta),\ \theta > 0 \right) \Rightarrow \alpha^{-1} G_\delta(\cdot)$$

weakly in $D(0, \infty)$.

Theorem

Under second-order regular variation,

$$\left(\sqrt{k_n} \left(H_n(k_n; \theta) - \alpha^{-1} g_\delta(\theta) \right) - \lambda b_{\delta,\rho}(\theta),\ \theta > 0 \right) \Rightarrow \alpha^{-1} G_\delta(\cdot)$$

weakly in $D(0, \infty)$.

$$g_\delta(\theta) = \begin{cases} 1, & \delta = 0, \\ 1 - (\delta/\theta) \log\left((\theta/\delta) + 1\right), & \delta > 0, \end{cases}$$

$$b_{\delta,\rho}(\theta) = \begin{cases} \frac{1}{1-\rho}\frac{1}{\theta^{\rho}}, & \delta = 0, \\ \frac{1+(\theta/\delta)\rho-(\theta/\delta+1)^{\rho}}{(\theta/\delta)(1-\rho)\rho}\frac{1}{(\delta+\theta)^{\rho}}, & \delta > 0, \end{cases}$$

$$b_{\delta,\rho}(\theta) = \begin{cases} \frac{1}{1-\rho}\frac{1}{\theta^\rho}, & \delta = 0, \\ \frac{1+(\theta/\delta)\rho-(\theta/\delta+1)^\rho}{(\theta/\delta)(1-\rho)\rho}\frac{1}{(\delta+\theta)^\rho}, & \delta > 0, \end{cases}$$

$$G_\delta(\theta) = \frac{1}{\theta}\int_\delta^{\delta+\theta}(1-\delta/x)dW(x),\ \theta > 0\,.$$

W the standard Brownian motion.

- For $\theta_i > 0$, $i = 1, \ldots, m$:
  $\big(H_n(k_n; \theta_i),\ i = 1, \ldots, m\big)$ asymptotically normal.

- For $\theta_i > 0$, $i = 1, \ldots, m$:
  $\big(H_n(k_n; \theta_i),\ i = 1, \ldots, m\big)$ asymptotically normal.

- Means and covariances depend on $\alpha, \delta, \rho, \lambda$.

- For $\theta_i > 0$, $i = 1, \ldots, m$:
  $\big(H_n(k_n; \theta_i),\ i = 1, \ldots, m\big)$ asymptotically normal.

- Means and covariances depend on $\alpha, \delta, \rho, \lambda$.

- α, δ: parameters of interest;

- For $\theta_i > 0$, $i = 1, \ldots, m$:
  $\big(H_n(k_n; \theta_i),\ i = 1, \ldots, m\big)$ asymptotically normal.

- Means and covariances depend on $\alpha, \delta, \rho, \lambda$.

- α, δ: parameters of interest; ρ, λ: nuisance parameters.

- For $\theta_i > 0$, $i = 1, \ldots, m$:
  $\big(H_n(k_n; \theta_i),\ i = 1, \ldots, m\big)$ asymptotically normal.

- Means and covariances depend on $\alpha, \delta, \rho, \lambda$.

- α, δ: parameters of interest; ρ, λ: nuisance parameters.

- We use Gaussian MLE assuming $\lambda = 0$.

- For $\theta_i > 0$, $i = 1, \ldots, m$:
  $\big(H_n(k_n; \theta_i),\ i = 1, \ldots, m\big)$ asymptotically normal.

- Means and covariances depend on $\alpha, \delta, \rho, \lambda$.

- α, δ: parameters of interest; ρ, λ: nuisance parameters.

- We use Gaussian MLE assuming $\lambda = 0$. This eliminates dependens on ρ as well.

- The resulting estimators are consistent and asymptotically normal even if true $\lambda \neq 0$.

- The resulting estimators are consistent and asymptotically normal even if true $\lambda \neq 0$.

- The nusiance parameters ρ, λ affect asymptotic bias.

- The resulting estimators are consistent and asymptotically normal even if true $\lambda \neq 0$.
- The nusiance parameters ρ, λ affect asymptotic bias.
- In the limiting case $\delta \to 0$ the estimator is as efficient as the Hill estimator,

- The resulting estimators are consistent and asymptotically normal even if true $\lambda \neq 0$.
- The nusiance parameters ρ, λ affect asymptotic bias.
- In the limiting case $\delta \to 0$ the estimator is as efficient as the Hill estimator, even though we are estimating δ.

- The resulting estimators are consistent and asymptotically normal even if true $\lambda \neq 0$.
- The nusiance parameters ρ, λ affect asymptotic bias.
- In the limiting case $\delta \to 0$ the estimator is as efficient as the Hill estimator, even though we are estimating δ.
- The results hold for any fixed number $m \geq 2$ of $\theta_1, \ldots, \theta_m$.

- If $X_1, \ldots, X_n$ are i.i.d. Pareto, we can take $\theta_i = \varepsilon + i/k_n$, $i = 1, \ldots, k_n$, $\varepsilon > 0$.

- If $X_1, \ldots, X_n$ are i.i.d. Pareto, we can take $\theta_i = \varepsilon + i/k_n$, $i = 1, \ldots, k_n$, $\varepsilon > 0$.

- The Gaussian MLE estimator is again consistent and asymptotically normal.

- If $X_1, \ldots, X_n$ are i.i.d. Pareto, we can take $\theta_i = \varepsilon + i/k_n$, $i = 1, \ldots, k_n$, $\varepsilon > 0$.

- The Gaussian MLE estimator is again consistent and asymptotically normal.

- Numerically, the estimator performs well even if $X_1, \ldots, X_n$ are not Pareto.

Numerical results

Numerical results

- We generate $n = 5000$ i.i.d. observations from the standard Pareto and standard Fréchet distributions.

Numerical results

- We generate $n = 5000$ i.i.d. observations from the standard Pareto and standard Fréchet distributions.

- $\alpha = 1$ in all cases.

Numerical results

- We generate $n = 5000$ i.i.d. observations from the standard Pareto and standard Fréchet distributions.
- $\alpha = 1$ in all cases.
- We choose $k_n = 200$.

Numerical results

- We generate $n = 5000$ i.i.d. observations from the standard Pareto and standard Fréchet distributions.
- $\alpha = 1$ in all cases.
- We choose $k_n = 200$.
- We remove 20, 40 and 100 extremes; $\delta = 0.1, 0.2, 0.5$.

Numerical results

- We generate $n = 5000$ i.i.d. observations from the standard Pareto and standard Fréchet distributions.
- $\alpha = 1$ in all cases.
- We choose $k_n = 200$.
- We remove 20, 40 and 100 extremes; $\delta = 0.1, 0.2, 0.5$.
- We used Procedure 1 with $m = 10$ (equally spaced θ_i) and Procedure 2 with $m = k_n$ (equally spaced θ_i)

Table: Pareto distribution, $n = 5000$, $k_n = 200$

	$\hat{\delta}_a$		$\hat{\gamma}_a$		$\rho_{\hat{\delta}_a,\hat{\gamma}_a}$	
δ_0	mean	(sd)	mean	(sd)	corr	(asy)
0.1	0.113	(0.057)	1.015	(0.143)	0.858	(0.829)
0.2	0.222	(0.104)	1.025	(0.187)	0.915	(0.894)
0.5	0.547	(0.285)	1.040	(0.309)	0.965	(0.956)

	$\hat{\delta}_b$		$\hat{\gamma}_b$		$\rho_{\hat{\delta}_b,\hat{\gamma}_b}$	
δ_0	mean	(sd)	mean	(sd)	corr	(asy)
0.1	0.104	(0.049)	1.006	(0.129)	0.841	(0.796)
0.2	0.207	(0.096)	1.010	(0.177)	0.915	(0.878)
0.5	0.515	(0.254)	1.014	(0.282)	0.962	(0.951)

Table: Fréchet distribution, $n = 5000$, $k_n = 200$

	$\hat{\delta}_a$	$\hat{\gamma}_a$	$\rho_{\hat{\delta}_a,\hat{\gamma}_a}$
δ_0	mean (sd)	mean (sd)	corr (asy)
0.1	0.106 (0.050)	0.992 (0.130)	0.829 (0.829)
0.2	0.208 (0.094)	0.993 (0.176)	0.906 (0.894)
0.5	0.535 (0.287)	1.011 (0.300)	0.961 (0.956)

	$\hat{\delta}_b$	$\hat{\gamma}_b$	$\rho_{\hat{\delta}_b,\hat{\gamma}_b}$
δ_0	mean (sd)	mean (sd)	corr (asy)
0.1	0.101 (0.045)	0.988 (0.122)	0.826 (0.796)
0.2	0.196 (0.085)	0.981 (0.165)	0.904 (0.878)
0.5	0.502 (0.252)	0.985 (0.274)	0.961 (0.951)

- Missing extremes may not be consecutive, from the largest.

- Missing extremes may not be consecutive, from the largest.
- We can still estimate number of the missing extremes.

- Missing extremes may not be consecutive, from the largest.
- We can still estimate number of the missing extremes.
- **Example** 10 out of the top 50 extremes are missing.

- Missing extremes may not be consecutive, from the largest.
- We can still estimate number of the missing extremes.
- **Example** 10 out of the top 50 extremes are missing.
- Remove artificially 40 top extremes and estimate now number of the missing extremes.

- Missing extremes may not be consecutive, from the largest.
- We can still estimate number of the missing extremes.
- **Example** 10 out of the top 50 extremes are missing.
- Remove artificially 40 top extremes and estimate now number of the missing extremes.
- Top 50 extremes now missing, estimate should be around 50.

- Missing extremes may not be consecutive, from the largest.
- We can still estimate number of the missing extremes.
- **Example** 10 out of the top 50 extremes are missing.
- Remove artificially 40 top extremes and estimate now number of the missing extremes.
- Top 50 extremes now missing, estimate should be around 50.
- Conclude that around 10 extremes were originally missing.

- In general: suppose that $\delta_0 k_n$ extremes are missing among the top $(\delta_0 + \delta_1)k_n$ extremes.

- In general: suppose that $\delta_0 k_n$ extremes are missing among the top $(\delta_0 + \delta_1)k_n$ extremes.

- Remove artificially i top remaining extremes, $i = 1, 2, \ldots$.

- In general: suppose that $\delta_0 k_n$ extremes are missing among the top $(\delta_0 + \delta_1)k_n$ extremes.
- Remove artificially i top remaining extremes, $i = 1, 2, \ldots$.
- Estimate δ (from δk_n missing top extremes).

- In general: suppose that $\delta_0 k_n$ extremes are missing among the top $(\delta_0 + \delta_1)k_n$ extremes.
- Remove artificially i top remaining extremes, $i = 1, 2, \ldots$.
- Estimate δ (from δk_n missing top extremes).
- If initially only the top $\delta_0 k_n$ extremes were missing ($\delta_1 = 0$), the plot would be close to linear.

- If the missing $\delta_0 k_n$ extremes not top consecutive extremes:

- If the missing $\delta_0 k_n$ extremes not top consecutive extremes:

  the plot close to linear once $\delta_1 k_n$ extremes are removed.

- If the missing $\delta_0 k_n$ extremes not top consecutive extremes:

  the plot close to linear once $\delta_1 k_n$ extremes are removed.

- This can be used to estimate number of original missing extremes.

- There is very high correlation in estimators of α and δ.

- There is very high correlation in estimators of α and δ.
- This makes it difficult to detect linearity after repeated estimation.

- There is very high correlation in estimators of α and δ.
- This makes it difficult to detect linearity after repeated estimation.
- It is better to fix α and estimate only δ.

- There is very high correlation in estimators of α and δ.
- This makes it difficult to detect linearity after repeated estimation.
- It is better to fix α and estimate only δ.
- This works reasonably well even when the fixed α is not quite correct.

Numerical experiments

Numerical experiments

- $n = 5000$ observations from standard Pareto and Fréchet, $\alpha = 1$, $k_n = 200$.

Numerical experiments

- $n = 5000$ observations from standard Pareto and Fréchet, $\alpha = 1$, $k_n = 200$.

- 3 setups:

Numerical experiments

- $n = 5000$ observations from standard Pareto and Fréchet, $\alpha = 1$, $k_n = 200$.
- 3 setups:
    1. No missing observations; $\delta_0 = 0$.

Numerical experiments

- $n = 5000$ observations from standard Pareto and Fréchet, $\alpha = 1$, $k_n = 200$.
- 3 setups:
  1. No missing observations; $\delta_0 = 0$.
  2. Consecutive top missing observations, $\delta_0 = 0.25$.

Numerical experiments

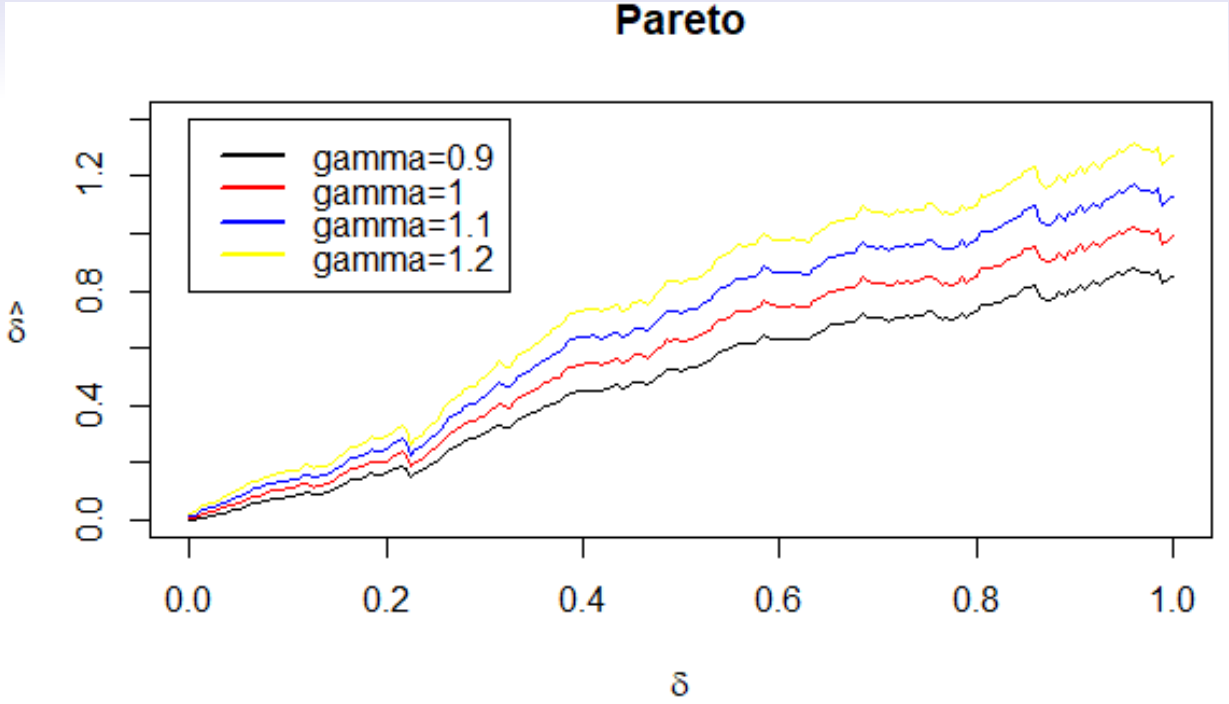
- $n = 5000$ observations from standard Pareto and Fréchet, $\alpha = 1$, $k_n = 200$.

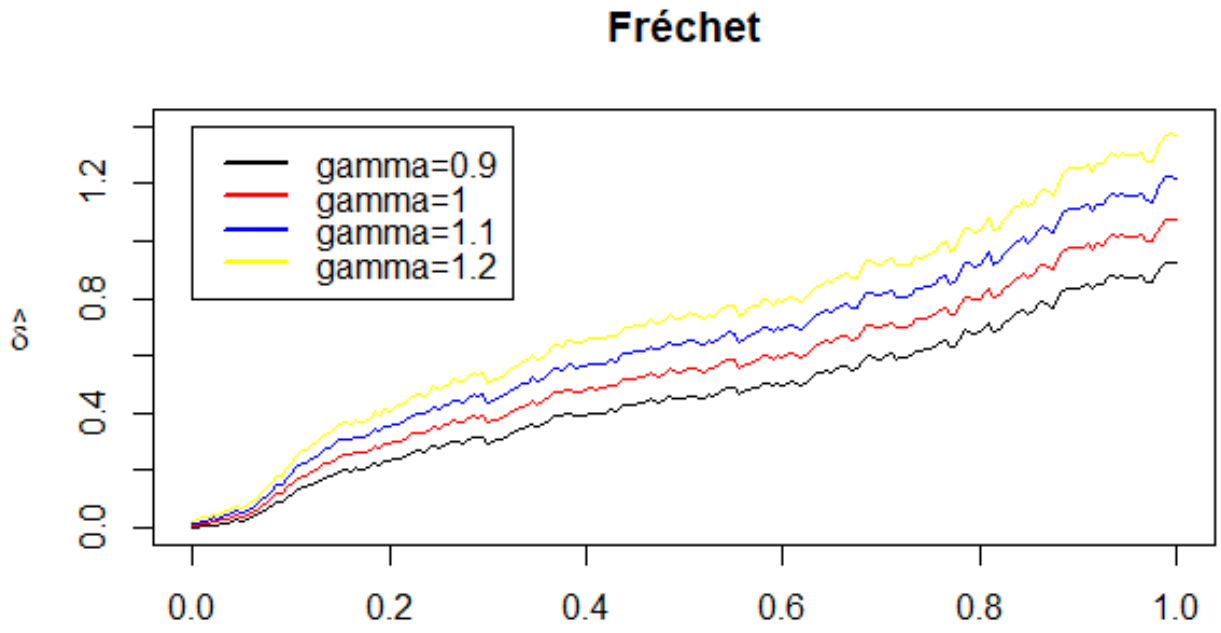
- 3 setups:

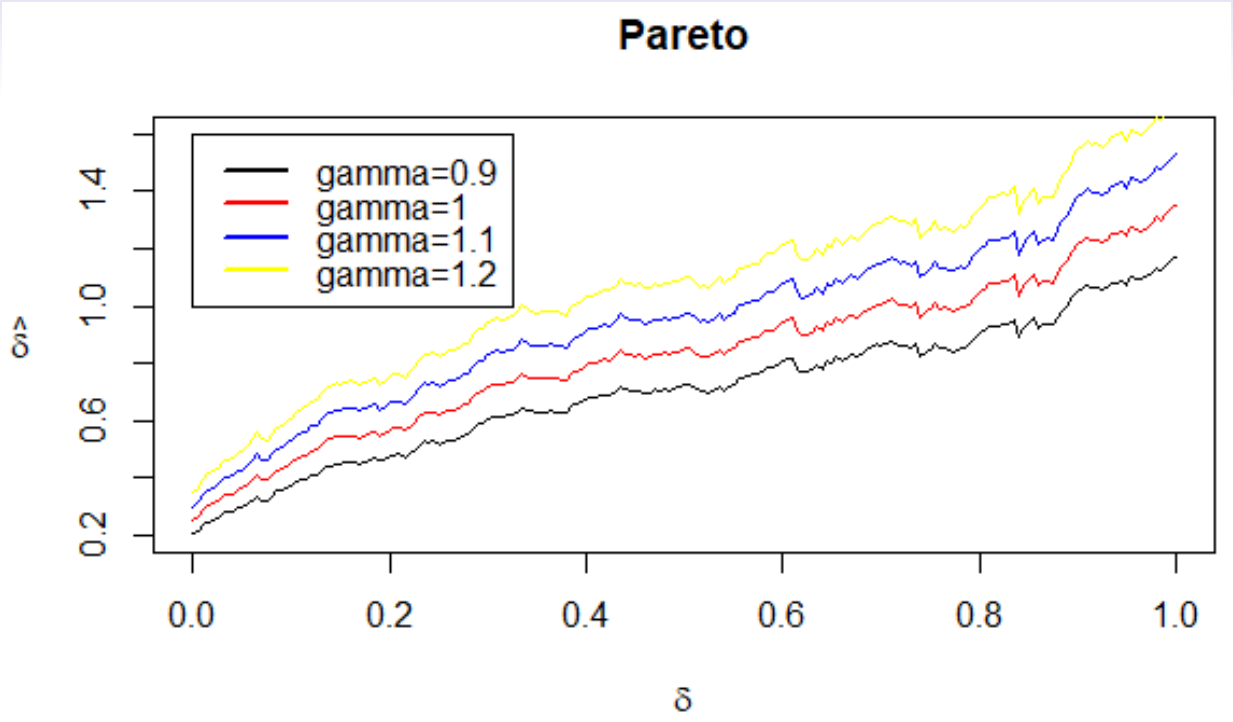
  1. No missing observations; $\delta_0 = 0$.

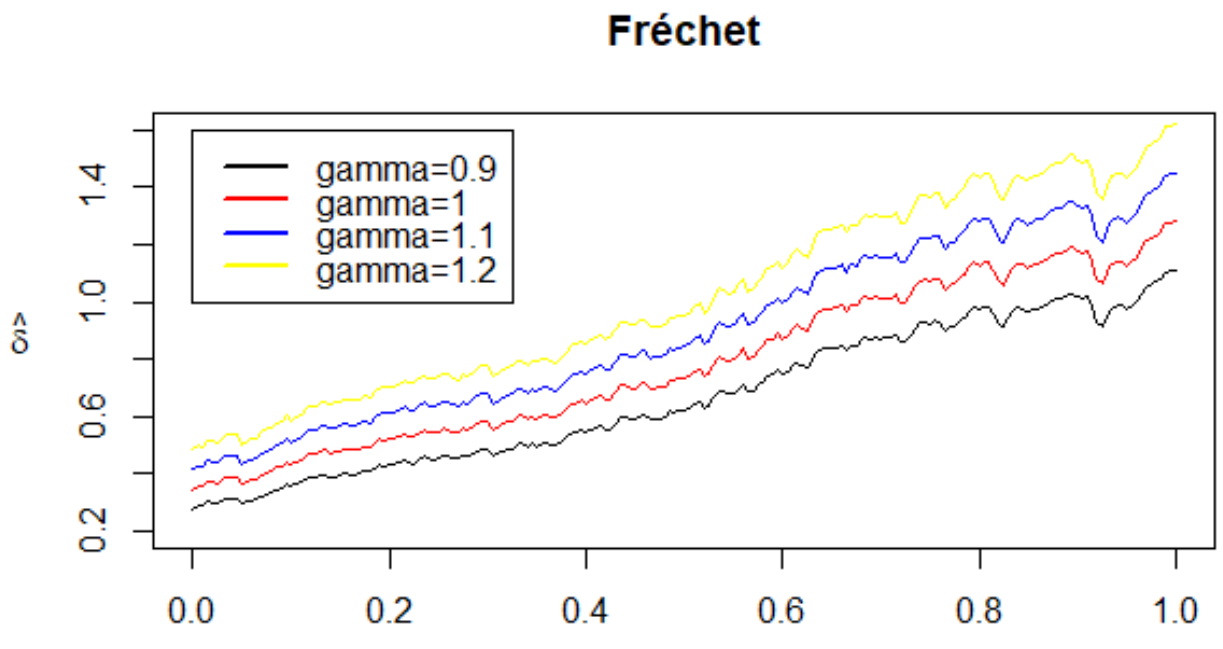
  2. Consecutive top missing observations, $\delta_0 = 0.25$.

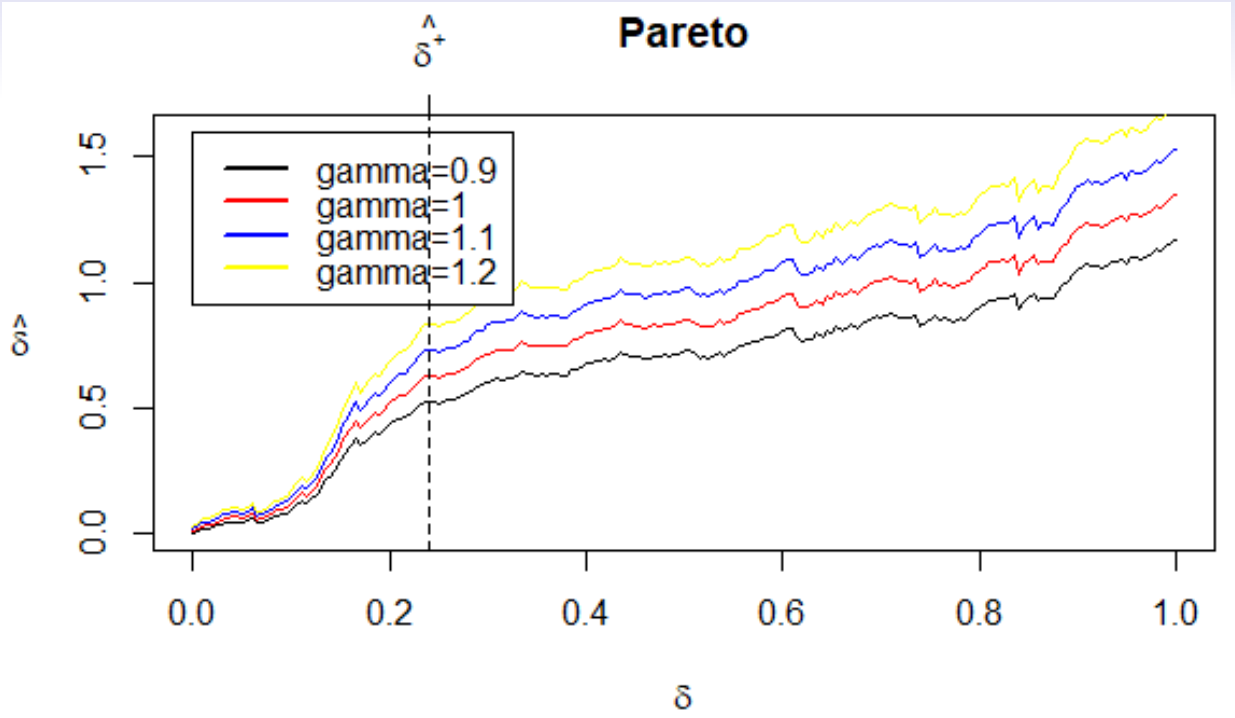
  3. $\delta_0 = 0.25$, the missing $\delta_0 k_n = 50$ missing extremes are uniformly chosen among top 100 extremes.

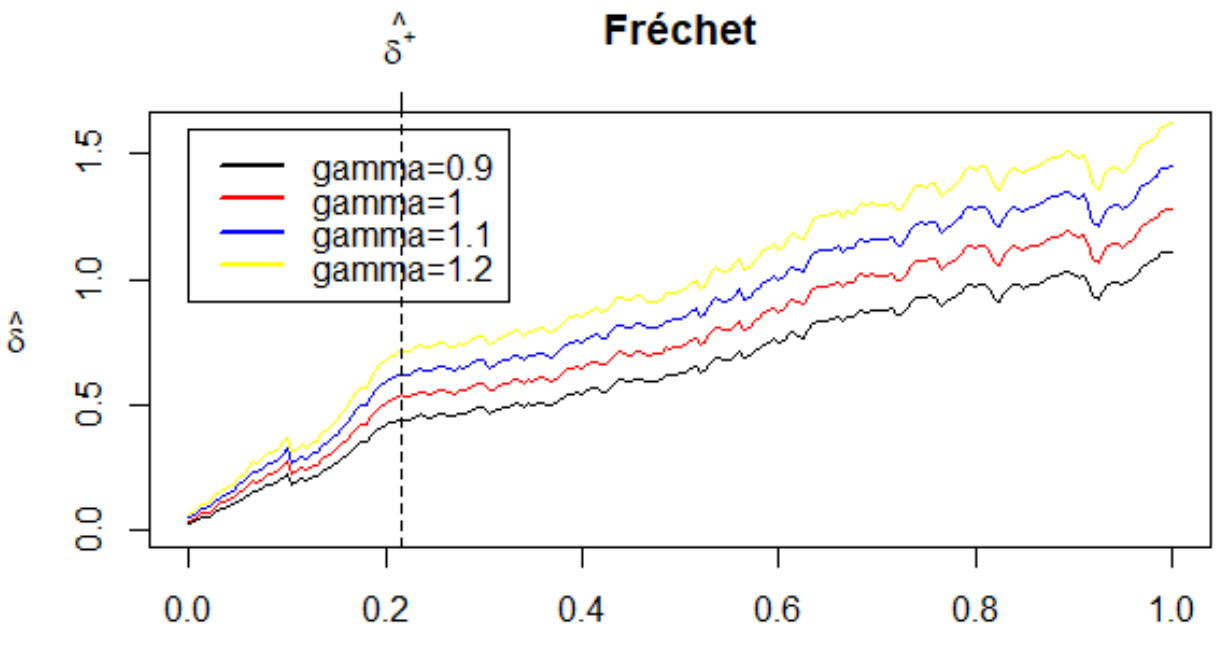

Pareto
gamma=0.9
gamma=1
gamma=1.1
gamma=1.2
0.0
0.4
0.8
1.2
0.0
0.2
0.4
0.6
0.8
1.0
δ


Fréchet
gamma=0.9
gamma=1
gamma=1.1
gamma=1.2
0.0
0.4
0.8
1.2
0.0
0.2
0.4
0.6
0.8
1.0


Pareto
gamma=0.9
gamma=1
gamma=1.1
gamma=1.2
0.2
0.6
1.0
1.4
0.0
0.2
0.4
0.6
0.8
1.0
δ


Fréchet
gamma=0.9
gamma=1
gamma=1.1
gamma=1.2
0.2
0.6
1.0
1.4
0.0
0.2
0.4
0.6
0.8
1.0


Pareto
$\hat{\delta}^+$
gamma=0.9
gamma=1
gamma=1.1
gamma=1.2
$\hat{\delta}$
0.0
0.5
1.0
1.5
0.0
0.2
0.4
0.6
0.8
1.0
δ


Fréchet
$\hat{\delta}^+$
gamma=0.9
gamma=1
gamma=1.1
gamma=1.2
$\hat{\delta}$
0.0
0.5
1.0
1.5
0.0
0.2
0.4
0.6
0.8
1.0

Support of the spectral measure and clustering

Support of the spectral measure and clustering

- $\mathbf{X}$ random vector with regularly varying tails.

Support of the spectral measure and clustering

- $\mathbf{X}$ random vector with regularly varying tails.
- Distribution of the direction of the extremes: **spectral measure**:

Support of the spectral measure and clustering

- $\mathbf{X}$ random vector with regularly varying tails.
- Distribution of the direction of the extremes: **spectral measure**:

$$P\big(\mathbf{X}/\|\mathbf{X}\| \in \cdot \big| \|\mathbf{X}\| > x\big) \Rightarrow \Gamma(\cdot) \text{ weakly on } S_{d-1}.$$

Support of the spectral measure and clustering

- $\mathbf{X}$ random vector with regularly varying tails.
- Distribution of the direction of the extremes: **spectral measure**:

$$P\big(\mathbf{X}/\|\mathbf{X}\| \in \cdot \big| \|\mathbf{X}\| > x\big) \Rightarrow \Gamma(\cdot) \text{ weakly on } S_{d-1}.$$

- Learning the spectral measure is crucial.

A generic procedure for estimating spectral measure

A generic procedure for estimating spectral measure

- Observations $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(n)}$.

A generic procedure for estimating spectral measure

- Observations $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(n)}$.
- Select threshold x and declare any $\mathbf{X}^{(i)}$ with $\|\mathbf{X}^{(i)}\| > x$ as extreme.

A generic procedure for estimating spectral measure

- Observations $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(n)}$.
- Select threshold x and declare any $\mathbf{X}^{(i)}$ with $\|\mathbf{X}^{(i)}\| > x$ as extreme.
- Random set $I_n \subset \{1, \ldots, n\}$ of extremes;

A generic procedure for estimating spectral measure

- Observations $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(n)}$.
- Select threshold x and declare any $\mathbf{X}^{(i)}$ with $\|\mathbf{X}^{(i)}\| > x$ as extreme.
- Random set $I_n \subset \{1, \ldots, n\}$ of extremes; $\text{card}(I_n) = N_n$.

A generic procedure for estimating spectral measure

- Observations $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(n)}$.
- Select threshold x and declare any $\mathbf{X}^{(i)}$ with $\|\mathbf{X}^{(i)}\| > x$ as extreme.
- Random set $I_n \subset \{1, \ldots, n\}$ of extremes; $\text{card}(I_n) = N_n$.
- Use $\mathbf{X}^{(i)}/\|\mathbf{X}^{(i)}\|$, $i \in I_n$, for estimating spectral measure.

- Parametric models of the spectral measure are restrictive.

- Parametric models of the spectral measure are restrictive.
- Estimating a measure nonparametrically is hard.

- Parametric models of the spectral measure are restrictive.
- Estimating a measure nonparametrically is hard.
- It is very hard to do in high dimensions.

- Parametric models of the spectral measure are restrictive.
- Estimating a measure nonparametrically is hard.
- It is very hard to do in high dimensions.
- Only a small part of the sample can be used (N_n out of n observations).

- Parametric models of the spectral measure are restrictive.
- Estimating a measure nonparametrically is hard.
- It is very hard to do in high dimensions.
- Only a small part of the sample can be used (N_n out of n observations).
- Normalized extremes do not have the exact spectral measure as their law.

- If the extremes are high-dimensional, the only hope is sparsity.

- If the extremes are high-dimensional, the only hope is sparsity.
- If the spectral measure lives on low-dimensional parts of S_{d-1},

- If the extremes are high-dimensional, the only hope is sparsity.
- If the spectral measure lives on low-dimensional parts of S_{d-1}, **and we could identify these low-dimensional parts,**

- If the extremes are high-dimensional, the only hope is sparsity.
- If the spectral measure lives on low-dimensional parts of S_{d-1}, **and we could identify these low-dimensional parts,** estimation would be easier.

- If the extremes are high-dimensional, the only hope is sparsity.
- If the spectral measure lives on low-dimensional parts of S_{d-1}, **and we could identify these low-dimensional parts,** estimation would be easier.
- A related issue: clustering.

- Extremes often cluster.

- Extremes often cluster.
- **If we could identify cluster centers,**

- Extremes often cluster.
- **If we could identify cluster centers,** we would only need to estimate the scatter within each cluster.

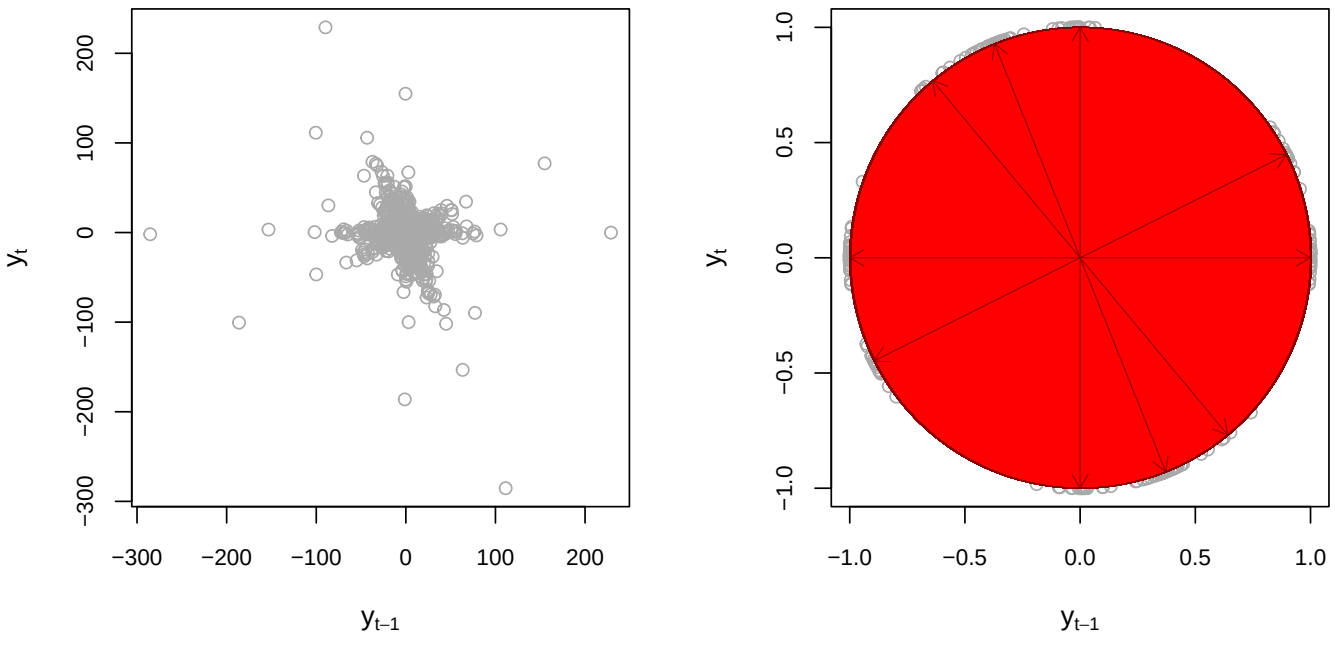
- Extremes often cluster.
- **If we could identify cluster centers,** we would only need to estimate the scatter within each cluster.
- This would make estimation of the spectral measure easier.

- Extremes often cluster.
- **If we could identify cluster centers,** we would only need to estimate the scatter within each cluster.
- This would make estimation of the spectral measure easier.
- How do we find lower-dimensional support and clustering in the spectral measure?

Clustering of extremes

Clustering of extremes

A 2-dim example with 10 clusters:



- The most natural procedure to identify clusters:

- The most natural procedure to identify clusters:
    1. Choose the extreme observations

- The most natural procedure to identify clusters:
  1. Choose the extreme observations
  2. Project the extremes onto the unit sphere

- The most natural procedure to identify clusters:
  1. Choose the extreme observations
  2. Project the extremes onto the unit sphere
  3. Apply a clustering k-means procedure on the sphere

- The most natural procedure to identify clusters:
  1. Choose the extreme observations
  2. Project the extremes onto the unit sphere
  3. Apply a clustering k-means procedure on the sphere
  4. The procedure chooses cluster centers to minimize certain average "dissimilarity" .

- The most natural procedure to identify clusters:
  1. Choose the extreme observations
  2. Project the extremes onto the unit sphere
  3. Apply a clustering k-means procedure on the sphere
  4. The procedure chooses cluster centers to minimize certain average "dissimilarity" .
- This was investigated in Janssen and Wan (2020).

Spectral clustering analysis
(Avella, Davis and S. (2924))

Spectral clustering analysis (Avella, Davis and S. (2924))

- Two main stages:

Spectral clustering analysis (Avella, Davis and S. (2924))

- Two main stages:
  1. Construct a graph with scaled extremes as verteces;

Spectral clustering analysis (Avella, Davis and S. (2924))

- Two main stages:

  1. Construct a graph with scaled extremes as verteces;

     connected components of the graph should correspond to clusters of extremal directions.

Spectral clustering analysis (Avella, Davis and S. (2924))

- Two main stages:

  1. Construct a graph with scaled extremes as verteces;

     connected components of the graph should correspond to clusters of extremal directions.

  2. Connected components of the graph can be detected using spectrum of the graph Laplacian.

- The first steps are the same:

- The first steps are the same:

  Choose the extreme observations and project them onto the unit sphere

- The first steps are the same:

  Choose the extreme observations and project them onto the unit sphere

- Two ways to construct a graph on scaled extremes $\mathbf{w}_1, \ldots, \mathbf{w}_{N_n}$:

- The first steps are the same:

  Choose the extreme observations and project them onto the unit sphere

- Two ways to construct a graph on scaled extremes $\mathbf{w}_1, \ldots, \mathbf{w}_{N_n}$:

  1. Choose a threshold $\varepsilon > 0$ and connect $\mathbf{w}_i, \mathbf{w}_j$ if $d(\mathbf{w}_i, \mathbf{w}_j) \leq \varepsilon$.

- The first steps are the same:

  Choose the extreme observations and project them onto the unit sphere

- Two ways to construct a graph on scaled extremes $\mathbf{w}_1, \ldots, \mathbf{w}_{N_n}$:

  1. Choose a threshold $\varepsilon > 0$ and connect $\mathbf{w}_i, \mathbf{w}_j$ if $d(\mathbf{w}_i, \mathbf{w}_j) \leq \varepsilon$.
  2. Choose $k \geq 1$ and connect $\mathbf{w}_i$ to $\mathbf{w}_j$ if $\mathbf{w}_j$ is among k-nearest neighbours of $\mathbf{w}_i$.

- We use the k-nearest neighbour approach as is reflects clustering better.

- We use the k-nearest neighbour approach as is reflects clustering better.

- Lack of symmetry: $\mathbf{w}_i$ may be among k-nearest neighbours of $\mathbf{w}_i$, but not vice versa.

- We use the k-nearest neighbour approach as is reflects clustering better.
- Lack of symmetry: $\mathbf{w}_i$ may be among k-nearest neighbours of $\mathbf{w}_i$, but not vice versa.
- This leads to undesirable directed graph.

- Two ways to obtain indirected graph:

- Two ways to obtain indirected graph:
  1. connect $\mathbf{w}_i$ and $\mathbf{w}_j$ if $\mathbf{w}_j$ is among k-nearest neighbours of $\mathbf{w}_i$ OR $\mathbf{w}_i$ is among k-nearest neighbours of $\mathbf{w}_j$

- Two ways to obtain indirected graph:

  1. connect $\mathbf{w}_i$ and $\mathbf{w}_j$ if $\mathbf{w}_j$ is among k-nearest neighbours of $\mathbf{w}_i$ OR $\mathbf{w}_i$ is among k-nearest neighbours of $\mathbf{w}_j$

  2. connect $\mathbf{w}_i$ and $\mathbf{w}_j$ if $\mathbf{w}_j$ is among k-nearest neighbours of $\mathbf{w}_i$ AND $\mathbf{w}_i$ is among k-nearest neighbours of $\mathbf{w}_j$

- Two ways to obtain indirected graph:
  1. connect $\mathbf{w}_i$ and $\mathbf{w}_j$ if $\mathbf{w}_j$ is among k-nearest neighbours of $\mathbf{w}_i$ OR $\mathbf{w}_i$ is among k-nearest neighbours of $\mathbf{w}_j$
  2. connect $\mathbf{w}_i$ and $\mathbf{w}_j$ if $\mathbf{w}_j$ is among k-nearest neighbours of $\mathbf{w}_i$ AND $\mathbf{w}_i$ is among k-nearest neighbours of $\mathbf{w}_j$
- The results are similar in the two cases.

- We use the graph Laplacian of the graph.

- We use the graph Laplacian of the graph.
- The weighted version seems to work a bit better.

- We use the graph Laplacian of the graph.
- The weighted version seems to work a bit better.
- The weighted adjacency matrix $W = [w_{i_1,i_2}]$:

- We use the graph Laplacian of the graph.
- The weighted version seems to work a bit better.
- The weighted adjacency matrix $W = [w_{i_1,i_2}]$:

$$w_{i_1,i_1} = \begin{cases} k(\mathbf{w}_{i_1}, \mathbf{w}_{i_2}) & \text{if } \mathbf{w}_{i_1}, \mathbf{w}_{i_2} \text{ are connected} \\ 0 & \text{otherwise.} \end{cases}$$

- We use the graph Laplacian of the graph.
- The weighted version seems to work a bit better.
- The weighted adjacency matrix $W = [w_{i_1,i_2}]$:

$$w_{i_1,i_1} = \begin{cases} k(\mathbf{w}_{i_1}, \mathbf{w}_{i_2}) & \text{if } \mathbf{w}_{i_1}, \mathbf{w}_{i_2} \text{ are connected} \\ 0 & \text{otherwise.} \end{cases}$$

k a similarity (positive kernel).

- We use the graph Laplacian of the graph.
- The weighted version seems to work a bit better.
- The weighted adjacency matrix $W = [w_{i_1,i_2}]$:

$$w_{i_1,i_1} = \begin{cases} k(\mathbf{w}_{i_1}, \mathbf{w}_{i_2}) & \text{if } \mathbf{w}_{i_1}, \mathbf{w}_{i_2} \text{ are connected} \\ 0 & \text{otherwise.} \end{cases}$$

k a similarity (positive kernel). We use $k(\mathbf{x}, \mathbf{y}) = \exp\{-\|\mathbf{x} - \mathbf{y}\|\}$.

- Weighted degree of a vertex: $d_i = \sum_{j=1}^{N_n} w_{i,j}$.

- Weighted degree of a vertex: $d_i = \sum_{j=1}^{N_n} w_{i,j}$.
- **The degree matrix**: D diagonal, with entries (d_i).

- Weighted degree of a vertex: $d_i = \sum_{j=1}^{N_n} w_{i,j}$.
- **The degree matrix**: D diagonal, with entries (d_i).
- The normalized symmetric graph Laplacian matrix:

$$L = I - D^{-1/2} W D^{-1/2},$$

I the identity matrx.

The key facts

The key facts

1. L is a symmetric nonnegative definite matrix.

The key facts

1. L is a symmetric nonnegative definite matrix.

2. The multiplicity m of the eigenvalue 0 of L equals the number of connected components $\mathcal{A}_1, \ldots, \mathcal{A}_m$ of the graph.

The key facts

1. L is a symmetric nonnegative definite matrix.
2. The multiplicity m of the eigenvalue 0 of L equals the number of connected components $\mathcal{A}_1, \ldots, \mathcal{A}_m$ of the graph.
3. The eigenspace of the eigenvalue 0 is spanned by the indicator functions of $\delta_{\mathcal{A}_1}, \ldots, \delta_{\mathcal{A}_m}$ of these components.

Comments

Comments

- The choice of $k = k_n$ is important.

Comments

- The choice of $k = k_n$ is important.
- Connected components of the graph form a noisy approximation to the true clusters.

Comments

- The choice of $k = k_n$ is important.
- Connected components of the graph form a noisy approximation to the true clusters.
- There is certain robustness of eigenvalues and eigenvectors under "modest perturbation" of a matrix.

Comments

- The choice of $k = k_n$ is important.
- Connected components of the graph form a noisy approximation to the true clusters.
- There is certain robustness of eigenvalues and eigenvectors under "modest perturbation" of a matrix.
- One looks for "small" (not only zero) eigenvalues of the Laplacian matrix.

Spectral clustering algorithm

Spectral clustering algorithm

- Compute the graph Laplacian.

Spectral clustering algorithm

- Compute the graph Laplacian.
- Decide on the number m of clusters by inspecting the smallest eigenvalues.

Spectral clustering algorithm

- Compute the graph Laplacian.
- Decide on the number m of clusters by inspecting the smallest eigenvalues.
- Construct a matrix U whose m columns are the corresponding eigenvectors.

- $\mathbf{u}_1, \ldots, \mathbf{u}_{N_n}$: the rows of this matrix, normalized to norm 1.

- $\mathbf{u}_1, \ldots, \mathbf{u}_{N_n}$: the rows of this matrix, normalized to norm 1.
- Use the m-means clustering algorithm to assign the rows to m clusters.

- $\mathbf{u}_1, \ldots, \mathbf{u}_{N_n}$: the rows of this matrix, normalized to norm 1.
- Use the m-means clustering algorithm to assign the rows to m clusters.
- Assign original points $\mathbf{w}_{i_1}, \mathbf{w}_{i_2}$ to the same cluster if $\mathbf{u}_{i_1}, \mathbf{u}_{i_2}$ are assigned to the same cluster.

- We prove that the spectral clustering algorithm correctly identifies extremal clusters in a particular model.

- We prove that the spectral clustering algorithm correctly identifies extremal clusters in a particular model.

- Numerical experiments indicate good results in a wide variety of situations.

- We prove that the spectral clustering algorithm correctly identifies extremal clusters in a particular model.
- Numerical experiments indicate good results in a wide variety of situations.
- The model: linear factor model.

Linear factor model

Linear factor model

- The model:

$$\mathbf{X}_i = A\mathbf{Z}_i, \quad i = 1, \ldots, n$$

Linear factor model

- The model:

$$\mathbf{X}_i = A\mathbf{Z}_i, \quad i = 1, \ldots, n$$

A: $d \times p$ matrix with nonnegative entries;

Linear factor model

- The model:

$$\mathbf{X}_i = A\mathbf{Z}_i, \quad i = 1, \ldots, n$$

A: $d \times p$ matrix with nonnegative entries;

$\mathbf{Z}$: p-dimensional with i.i.d. nonnegative random variables with asymptotically power tails.

- For the linear factor model the spectral measure is discrete.

- For the linear factor model the spectral measure is discrete.
- The atoms: nonzero columns of A normalized to norm 1.

- For the linear factor model the spectral measure is discrete.
- The atoms: nonzero columns of A normalized to norm 1.
- An atom corresponds to a very large component of $\mathbf{Z}$.

- For the linear factor model the spectral measure is discrete.
- The atoms: nonzero columns of A normalized to norm 1.
- An atom corresponds to a very large component of $\mathbf{Z}$.
- An extremal cluster corresponds to a large component of $\mathbf{Z}$.

- For the linear factor model the spectral measure is discrete.
- The atoms: nonzero columns of A normalized to norm 1.
- An atom corresponds to a very large component of $\mathbf{Z}$.
- An extremal cluster corresponds to a large component of $\mathbf{Z}$.
- Spectral clustering is proven to work asymptotically when $d = 2$ if $k = k_n > G \log n$, for large $G > 0$.

- In applications linear factor model is an approximation.

- In applications linear factor model is an approximation.
- A robust way to decide on a good number m of clusters:

- In applications linear factor model is an approximation.
- A robust way to decide on a good number m of clusters: *largest eigenvalues of fully connected weighted adjacency matrix*.

- In applications linear factor model is an approximation.
- A robust way to decide on a good number m of clusters: *largest eigenvalues of fully connected weighted adjacency matrix*.
- Then use this m in the spectral clustering algorithm.

Contaminated linear factor model

Contaminated linear factor model

- The model:

$$\mathbf{X}_i = A\mathbf{Z}_i + \sigma\varepsilon_i, \quad i = 1, \ldots, n$$

Contaminated linear factor model

- The model:

$$\mathbf{X}_i = A\mathbf{Z}_i + \sigma\varepsilon_i, \quad i = 1, \ldots, n$$

A: $d \times p$ matrix with nonnegative entries;

Contaminated linear factor model

- The model:

$$\mathbf{X}_i = A\mathbf{Z}_i + \sigma\varepsilon_i, \quad i = 1, \ldots, n$$

A: $d \times p$ matrix with nonnegative entries;

$\mathbf{Z}$: p-dim with i.i.d. nonnegative random variables with asymptotically power tails.

Contaminated linear factor model

- The model:

$$\mathbf{X}_i = A\mathbf{Z}_i + \sigma\varepsilon_i, \quad i = 1, \ldots, n$$

  A: $d \times p$ matrix with nonnegative entries;

  $\mathbf{Z}$: p-dim with i.i.d. nonnegative random variables with asymptotically power tails.

- $\sigma > 0$, (ε_i): i.i.d., d-dim, $\varepsilon \stackrel{\mathrm{d}}{=} Y\mathbf{G}$,
  $\mathbf{G}$ d-dim $N(0, I)$, independent of Pareto (1) Y.

- Contamination introduces a continuous (uniform) component in the spectral measure.

- Contamination introduces a continuous (uniform) component in the spectral measure.

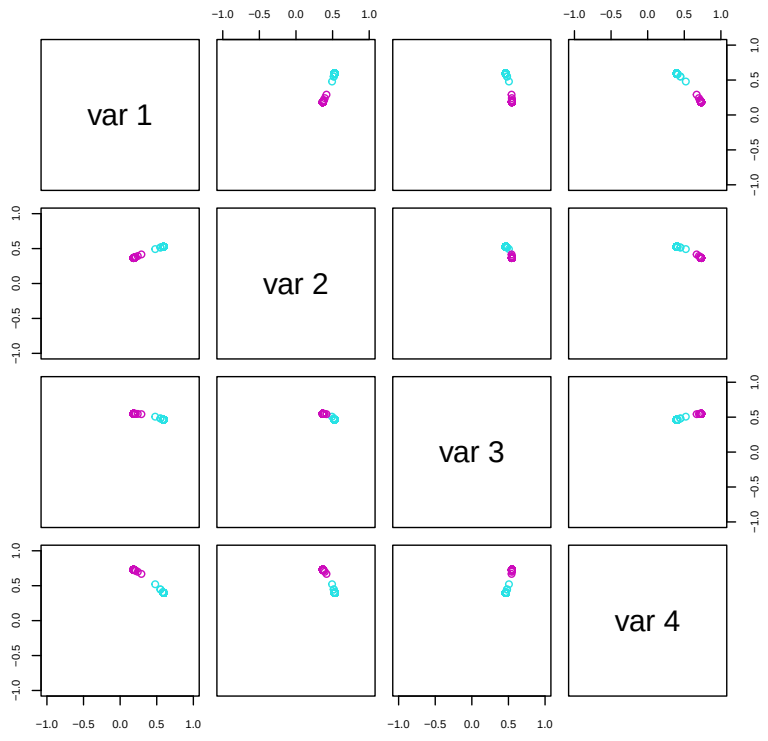
- The larger σ, the larger contamination.

- Contamination introduces a continuous (uniform) component in the spectral measure.
- The larger σ, the larger contamination.
- Small σ does not change smallest eigenvalues and corresponding eigenvectors too much.

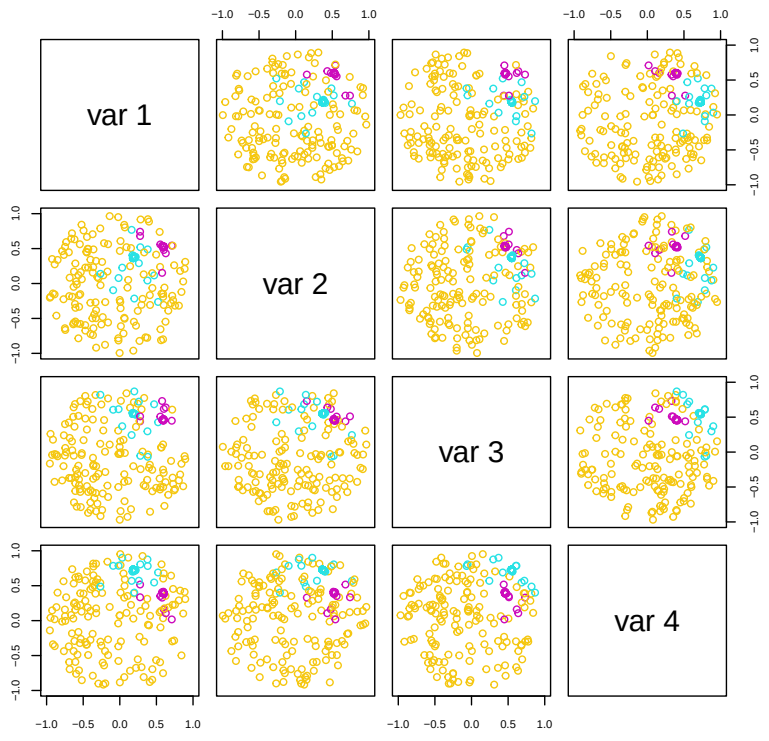
- Contamination introduces a continuous (uniform) component in the spectral measure.
- The larger σ, the larger contamination.
- Small σ does not change smallest eigenvalues and corresponding eigenvectors too much.

  We expect the algorithm to continue to work well.

4-dim data, 2 clusters, $\alpha = 1, n = 125000, N_n = 400, k_n = 15, \sigma \in \{0, 1\}$



(a) Pure signal LFM



(b) Noisy LFM

Back to detecting low-dimensional sets supporting spectral measure

Back to detecting low-dimensional sets supporting spectral measure

- Detecting these low-dimensional sets is crucial.

Back to detecting low-dimensional sets supporting spectral measure

- Detecting these low-dimensional sets is crucial.
- Difficulties:

Back to detecting low-dimensional sets supporting spectral measure

- Detecting these low-dimensional sets is crucial.
- Difficulties:

  potentially large number of these low-dimensional sets;

Back to detecting low-dimensional sets supporting spectral measure

- Detecting these low-dimensional sets is crucial.
- Difficulties:

  potentially large number of these low-dimensional sets;

  the extremes may be "contaminated" by noise ;

Back to detecting low-dimensional sets supporting spectral measure

- Detecting these low-dimensional sets is crucial.
- Difficulties:

  potentially large number of these low-dimensional sets;

  the extremes may be "contaminated" by noise ;
- It is easier to search for "linear sets".

- One can search for “subspaces”:

- One can search for "subspaces":
  for a small subset $I \subset \{1, \ldots, d\}$,

$$\left\{ \mathbf{s} = (s_1, \ldots, s_d) \in S_{d-1} : \; s_i = 0 \;\; \text{for all } i \notin I \right\}$$

- One can search for "subspaces":
  for a small subset $I \subset \{1, \ldots, d\}$,

$$\left\{ \mathbf{s} = (s_1, \ldots, s_d) \in S_{d-1} : \; s_i = 0 \;\; \text{for all } i \notin I \right\}$$

- PCA: a natural idea to detect "linear" low-dimensional support.

- One can search for "subspaces":
  for a small subset $I \subset \{1, \ldots, d\}$,

$$\left\{ \mathbf{s} = (s_1, \ldots, s_d) \in S_{d-1} : \; s_i = 0 \;\; \text{for all } i \notin I \right\}$$

- PCA: a natural idea to detect "linear" low-dimensional support.

  Finite variance needed;

- One can search for "subspaces":
  for a small subset $I \subset \{1, \ldots, d\}$,

$$\left\{ \mathbf{s} = (s_1, \ldots, s_d) \in S_{d-1} : \; s_i = 0 \;\; \text{for all } i \notin I \right\}$$

- PCA: a natural idea to detect "linear" low-dimensional support.

  Finite variance needed; this can be arranged.

- Drees and Sabourin (2019):

- Drees and Sabourin (2019): use a PCA method to search for ‘linear” low-dimensional support of the spectral measure.

- Drees and Sabourin (2019): use a PCA method to search for ‘linear” low-dimensional support of the spectral measure.

- Lower-dimensional support of the spectral measure may be “nonlinear”.

- Drees and Sabourin (2019): use a PCA method to search for 'linear" low-dimensional support of the spectral measure.

- Lower-dimensional support of the spectral measure may be "nonlinear".

- Avella, Davis and S. (2024a):

- Drees and Sabourin (2019): use a PCA method to search for 'linear" low-dimensional support of the spectral measure.
- Lower-dimensional support of the spectral measure may be "nonlinear".
- Avella, Davis and S. (2024a): propose a different PCA approach that allows search for "nonlinear" sets.

The idea of kernel PCA

The idea of kernel PCA

- View the unit sphere $\mathcal{S}_{d-1} \subset \mathbb{R}^d$.

The idea of kernel PCA

- View the unit sphere $\mathcal{S}_{d-1} \subset \mathbb{R}^d$.
- $(G(\mathbf{x}), \mathbf{x} \in \mathbb{R}^d)$: zero mean continuous Gaussian field, covariance function $R(\cdot, \cdot)$.

The idea of kernel PCA

- View the unit sphere $\mathcal{S}_{d-1} \subset \mathbb{R}^d$.
- $(G(\mathbf{x}), \mathbf{x} \in \mathbb{R}^d)$: zero mean continuous Gaussian field, covariance function $R(\cdot, \cdot)$.
- An inner product space $\mathcal{H}_0$: all finite linear combinations of continuous functions $\phi(\mathbf{x}) = R(\mathbf{x}, \cdot)$, $\mathbf{x} \in \mathbb{R}^d$,

The idea of kernel PCA

- View the unit sphere $\mathcal{S}_{d-1} \subset \mathbb{R}^d$.
- $(G(\mathbf{x}),\, \mathbf{x} \in \mathbb{R}^d)$: zero mean continuous Gaussian field, covariance function $R(\cdot, \cdot)$.
- An inner product space $\mathcal{H}_0$: all finite linear combinations of continuous functions $\phi(\mathbf{x}) = R(\mathbf{x}, \cdot)$, $\mathbf{x} \in \mathbb{R}^d$,

$$\big(\phi(\mathbf{x}_1), \phi(\mathbf{x}_2)\big) = R(\mathbf{x}_1, \mathbf{x}_2).$$

- Reproducing kernel Hilbert space (RKHS) $\mathcal{H}$: completion of $\mathcal{H}_0$.

- Reproducing kernel Hilbert space (RKHS) $\mathcal{H}$: completion of $\mathcal{H}_0$.

- $\mathbf{w}_1, \ldots, \mathbf{w}_{N_n}$: projections of the extremes in the sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ onto S_{d-1}.

- Reproducing kernel Hilbert space (RKHS) $\mathcal{H}$: completion of $\mathcal{H}_0$.

- $\mathbf{w}_1, \ldots, \mathbf{w}_{N_n}$: projections of the extremes in the sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ onto S_{d-1}.

- Map $\mathbf{w}_1, \ldots, \mathbf{w}_{N_n}$ into $\mathcal{H}$ by $\mathbf{w}_i \mapsto \phi(\mathbf{w}_i)$

- Reproducing kernel Hilbert space (RKHS) $\mathcal{H}$: completion of $\mathcal{H}_0$.

- $\mathbf{w}_1, \ldots, \mathbf{w}_{N_n}$: projections of the extremes in the sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ onto S_{d-1}.

- Map $\mathbf{w}_1, \ldots, \mathbf{w}_{N_n}$ into $\mathcal{H}$ by $\mathbf{w}_i \mapsto \phi(\mathbf{w}_i)$ (the feature map).

- Reproducing kernel Hilbert space (RKHS) $\mathcal{H}$: completion of $\mathcal{H}_0$.

- $\mathbf{w}_1, \ldots, \mathbf{w}_{N_n}$: projections of the extremes in the sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ onto S_{d-1}.

- Map $\mathbf{w}_1, \ldots, \mathbf{w}_{N_n}$ into $\mathcal{H}$ by $\mathbf{w}_i \mapsto \phi(\mathbf{w}_i)$ (the feature map).

- Functions $\phi(\mathbf{w}_1), \ldots, \phi(\mathbf{w}_{N_n})$ define nonnegative definite covariance kernel $\mathcal{C}_n : \mathcal{H} \to \mathcal{H}$

$$\mathcal{C}_n(f) = \frac{1}{N_n} \sum_{i=1}^{N_n} f(\mathbf{w}_i)\phi(\mathbf{w}_i).$$

- Perform PCA in $\mathcal{H}$.

- Perform PCA in $\mathcal{H}$.
- The eigenvalues of $\mathcal{C}_n$ coincide with the eigenvalues of $N_n^{-1} R(\mathbf{w}_i, \mathbf{w}_j)$, $i, j = 1, \ldots, N_n$.

- Perform PCA in $\mathcal{H}$.
- The eigenvalues of $\mathcal{C}_n$ coincide with the eigenvalues of $N_n^{-1} R(\mathbf{w}_i, \mathbf{w}_j)$, $i, j = 1, \ldots, N_n$.
- Take $m < N_n$ largest eigenvalues.

- Perform PCA in $\mathcal{H}$.
- The eigenvalues of $\mathcal{C}_n$ coincide with the eigenvalues of $N_n^{-1} R(\mathbf{w}_i, \mathbf{w}_j)$, $i, j = 1, \ldots, N_n$.
- Take $m < N_n$ largest eigenvalues.
- $\mathcal{P}_m \phi(\mathbf{w}_i)$: the projection of $\phi(\mathbf{w}_i)$ onto the subspace of $\mathcal{H}$ spanned by the m eigenfunctions of $\mathcal{C}_n$ corresponding to the largest eigenvalues.

- Map each $\mathcal{P}_m\phi(\mathbf{w}_i)$ onto the unit sphere $\mathcal{S}_{d-1}$

- Map each $\mathcal{P}_m\phi(\mathbf{w}_i)$ onto the unit sphere $\mathcal{S}_{d-1}$ by solving

$$T(w_i) = \text{argmin}_{\mathbf{v}\in S_{d-1}} \left\| \phi(\mathbf{v}) - \mathcal{P}_m\phi(\mathbf{w}_i) \right\|.$$

- Map each $\mathcal{P}_m\phi(\mathbf{w}_i)$ onto the unit sphere $\mathcal{S}_{d-1}$ by solving

$$T(w_i) = \text{argmin}_{\mathbf{v}\in S_{d-1}} \left\| \phi(\mathbf{v}) - \mathcal{P}_m\phi(\mathbf{w}_i) \right\|.$$

- If many of the points $\mathbf{w}_1, \ldots, \mathbf{w}_{N_n}$ lie near a small subset $S_0 \subset \mathcal{S}_{d-1}$,

- Map each $\mathcal{P}_m\phi(\mathbf{w}_i)$ onto the unit sphere $\mathcal{S}_{d-1}$ by solving

$$T(w_i) = \text{argmin}_{\mathbf{v}\in S_{d-1}} \left\| \phi(\mathbf{v}) - \mathcal{P}_m\phi(\mathbf{w}_i) \right\|.$$

- If many of the points $\mathbf{w}_1, \ldots, \mathbf{w}_{N_n}$ lie near a small subset $S_0 \subset \mathcal{S}_{d-1}$,

  then most of the points $T(\mathbf{w}_1), \ldots, T(\mathbf{w}_{N_n})$ lie near $S_0 \subset \mathcal{S}_{d-1}$.

- We justify the procedure using a version of the Davis-Kahan theorem on eigenvectors of perturbed matrices.

- We justify the procedure using a version of the Davis-Kahan theorem on eigenvectors of perturbed matrices.

- Our argument is designed for the linear factor model $\mathbf{X} = A\mathbf{Z}$:

- We justify the procedure using a version of the Davis-Kahan theorem on eigenvectors of perturbed matrices.

- Our argument is designed for the linear factor model $\mathbf{X} = A\mathbf{Z}$:

  A: $d \times p$ matrix with nonnegative entries;

- We justify the procedure using a version of the Davis-Kahan theorem on eigenvectors of perturbed matrices.

- Our argument is designed for the linear factor model $\mathbf{X} = A\mathbf{Z}$:

  A: $d \times p$ matrix with nonnegative entries;

  $\mathbf{Z}$: p-dimensional with i.i.d. nonnegative random variables with asymptotically power tails.

- The small set: the atoms of the spectral measure.

- The small set: the atoms of the spectral measure.
- $\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(p)}$: columns of A;

- The small set: the atoms of the spectral measure.
- $\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(p)}$: columns of A;
  the atoms are $\mathbf{a}^{(i)}/\|\mathbf{a}^{(i)}\|$, $i = 1, \ldots, p$.

- The small set: the atoms of the spectral measure.
- $\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(p)}$: columns of A;
  the atoms are $\mathbf{a}^{(i)}/\|\mathbf{a}^{(i)}\|$, $i = 1, \ldots, p$.
- Suppose that the directions of extremes are exactly
  $\mathbf{a}^{(i)}/\|\mathbf{a}^{(i)}\|$, $i = 1, \ldots, p$,

- The small set: the atoms of the spectral measure.
- $\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(p)}$: columns of A;
  the atoms are $\mathbf{a}^{(i)}/\|\mathbf{a}^{(i)}\|$, $i = 1, \ldots, p$.
- Suppose that the directions of extremes are exactly
  $\mathbf{a}^{(i)}/\|\mathbf{a}^{(i)}\|$, $i = 1, \ldots, p$,

  and the directions are well separated.

- Suppose the Gaussian filed is stationary.

- Suppose the Gaussian filed is stationary.
- The optimization problem:

- Suppose the Gaussian filed is stationary.
- The optimization problem:

$$T(w_i) = \text{argmin}_{\mathbf{v} \in S_{d-1}} \left\| \phi(\mathbf{v}) - \mathcal{P}_m \phi(\mathbf{w}_i) \right\|^2$$

- This is a linear combination of the terms $R\left(\mathbf{v} - \mathbf{a}^{(i)}/\|\mathbf{a}^{(i)}\|\right)$, $i = 1, \ldots, p$;

- Suppose the Gaussian filed is stationary.
- The optimization problem:

$$\begin{aligned} T(w_i) =& \text{argmin}_{\mathbf{v}\in S_{d-1}} \left\| \phi(\mathbf{v}) - \mathcal{P}_m \phi(\mathbf{w}_i) \right\|^2 \\ =& \text{argmax}_{\mathbf{v}\in S_{d-1}} \langle \phi(\mathbf{v}), \mathcal{P}_m \phi(\mathbf{w}_i) \rangle \end{aligned}$$

- This is a linear combination of the terms $R\big(\mathbf{v} - \mathbf{a}^{(i)}/\|\mathbf{a}^{(i)}\|\big)$, $i = 1, \ldots, p$;
- the max is achieved close to one of the points $\mathbf{a}^{(i)}/\|\mathbf{a}^{(i)}\|$, $i = 1, \ldots, p$.

- Suppose the Gaussian filed is stationary.
- The optimization problem:

$$\begin{aligned} T(w_i) =& \text{argmin}_{\mathbf{v}\in S_{d-1}} \left\| \phi(\mathbf{v}) - \mathcal{P}_m \phi(\mathbf{w}_i) \right\|^2 \\ =& \text{argmax}_{\mathbf{v}\in S_{d-1}} \langle \phi(\mathbf{v}), \mathcal{P}_m \phi(\mathbf{w}_i) \rangle. \end{aligned}$$

- This is a linear combination of the terms $R\big(\mathbf{v} - \mathbf{a}^{(i)}/\|\mathbf{a}^{(i)}\|\big)$, $i = 1, \ldots, p$;
- the max is achieved close to one of the points $\mathbf{a}^{(i)}/\|\mathbf{a}^{(i)}\|$, $i = 1, \ldots, p$.

- The directions of extremes are “contaminated” due to only approximately correct distribution

- The directions of extremes are "contaminated" due to only approximately correct distribution
- If the contamination is modest: the covariance kernel changes moderately.

- The directions of extremes are "contaminated" due to only approximately correct distribution
- If the contamination is modest: the covariance kernel changes moderately.
- The Davis-Kahan theorem guarantees that eigenvectors change modestly.

- The directions of extremes are "contaminated" due to only approximately correct distribution
- If the contamination is modest: the covariance kernel changes moderately.
- The Davis-Kahan theorem guarantees that eigenvectors change modestly.
- The kernel PCA procedure still clarifies the picture.

- The procedure seems to work well for many other models.

- The procedure seems to work well for many other models.
- The “small sets” no longer discrete.

- The procedure seems to work well for many other models.
- The "small sets" no longer discrete.
- Choice of the covariance function R does not seem to matter.

- The procedure seems to work well for many other models.
- The "small sets" no longer discrete.
- Choice of the covariance function R does not seem to matter.

  We use $R(\mathbf{x}_1, \mathbf{x}_2) = \exp\{-\|\mathbf{x}_1 - \mathbf{x}_2\|^2\}$.

- The procedure seems to work well for many other models.
- The "small sets" no longer discrete.
- Choice of the covariance function R does not seem to matter.

  We use $R(\mathbf{x}_1, \mathbf{x}_2) = \exp\{-\|\mathbf{x}_1 - \mathbf{x}_2\|^2\}$.
- We use the screeplot of the covariance matrix to choose the number m of largest eigenvalues.

- The procedure seems to work well for many other models.
- The "small sets" no longer discrete.
- Choice of the covariance function R does not seem to matter.

  We use $R(\mathbf{x}_1, \mathbf{x}_2) = \exp\{-\|\mathbf{x}_1 - \mathbf{x}_2\|^2\}$.
- We use the screeplot of the covariance matrix to choose the number m of largest eigenvalues.
- In most examples this identifies m correctly.

- Once again we allow the contaminated linear factor model

$$\mathbf{X}_i = A\mathbf{Z}_i + \sigma\varepsilon_i, \quad i = 1, \ldots, n$$

- Once again we allow the contaminated linear factor model

$$\mathbf{X}_i = A\mathbf{Z}_i + \sigma\varepsilon_i, \quad i = 1, \ldots, n$$

- $(\mathbf{Z}_i)$: i.i.d. 2-dim, i.i.d. Pareto(1) components;

- Once again we allow the contaminated linear factor model

$$\mathbf{X}_i = A\mathbf{Z}_i + \sigma\varepsilon_i, \quad i = 1, \ldots, n$$

- $(\mathbf{Z}_i)$: i.i.d. 2-dim, i.i.d. Pareto(1) components;

$$A = \begin{pmatrix} 0.1 & 0.9 \\ 0.2 & 0.8 \\ 0.3 & 0.7 \\ 0.4 & 0.6 \end{pmatrix}$$

- Once again we allow the contaminated linear factor model

$$\mathbf{X}_i = A\mathbf{Z}_i + \sigma\varepsilon_i, \quad i = 1, \ldots, n$$

- $(\mathbf{Z}_i)$: i.i.d. 2-dim, i.i.d. Pareto(1) components;

$$A = \begin{pmatrix} 0.1 & 0.9 \\ 0.2 & 0.8 \\ 0.3 & 0.7 \\ 0.4 & 0.6 \end{pmatrix}$$

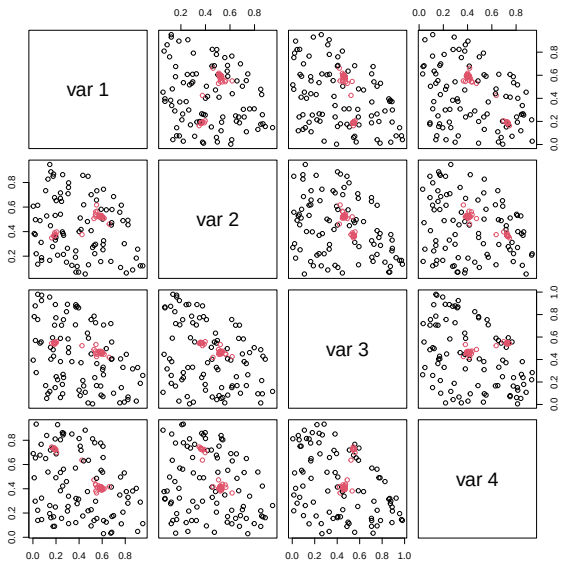
- $\sigma > 0$, (ε_i): i.i.d., 4-dim, $\varepsilon \stackrel{\mathrm{d}}{=} Y\mathbf{G}$,
  $\mathbf{G}$ 4-dim $N(0, I)$, independent of Pareto (1) Y.

- The spectral measure has a discrete component and a uniform component.

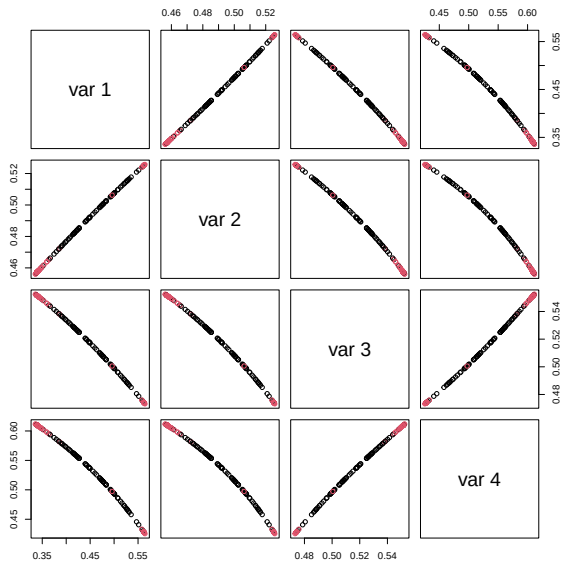
- The spectral measure has a discrete component and a uniform component.
- We use $n = 10,000$, $N_n \approx 200$.

- The spectral measure has a discrete component and a uniform component.
- We use $n = 10,000$, $N_n \approx 200$.
- Roughly 50% of the extremes come from the noise.

- The spectral measure has a discrete component and a uniform component.
- We use $n = 10,000$, $N_n \approx 200$.
- Roughly 50% of the extremes come from the noise.



(m) Contaminated linear factor model data



(n) Preimages

Spiked Gaussian model

Spiked Gaussian model

- The model:

$$\mathbf{X}_i = u_i \mathbf{N}_i + \sigma \varepsilon_i, \quad i = 1, \ldots, n,$$

Spiked Gaussian model

- The model:

$$\mathbf{X}_i = u_i \mathbf{N}_i + \sigma \varepsilon_i, \quad i = 1, \ldots, n,$$

(u_i) i.i.d. Fréchet(1);

Spiked Gaussian model

- The model:

$$\mathbf{X}_i = u_i \mathbf{N}_i + \sigma \varepsilon_i, \quad i = 1, \ldots, n,$$

(u_i) i.i.d. Fréchet(1); $(\mathbf{N}_i)$ i.i.d. d-dim centered normal, covariance matrix

$$\Sigma = \sum_{k=1}^{p} \lambda_k \mathbf{v}_k \mathbf{v}_k^\top + \sigma_0^2 I_d,$$

$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$, $\mathbf{v}_1, \ldots, \mathbf{v}_p$ orthonormal.

Spiked Gaussian model

- The model:

$$\mathbf{X}_i = u_i \mathbf{N}_i + \sigma \varepsilon_i, \quad i = 1, \ldots, n,$$

(u_i) i.i.d. Fréchet(1); $(\mathbf{N}_i)$ i.i.d. d-dim centered normal, covariance matrix

$$\Sigma = \sum_{k=1}^{p} \lambda_k \mathbf{v}_k \mathbf{v}_k^\top + \sigma_0^2 I_d,$$

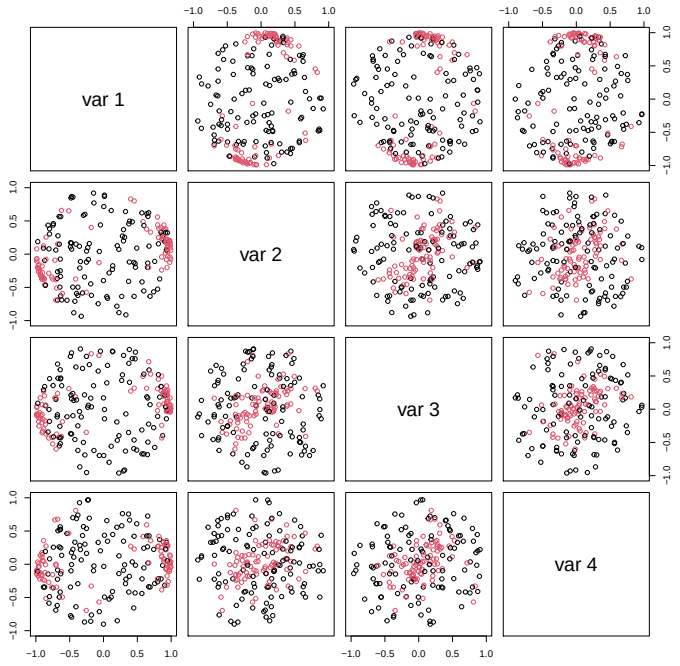
$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$, $\mathbf{v}_1, \ldots, \mathbf{v}_p$ orthonormal.

$(\sigma \varepsilon_i)$: contamination noise, as before.

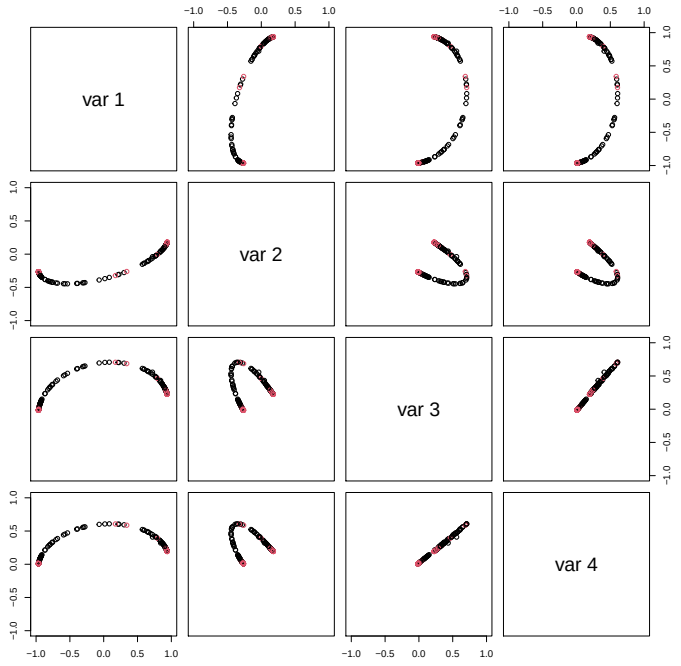
- We choose $d = 4$, $p = 2$, $\sigma = 1$.

- We choose $d = 4$, $p = 2$, $\sigma = 1$. Small set spanned by $\mathbf{v}_1, \mathbf{v}_2$.

- We choose $d = 4$, $p = 2$, $\sigma = 1$. Small set spanned by $\mathbf{v}_1, \mathbf{v}_2$.



(s) Spiked Gaussian model data



(t) Preimages