Modeling and Estimation of Bivariate Tails

Michaël Lalancette[1]

Joint work with Sebastian Engelke[2] and Stanislav Volgushev[1]

[1]Department of Statistical Sciences, University of Toronto

[2]Research Center for Statistics, University of Geneva

May 22, 2020

1. Extreme Value Theory
2. Extremal Dependence
3. Our Work: Estimation of c
4. Bonus: Application to Spatial Extremes (Some Ideas)

Extreme Value Theory

- Very broadly, EVT is the field of statistics that studies how much we can extrapolate to understand the data-generating mechanism *outside of the range of the data*

Extreme Value Theory

- Very broadly, EVT is the field of statistics that studies how much we can extrapolate to understand the data-generating mechanism *outside of the range of the data*
- Given, say, 10 years of rainfall data, EVT tries to answer to questions like

Extreme Value Theory

- Very broadly, EVT is the field of statistics that studies how much we can extrapolate to understand the data-generating mechanism *outside of the range of the data*
- Given, say, 10 years of rainfall data, EVT tries to answer to questions like
  - What is a 1-in-100-years rainfall (or even sometimes 1-in-10 000 years)?

Extreme Value Theory

- Very broadly, EVT is the field of statistics that studies how much we can extrapolate to understand the data-generating mechanism *outside of the range of the data*
- Given, say, 10 years of rainfall data, EVT tries to answer to questions like
  - What is a 1-in-100-years rainfall (or even sometimes 1-in-10 000 years)?
  - What is the probability that on a given day at least $x$mm or rain happen (where x is possibly larger than anything observed we have oberved)?

The Problem

- Consider a sample $X_1, \dots X_n \sim X$ and the problem of estimating $P(X > x)$

The Problem

- Consider a sample $X_1, \ldots X_n \sim X$ and the problem of estimating $P(X > x)$
- For simplicity, X unbounded

The Problem

- Consider a sample $X_1, \dots X_n \sim X$ and the problem of estimating $P(X > x)$
- For simplicity, X unbounded
- Two approaches

The Problem

- Consider a sample $X_1, \ldots X_n \sim X$ and the problem of estimating $P(X > x)$
- For simplicity, X unbounded
- Two approaches
  1. Use empirical probability $\frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{X_i > x\}$

The Problem

- Consider a sample $X_1, \dots X_n \sim X$ and the problem of estimating $P(X > x)$
- For simplicity, X unbounded
- Two approaches
  1. Use empirical probability $\frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left\{X_i > x\right\}$
  2. Fit parametric model $\{F_\theta\}$, get estimator $\hat{\theta}$, and use $1 - F_{\hat{\theta}}(x)$

The Problem

- Consider a sample $X_1, \ldots X_n \sim X$ and the problem of estimating $P(X > x)$
- For simplicity, X unbounded
- Two approaches
  1. Use empirical probability $\frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{X_i > x\}$
  2. Fit parametric model $\{F_\theta\}$, get estimator $\hat{\theta}$, and use $1 - F_{\hat{\theta}}(x)$
- But what if x is out of sample range (i.e. $x > \max_i X_i$)

The Problem

- Consider a sample $X_1, \dots X_n \sim X$ and the problem of estimating $P(X > x)$
- For simplicity, X unbounded
- Two approaches
  1. Use empirical probability $\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left\{X_i > x\right\}$
  2. Fit parametric model $\{F_\theta\}$, get estimator $\hat{\theta}$, and use $1 - F_{\hat{\theta}}(x)$
- But what if x is out of sample range (i.e. $x > \max_i X_i$)
  1. Nonparametric estimator is 0

The Problem

- Consider a sample $X_1, \ldots X_n \sim X$ and the problem of estimating $P(X > x)$
- For simplicity, X unbounded
- Two approaches
  1. Use empirical probability $\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{X_i > x\}$
  2. Fit parametric model $\{F_\theta\}$, get estimator $\hat{\theta}$, and use $1 - F_{\hat{\theta}}(x)$
- But what if x is out of sample range (i.e. $x > \max_i X_i$)
  1. Nonparametric estimator is 0
  2. Parametric estimator is based on assumption that tails of the parametric model are correct (uncheckable)

Conditional Tail

- The problem: tail of the distribution $S(x) := \mathbb{P}(X > x)$ can *a priori* be anything, independently of the "central part" that is observed

Conditional Tail

- The problem: tail of the distribution $S(x) := \mathbb{P}(X > x)$ can *a priori* be anything, independently of the "central part" that is observed
- However, not true for the *conditional tail*

$$S(y \mid u) := \mathbb{P}(X > u + y \mid X > u)$$

Conditional Tail

- The problem: tail of the distribution $S(x) := \mathbb{P}(X > x)$ can *a priori* be anything, independently of the "central part" that is observed
- However, not true for the *conditional tail*

$$S(y \mid u) := \mathbb{P}(X > u + y \mid X > u)$$

Theorem (Balkema, de Haan (1974), Pickands (1975))

For a very large class of distributions, as $u \to \infty$,

$$S(y \mid u) \longrightarrow \left(1 + \frac{\gamma y}{\sigma}\right)^{-1/\gamma}, \quad y > 0,$$

for some $\sigma > 0$ and $\gamma \in \mathbb{R}$. That is, $X - u \mid X > u$ approximately has a $GP(\sigma, \gamma)$ distribution.

Conditional Tail

- The problem: tail of the distribution $S(x) := \mathbb{P}(X > x)$ can *a priori* be anything, independently of the "central part" that is observed
- However, not true for the *conditional tail*

$$S(y \mid u) := \mathbb{P}(X > u + y \mid X > u)$$

Theorem (Balkema, de Haan (1974), Pickands (1975))

For a very large class of distributions, as $u \to \infty$,

$$S(y \mid u) \longrightarrow \left(1 + \frac{\gamma y}{\sigma}\right)^{-1/\gamma}, \quad y > 0,$$

for some $\sigma > 0$ and $\gamma \in \mathbb{R}$. That is, $X - u \mid X > u$ approximately has a $GP(\sigma, \gamma)$ distribution.

- For $\gamma = 0$, $(1 + \gamma y/\sigma)^{-1/\gamma}$ understood as $e^{-y/\sigma}$

Peaks-over-threshold Method

- Choose a threshold u that is large but in the sample range (say around 80th sample percentile)

Peaks-over-threshold Method

- Choose a threshold u that is large but in the sample range (say around 80th sample percentile)
- Write $\mathbb{P}(X > x) = S(x) = S(u)S(x - u \mid u)$

Peaks-over-threshold Method

- Choose a threshold u that is large but in the sample range (say around 80th sample percentile)
- Write $\mathbb{P}(X > x) = S(x) = S(u)S(x - u \mid u)$
- $S(u)$ can be estimated by sample proportion

Peaks-over-threshold Method

- Choose a threshold u that is large but in the sample range (say around 80th sample percentile)
- Write $\mathbb{P}(X > x) = S(x) = S(u)S(x - u \mid u)$
- $S(u)$ can be estimated by sample proportion
- $S(x - u \mid u) \approx (1 + \gamma(x - u)/\sigma)^{-1/\gamma}$

Peaks-over-threshold Method

- Choose a threshold u that is large but in the sample range (say around 80th sample percentile)
- Write $\mathbb{P}(X > x) = S(x) = S(u)S(x - u \mid u)$
- $S(u)$ can be estimated by sample proportion
- $S(x - u \mid u) \approx (1 + \gamma(x - u)/\sigma)^{-1/\gamma}$
- Parameters σ, γ are estimated by assuming that for every observation X_i above u, $X_i - u$ is approximately $\mathrm{GP}(\sigma, \gamma)$ distributed

Bivariate Tail

- Consider random vector (X, Y) (for simplicity, X and Y are unbounded)

Bivariate Tail

- Consider random vector (X, Y) (for simplicity, X and Y are unbounded)
- What does "tail of (X, Y)" even mean? Equivalently, what is a bivariate extreme event?

Bivariate Tail

- Consider random vector (X, Y) (for simplicity, X and Y are unbounded)
- What does "tail of (X, Y)" even mean? Equivalently, what is a bivariate extreme event?
- Most common definitions are probabilities of the form

$$\mathbb{P}\left(X > x \text{ or } Y > y\right), \quad \mathbb{P}\left(X > x, Y > y\right), \quad x, y \text{ large}$$

Bivariate Tail

- Consider random vector (X, Y) (for simplicity, X and Y are unbounded)
- What does "tail of (X, Y)" even mean? Equivalently, what is a bivariate extreme event?
- Most common definitions are probabilities of the form

$$\mathbb{P}\left(X > x \text{ or } Y > y\right), \quad \mathbb{P}\left(X > x, Y > y\right), \quad x, y \text{ large}$$

- The marginal tails of X and Y can easily be modeled, but unfortunately there exists no unique parametric model for dependence structure in the tail

Tail Dependence Modeling

- Use copula approach: suppose X and Y have continuous marginal cdf F_1 and F_2

Tail Dependence Modeling

- Use copula approach: suppose X and Y have continuous marginal cdf F_1 and F_2
- Under regularity conditions in the tails (satisfied if (X, Y) is in a max-domain of attraction),

$$\ell(x, y) := \lim_{t \to 0} \frac{1}{t} \mathbb{P}\left(F_1(X) \geq 1 - tx \quad \text{or} \quad F_2(Y) \geq 1 - ty\right) \tag{1}$$

exists for every $(x, y) \in [0, \infty)^2$

Tail Dependence Modeling

- Use copula approach: suppose X and Y have continuous marginal cdf F_1 and F_2
- Under regularity conditions in the tails (satisfied if (X, Y) is in a max-domain of attraction),

$$\ell(x, y) := \lim_{t \to 0} \frac{1}{t} \mathbb{P}\left(F_1(X) \geq 1 - tx \quad \text{or} \quad F_2(Y) \geq 1 - ty\right) \tag{1}$$

  exists for every $(x, y) \in [0, \infty)^2$
- Idea: model and estimate ℓ

Tail Dependence Modeling

- Use copula approach: suppose X and Y have continuous marginal cdf F_1 and F_2
- Under regularity conditions in the tails (satisfied if (X, Y) is in a max-domain of attraction),

$$\ell(x, y) := \lim_{t \to 0} \frac{1}{t} \mathbb{P}\left(F_1(X) \geq 1 - tx \quad \text{or} \quad F_2(Y) \geq 1 - ty\right) \qquad (1)$$

  exists for every $(x, y) \in [0, \infty)^2$
- Idea: model and estimate ℓ
- Then, for t arbitrarily small, use approximations

$$\mathbb{P}\left(F_1(X) \geq 1 - tx \quad \text{or} \quad F_2(Y) \geq 1 - ty\right) \approx t\ell(x, y)$$
$$\mathbb{P}\left(F_1(X) \geq 1 - tx, F_2(Y) \geq 1 - ty\right) \approx t(x + y - \ell(x, y))$$

- Example 1: If X, Y are independent,

Asymptotic Independence

- Example 1: If X, Y are independent,

$$\ell(x, y) := \lim_{t \to 0} \frac{1}{t}(tx + ty - t^2 xy) = x + y$$

Asymptotic Independence

- Example 1: If X, Y are independent,

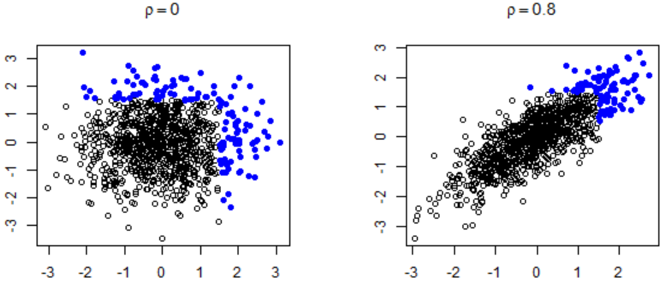
$$\ell(x, y) := \lim_{t \to 0} \frac{1}{t}(tx + ty - t^2 xy) = x + y$$

- Example 2: If X, Y are Gaussian with $|\rho| < 1$, can be shown that $\ell(x, y) = x + y$

- Example 1: If X, Y are independent,

$$\ell(x, y) := \lim_{t \to 0} \frac{1}{t}(tx + ty - t^2 xy) = x + y$$

- Example 2: If X, Y are Gaussian with $|\rho| < 1$, can be shown that $\ell(x, y) = x + y$



Asymptotic Independence

- Those distributions with $\ell(x, y) = x + y$ are called *asymptotically independent*

Asymptotic Independence

- Those distributions with $\ell(x, y) = x + y$ are called *asymptotically independent*
- Asymptotic independence is equivalent to

$$\lim_{t \to 0} \frac{1}{t} \mathbb{P}\left(F_1(X) \geq 1 - tx, F_2(Y) \geq 1 - ty\right) = x + y - \ell(x, y) = 0$$

Asymptotic Independence

- Those distributions with $\ell(x, y) = x + y$ are called *asymptotically independent*
- Asymptotic independence is equivalent to

$$\lim_{t\to 0} \frac{1}{t}\mathbb{P}\left(F_1(X) \geq 1 - tx, F_2(Y) \geq 1 - ty\right) = x + y - \ell(x, y) = 0$$

- In particular, if $x = y = 1$, it means

$$\lim_{t\to 0} \mathbb{P}\left(F_1(X) \geq 1 - t \middle| F_2(Y) \geq 1 - t\right) = 0$$

Asymptotic Independence

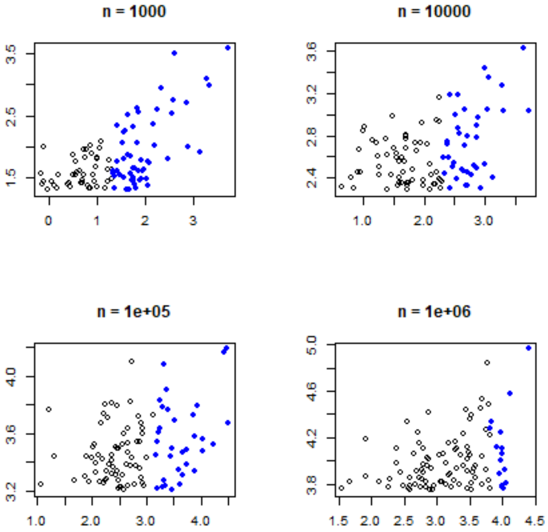
- Those distributions with $\ell(x, y) = x + y$ are called *asymptotically independent*
- Asymptotic independence is equivalent to

$$\lim_{t\to 0} \frac{1}{t}\mathbb{P}\left(F_1(X) \geq 1 - tx, F_2(Y) \geq 1 - ty\right) = x + y - \ell(x, y) = 0$$

- In particular, if $x = y = 1$, it means

$$\lim_{t\to 0} \mathbb{P}\left(F_1(X) \geq 1 - t \middle| F_2(Y) \geq 1 - t\right) = 0$$

- That is, extremes do not occur simultaneously


n = 1000
n = 10000
n = 1e+05
n = 1e+06
1.5
2.5
3.5
0
1
2
3
2.4
2.8
3.2
3.6
1.0
2.0
3.0
3.2
3.6
4.0
1.0
2.0
3.0
4.0
3.8
4.2
4.6
5.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5

Alternative Characterization of Tail Dependence

- The possibility of AI implies two issues with the use of ℓ to identify tail dependence structure:

Alternative Characterization of Tail Dependence

- The possibility of AI implies two issues with the use of ℓ to identify tail dependence structure:
  1. Very different distributions become indistinguishable

Alternative Characterization of Tail Dependence

- The possibility of AI implies two issues with the use of ℓ to identify tail dependence structure:
  1. Very different distributions become indistinguishable
  2. For joint exceedances, ℓ gives the approximation

$$\mathbb{P}\left(F_1(X) \geq 1 - tx, F_2(Y) \geq 1 - ty\right) \approx 0$$

Alternative Characterization of Tail Dependence

- The possibility of AI implies two issues with the use of ℓ to identify tail dependence structure:
  1. Very different distributions become indistinguishable
  2. For joint exceedances, ℓ gives the approximation

$$\mathbb{P}\left(F_1(X) \geq 1 - tx, F_2(Y) \geq 1 - ty\right) \approx 0$$

- Instead, what if we directly model the probability of joint threshold exceedance?

Alternative Characterization of Tail Dependence

- Assume the existence of a scaling function q such that

$$c(x, y) := \lim_{t \to 0} \frac{1}{q(t)} \mathbb{P}\left(F_1(X) \geq 1 - tx, F_2(Y) \geq 1 - ty\right)$$

exists and is non-trivial

Alternative Characterization of Tail Dependence

- Assume the existence of a scaling function q such that

$$c(x, y) := \lim_{t \to 0} \frac{1}{q(t)} \mathbb{P}\left(F_1(X) \geq 1 - tx, F_2(Y) \geq 1 - ty\right)$$

  exists and is non-trivial

- Under AI, $q(t) = o(t)$. Under *asymptotic dependence*, $\lim q(t)/t \in (0, 1]$

Alternative Characterization of Tail Dependence

- Assume the existence of a scaling function q such that

$$c(x, y) := \lim_{t \to 0} \frac{1}{q(t)} \mathbb{P}\left(F_1(X) \geq 1 - tx, F_2(Y) \geq 1 - ty\right)$$

  exists and is non-trivial
- Under AI, $q(t) = o(t)$. Under *asymptotic dependence*, $\lim q(t)/t \in (0, 1]$
- Essentially, q describes the strength of tail dependence and c describes the shape of the joint tail, *but they are not completely unrelated*

Alternative Characterization of Tail Dependence

- Assume the existence of a scaling function q such that

$$c(x, y) := \lim_{t \to 0} \frac{1}{q(t)} \mathbb{P}\left(F_1(X) \geq 1 - tx, F_2(Y) \geq 1 - ty\right)$$

  exists and is non-trivial

- Under AI, $q(t) = o(t)$. Under *asymptotic dependence*, $\lim q(t)/t \in (0, 1]$
- Essentially, q describes the strength of tail dependence and c describes the shape of the joint tail, *but they are not completely unrelated*
- Under AD, c and ℓ are almost equivalent since $\ell(x, y) = x + y - (2 - \ell(1, 1))c(x, y)$

Alternative Characterization of Tail Dependence

- Assume the existence of a scaling function q such that

$$c(x, y) := \lim_{t \to 0} \frac{1}{q(t)} \mathbb{P}\left(F_1(X) \geq 1 - tx, F_2(Y) \geq 1 - ty\right)$$

  exists and is non-trivial
- Under AI, $q(t) = o(t)$. Under *asymptotic dependence*, $\lim q(t)/t \in (0, 1]$
- Essentially, q describes the strength of tail dependence and c describes the shape of the joint tail, *but they are not completely unrelated*
- Under AD, c and ℓ are almost equivalent since $\ell(x, y) = x + y - (2 - \ell(1, 1))c(x, y)$
- Under AI however, c contains much information on dependence structure

Examples

- If X, Y are independent, then $c(x, y) = xy$

Examples

- If X, Y are independent, then $c(x, y) = xy$
- If X, Y are Gaussian with correlation $\rho \in (-1, 1)$, then $c(x, y) = (xy)^{1/(1+\rho)}$

Examples

- If X, Y are independent, then $c(x, y) = xy$
- If X, Y are Gaussian with correlation $\rho \in (-1, 1)$, then $c(x, y) = (xy)^{1/(1+\rho)}$
- Notice that c is homogeneous: for $r > 0$, $c(rx, ry) = r^{\alpha} c(x, y)$, for some $\alpha \geq 1$

Examples

- If X, Y are independent, then $c(x, y) = xy$
- If X, Y are Gaussian with correlation $\rho \in (-1, 1)$, then $c(x, y) = (xy)^{1/(1+\rho)}$
- Notice that c is homogeneous: for $r > 0$, $c(rx, ry) = r^\alpha c(x, y)$, for some $\alpha \geq 1$
- This is always true, and α relates to strength of dependence

Examples

- If X, Y are independent, then $c(x, y) = xy$
- If X, Y are Gaussian with correlation $\rho \in (-1, 1)$, then $c(x, y) = (xy)^{1/(1+\rho)}$
- Notice that c is homogeneous: for $r > 0$, $c(rx, ry) = r^{\alpha} c(x, y)$, for some $\alpha \geq 1$
- This is always true, and α relates to strength of dependence
  - $\alpha = 1 \Rightarrow$ AD or almost AD

Examples

- If X, Y are independent, then $c(x, y) = xy$
- If X, Y are Gaussian with correlation $\rho \in (-1, 1)$, then $c(x, y) = (xy)^{1/(1+\rho)}$
- Notice that c is homogeneous: for $r > 0$, $c(rx, ry) = r^{\alpha} c(x, y)$, for some $\alpha \geq 1$
- This is always true, and α relates to strength of dependence
  - $\alpha = 1 \Rightarrow$ AD or almost AD
  - $\alpha \in (1, 2) \Rightarrow$ AI, positive pre-asymptotic dependence

Examples

- If X, Y are independent, then $c(x, y) = xy$
- If X, Y are Gaussian with correlation $\rho \in (-1, 1)$, then $c(x, y) = (xy)^{1/(1+\rho)}$
- Notice that c is homogeneous: for $r > 0$, $c(rx, ry) = r^{\alpha} c(x, y)$, for some $\alpha \geq 1$
- This is always true, and α relates to strength of dependence
  - $\alpha = 1 \Rightarrow$ AD or almost AD
  - $\alpha \in (1, 2) \Rightarrow$ AI, positive pre-asymptotic dependence
  - $\alpha = 2 \Rightarrow$ AI, perfect or near-perfect independence

Examples

- If X, Y are independent, then $c(x, y) = xy$
- If X, Y are Gaussian with correlation $\rho \in (-1, 1)$, then $c(x, y) = (xy)^{1/(1+\rho)}$
- Notice that c is homogeneous: for $r > 0$, $c(rx, ry) = r^{\alpha} c(x, y)$, for some $\alpha \geq 1$
- This is always true, and α relates to strength of dependence
  - $\alpha = 1 \Rightarrow$ AD or almost AD
  - $\alpha \in (1, 2) \Rightarrow$ AI, positive pre-asymptotic dependence
  - $\alpha = 2 \Rightarrow$ AI, perfect or near-perfect independence
  - $\alpha > 2 \Rightarrow$ AI, negative association between extremes (rare)

An Example that Connects AI and AD

- Let $R \sim \text{Pareto}\,(\lambda)$, $\lambda \in (0,2]$, $W_j \sim \text{Pareto}\,(1)$, $i = 1,2$, and R, W_1, W_2 are independent. Then $(X, Y) = R(W_1, W_2)$ satisfies our expansion

An Example that Connects AI and AD

- Let $R \sim \text{Pareto}\,(\lambda)$, $\lambda \in (0,2]$, $W_j \sim \text{Pareto}\,(1)$, $i = 1,2$, and R, W_1, W_2 are independent. Then $(X, Y) = R(W_1, W_2)$ satisfies our expansion
- Function $c_\lambda(x, y)$ is ugly, but take-home message is

An Example that Connects AI and AD

- Let $R \sim \text{Pareto}\,(\lambda)$, $\lambda \in (0, 2]$, $W_j \sim \text{Pareto}\,(1)$, $i = 1, 2$, and R, W_1, W_2 are independent. Then $(X, Y) = R(W_1, W_2)$ satisfies our expansion
- Function $c_\lambda(x, y)$ is ugly, but take-home message is
  - $\lambda < 1 \Rightarrow$ AD
  - $\lambda \geq 1 \Rightarrow$ AI

An Example that Connects AI and AD

- Let $R \sim \text{Pareto}\,(\lambda)$, $\lambda \in (0, 2]$, $W_j \sim \text{Pareto}\,(1)$, $i = 1, 2$, and R, W_1, W_2 are independent. Then $(X, Y) = R(W_1, W_2)$ satisfies our expansion
- Function $c_\lambda(x, y)$ is ugly, but take-home message is
  - $\lambda < 1 \Rightarrow$ AD
  - $\lambda \geq 1 \Rightarrow$ AI
- Motivates inference for tail dependence that is based on c

Nonparametric Estimation of c

- Let $(X_1, Y_1), ..., (X_n, Y_n)$ be independent copies of (X, Y)

Nonparametric Estimation of c

- Let $(X_1, Y_1), ..., (X_n, Y_n)$ be independent copies of (X, Y)
- The definition of c suggests the "estimator"

$$\hat{c}_n(x, y) := \frac{1}{q(k/n)} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\left\{\hat{F}_1(X_i) \geq 1 - \frac{k}{n}x, \hat{F}_2(Y_i) \geq 1 - \frac{k}{n}y\right\},$$

where $\hat{F}_j$ are the empirical CDF's

Nonparametric Estimation of c

- Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent copies of (X, Y)
- The definition of c suggests the "estimator"

$$\hat{c}_n(x,y) := \frac{1}{q(k/n)} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\left\{ \hat{F}_1(X_i) \geq 1 - \frac{k}{n}x, \hat{F}_2(Y_i) \geq 1 - \frac{k}{n}y \right\},$$

  where $\hat{F}_j$ are the empirical CDF's
- This is a rank-based estimator (can be rewritten as a function of ranks)

Nonparametric Estimation of c

- Let $(X_1, Y_1), ..., (X_n, Y_n)$ be independent copies of (X, Y)
- The definition of c suggests the "estimator"

$$\hat{c}_n(x, y) := \frac{1}{q(k/n)} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\left\{ \hat{F}_1(X_i) \geq 1 - \frac{k}{n}x, \hat{F}_2(Y_i) \geq 1 - \frac{k}{n}y \right\},$$

  where $\hat{F}_j$ are the empirical CDF's
- This is a rank-based estimator (can be rewritten as a function of ranks)
- It appears in [Draisma et al., 2004], but just as a tool in their proofs. Never used directly for inference before

The Most Difficult Theorem I Ever Proved

- Assume that as $t \to 0$,

$$\frac{1}{q(t)} \mathbb{P} \left(F_1(X) \geq 1 - tx, F_2(Y) \geq 1 - ty \right) = c(x, y) + O\left(\frac{1}{\log(1/t)} \right)$$

locally uniformly over $(x, y) \in [0, \infty)^2$

The Most Difficult Theorem I Ever Proved

- Assume that as $t \to 0$,

$$\frac{1}{q(t)} \mathbb{P}\left(F_1(X) \geq 1 - tx, F_2(Y) \geq 1 - ty\right) = c(x, y) + O\left(\frac{1}{\log(1/t)}\right)$$

  locally uniformly over $(x, y) \in [0, \infty)^2$

- For an suitably chosen intermediate sequence $k = k_n$, define $m = m_n := nq(k/n)$

The Most Difficult Theorem I Ever Proved

- Assume that as $t \to 0$,

$$\frac{1}{q(t)} \mathbb{P}\left(F_1(X) \geq 1 - tx, F_2(Y) \geq 1 - ty\right) = c(x,y) + O\left(\frac{1}{\log(1/t)}\right)$$

  locally uniformly over $(x,y) \in [0,\infty)^2$

- For an suitably chosen intermediate sequence $k = k_n$, define $m = m_n := nq(k/n)$

Theorem (L, Engelke and Volgushev (2020))

There exist Gaussian processes $W^{(1)}$ and $W^{(2)}$ on $[0,\infty)^2$ such that

The Most Difficult Theorem I Ever Proved

- Assume that as $t \to 0$,

$$\frac{1}{q(t)} \mathbb{P}\left(F_1(X) \geq 1 - tx, F_2(Y) \geq 1 - ty\right) = c(x, y) + O\left(\frac{1}{\log(1/t)}\right)$$

  locally uniformly over $(x, y) \in [0, \infty)^2$

- For an suitably chosen intermediate sequence $k = k_n$, define $m = m_n := nq(k/n)$

Theorem (L, Engelke and Volgushev (2020))

There exist Gaussian processes $W^{(1)}$ and $W^{(2)}$ on $[0, \infty)^2$ such that

1. *Under AI, $\sqrt{m}\,(\hat{c}_n - c) \rightsquigarrow W^{(1)}$ (in $\ell^\infty\left([0, T]^2\right)$).*

The Most Difficult Theorem I Ever Proved

- Assume that as $t \to 0$,

$$\frac{1}{q(t)} \mathbb{P}\left(F_1(X) \geq 1 - tx, F_2(Y) \geq 1 - ty\right) = c(x, y) + O\left(\frac{1}{\log(1/t)}\right)$$

  locally uniformly over $(x, y) \in [0, \infty)^2$

- For an suitably chosen intermediate sequence $k = k_n$, define $m = m_n := nq(k/n)$

Theorem (L, Engelke and Volgushev (2020))

There exist Gaussian processes $W^{(1)}$ and $W^{(2)}$ on $[0, \infty)^2$ such that

1. *Under AI, $\sqrt{m}\left(\hat{c}_n - c\right) \rightsquigarrow W^{(1)}$ (in $\ell^\infty\left([0, T]^2\right)$).*
2. *Under AD, $\sqrt{m}\left(\hat{c}_n - c\right) \rightsquigarrow W^{(2)}$ (in the topo. of hypi-convergence for locally bounded functions ([Bücher et al., 2014])).*

Important Remarks

- Weak assumptions (no smoothness on c, very slow bias rate allowed)

Important Remarks

- Weak assumptions (no smoothness on c, very slow bias rate allowed)
- Basically,

$$\sqrt{m}\left(\hat{c}_n(x,y) - c(x,y)\right) = \underbrace{\text{Something}}_{\rightsquigarrow W^{(1)}} + \sqrt{m}\left(c(\hat{x}_n, \hat{y}_n) - c(x,y)\right),$$

where $\hat{x}_n$ and $\hat{y}_n$ are based on the empirical quantiles of X and of Y

Important Remarks

- Weak assumptions (no smoothness on c, very slow bias rate allowed)
- Basically,

$$\sqrt{m}\left(\hat{c}_n(x,y) - c(x,y)\right) = \underbrace{\text{Something}}_{\rightsquigarrow W^{(1)}} + \sqrt{m}\left(c(\hat{x}_n,\hat{y}_n) - c(x,y)\right),$$

  where $\hat{x}_n$ and $\hat{y}_n$ are based on the empirical quantiles of X and of Y
- "Something" is what one would obtain with known marginal distributions F_1, F_2. It is a fairly standard empirical process

Important Remarks

- Weak assumptions (no smoothness on c, very slow bias rate allowed)
- Basically,

$$\sqrt{m}\left(\hat{c}_n(x,y) - c(x,y)\right) = \underbrace{\text{Something}}_{\rightsquigarrow W^{(1)}} + \sqrt{m}\left(c(\hat{x}_n,\hat{y}_n) - c(x,y)\right),$$

  where $\hat{x}_n$ and $\hat{y}_n$ are based on the empirical quantiles of X and of Y
- "Something" is what one would obtain with known marginal distributions F_1, F_2. It is a fairly standard empirical process
- The other term comes from the error in estimating the marginals

Important Remarks

- Weak assumptions (no smoothness on c, very slow bias rate allowed)
- Basically,

$$\sqrt{m}\left(\hat{c}_n(x,y) - c(x,y)\right) = \underbrace{\text{Something}}_{\rightsquigarrow W^{(1)}} + \sqrt{m}\left(c(\hat{x}_n, \hat{y}_n) - c(x,y)\right),$$

  where $\hat{x}_n$ and $\hat{y}_n$ are based on the empirical quantiles of X and of Y
- "Something" is what one would obtain with known marginal distributions F_1, F_2. It is a fairly standard empirical process
- The other term comes from the error in estimating the marginals
- Under AD, it converges to a non trivial limit

Important Remarks

- Weak assumptions (no smoothness on c, very slow bias rate allowed)
- Basically,

$$\sqrt{m}\left(\hat{c}_n(x,y) - c(x,y)\right) = \underbrace{\text{Something}}_{\rightsquigarrow W^{(1)}} + \sqrt{m}\left(c(\hat{x}_n, \hat{y}_n) - c(x,y)\right),$$

  where $\hat{x}_n$ and $\hat{y}_n$ are based on the empirical quantiles of X and of Y
- "Something" is what one would obtain with known marginal distributions F_1, F_2. It is a fairly standard empirical process
- The other term comes from the error in estimating the marginals
- Under AD, it converges to a non trivial limit
- Under AI, it disappears because convergence of $\hat{x}_n$ and $\hat{y}_n$ is faster than convergence of "Something" to $W^{(1)}$ (based on more data)

Why a Parametric Estimator?

- Parametric models often allow for a nice interpretation

Why a Parametric Estimator?

- Parametric models often allow for a nice interpretation
- The non-parametric estimator $\hat{c}_n$ is not a proper function c

Why a Parametric Estimator?

- Parametric models often allow for a nice interpretation
- The non-parametric estimator $\hat{c}_n$ is not a proper function c
- More importantly, recall that $\hat{c}_n$ depends on the unknown scaling function q (through $m = nq(k/n)$)

Why a Parametric Estimator?

- Parametric models often allow for a nice interpretation
- The non-parametric estimator $\hat{c}_n$ is not a proper function c
- More importantly, recall that $\hat{c}_n$ depends on the unknown scaling function q (through $m = nq(k/n)$)
- The following parametric estimation procedure fixes this problem

The M-Estimator We Need

- Assume parametric family $\{c_\theta : \theta \in \Theta \subset \mathbb{R}^p\}$

The M-Estimator We Need

- Assume parametric family $\{c_\theta : \theta \in \Theta \subset \mathbb{R}^p\}$
- Idea: Choose θ as to minimize

$$\left\| \int_{[0,T]^2} g(x,y) c_\theta(x,y)\, dx\, dy - \int_{[0,T]^2} g(x,y) \hat{c}_n(x,y)\, dx\, dy \right\|,$$

where $g : [0, T]^2 \to \mathbb{R}^q$ is a vector of arbitrary weight functions

The M-Estimator We Need

- Assume parametric family $\{c_\theta : \theta \in \Theta \subset \mathbb{R}^p\}$
- Idea: Choose θ as to minimize

$$\left\| \int_{[0,T]^2} g(x,y) c_\theta(x,y)\, dx\, dy - \int_{[0,T]^2} g(x,y) \hat{c}_n(x,y)\, dx\, dy \right\|,$$

  where $g : [0, T]^2 \to \mathbb{R}^q$ is a vector of arbitrary weight functions
- Problem: $\hat{c}_n$ can only be calculated up to the unknown scaling m

The M-Estimator We Need

- Assume parametric family $\{c_\theta : \theta \in \Theta \subset \mathbb{R}^p\}$
- Idea: Choose θ as to minimize

$$\left\| \int_{[0,T]^2} g(x,y) c_\theta(x,y)\, dx\, dy - \int_{[0,T]^2} g(x,y) \hat{c}_n(x,y)\, dx\, dy \right\|,$$

  where $g : [0, T]^2 \to \mathbb{R}^q$ is a vector of arbitrary weight functions
- Problem: $\hat{c}_n$ can only be calculated up to the unknown scaling m
- Solution: Since

$$m\hat{c}_n(x,y) = \sum_{i=1}^{n} \mathbb{1}\left\{ \hat{F}_1(X_i) \geq 1 - \frac{k}{n}x, \hat{F}_2(Y_i) \geq 1 - \frac{k}{n}y \right\}$$

  can be calculated, simply multiply the second integral by m

The M-Estimator We Need

- Assume parametric family $\{c_\theta : \theta \in \Theta \subset \mathbb{R}^p\}$
- Idea: Choose θ as to minimize

$$\left\| \int_{[0,T]^2} g(x,y) c_\theta(x,y)\, dx\, dy - \int_{[0,T]^2} g(x,y) \hat{c}_n(x,y)\, dx\, dy \right\|,$$

  where $g : [0, T]^2 \to \mathbb{R}^q$ is a vector of arbitrary weight functions
- Problem: $\hat{c}_n$ can only be calculated up to the unknown scaling m
- Solution: Since

$$m\hat{c}_n(x,y) = \sum_{i=1}^{n} \mathbb{1}\left\{ \hat{F}_1(X_i) \geq 1 - \frac{k}{n}x, \hat{F}_2(Y_i) \geq 1 - \frac{k}{n}y \right\}$$

  can be calculated, simply multiply the second integral by m
- To adjust, multiply left integral by a new unknown parameter

The M-Estimator We Deserve

- We obtain the following objective function:

$$\Psi_n(\theta, \sigma) := \left\| \sigma \int_{[0,T]^2} g(x,y) c_\theta(x,y) \, dx \, dy - m \int_{[0,T]^2} g(x,y) \hat{c}_n(x,y) \, dx \, dy \right\|$$

The M-Estimator We Deserve

- We obtain the following objective function:

$$\Psi_n(\theta,\sigma) := \left\| \sigma \int_{[0,T]^2} g(x,y) c_\theta(x,y)\, dx\, dy - m \int_{[0,T]^2} g(x,y) \hat{c}_n(x,y)\, dx\, dy \right\|$$

- By minimizing this objective function, we hope that c_θ will estimate c and σ will estimate m

A Much Easier Theorem

- Suppose that the true function generating the data is c_{θ_0}, $\theta_0 \in \Theta$, and that the map

$$(\theta, \sigma) \mapsto \sigma \int_{[0,T]^2} g(x, y) c_\theta(x, y)\, dx\, dy$$

is cool enough

A Much Easier Theorem

- Suppose that the true function generating the data is c_{θ_0}, $\theta_0 \in \Theta$, and that the map

$$(\theta, \sigma) \mapsto \sigma \int_{[0,T]^2} g(x,y) c_\theta(x,y)\, dx\, dy$$

  is cool enough
- Assume the setting of the previous theorem

A Much Easier Theorem

- Suppose that the true function generating the data is c_{θ_0}, $\theta_0 \in \Theta$, and that the map

$$(\theta, \sigma) \mapsto \sigma \int_{[0,T]^2} g(x,y) c_\theta(x,y)\, dx\, dy$$

  is cool enough
- Assume the setting of the previous theorem

Theorem (L, Engelke and Volgushev (2020))

If $(\hat{\theta}_n, \hat{\sigma}_n) = argmin_{\theta,\sigma} \Psi_n(\theta, \sigma)$,

$$\sqrt{m}\left(\left(\hat{\theta}_n, \frac{\hat{\sigma}_n}{m}\right) - (\theta_0, 1)\right) \rightsquigarrow N\left(0, \Sigma(\theta_0)\right).$$

Summary

- We use a higher order representation of the tail dependence that naturally encompasses AD and AI

Summary

- We use a higher order representation of the tail dependence that naturally encompasses AD and AI
- It generalizes ℓ in some sort

Summary

- We use a higher order representation of the tail dependence that naturally encompasses AD and AI
- It generalizes ℓ in some sort
- We obtain asymptotically normal estimators of the shape of tail dependence (represented by c)

Summary

- We use a higher order representation of the tail dependence that naturally encompasses AD and AI
- It generalizes ℓ in some sort
- We obtain asymptotically normal estimators of the shape of tail dependence (represented by c)

Thank you! Questions?

Spatial Tail Dependence

- We are interested in the extremal behavior of a process $Y = \{Y(u) : u \in \mathcal{T}\}$

Spatial Tail Dependence

- We are interested in the extremal behavior of a process $Y = \{Y(u) : u \in \mathcal{T}\}$
- Usually, extremal dependence of Y is characterized by all the functions

$$\ell^{(u_1,\ldots,u_d)}(x) := \lim_{t \to 0} \frac{1}{t} \mathbb{P}\left(\bigcup_{1 \le j \le d} \left\{ F^{(u_j)}(Y(u_j)) > 1 - t x_j \right\} \right),$$

$d \in \mathbb{N}$, $u_j \in \mathcal{T}$, $x \in [0, \infty)^d$

Spatial Tail Dependence

- We are interested in the extremal behavior of a process $Y = \{Y(u) : u \in \mathcal{T}\}$
- Usually, extremal dependence of Y is characterized by all the functions

$$\ell^{(u_1,\dots,u_d)}(x) := \lim_{t\to 0} \frac{1}{t}\mathbb{P}\left(\bigcup_{1\le j\le d}\left\{F^{(u_j)}(Y(u_j)) > 1 - tx_j\right\}\right),$$

  $d \in \mathbb{N}$, $u_j \in \mathcal{T}$, $x \in [0,\infty)^d$
- Same problem as before: If for two locations u_1, u_2, $Y(u_1)$ and $Y(u_2)$ are AI, then $\ell^{(u_1,u_2)}$ is trivial

A Partial Solution

- Can characterize the extremal dependence of Y by functions

$$c^{(u_1,u_2)}(x,y) := \lim_{t \to 0} \frac{1}{q^{(u_1,u_2)}(t)} \mathbb{P}\left(F^{(u_j)}(Y(u_j)) > 1 - tx_j, \quad j = 1, 2\right),$$

$u_j \in \mathcal{T}$

A Partial Solution

- Can characterize the extremal dependence of Y by functions

$$c^{(u_1,u_2)}(x,y) := \lim_{t \to 0} \frac{1}{q^{(u_1,u_2)}(t)} \mathbb{P}\left(F^{(u_j)}(Y(u_j)) > 1 - tx_j, \quad j = 1,2 \right),$$

  $u_j \in \mathcal{T}$

- Advantage: contains more information on the pairwise dependencies under AI

A Partial Solution

- Can characterize the extremal dependence of Y by functions

$$c^{(u_1,u_2)}(x,y) := \lim_{t\to 0} \frac{1}{q^{(u_1,u_2)}(t)} \mathbb{P}\left(F^{(u_j)}(Y(u_j)) > 1 - tx_j, \quad j = 1,2\right),$$

$u_j \in \mathcal{T}$

- Advantage: contains more information on the pairwise dependencies under AI
- Disadvantage: only contains information on pairs. Luckily, currently used tail models are completely characterize by pairwise structure

Estimation of Functions $c^{(u_1,u_2)}$

- Find a parametric model $\left\{\left\{c_\theta^{(u_i,u_j)} : 1 \leq i,j \leq d\right\} : \theta \in \Theta\right\}$

Estimation of Functions $c^{(u_1,u_2)}$

- Find a parametric model $\left\{\left\{c_\theta^{(u_i,u_j)} : 1 \leq i,j \leq d\right\} : \theta \in \Theta\right\}$
- Given observations of Y at locations $u_1, \ldots, u_d$, estimate each $c^{(u_i,u_j)}$ using only the bivariate data from locations u_i, u_j

Estimation of Functions $c^{(u_1,u_2)}$

- Find a parametric model $\left\{\left\{c_\theta^{(u_i,u_j)} : 1 \leq i,j \leq d\right\} : \theta \in \Theta\right\}$
- Given observations of Y at locations $u_1, \ldots, u_d$, estimate each $c^{(u_i,u_j)}$ using only the bivariate data from locations u_i, u_j
- Combine all nonparametric estimators $\hat{c}_n^{(u_i,u_j)}$ to estimate θ by minimizing some global distance, e.g.

$$\hat{\theta} = \arg\min_\theta \sum_{1 \leq i,j \leq d} \left\| f\left(\hat{c}_n^{(u_i,u_j)}\right) - f\left(c_\theta^{(u_i,u_j)}\right)\right\|^2,$$

for some vector-valued functional

Bücher, A., Segers, J., and Volgushev, S. (2014).
When Uniform Weak Convergence Fails: Empirical Processes for Dependence Functions and Residuals via Epi- and Hypographs.
Ann. Stat., 42:1598–1634.

Draisma, G., Drees, H., Ferreira, A., and de Haan, L. (2004).
Bivariate Tail Estimation: Dependence in Asymptotic Independence.
Bernoulli, 10:251–280.