An M-estimator for tail dependence in spatial extremes

Anna Kiriliouk Johan Segers

ISBA
Université Catholique de Louvain

Young Researchers Day
20 September, 2013

An M-estimator for tail dependence

1. Motivation
2. Tail dependence
   - Stable tail dependence function
   - Spatial Models
3. Estimation
   - Nonparametric inference
   - Semiparametric inference
4. Applications
   - Simulation study
   - Wind speed data

An M-estimator for tail dependence

1. Motivation
2. Tail dependence
   - Stable tail dependence function
   - Spatial Models
3. Estimation
   - Nonparametric inference
   - Semiparametric inference
4. Applications
   - Simulation study
   - Wind speed data

Spatial data

Spatial data could consist of

- Max weekly temperatures
- Max amount of hourly rainfall in one day
- Max daily sea levels (coastal stations)
- Max daily windspeeds
- ...

Extreme value theory tries to answer questions about rare events.

- How high should dikes be such that a flood occurs less than once in 10,000 years? Where should wind turbines be built so that they are not damaged in a windstorm?
- We are often interested in events that are more extreme than the ones that we have encountered in the past.
- The extreme event "flood" is a high quantile of the distribution of sea levels.

Spatial dependence among extremes

- We consider an event "extreme" if it is extreme at at least one location.
- Locations close to each other might be highly dependent.
- Spatial dependence models are needed to incorporate this difference in dependence structure between locations.

An M-estimator for tail dependence

1. Motivation
2. Tail dependence
   - Stable tail dependence function
   - Spatial Models
3. Estimation
   - Nonparametric inference
   - Semiparametric inference
4. Applications
   - Simulation study
   - Wind speed data

Stable tail dependence function

Suppose we have n iid random vectors in $\mathbb{R}^d$, $\mathbf{X}_i = (X_{i1}, \ldots, X_{id})$ for $i = 1, \ldots, n$ with marginal distribution functions $F_1, \ldots, F_d$.

The dimension d represents the number of stations and n represents the number of days (or weeks, months, ...).

We want to estimate the **stable tail dependence function**

$$\ell(x_1, \ldots, x_d) = \lim_{t \downarrow 0} \frac{\mathbb{P}[1 - F_1(X_{11}) \leq x_1 t \text{ or } \ldots \text{ or } 1 - F_d(X_{1d}) \leq x_d t]}{t}.$$

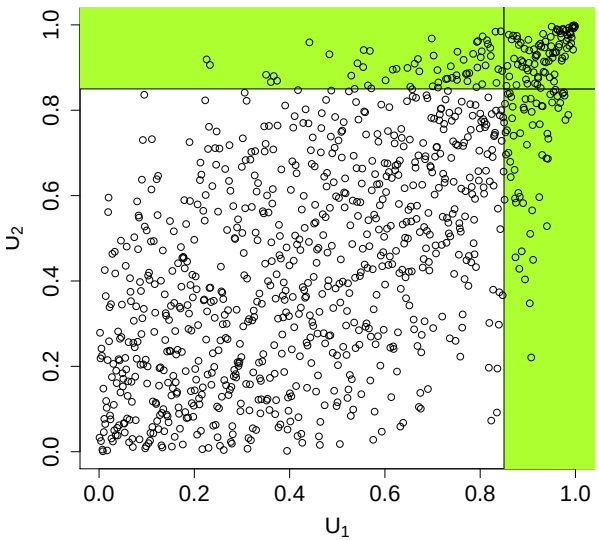
Stable tail dependence function

In two dimensions:

$$\ell(x,y) = \lim_{t\downarrow 0} \frac{\mathbb{P}[1-F_1(X) \le xt \text{ or } 1-F_2(Y) \le yt]}{t}$$
$$= \lim_{t\downarrow 0} \frac{\mathbb{P}[U_1 \ge 1-xt \text{ or } U_2 \ge 1-yt]}{t}$$
$$= \lim_{t\downarrow 0} \frac{1-\mathbb{P}[U_1 < 1-xt, U_2 < 1-yt]}{t}$$

where U_1 and U_2 are $U(0,1)$ random variables.

At least one variable is large.



Smith model

Smith (1990) introduced a parametric model for spatial extremes. A simulation from it represents extreme rainfall "storms" with a fixed Gaussian shape.

It has bivariate stable tail dependence function

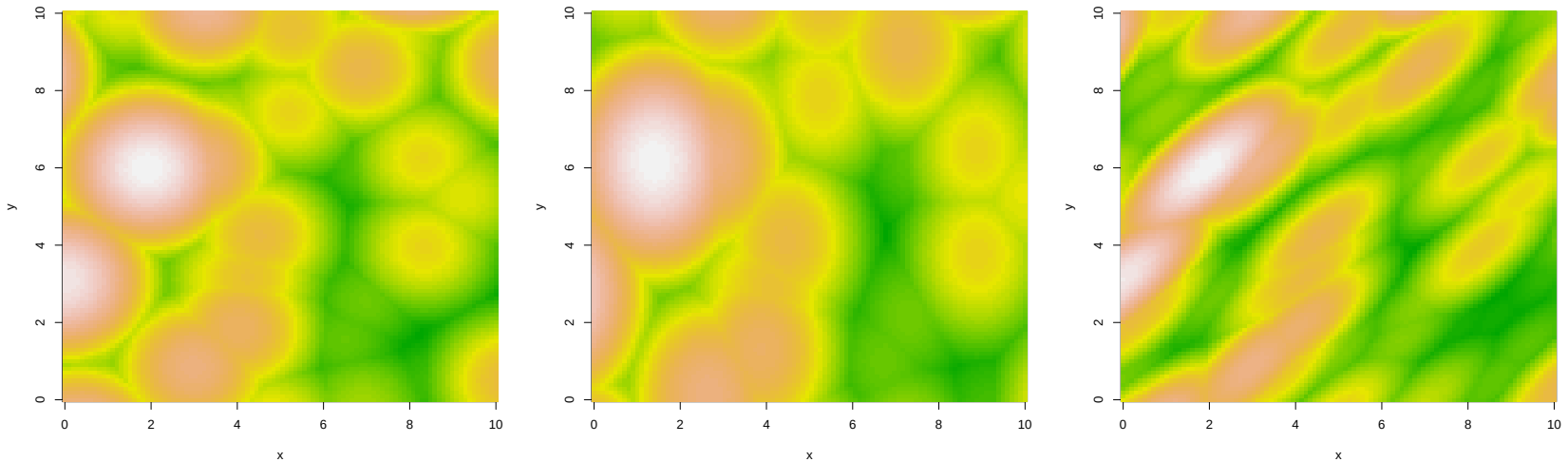
$$\ell_{st}(x_s, x_t) = x_s \, \Phi\left(\frac{a_{st}}{2} + \frac{1}{a_{st}} \log \frac{x_s}{x_t}\right) + x_t \, \Phi\left(\frac{a_{st}}{2} + \frac{1}{a_{st}} \log \frac{x_t}{x_s}\right),$$

for $s, t \in \mathbb{R}^2$ where

$$a_{st} = \sqrt{(s-t)^T \Sigma^{-1} (s-t)}, \qquad \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}.$$

a_{st} represents a distance between two locations s and t. If Σ is the identity matrix then the distance is Euclidian.

Simulating storms by the Smith model



- $\sigma_{11} = 1$, $\sigma_{22} = 1$, and $\sigma_{12} = 0$ (left).
- $\sigma_{11} = 1$, $\sigma_{22} = 2$, and $\sigma_{12} = 0$ (middle).
- $\sigma_{11} = 1$, $\sigma_{22} = 1$, and $\sigma_{12} = 0.75$ (right).

An M-estimator for tail dependence

1. Motivation
2. Tail dependence
   - Stable tail dependence function
   - Spatial Models
3. Estimation
   - Nonparametric inference
   - Semiparametric inference
4. Applications
   - Simulation study
   - Wind speed data

Empirical stable tail dependence function

Let $R_i^j = (n+1)\widehat{F}_j(X_{ij})$ denote the rank of X_{ij} among $X_{1j}, \ldots, X_{nj}$. If we write $t = k/n$ then we can estimate

$$\lim_{t \downarrow 0} \frac{\mathbb{P}[F_1(X_{11}) \geq 1 - x_1 t \text{ or } \ldots \text{ or } F_d(X_{1d}) \geq 1 - x_d t]}{t}$$

by

$$\frac{1}{k} \sum_{i=1}^{n} \mathbb{1} \left\{ R_i^1 > n + 1 - k x_1 \text{ or } \cdots \text{ or } R_i^d > n + 1 - k x_d \right\}$$

where $k \to \infty$ and $k/n \to 0$ as $n \to \infty$.

Non-parametric stable tail dependence function

Define the **non-parametric estimator** of ℓ as

$$\widehat{\ell}_n(x) = \frac{1}{k} \sum_{i=1}^{n} \mathbb{1}\left\{ R_i^1 > n + \frac{1}{2} - kx_1 \text{ or } \cdots \text{ or } R_i^d > n + \frac{1}{2} - kx_d \right\},$$

where $k = k_n \in \{1, \ldots, n\}$ such that $k \to \infty$ and $k/n \to 0$ as $n \to \infty$.

We use $(n + 1/2)$ instead of $(n + 1)$ since it makes no difference asymptotically but $\widehat{\ell}_n$ has better finite-sample properties.

A distance-based estimator

ℓ belongs to some **parametric family** $\{\ell(\cdot;\theta) \mid \theta \in \Theta\}$, $\Theta \subset \mathbb{R}^p$.

Einmahl, Krajina and Segers (2012) propose an estimator based on the distance between $\ell(x;\theta)$ and $\widehat{\ell}_n(x)$, $x \in \mathbb{R}^d$.

This is a **semi-parametric estimator**: we do not assume a parametric models for the marginal distribution but use rank-based methods.

Reduction to pairs

In spatial models $\ell(x_1, \ldots, x_d; \theta)$ often lacks an analytical expression for $d > 2$ and can be slow to work with.

Idea: Use only bivariate stable tail dependence functions. Write

$$\ell_{st}(x_s, x_t; \theta) = \ell(0, \ldots, 0, x_s, 0, \ldots, 0, x_t, 0, \ldots, 0; \theta)$$

Let $P = \{P_1, \ldots, P_q\}$ with $q = d(d-1)/2$ denote the list of all pairs of the set $\{1, \ldots, d\}$.

A homeomorphism from Θ to $\mathbb{R}^q$

Define the mapping $\psi : \Theta \to \mathbb{R}^q$ as

$$\psi(\theta) = \left(\int_{[0,1]^2} w_{st}\, \ell_{st}(x_s, x_t; \theta)\, \mathrm{d}x_s \mathrm{d}x_t \right)_{(s,t)\in P},$$

where $w_{st} \geq 0$ for all $(s,t) \in P$ and $\sum_{(s,t)\in P} w_{st} = 1$.

Its empirical counterpart is defined similarly as

$$\widehat{\psi} = \left(\int_{[0,1]^2} w_{st}\, \widehat{\ell}_{n,st}(x_s, x_t)\, \mathrm{d}x_s \mathrm{d}x_t \right)_{(s,t)\in P}.$$

An example

- $d = 36$ stations
- $P_1, \ldots, P_5$ first five pairs
- Dependence in P_1 is higher than in P_4
- Weights w can give more importance to certain pairs

M-estimator

Define the **M-estimator** $\widehat{\theta}_n$ as

$$\widehat{\theta}_n = \arg\min_{\theta \in \Theta} \left\| \psi(\theta) - \widehat{\psi} \right\|^2$$

$$= \arg\min_{\theta \in \Theta} \sum_{(s,t) \in P} w_{st}^2 \left(\int_{[0,1]^2} \left\{ \widehat{\ell}_{n,st}(x_s, x_t) - \ell_{st}(x_s, x_t; \theta) \right\} \mathrm{d}x_s \mathrm{d}x_t \right)^2 .$$

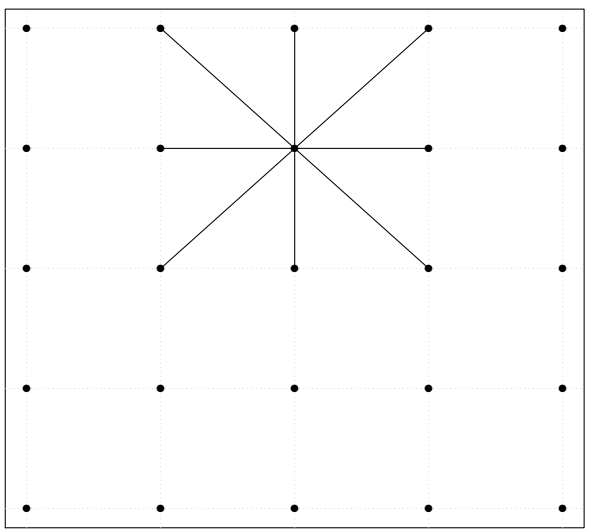
Under suitable conditions $\widehat{\theta}_n$ is **consistent** and **asymptotically normal**. The proof follows almost directly from Einmahl, Krajina and Segers (2012).

An M-estimator for tail dependence

1. Motivation
2. Tail dependence
   - Stable tail dependence function
   - Spatial Models
3. Estimation
   - Nonparametric inference
   - Semiparametric inference
4. Applications
   - Simulation study
   - Wind speed data

Simulation study of the Smith model

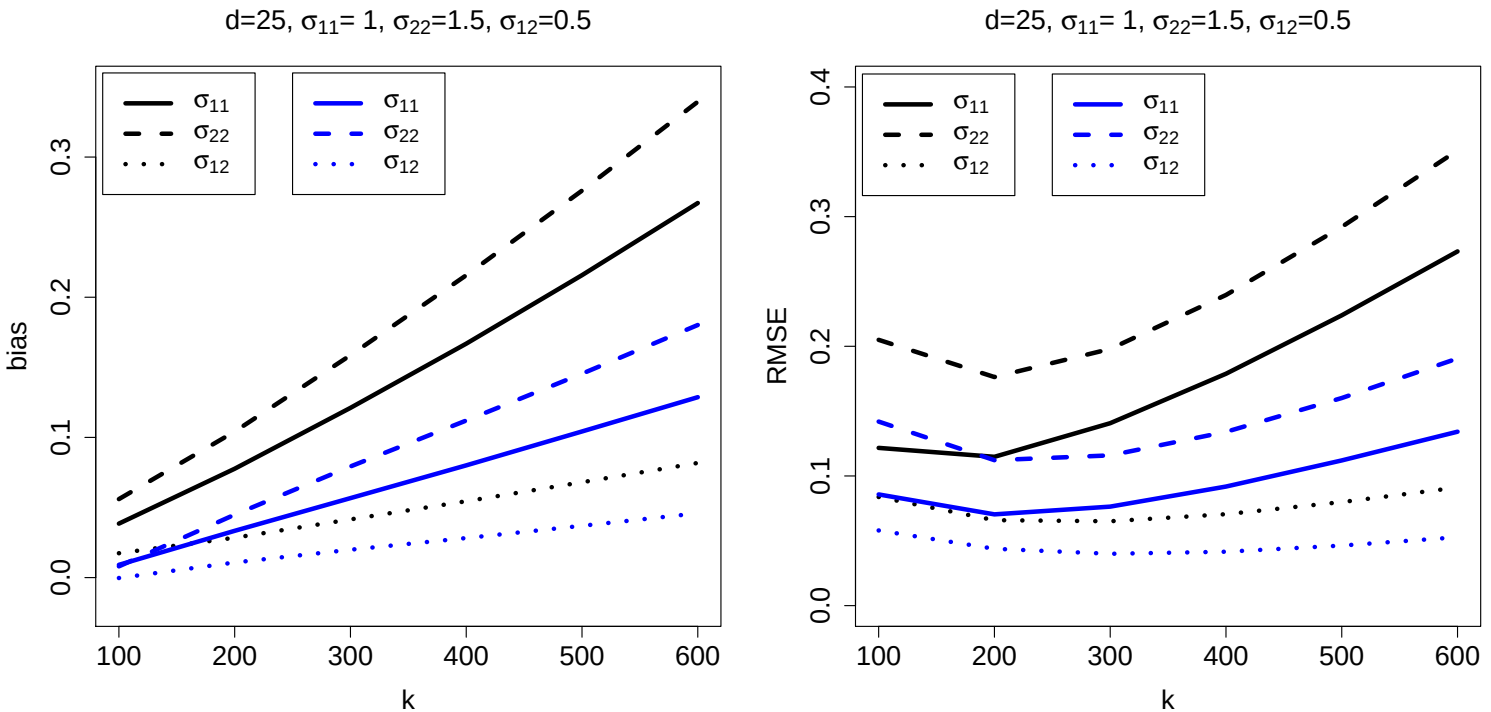
- equally spaced 5×5 grid of weather stations: $d = 25$
- 500 samples of size $n = 5000$
- $\theta = (\sigma_{11}, \sigma_{22}, \sigma_{12}) = (1.0, 1.5, 0.5)$
- Number of pairs $= q = 300$.



We study two cases:

1. $w_{st} = 1/300$ for all s, t.
2. $w_{st} = 1/72$ for surrounding pairs and $w_{st} = 0$ for non-surrounding pairs.

Simulation study of the Smith model



Bias and RMSE for **all pairs** and **surrounding pairs**

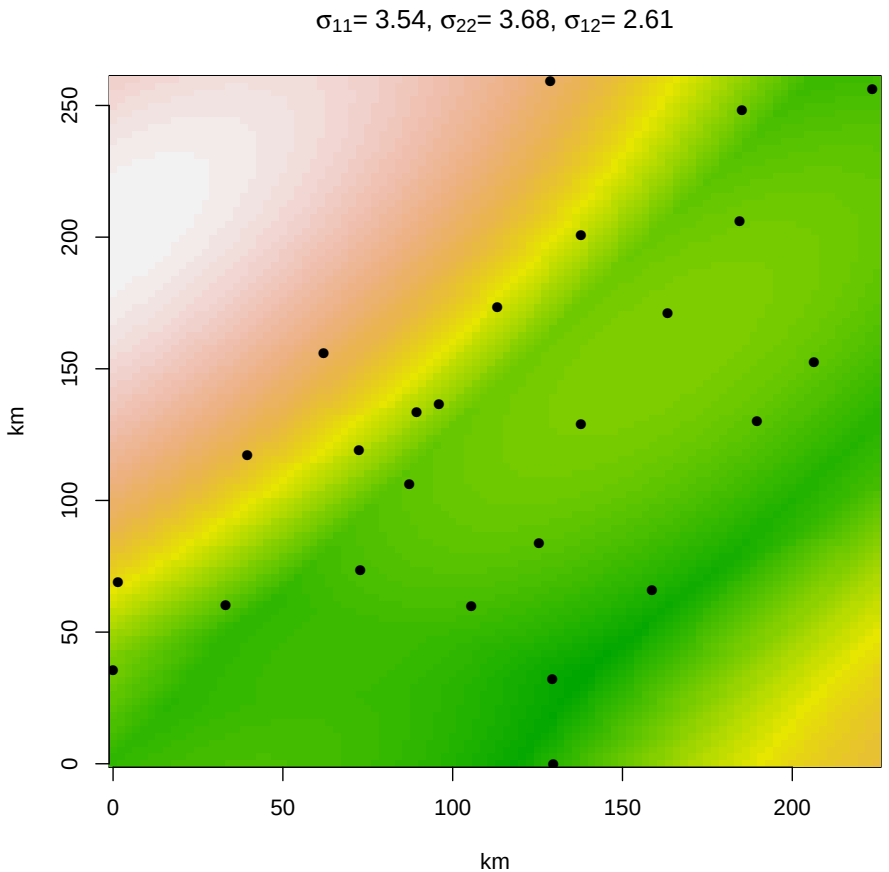
Wind speed data

Maximum daily windspeeds

- $d = 25$ locations
- $n = 1191$ observations (from 01-01-2000 to 05-12-2008)
- assume data are iid through time

Wind speed data: results

- $\widehat{\theta} = (3.54, 3.68, 2.61)$
- high value for σ_{12} indicates strong dependence between southwest—northeast pairs
- the prevailing wind direction in the Netherlands is southwest (KNMI website)



Conclusions and future work

Conclusions:

- The pairwise M-estimator functions well in terms of bias and RMSE
- It is an improvement on the original M-estimator since it makes estimation of high dimensional models possible, which is necessary in the spatial setting

Future work

- To find an expression for the weights that minimize the asymptotic variance
- To extend the results to more general spatial processes
- To do asymptotics for $d \to \infty$

References

Einmahl, J. H. , A. Krajina, and J. Segers (2012). An M-estimator for tail dependence in arbitrary dimensions. The Annals of Statistics 40(3), 1764-1793.

Smith, R. L. (1990). Max-stable processes and spatial extremes. Unpublished manuscript.

Data can be found on `http://climexp.knmi.nl` and `http://www.knmi.nl/climatology/daily_data`.