MULTIVARIATE MODELING OF LOW, MODERATE, AND LARGE POSITIVE VALUES WITHOUT THRESHOLD SELECTION STEPS

Carlo Gaetan,

Università Ca' Foscari di Venezia, Italy

gaetan@unive.it

SÉMINAIRE GEOLEARNING, Fréjus, 31/02/2025 - 03/04/2025

Joint work with

- Pierre Ailliot (UBO, France)
- Philippe Naveau (LSCE, France)

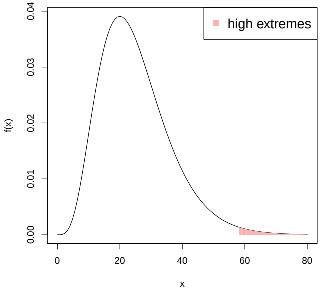
Warm up

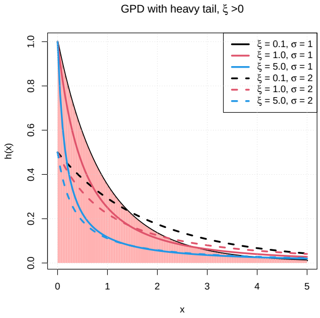
GENERALIZED PARETO DISTRIBUTION (GPD)

The distribution of X, when X exceeds a high threshold u, can be approximated by a GPD

$$H_{\xi}((x-u)/\sigma) = \begin{cases} 1-(1+\xi(x-u)/\sigma)_{+}^{-1/\xi} & \text{for } \xi \neq 0 \\ 1-\exp(-(x-u)/\sigma) & \text{for } \xi = 0 \end{cases}$$

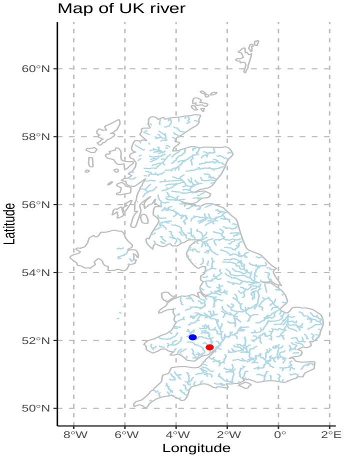
ξ shape parameter, $\sigma > 0$ scale parameter and $a_{+} = \max(a, 0)$.

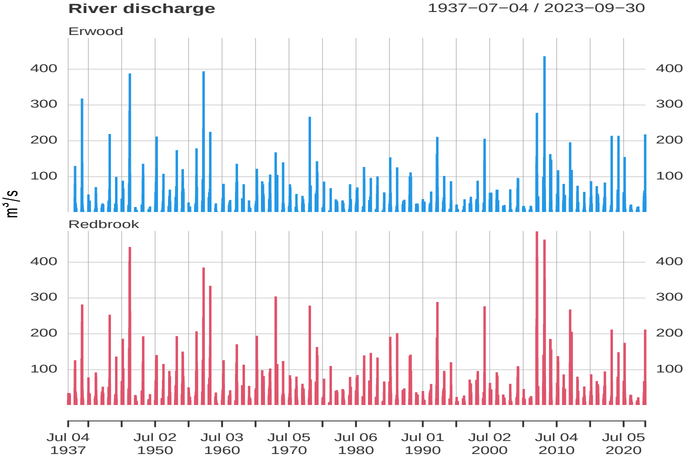




Motivating example

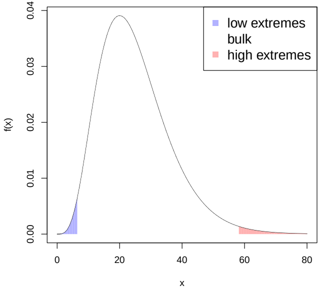
EXAMPLE: WEEKLY MAXIMUM SUMMER RIVER DISCHARGES OF WYE RIVER





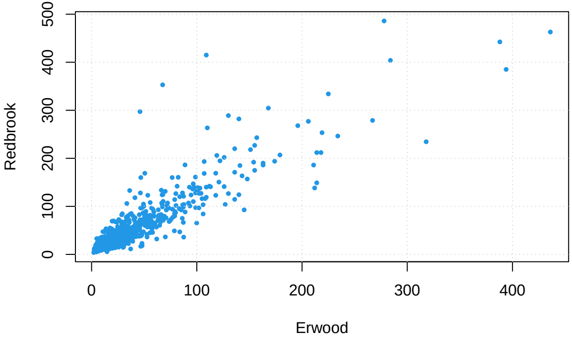
RIVER DISCHARGES

- Flood risk managers often focus on the analysis of high river flows
- Farmers may be interested in periods of low river runoffs to prevent food production shortages
- Energy producers in charge of electrical dams can be concerned by the full range of the variable of interest



RIVER DISCHARGES: DEPENDENCE

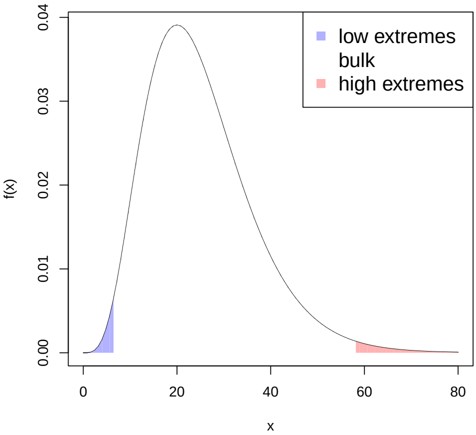
Sites along the same river basin as nearby measurements can be strongly dependent



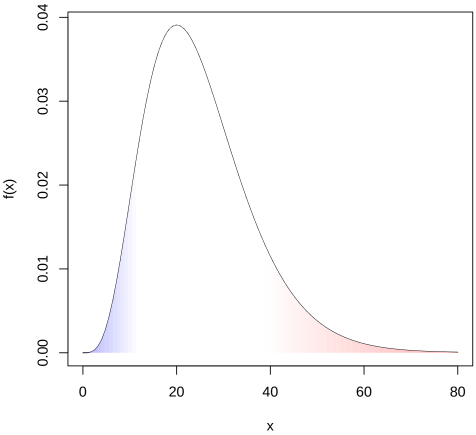
What is an Extended Generalized Pareto Distribution ?

Naveau, P., Huser, R., Ribereau, P., & Hannard, A. (2016). Modeling jointly low, moderate, and heavy rainfall intensities without a threshold selection. Water Resources Research, 52(4), 2753-2769.

TWO STRATEGIES TO MODEL JOINTLY EXTREMES AND BULK



Classical approach



EGPD approach

Low extremes

$1/\kappa$ the GPD parameter of $1/X$

Bulk

B a CDF on $[0, 1]$ with a pdf b

High extremes

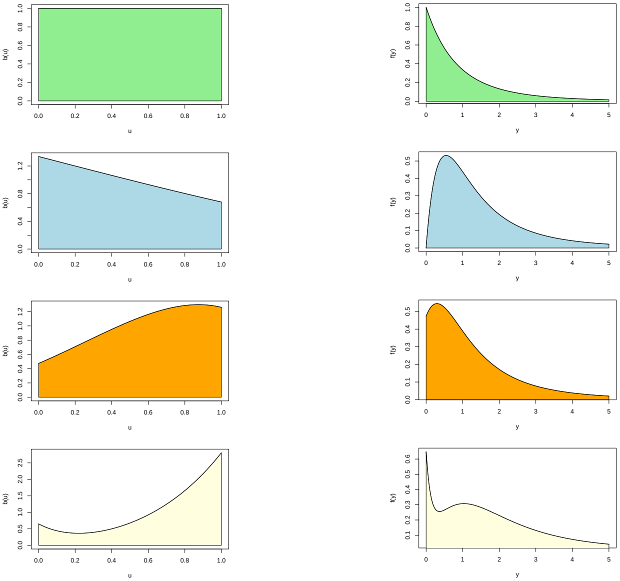
ξ the GPD parameter of X

$$\Pr(X \leq x) = B(H_{\xi}(x)^{\kappa})$$

where the pdf $b(u)$ is such that

$$0 < b(0) < \infty, \; 0 < b(1) < \infty$$

EGPD EXAMPLES



If

$$X \sim EGPD(\kappa, \xi, B),$$

then

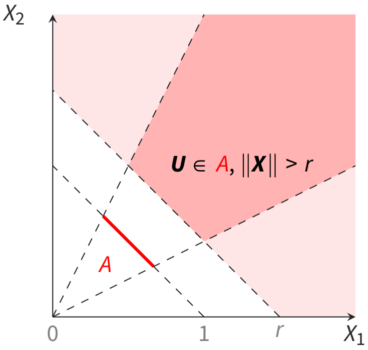
$$1/X \sim \mathrm{EGPD}(1/\xi, 1/\kappa, \tilde{B})$$

with

$$\tilde{b}(0) = \kappa\ \xi^{1/\xi}\ b(1) \text{ and } \tilde{b}(1) = \xi\ b(0)$$

A multivariate EGPD

POLAR EXTREMES COORDINATES



pseudo-radius $\|\boldsymbol{X}\| = X_1 + X_2$ and pseudo-angle $\boldsymbol{U} = (X_1/\|\boldsymbol{X}\|, X_2/\|\boldsymbol{X}\|)$

$$\mathbf{X} = \|\boldsymbol{X}\| \times \boldsymbol{U}$$

DISTRIBUTION OF $\|\boldsymbol{X}\|$ AND $1/\|\boldsymbol{X}\|$

Let $X_i \sim \mathrm{EGPD}(\kappa, \xi, B)$ for all $i = 1, \ldots, d$.

If there exist three positive and finite constants, a, c_0 and c_1 such that

$$\lim_{x\to\infty} \frac{\Pr(\|\mathbf{X}\| > x)}{\Pr(X_i > x)} = c_1 \tag{1}$$

and

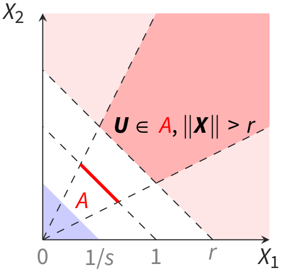
$$\lim_{x\to 0^+} \frac{\Pr(\|\mathbf{X}\| \leq x)}{[\Pr(X_i \leq x)]^a} = c_0, \tag{2}$$

then there exists a CDF B_d such that $\|\mathbf{X}\| \sim \mathrm{EGPD}(\kappa, \xi, B_d)$ with

$$b_d(0) = c_0\ b(0)^a \text{ and } b_d(1) = c_1\ b(1)/a.$$

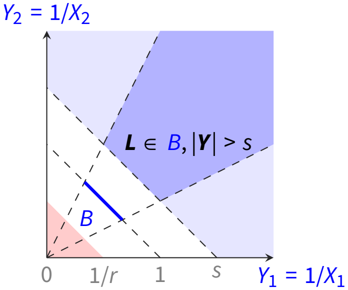
Moreover $1/\|\boldsymbol{X}\|$ is also $\mathrm{EGPD}(1/\xi, 1/\kappa, \tilde{B}_d)$

Upper extreme's representation



Radius $\|\boldsymbol{X}\| = X_1 + X_2$ and $\boldsymbol{U} = \frac{\boldsymbol{X}}{\|\boldsymbol{X}\|}$

Lower extreme's representation



$|\boldsymbol{Y}| = \frac{1}{\|\boldsymbol{1}/\boldsymbol{Y}\|}$ and $\boldsymbol{L} = \frac{\boldsymbol{Y}}{|\boldsymbol{Y}|}$

$$\mathbf{X} = \|\boldsymbol{X}\| \times \boldsymbol{U}$$

- $\|\mathbf{X}\|$ independent of $\boldsymbol{U}$ when $\|\mathbf{X}\|$ gets large
- $\Pr(\boldsymbol{U} \in A \mid \|\mathbf{X}\| > r)$ has a non-degenerate limit as $r \to \infty$, i.e.

$$\lim_{r \to \infty} \Pr(\boldsymbol{U} \in A \mid \|\mathbf{X}\| > r) = \Pr(\boldsymbol{U} \in A)$$

The three main differences with classical EVT modelling are that

1. our interest is not only on the upper extremal behaviour of $\boldsymbol{X}$, but also its lower extremal behaviour;
2. the radial component is assumed to follow a EGPD, and consequently be in compliance with EVT for both small and large values of $\|\mathbf{X}\|$;
3. in contrast to classical regular variation principles, the radial component is not necessarily assumed independent of the angular component. In particular, the degree of dependence will change according to the value $\|\mathbf{X}\|$.

BIVARIATE EGPD WITH FOUR INGREDIENTS

Low extremes

κ_d the GPD parameter of $1/\|\boldsymbol{X}\|$

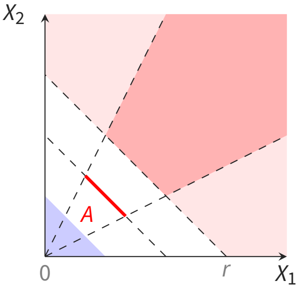
Bulk

B_d a CDF function on $[0,1]$ (with PDF b_d)

High extremes

ξ the GPD parameter of $\|\boldsymbol{X}\|$

$$\|\boldsymbol{X}\| \sim EGPD(\kappa, \xi, B), \qquad \boldsymbol{U} = \boldsymbol{X}/\|\boldsymbol{X}\|$$



Bivariate conditional model

$$\left[\log\left(\frac{U_1}{1-U_1}\right) \middle| \|\boldsymbol{X}\| = r\right] \stackrel{d}{=} \delta(r) Z$$

with Z standard normal $\perp \|\boldsymbol{X}\|$

Bivariate conditional model

$$\left[\log\left(\frac{U_1}{1-U_1}\right)\Bigg|\|\boldsymbol{X}\| = r\right] \stackrel{d}{=} \delta(r)Z$$

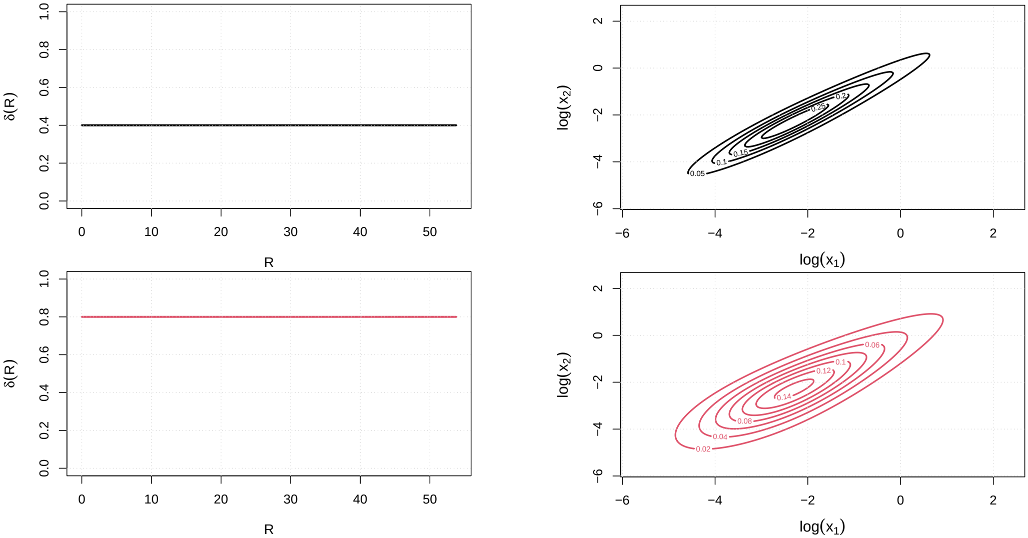
with Z standard normal $\perp \|\boldsymbol{X}\|$

- If $\delta(R)$ remains constant for large values of R, then we are in the multivariate regular variation framework
- We can specify other conditional distribution

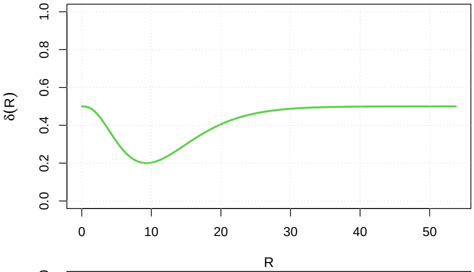
$$U_1 \,|\, \|\boldsymbol{X}\| = r \sim f(\cdot; \delta(r))$$

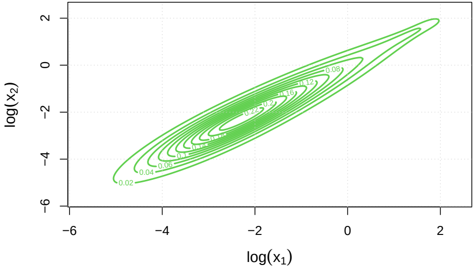
  under the constraint $\mathbb{E}\left(U_1 \,|\, \|\boldsymbol{X}\| = r\right) = 1/2$
- Why Gaussian model? Flexible multivariate distribution that is easy to specify and estimate!

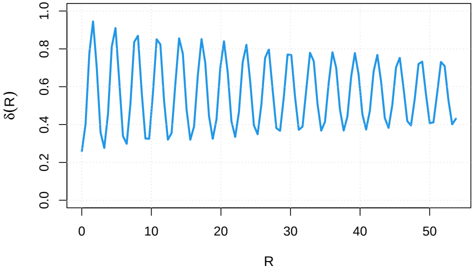
FLEXIBILITY (I)

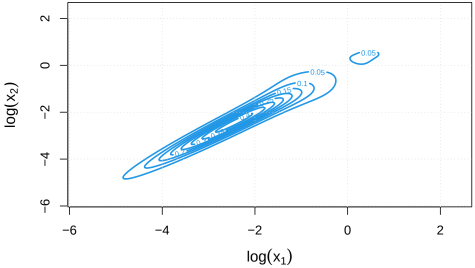


FLEXIBILITY (II)









Does this work in practice?

ESTIMATION IN TWO STEPS: FIRST STEP

Transform $\boldsymbol{x}_i = (x_{i,1}, x_{i,2})^T$ into $r_i = \|\boldsymbol{x}_i\|$ and $v_i = \log(x_{i,1}) - \log(x_{i,2})$

1. Maximize the EGPD log-likelihood

$$l_R(\kappa, \xi) = \sum_{i=1}^{n} \left\{ \log \kappa + (\kappa - 1) \log H_{\xi}(r_i) + \log h_{\xi}(r_i) + \log \widehat{b}(H_{\xi}(r_i)^{\kappa}) \right\}.$$

- Density $b(u)$ is approximated with Bernstein polynomials

$$\widehat{b}(u) = \sum_{k=1}^{m} \omega_{k,m} \beta_{k,m-k+1}(u)$$

with $\beta_{i,j}(u) = \binom{j}{i} u^i (1-u)^{j-i}$.

- **Ad-hoc R program but simple to write!**

```
evd::dgpd(x = x,loc = 0,scale = 1,shape = xi)
evd::pgpd(x = x,loc = 0,scale = 1,shape = xi)
ecdf(u)
dbeta(u,shape1 = i, shape2 = j-i+1)
```

2. Maximize the penalized Gaussian log-likelihood

$$PL_V(\boldsymbol{\gamma}) = -\sum_{i=1}^{n}\left\{\log(\delta(r_i)) + 0.5\left(\frac{v_i}{\delta(r_i)}\right)^2\right\} + \lambda\boldsymbol{\gamma}^\top \boldsymbol{P}\boldsymbol{\gamma}.$$

- Linear combination of K basis functions (cubic splines)

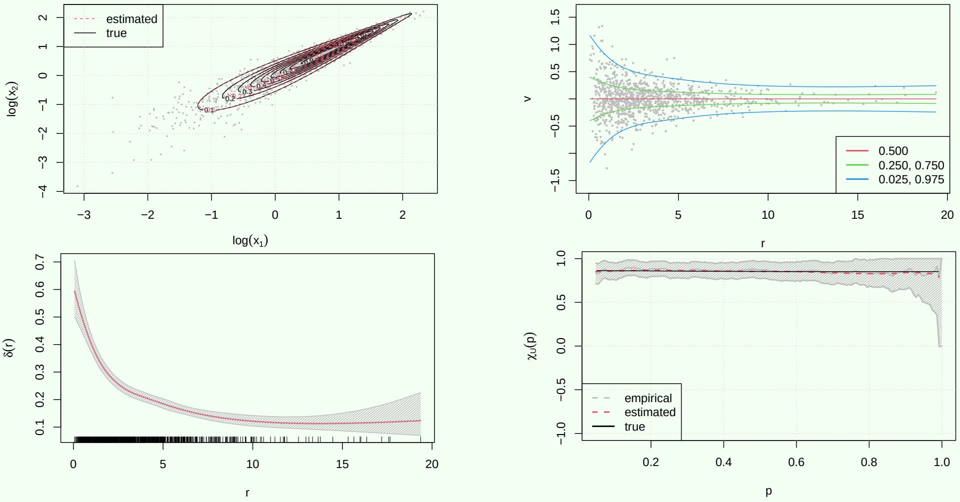
$$\log\delta(r) = \gamma_0 + \sum_{j=1}^{K}\gamma_j S_j(r), \qquad \boldsymbol{\gamma} = (\gamma_0, \ldots, \gamma_K)^\top$$

- $\lambda > 0$ smoothing parameter and $\boldsymbol{P}$ is a positive semi-definite matrix.
- **R code**

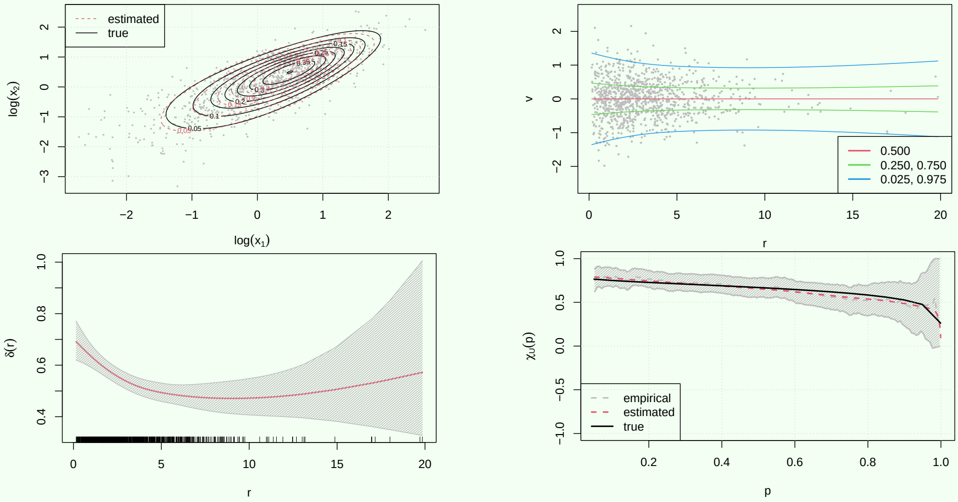
```
mgcv::gam(list(v~1,~s(r),method = "REML",family=gaulss())
```

Can we approximate common copula models ?

MRV: SYMMETRIC LOGISTIC COPULA

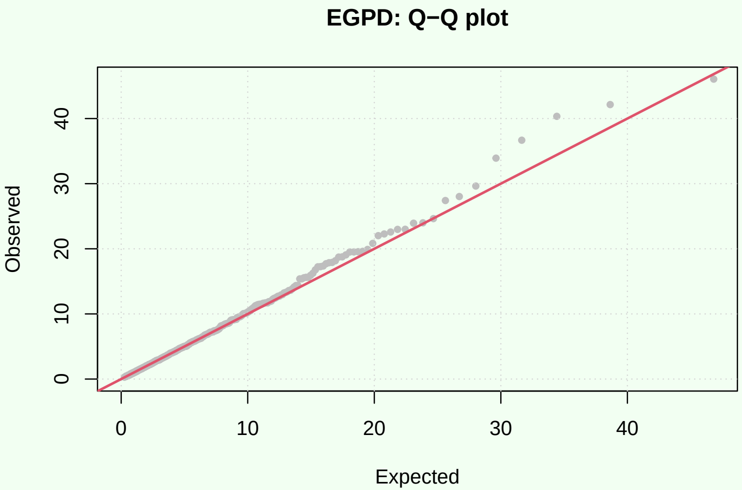


NO MRV: GAUSSIAN COPULA



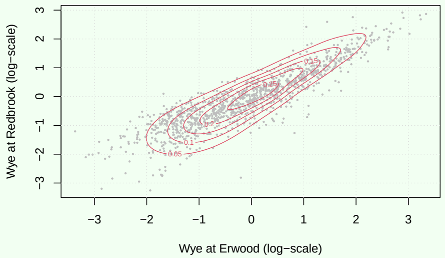
Coming back to the motivating example …

RIVER DISCHARGES: DISTRIBUTION OF $\|\boldsymbol{X}\|$

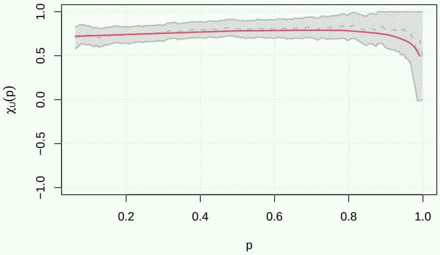


RIVER DISCHARGES: DISTRIBUTION OF $\boldsymbol{X} = (X_1, X_2)^T$

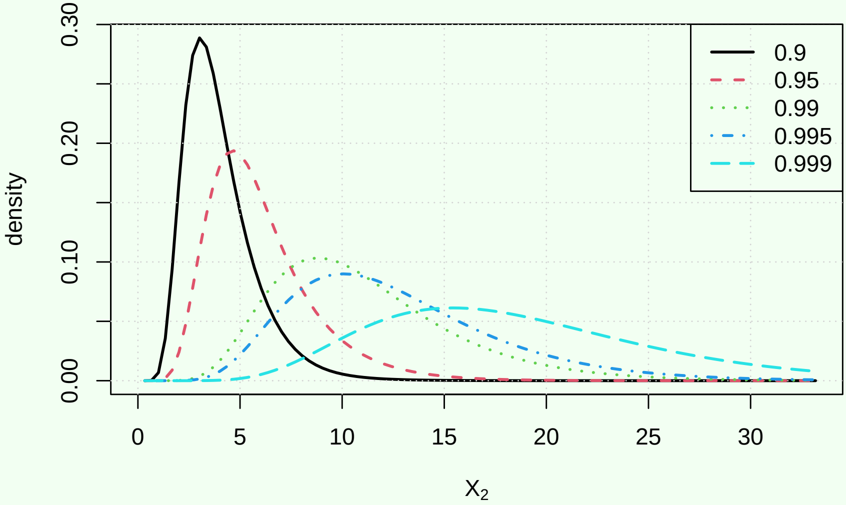
Bivariate density



Goodness of fit



RIVER DISCHARGES: DISTRIBUTION OF $X_2|X_1$

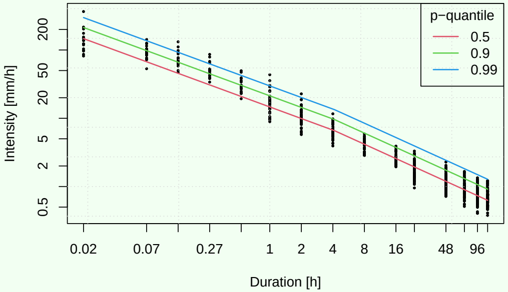


TAKE HOME MESSAGE

Sum of dependent EGPD are still EGPD and multivariate EGPD exists

FUTURE WORKS

- $X_1 + \ldots + X_d \sim EGPD(\kappa, \xi, B_d)$
- Duration d
- Intensity Duration Frequency (IDF) curve



- Rainfall measurement are discrete ...
- ...and most of the time, it is not raining, i.e. zero inflation !

Thanks !!!

Merci !!!