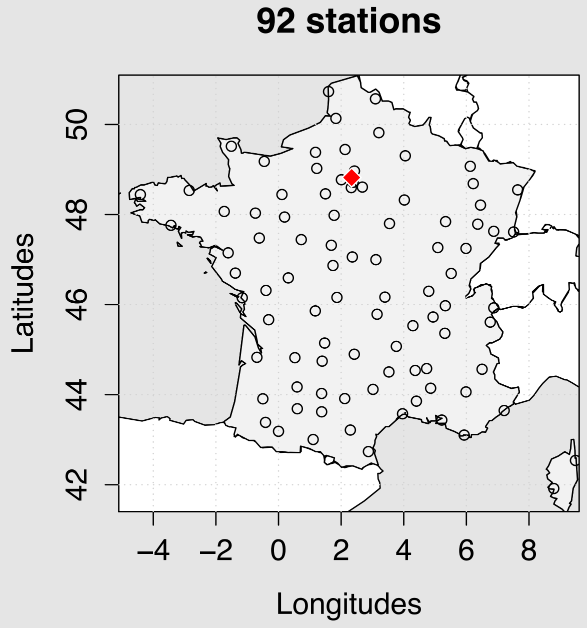
Heavy rainfall modeling in high dimensions

Philippe Naveau naveau@lsce.ipsl.fr
Laboratoire des Sciences du Climat et l'Environnement (LSCE)
Gif-sur-Yvette, France
joint work with A. Sabourin, E. Bernard, M. Vrac and O. Mestre

FP7–ACQWA, GIS–PEPER, MIRACLE & ANR–McSim, MOPERA

22 mai 2013

Hourly precipitation, 1992-2011 (Olivier Mestre)



Our game plan to handle extremes from this big rainfall dataset

	Spatial scale	
	Large (country)	Local (region)
Problem	Dimension reduction	Spectral density in moderate dimension
Data	Weekly maxima of hourly precipitation	Heavy hourly rainfall excesses
Method	Clustering algorithms for maxima	Mixture of Dirichlet

Without imposing a given parametric structure

Clustering of maxima (joint work with E. Bernard, M. Vrac and O. Mestre)

Task 1

Clustering 92 grid points into around 10-20 climatologically homogeneous groups wrt spatial dependence

Clusterings

Challenges

- Comparing apples and oranges
- An average of maxima (centroid of a cluster) is not a maximum
- variances have to be finite
- Difficult interpretation of clusters

Questions

- How to find an appropriate metric for maxima ?
- How to create cluster centroids that are maxima ?

A central question (assuming that $\mathbb{P}[M(x) < v] = \mathbb{P}[M(y) < u] = \exp(-1/u)$)

$$\mathbb{P}[M(x) < u, M(y) < v] = ??$$

Max-stable vector (de Haan, Resnick, and others)

Suppose $M(x)$ and $M(y)$ have unit Fréchet margins, we have under mild conditions

$$-\log \mathbb{P}\left[M(x) < u, M(y) < v\right] = 2\int_0^1 \max\left(\frac{w}{u}, \frac{1-w}{v}\right)\, dH(w)$$

where $H(.)$ a distribution function on $[0,1]$ such that $\int_0^1 w\, dH(w) = 0.5$.

θ = Extremal coefficient

$$\mathbb{P}\left[M(x) < u, M(y) < u\right] = \left(\mathbb{P}\left[M(x) < u\right]\right)^{\theta}$$

Interpretation

- Independence $\Rightarrow \theta = 2$
- $M(x) = M(y) \Rightarrow \theta = 1$
- Similar to correlation coefficients for Gaussian but ...
- No characterization of the **full** bivariate dependence

A L1 marginal free distance (Cooley, Poncet and N., 2005, N. and al., 2007)

$$d(x,y) = \frac{1}{2}\mathbb{E}\left|F_y(M(y)) - F_x(M(x))\right|$$

A L1 marginal free distance (Cooley, Poncet and N., 2005, N. and al., 2007)

$$d(x,y) = \frac{1}{2}\mathbb{E}\left|F_y(M(y)) - F_x(M(x))\right|$$

If $M(x)$ and $M(y)$ bivariate GEV, then

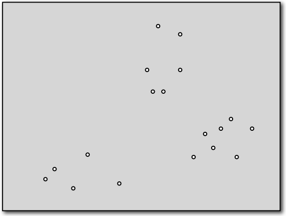
$$\text{extremal coefficient} = \frac{1+2d(x,y)}{1-2d(x,y)}$$

Clusterings

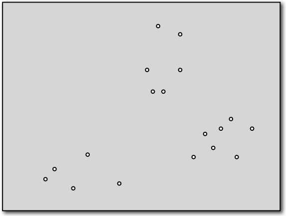
Questions

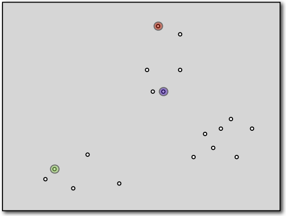
- How to find an appropriate metric for maxima ?
- **How to create cluster centroids that are maxima** ?

Partitioning Around Medoids (PAM) (Kaufman, L. and Rousseeuw, P.J. (1987))

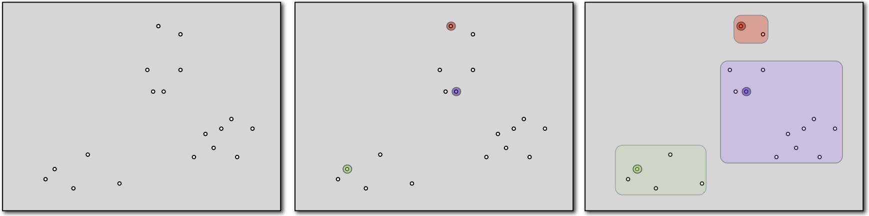


PAM : Choose K initial mediods

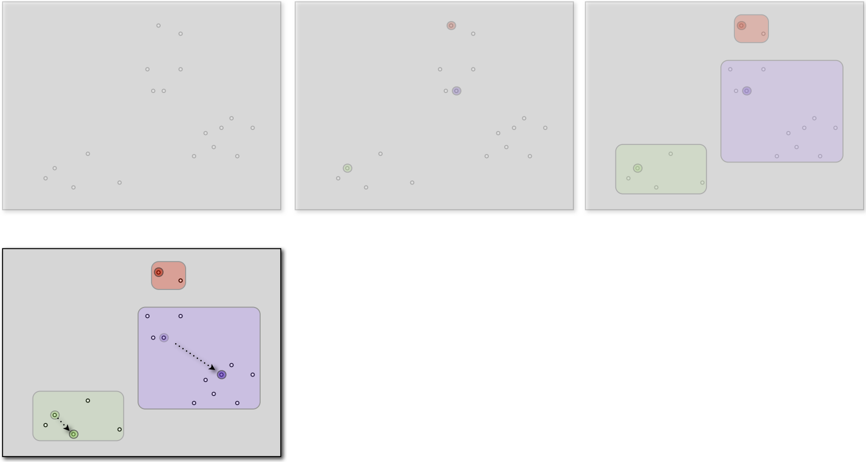




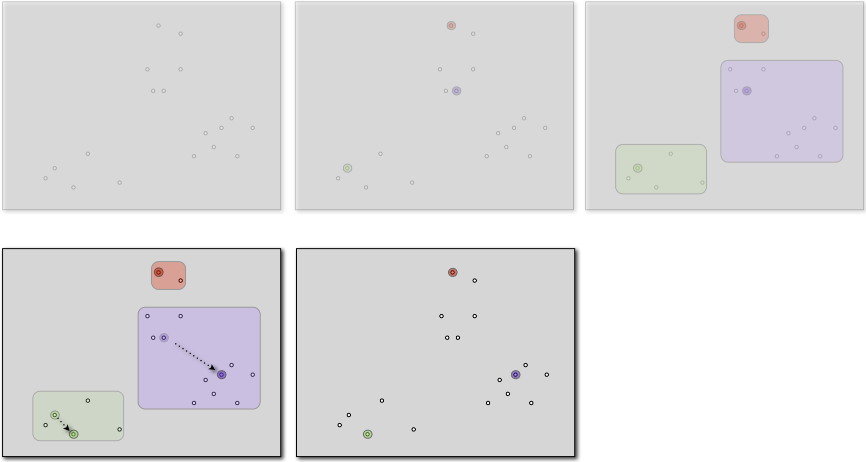
PAM : Assign each point to each closest mediod



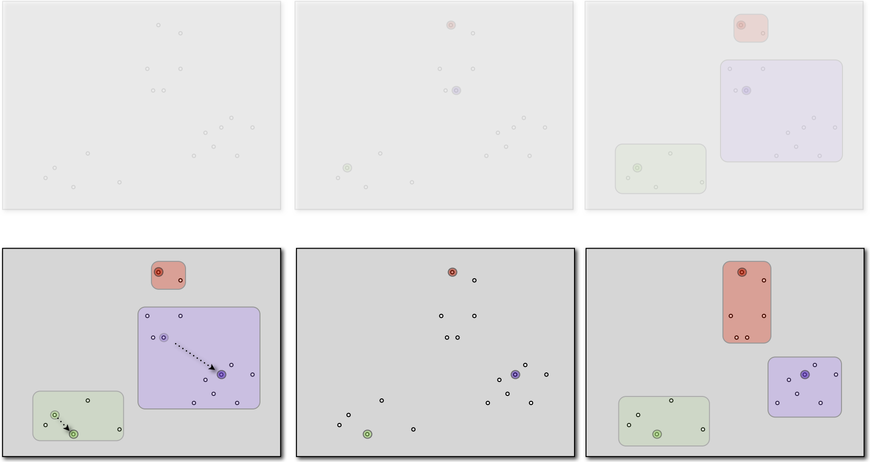
PAM : Recompute each mediod as the gravity center of each cluster



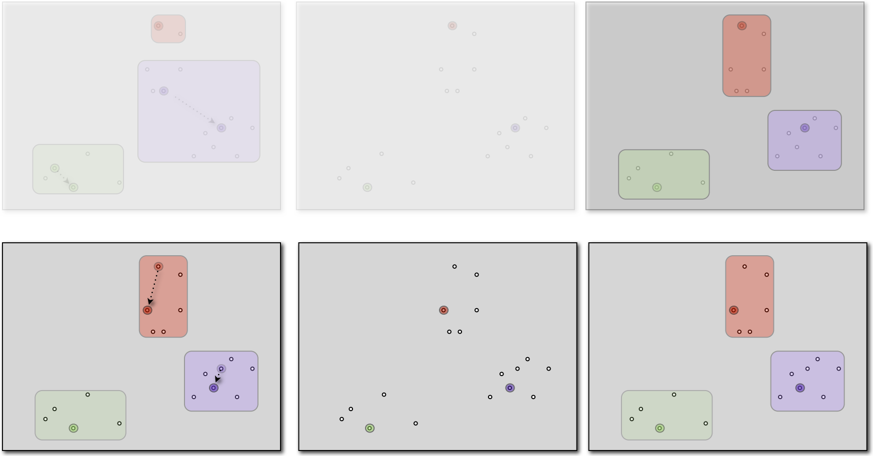
PAM : continue if a medoid has been moved

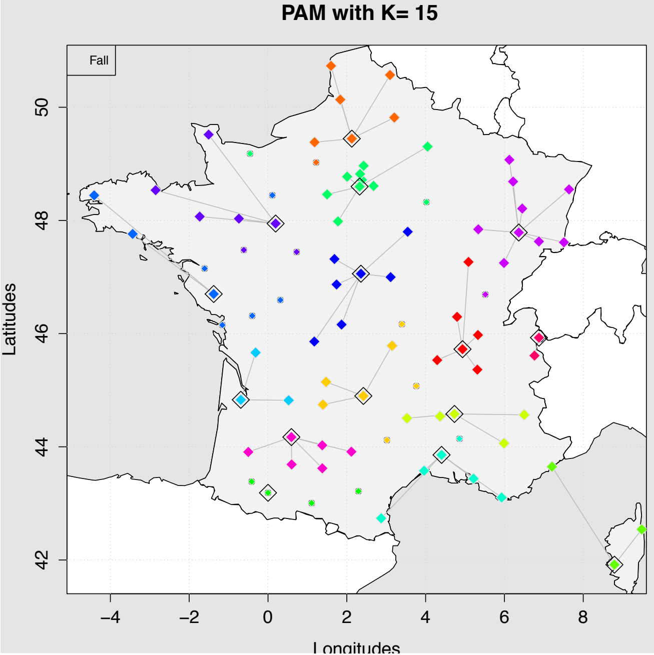


PAM : Assign each point to each closest mediod



PAM : Recompute each mediod as the gravity center of each cluster




PAM with K= 15
Fall
Latitudes
Longitudes
50
48
46
44
42
−4
−2
0
2
4
6
8

Summary on clustering of maxima

- Classical clustering algorithms (kmeans) are not in compliance with EVT
- Madogram provides a convenient distance that is marginal free and very fast to compute
- PAM applied with mado preserves maxima and gives interpretable results
- R package available on my web site

Our game plan to handle extremes from this rainfall dataset

	Spatial scale	
	Large (country)	Local (region)
Problem	Dimension reduction	Spectral density in moderate dimension
Data	Weekly maxima of hourly precipitation	Heavy hourly rainfall excesses
Method	Clustering algorithms for maxima	Mixture of Dirichlet

Bayesian Dirichlet mixture model for multivariate excesses (joint work with A. Sabourin)

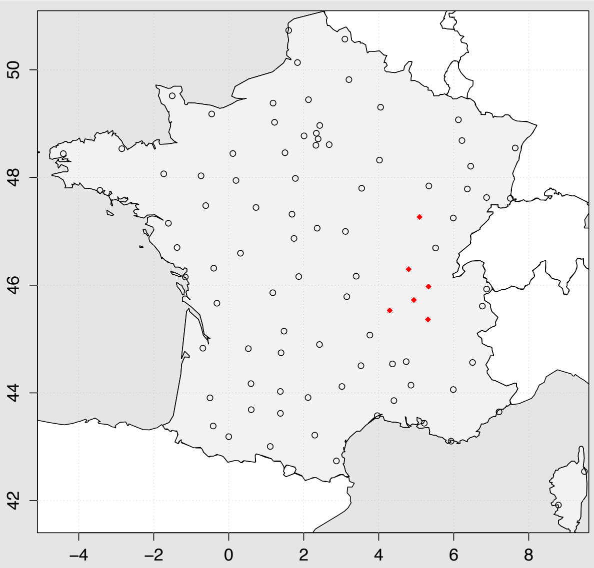
Meteo-France data

Wet hourly events at the regional scale (temporally declustered) of moderate dimensions (from 2 to 5)

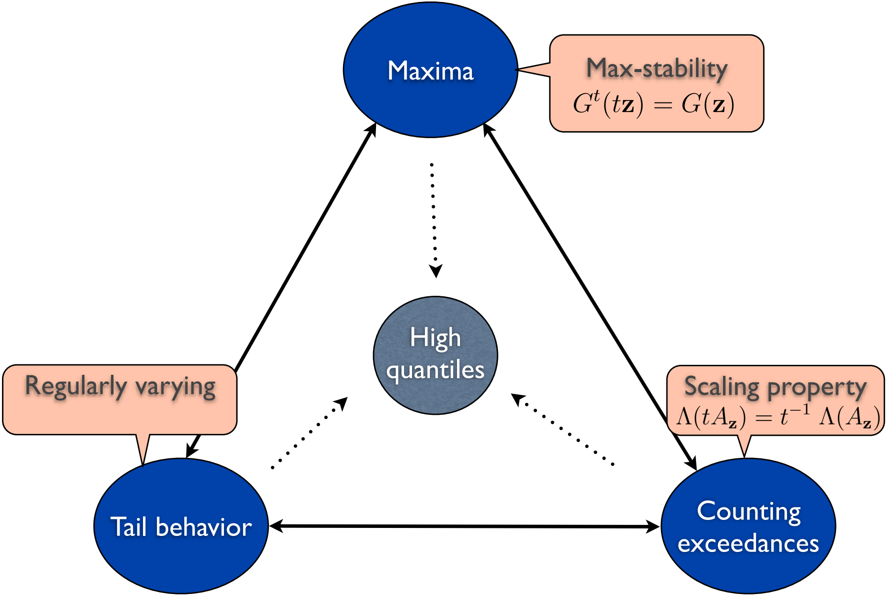
Task 2

Assessing the dependence among rainfall excesses

Focusing on the “Lyon” cluster



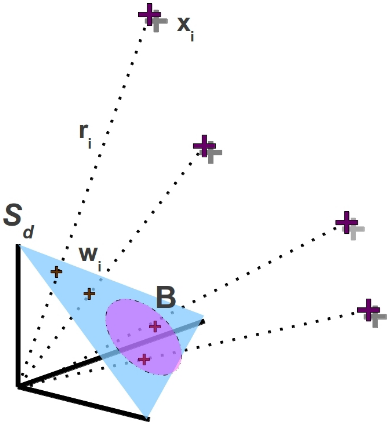
Multivariate Extreme Value Theory (de Haan, Resnick and others)



Defining radius and angular points

Example with $d = 3$ and $\mathbf{X} = (X_1, X_2, X_3)$ such that $\mathbf{P}(X_i < x) = e^{\frac{-1}{x}}$

$$\text{Simplex } \mathbf{S}_3 = \left\{ \mathbf{w} = (w_1, w_2, w_3) : \sum_{i=1}^{3} w_i = 1,\ w_i \geq 0 \right\}.$$



Mathematical constraints on the distribution of the angular points H

$$\mathbf{P}\left(\mathbf{W} \in B, R > r\right) \underset{r \to \infty}{\sim} \frac{1}{r} H(B)$$

Features of H

- H can be non-parametric
- The gravity center of H has to be centered on the simplex

$$\boxed{\forall i \in \{1, \ldots, d\}, \ \int_{\mathbf{S}_d} w_i \, \mathrm{d}H(\mathbf{w}) = \tfrac{1}{d}}$$

A few references on Bayesian non-parametric and semi-parametric spectral inference

M.-O. Boldi and A. C. Davison.
A mixture model for multivariate extremes.
JRSS : Series B (Statistical Methodology), 69(2) :217–229, 2007.

S. Guillotte, F. Perron, and J. Segers.
Non-parametric bayesian inference on bivariate extremes.
JRSS : Series B (Statistical Methodology), 2011.

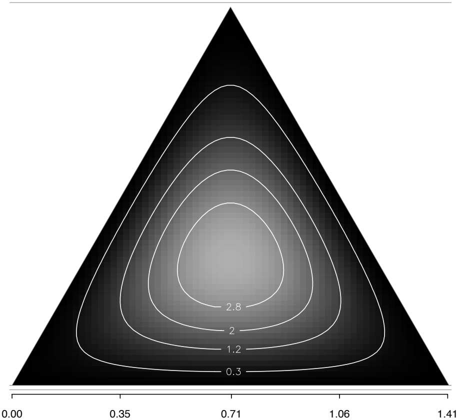
A. Sabourin and P. Naveau.
Bayesian Drichlet mixture model for multivariate extremes.
CSDA, 2013, in press.

P.J. Green.
Reversible jump Markov chain Monte Carlo computation and Bayesian model determination.
Biometrika, 82(4) :711, 1995.

Roberts, G.O. and Rosenthal, J.S.
Harris recurrence of Metropolis-within-Gibbs and trans-dimensional Markov chains
The Annals of Applied Probability,16,4,2123 :2139, 2006.

Dirichlet distribution

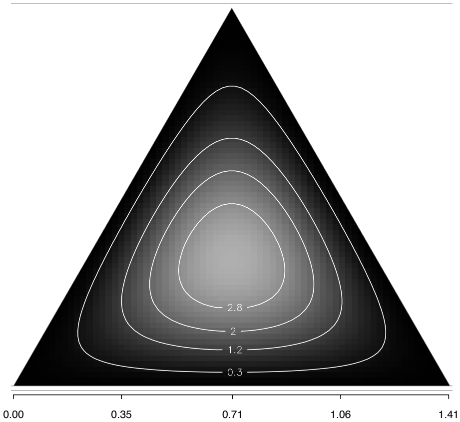
$$\forall \mathbf{w} \in \mathring{\mathbf{S}}_d, \ \mathrm{diri}(\mathbf{w} \mid \boldsymbol{\mu}, \nu) = \frac{\Gamma(\nu)}{\prod_{i=1}^{d} \Gamma(\nu\mu_i)} \prod_{i=1}^{d} w_i^{\nu\mu_i - 1} .$$



$\boldsymbol{\mu} = (1/3, 1/3, 1/3)$ and $\nu = 9$

Dirichlet distribution

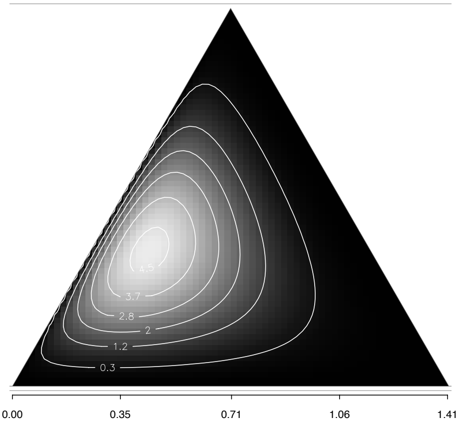
$$\forall \mathbf{w} \in \mathring{\mathbf{S}}_d, \ \mathrm{diri}(\mathbf{w} \mid \boldsymbol{\mu}, \nu) = \frac{\Gamma(\nu)}{\prod_{i=1}^{d} \Gamma(\nu\mu_i)} \prod_{i=1}^{d} w_i^{\nu\mu_i - 1} \, .$$



$\boldsymbol{\mu} = (1/3, 1/3, 1/3)$ and $\nu = 9$

Dirichlet distribution

$$\forall \mathbf{w} \in \mathring{\mathbf{S}}_d,\ \mathrm{diri}(\mathbf{w} \mid \boldsymbol{\mu}, \nu) = \frac{\Gamma(\nu)}{\prod_{i=1}^d \Gamma(\nu\mu_i)} \prod_{i=1}^d w_i^{\nu\mu_i - 1}.$$



$\boldsymbol{\mu} = (.15, .35, .05)$ and $\nu = 9$

But this one is not centered ! !

Mixture of Dirichlet distribution

Boldi and Davision, 2007

$$h_{(\boldsymbol{\mu},\mathbf{p},\boldsymbol{\nu})}(\mathbf{w}) = \sum_{m=1}^{k} p_m \,\mathrm{diri}(\mathbf{w} \mid \boldsymbol{\mu}_{\cdot\,,\,m}, \nu_m)$$

with $\boldsymbol{\mu} = \boldsymbol{\mu}_{\cdot\,,1:k}$, $\boldsymbol{\nu} = \nu_{1:k}$, $\mathbf{p} = p_{1:k}$

Mixture of Dirichlet distribution

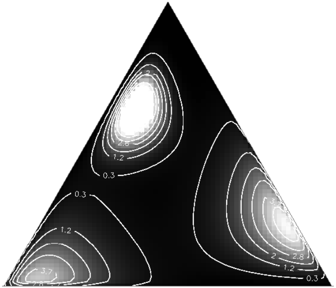
Boldi and Davision, 2007

$$h_{(\boldsymbol{\mu},\mathbf{p},\boldsymbol{\nu})}(\mathbf{w}) = \sum_{m=1}^{k} p_m \operatorname{diri}(\mathbf{w} \mid \boldsymbol{\mu}_{\cdot\,,\,m}, \nu_m)$$

with $\boldsymbol{\mu} = \boldsymbol{\mu}_{\cdot\,,1:k}$, $\boldsymbol{\nu} = \nu_{1:k}$, $\mathbf{p} = p_{1:k}$

Constraint on (μ, p)

$$\boxed{p_1\, \boldsymbol{\mu}_{\cdot,1} + \cdots + p_k\, \boldsymbol{\mu}_{\cdot,k} = (\tfrac{1}{d}, \ldots, \tfrac{1}{d})}$$



Inference of Dirichlet density mixtures

Boldi and Davison (2007)

Prior of $[\mu|\mathbf{p}]$ defined on the set

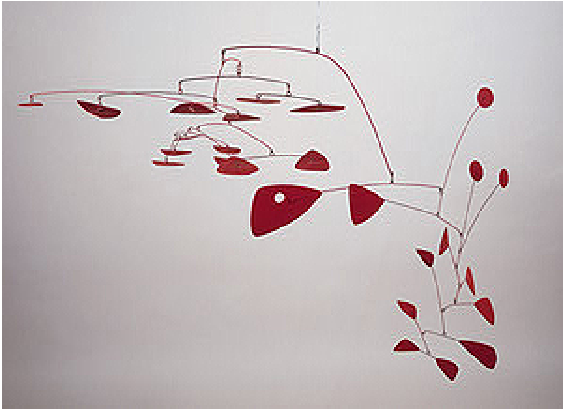
$$\boxed{p_1\, \boldsymbol{\mu}_{\cdot,1} + \cdots + p_k\, \boldsymbol{\mu}_{\cdot,k} = (\tfrac{1}{d}, \ldots, \tfrac{1}{d})}$$

- Sequential inference : first **p**, then $\boldsymbol{\mu}$ one coordinate after the other
- skewed, not interpretable, slow sampling
- Difficult inference in dimension > 3

Inference of Dirichlet density mixtures

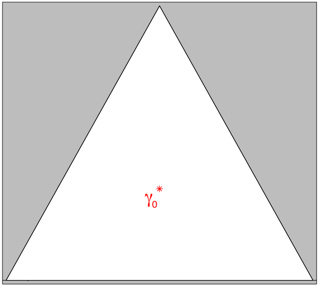
How to build priors for (p, μ) such that

$$\boxed{p_1\, \boldsymbol{\mu}_{\cdot,1} + \cdots + p_k\, \boldsymbol{\mu}_{\cdot,k} = \left(\tfrac{1}{d}, \ldots, \tfrac{1}{d}\right)}$$



New parametrisation

Ex : $k = 4$ and $d = 3$

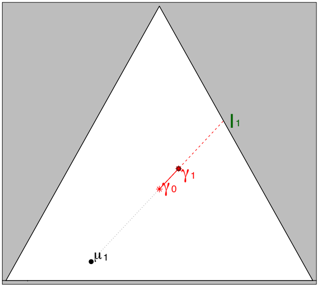


γ_m : "Equilibrium" centers built from $\boldsymbol{\mu}_{\cdot,m+1}, \ldots, \boldsymbol{\mu}_{\cdot,k}$.

$$\gamma_m = \sum_{j=m+1}^{k} \frac{p_j}{p_{m+1} + \cdots + p_k} \boldsymbol{\mu}_{\cdot,j}$$

New parametrisation

Ex : $k = 4$ and $d = 3$

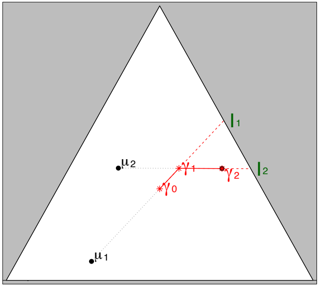


$$\boldsymbol{\mu}_{.,1},\ e_1 \quad \Rightarrow \gamma_1 : \frac{\overline{\gamma_0\,\gamma_1}}{\overline{\gamma_0\,I_1}} = e_1\ ;$$

$$\Rightarrow p_1$$

New parametrisation

Ex : $k = 4$ and $d = 3$

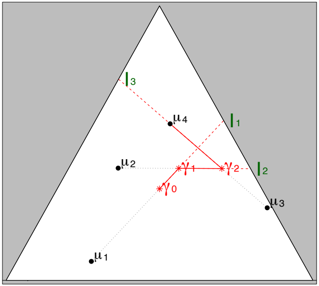


$$\mu_{.,2},\ e_2 \quad \Rightarrow \gamma_2 : \frac{\overline{\gamma_1\,\gamma_2}}{\overline{\gamma_1\, l_2}} = e_2\ ;$$

$$\Rightarrow p_2$$

New parametrisation

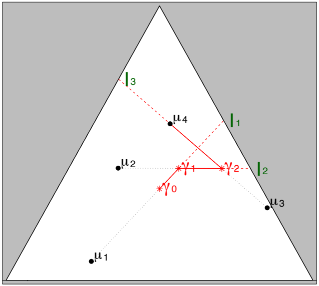
Ex : $k = 4$ and $d = 3$



$$\boldsymbol{\mu}_{.,3},\ e_3 \quad \Rightarrow \gamma_3 : \frac{\overline{\gamma_2\,\gamma_3}}{\overline{\gamma_2\,l_3}} = e_3\ ; \quad \boldsymbol{\mu}_{.,4} = \gamma_3.$$
$$\Rightarrow p_3, p_4$$

New parametrisation

Ex : $k = 4$ and $d = 3$



Parametrisation of h with $\theta = (\boldsymbol{\mu}_{\cdot,1:k-1}, \mathbf{e}_{1:k-1}, \nu_{1:k})$

$$(\boldsymbol{\mu}_{\cdot,1:k-1}, e_{1:k-1}) \ \text{gives} \ (\boldsymbol{\mu}_{\cdot,1:k}, p_{1:k})$$

Unconstrained Bayesian modeling for $\Theta = \coprod_{k=1}^{\infty} \Theta_k; \quad \Theta_k = \left\{ (\mathbf{S}_d)^{k-1} \times [0,1)^{k-1} \times (0,\infty]^{k-1} \right\}$

Prior

$$
\begin{aligned}
&k \sim \text{Truncated geometric} \\
&\boldsymbol{\mu}_{\cdot,m} | (\boldsymbol{\mu}_{\cdot,1:m-1}, e_{1:m-1}) \sim \text{Dirichlet} \\
&e_m | (\boldsymbol{\mu}_{\cdot,1:m}, e_{1:m-1}) \sim \text{Beta} \\
&\nu_m \sim \text{logN}
\end{aligned}
$$

Posterior sampling : MCMC reversible jumps

Summary of the Bayesian schemes

Boldi and Davison (2012) Our approach

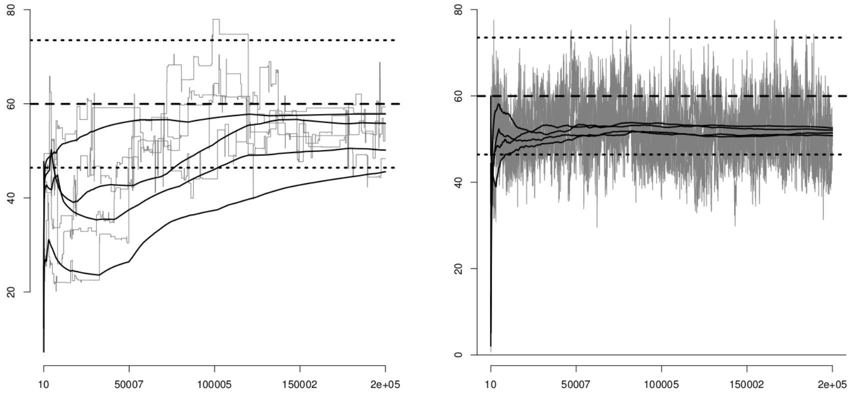
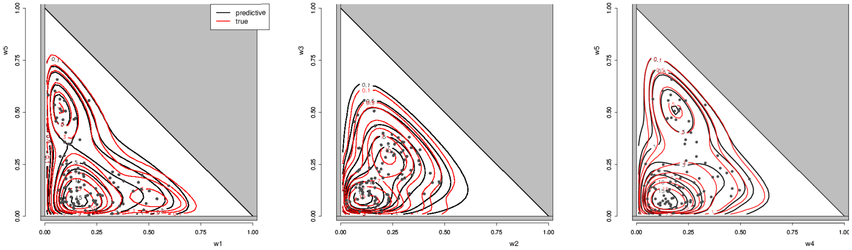


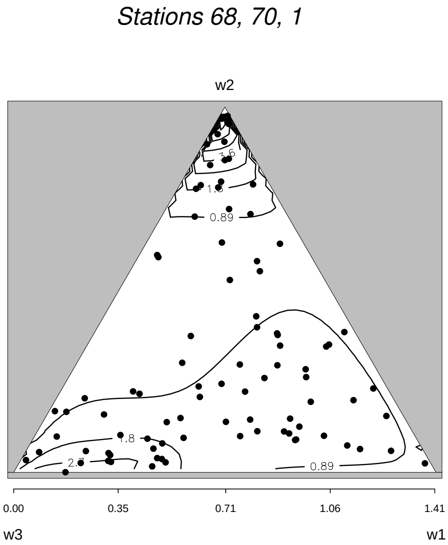
Figure 5: Convergence monitoring with five-dimensional data in the original DM model (left panel) and in the re-parametrized v
with four parallel chains in each model. Grey lines: Evolution of $\langle g, h_{\theta_t(j)} \rangle$. Black, solid lines: cumulative mean. Dashed line

Simulation example with $d = 5$ and $k = 3$



$$T_2 = 150\,10^3,\ T_1 = 50\,10^3.$$

Back to our excesses of the "Lyon" cluster



Take home messages

Conclusions

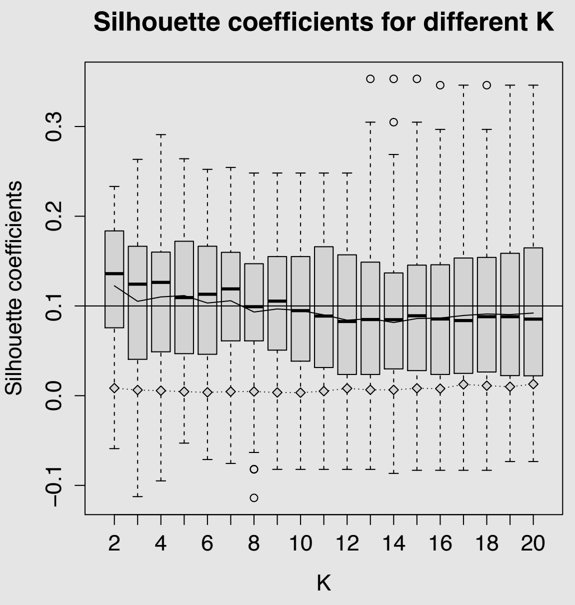
- Clustering of weekly maxima with PAM is fast and gives spatially coherent structures
- Bayesian semi-parametric mixture can handle moderate dimensions and provide credibility intervals

Statistical challenges

- Moving away from bivariate (extremal coefficient) to truly multivariate based clustering algorithms (with Vine ?)
- Moving from semi-parametric to truly parametric spectral models in high dimension (with uncertainty estimates)
- Handling asymptotically independence in geophysical data

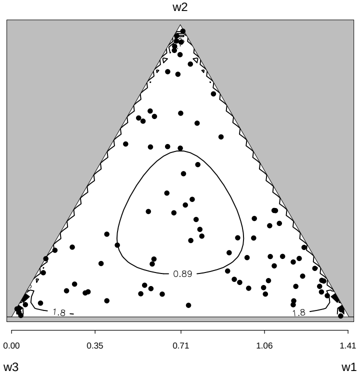
References

- Bernard, E., et al.. Clustering of maxima : Spatial dependencies among heavy rainfall in france. Journal of Climate, 2013, [**R package**].
- Sabourin, A. , Naveau, P. Dirichlet Mixture model for multivariate extremes. To appear in Computational Statistics and Data Analysis. [**R package**].
- Naveau P. et al., Modeling Pairwise Dependence of Maxima in Space. Biometrika, (2009)

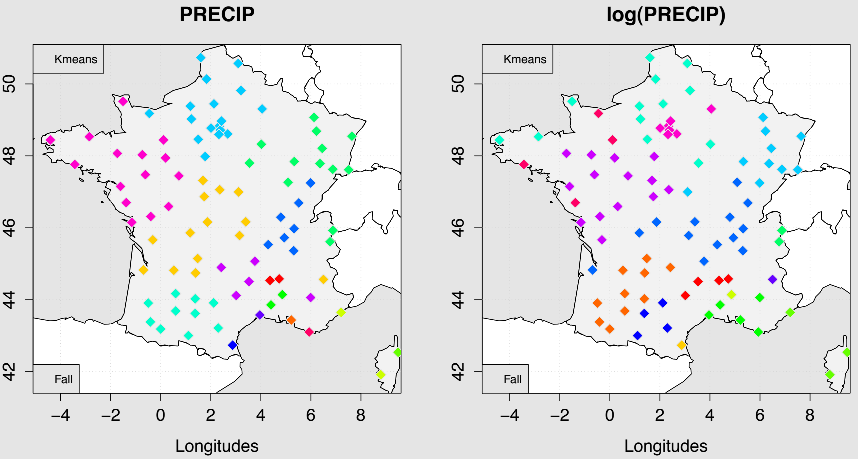

Silhouette coefficients for different K
Silhouette coefficients
0.3
0.2
0.1
0.0
−0.1
2
4
6
8
10
12
14
16
18
20
K

Different results from different Monte Carlo chains ?

Stations 68, 70, 42



Applying the kmeans algorithm to maxima (15 clusters)



Extension for the asymptotically independent case (Ramos and Ledford)

Guillou et al, 2012

η-Madogram

$$\begin{aligned}\nu(\eta) &= \frac{1}{2}\mathbb{E}\left[\left|F_\eta(M_X^{*1/\eta}) - F_\eta(M_Y^{*1/\eta})\right|\right] \\ &= \frac{1}{2}\mathbb{E}\left[\left|F(M_X^{*}) - F(M_Y^{*})\right|\right]\end{aligned}$$

where F_η (resp. F) is the df of $M_X^{*1/\eta}$ and $M_Y^{*1/\eta}$ (resp. of M_X^* and M_Y^*)

$$\nu(\eta) = \frac{V_\eta(1,1)/V_\eta(1,\infty)}{1 + V_\eta(1,1)/V_\eta(1,\infty)} - \frac{1}{2}$$

Extension for the asymptotically independent case (Ramos and Ledford)

Guillou et al, 2012

Estimation of the $\eta-$madogram

$\widehat{F}_X$, resp. $\widehat{F}_Y$, be the empirical df of $M^*_{X_i}$, resp. $M^*_{Y_i}$

$$\widehat{\nu}(\eta) = \frac{1}{2N} \sum_{i=1}^{N} \left| \widehat{F}_X(M^*_{X_i}) - \widehat{F}_Y(M^*_{Y_i}) \right|$$

Theorem 1. *Let* $\left(M^*_{X_i}, M^*_{Y_i}\right)$ *be a sample of* N *bivariate vectors such that*

$$\left(\frac{M^*_{X_i}}{b_n}, \frac{M^*_{Y_i}}{b_n}\right)$$

converges in distribution to a bivariate extreme value distribution with an $\eta-$*extremal function. Then as* $n \to \infty$ *and* $N \to \infty$

$$\sqrt{N}\left(\widehat{\nu}(\eta) - \frac{1}{2}\mathbb{E}|F(M^*_X) - F(M^*_Y)|\right) \xrightarrow{d} \int_{[0,1]^2} N_C(u,v)dJ(u,v)$$

Guillou et al, 2012

Dependence function V_η

$$R_\varepsilon = \{(x,y) : x > \varepsilon, y > \varepsilon\}$$

$M_{\bullet,n,\varepsilon}$ componentwise maxima such that (X_i, Y_i) occur within $R_{\varepsilon b_n}$

$$\lim_{\varepsilon \to 0} \lim_{n \to \infty} \mathbb{P}\left[\frac{M_{X,n,\varepsilon}}{b_n} \le x, \frac{M_{Y,n,\varepsilon}}{b_n} \le y\right] = G_\eta(x,y) = \exp\Big[-V_\eta(x,y)\Big]$$

$$V_\eta(x,y) = \eta \int_0^1 \left[\max\left(\frac{\omega}{x}, \frac{1-\omega}{y}\right)\right]^{\frac{1}{\eta}} dH_\eta(\omega)$$

$\Rightarrow$ V_η homogeneous of order $-1/\eta$: $V_\eta(tx,ty) = t^{-1/\eta} V_\eta(x,y)$

$\Rightarrow$ G_η max-stable: $G_\eta^n(n^\eta u, n^\eta v) = G(x,y)$

Guillou et al, 2012

η-Madogram (cont'd)

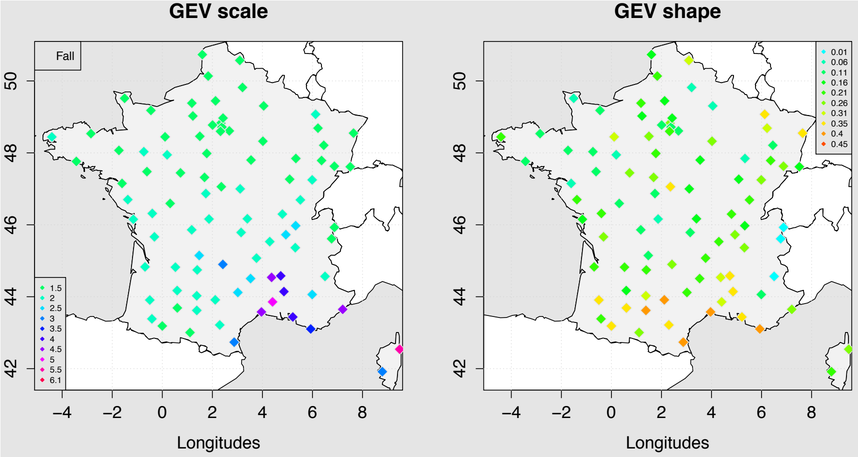
V_η symmetric $\Rightarrow$ extremal coefficient θ :

$$1 \leq \theta := \frac{V_\eta(1,1)}{V_\eta(1,+\infty)} \leq 2$$

$\Rightarrow$ independence ($\theta \to 2$) between the marginal distributions
$\Rightarrow$ dependence ($\theta = 1$)

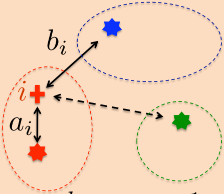
$$\nu(\eta) = \frac{\theta}{1+\theta} - \frac{1}{2}$$

The scale and shape GEV parameters



- **Clustering validation**

SILHOUETTE COEFFICIENT

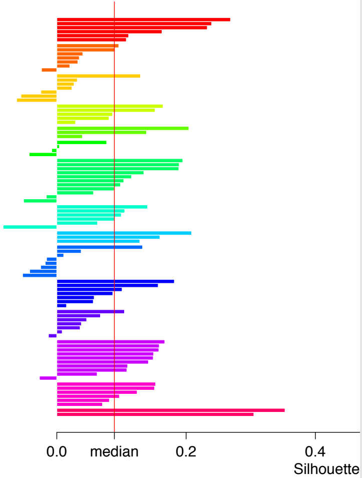


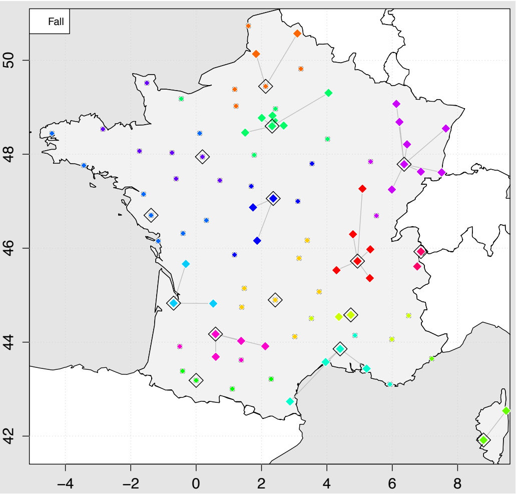
$$\boxed{s_i = \frac{b_i - a_i}{\max(a_i, b_i)}}$$

$$a_i \ll b_i, \quad s_i \approx 1 \quad \rightarrow \text{Well classified}$$

$$a_i \sim b_i, \quad s_i \approx 0 \quad \rightarrow \text{Neutral}$$

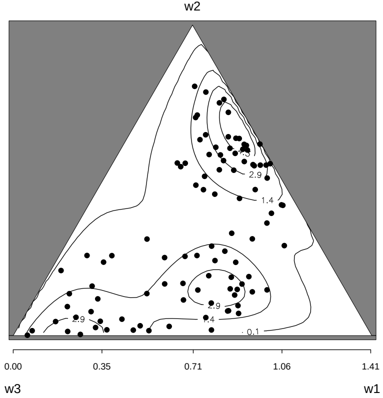
$$a_i \gg b_i, \quad s_i \approx -1 \quad \rightarrow \text{Badly classified}$$


0.0
median
0.2
0.4
Silhouette


Fall
50
48
46
44
42
−4
−2
0
2
4
6
8

Simulation example with $d = 3$ and $k = 3$

Simulated points with true density



Predictive density