Bayesian nonparametrics for multivariate extremes including censored data

Anne Sabourin

PhD advisors:

Anne-Laure Fougères (Lyon 1), Philippe Naveau (LSCE, Saclay).

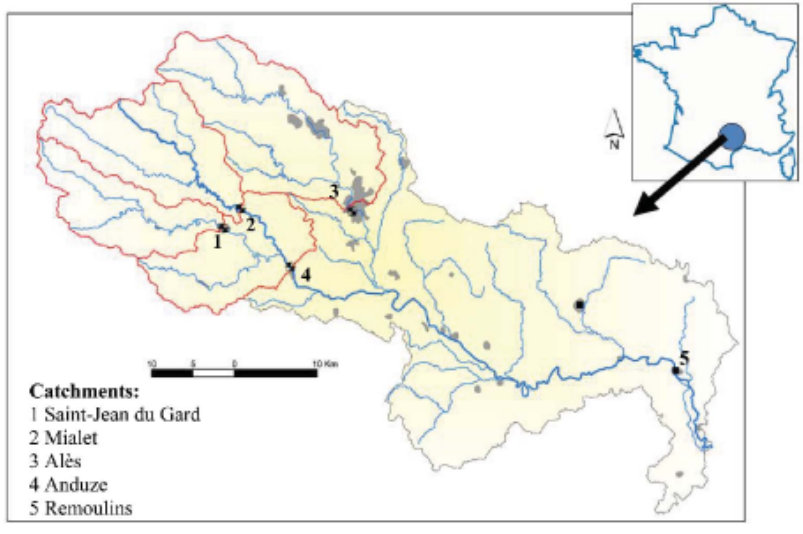
Joint work with Benjamin Renard, IRSTEA, France.

EVT 2013, Vimeiro

September 10, 2013

Censored Multivariate extremes: Why bother ?

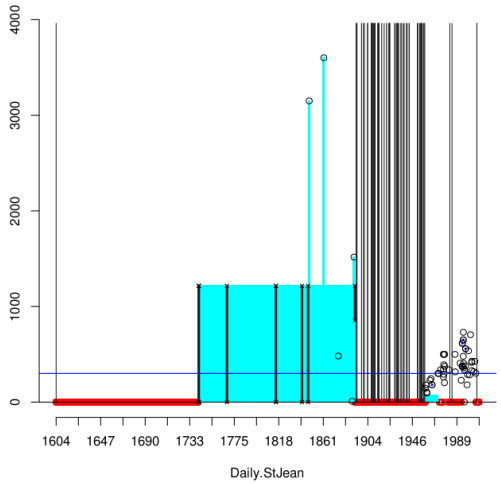
- ex: Daily streamflow at 4 neighbouring sites (Gardons, Cévennes, France).
- Return levels for jointly extreme events ?
- Recent, 'clean' series very short
- Censored historical data availble from archives, variable 'perception thresholds' for floods. Earliest: 1604.

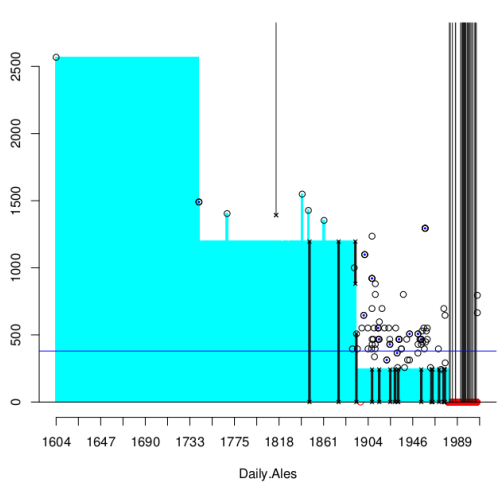


How to use all different kinds of data ?

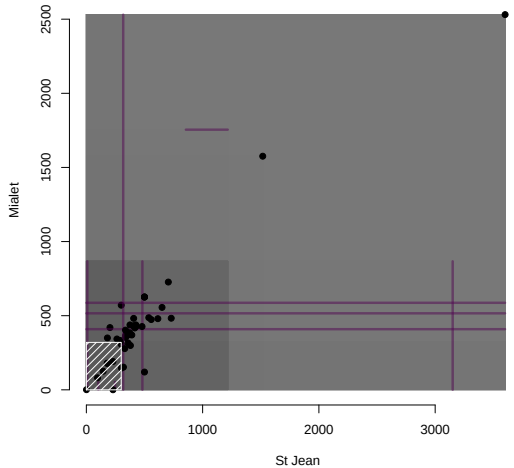
Censored data: pairwise plots

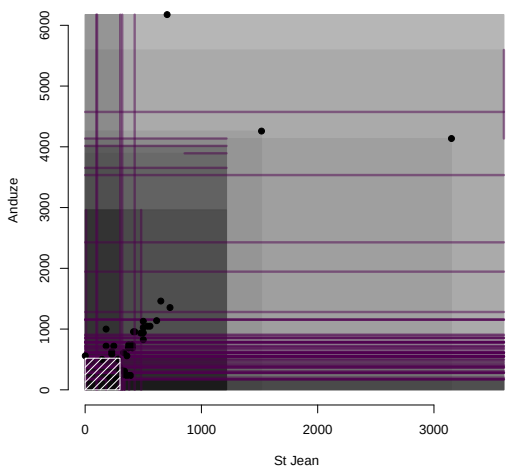
Univariate series:

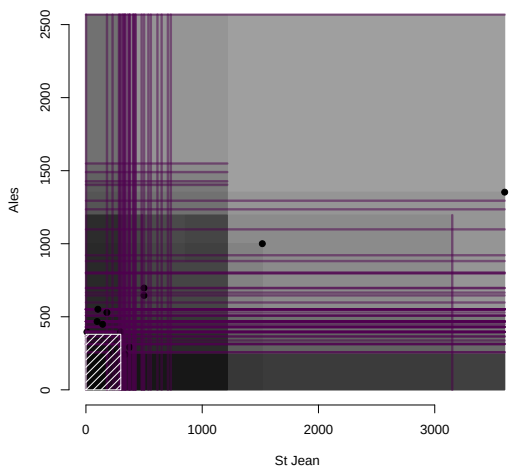




Bivariate:







Purposes

- Take into account as many data as possible
  $\rightarrow$ Censored likelihood, integration problems

- Estimate marginal GPD parameters + dependence structure jointly
  $\rightarrow$ additional Gibbs step in a MCMC algo.

- Probability of failure regions (return periods for jointly extreme events)
  $\rightarrow$ Model for excesses (POT-Poisson).

Approach: use Dirichlet consistency properties

- Angular measure model on the simplex.

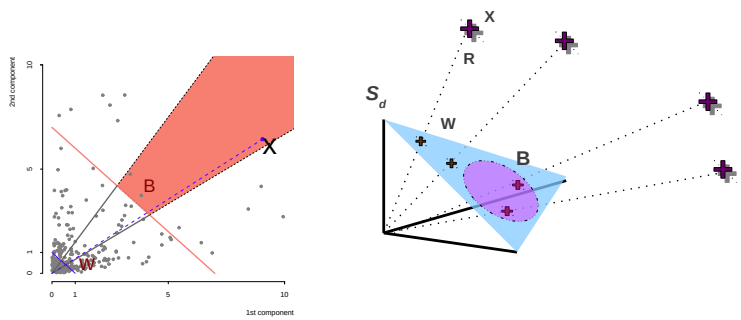
- Dirichlet distributions: Nice marginalization and conditioning properties.

- Dirichlet mixture model for angular measures
  Boldi & Davison, 2007 ; Sabourin & Naveau, 2013 :
  - Flexible, non parametric.
  - Reversible jumps - MCMC algorithm available.

Adaptation needed: censored likelihood and variable threshold
$\Rightarrow$ **Data augmentation**

Excess Data: polar decomposition

- Vectorial obs $\mathbf{Y}_t = (Y_{1,t}, \ldots, Y_{d,t})$; *i.i.d.*
- Margin $Y_{j,t} \sim F_j$: Pareto above threshold v_j.
- Dependence defined among the $X_{j,t} = -1/log\, F_{j,t}(Y_j, t)$: standard Fréchet
- Polar coordinates:
  $R = \|X\|_1 = \sum_j X_j$: 'radius' (norm), $\quad \mathbf{W} = \frac{\mathbf{X}}{R}$: 'angle'



$$\mathbf{W} \in \text{simplex } \mathbf{S}_d = \{(w_1, \ldots w_d) : \sum w_j = 1, w_j \geq 0\}\,.$$

Angular measure models

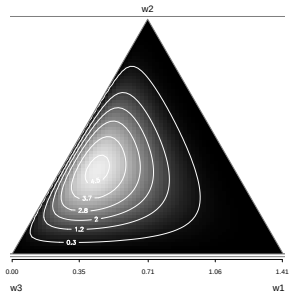
- Joint distribution of excesses: product measure in polar coordinates.
$$P(R > r, \mathbf{W} \in B) \propto \frac{1}{r} H(B)$$
- H: angular measure, distribution of angles. Non parametric: Valid *iff* $\int_{\mathbf{S}_d} w_i \, \mathrm{d}H(\mathbf{w}) = \frac{1}{d}$ $(1 \leq i \leq d)$

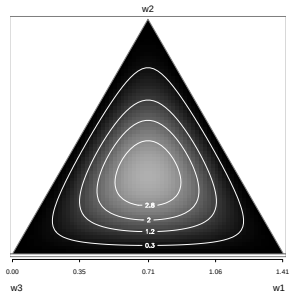
- *Inference*: **Empirically**, Einmahl et.al., 2001, Einmahl, Segers, 2009, Guillotte et al, 2011. : No explicit expression of asymptotic variance, Bayesian inference with $d = 2$ only.
Restriction to **parametric subclass** Gumbel, Coles&Tawn, Cooley *et.al.*. . . : logistic family, Pairwise-Beta . . .

- Preferred here: **Dirichlet Mixture model**: Compromise
  - Flexibility (weakly dense family)
  - Uncertainty quantification through parameters, Bayesian implementation.

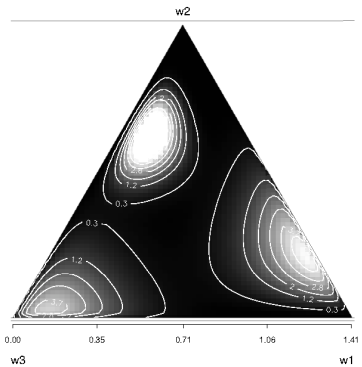
Dirichlet mixture model for angular measure

- Dirichlet distribution: generalization of Beta to higher dimensions.
- 1 location parameter (point on the simplex): 'center' + 1 concentration parameter.





- Dirichlet mixture model, k components: $h(\mathbf{w}) = \sum_{m=1}^{k} p_m \, \text{diri}_m(\mathbf{w})$. Boldi, Davison, 2007



- H valid $\Leftrightarrow$ center of mass of the location parameters = center of the simplex.

Issues: moments constraints and censored data

(i) Sampling the posterior distribution with MCMC methods. Constraints $\Rightarrow$ Sampling issues

- Re-parametrization: No more constraint, sampling is manageable for $d = 5$: **Sabourin, Naveau, 2013**

(ii) Inference with censored data:

- Variable threshold: changing normalizing constants.
- Censored data = segments or boxes. Integrating the likelihood (density $\frac{\mathrm{d}r}{r^2}\,\mathrm{d}H(\mathbf{w})$) on rectangles ?

Sabourin, ***under review*** ; **Sabourin**, **Renard**, ***in preparation***

Issues: moments constraints and censored data

(i) Sampling the posterior distribution with MCMC methods. Constraints $\Rightarrow$ Sampling issues

- Re-parametrization: No more constraint, sampling is manageable for $d = 5$: **Sabourin, Naveau, 2013**

(ii) Inference with censored data:

- Variable threshold: changing normalizing constants.
- Censored data = segments or boxes. Integrating the likelihood (density $\frac{\mathrm{d}r}{r^2}\,\mathrm{d}H(\mathbf{w})$) on rectangles ?

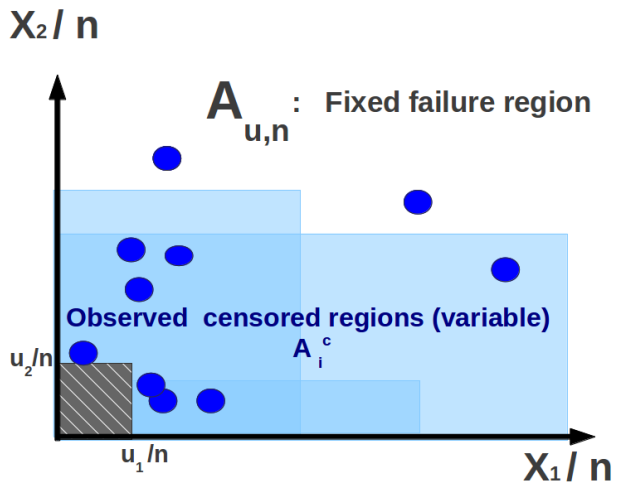
Sabourin, ***under review*** ; **Sabourin**, **Renard**, ***in preparation***

Poisson model for variable threshold

$$\left\{ \left(\frac{t}{n}, \frac{\mathbf{X}_t}{n} \right), \ 1 \leq t \leq n \right\} \sim \text{Poisson Process } (\text{Leb} \times \lambda) \quad \text{on } [0,1] \times A_{u,n}$$

λ: ' exponent measure', with Dirichlet Mixture angular component

$$\frac{\mathrm{d}\lambda}{\mathrm{d}r \times \mathrm{d}\mathbf{w}}(r, \mathbf{w}) = \frac{d}{r^2} h(\mathbf{w}) \, .$$



'A_i' data overlapping threshold: included in Poisson likelihood as

$$N \left\{ \left(\frac{t_2}{n} - \frac{t_1}{n} \right) \times \frac{1}{n} A_i \right\} = 0$$

'Censored' likelihood: model density integrated over boxes

- Ledford & Tawn, 1996: GEV parametric model, partially extreme data censored at threshold.
- Here: more general censoring framework, Poisson likelihood $\frac{\mathrm{d}\lambda}{\mathrm{d}x}$, non parametric, no explicit expression.
- Two terms without closed form:
  - Censored regions A_i overlapping threshold:

$$\exp\left\{-(t_2 - t_1)\lambda(A_i)\right\}$$

  - Classical censoring above threshold

$$\int_{\text{censored region}} \frac{\mathrm{d}\lambda}{\mathrm{d}\mathbf{x}}\,.$$

Data augmentation

One more Gibbs step, no more numerical integration.

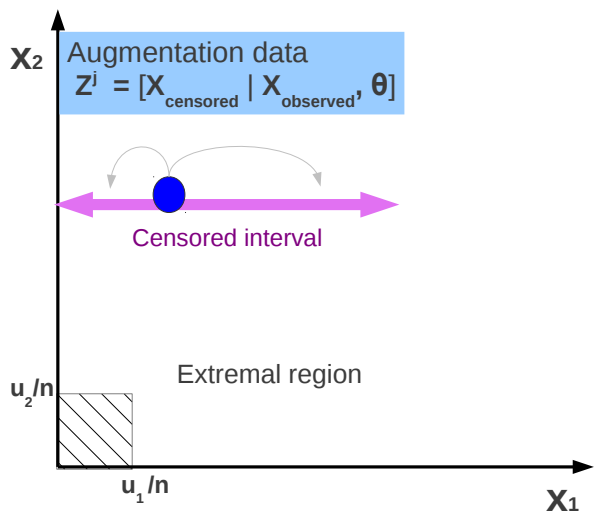
- Objective: sample $[\theta|Obs]$. (Parameter space: Θ)
- Additional variables (replace missing data component): $\mathcal{Z}$
- Full conditionals $[Z_i|Z_{j\neq j}, \theta, Obs]$ explicit (Thanks Dirichlet): $\rightarrow$ Gibbs sampling.
- Sample $[z, \theta|Obs]_+$ (augmented distribution) on $\Theta \times \mathcal{Z}$.

Censored regions above threshold

$$\int_{\text{Censored region}} \frac{d\lambda}{dx}\, dx_{j_1:j_r} \quad :$$

Generate missing components under univariate conditional distributions

$$\mathbf{Z}^j_{1:r} \sim [X_{\text{missing}} | X_{\text{obs}}, \theta]$$

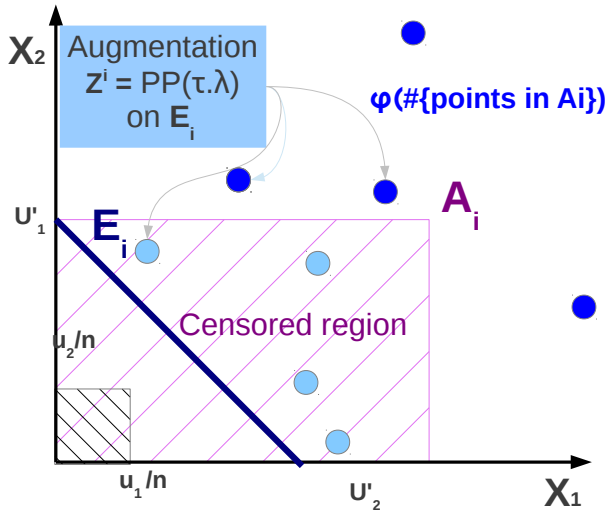


Dirichlet $\Rightarrow$ Explicit univariate conditionals

Exact sampling of censored data on censored interval

Censored regions overlapping threshold

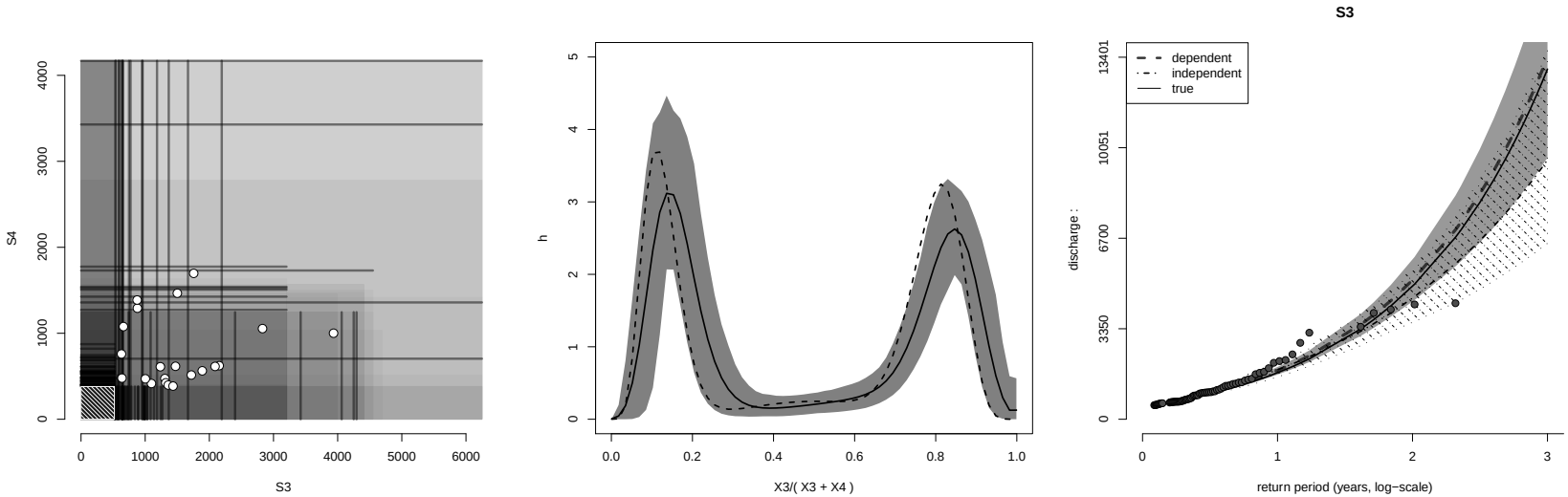
$$e^{-(t_{2,i}-t_{1,i})\lambda(A_i)} \Leftrightarrow \begin{cases} \text{augmentation Poisson process } N_i \text{ on } E_i \supset A_i. \\ + \\ \text{Functional } \varphi(N_i) \end{cases}$$



$$[z, \theta | Obs] \propto \underbrace{\ldots}_{\text{density terms, prior, augmented missing components}} [N_i]\varphi(N_i)$$

Results on simulated data

- Angular measure: Dirichlet, $d = 4$, $k = 3$ mixture components
- Censoring: same pattern as real data.



Pair $(S3, S4)$: data, predictive *a posteriori* of the angular measure, quantile curve for $S3$.

Conclusion

- Accounting for all types of censored data: manageable using Dirichlet consistency properties (marignalization, conditioning)

- High dimension (GCM grid, spatial fields)?
  - Impose a reasonable structure (sparse) on Dirichlet parameters
  - Dirichlet Process ? Challenges :
    Discrete random measure $\neq$ continuous framework for GPP's.

Bibliographie I

M.-O. Boldi and A. C. Davison.

A mixture model for multivariate extremes.
JRSS: Series B (Statistical Methodology), 69(2):217–229, 2007.

Sabourin, A., Naveau, P.

Bayesian Dirichlet mixture model for multivariate extremes: a re-parametrization.
CSDA, 2013.

Ledford, A. and Tawn, J. (1996).

Statistics for near independence in multivariate extreme values.
Biometrika, 83(1):169–187.

Schnedler, W. (2005).

Likelihood estimation for censored random vectors.
Econometric Reviews, 24(2):195–217.

Tanner, M. and Wong, W. (1987).

The calculation of posterior distributions by data augmentation.
Journal of the American Statistical Association, 82(398):528–540.

Van Dyk, D. and Meng, X. (2001).

The art of data augmentation.
Journal of Computational and Graphical Statistics, 10(1):1–50.

Gómez, G., Calle, M. L., and Oller, R.

Frequentist and bayesian approaches for interval-censored data.
Statistical Papers, 45(2):139–173, 2004.

Bibliographie II

Hosking, J.R.M. and Wallis, J.R..

Regional frequency analysis: an approach based on L-moments
Cambridge University Press, 2005.

Neppel, L., Renard, B., Lang, M., Ayral, P., Coeur, D., Gaume, E., Jacob, N., Payrastre, O., Pobanz, K., and Vinet, F. (2010).

Flood frequency analysis using historical data: accounting for random and systematic errors.
Hydrological Sciences Journal–Journal des Sciences Hydrologiques, 55(2):192–208.

Details: Augmentation Poisson process

$$e^{-(t_{2,i}-t_{1,i})\lambda(A_i)} \text{ ??}$$

Proposal $\mathbf{Z}^i \sim PP(\tilde{\lambda}^i)$; define f^i s.t.

$$\mathbb{E}\left\{e^{-\sum_j f^i(\mathbf{Z}^i{}_j)}\right\} = \mathcal{L}\mathrm{apl}_{\mathbf{Z}^i}(f^i) = e^{-\int 1-e^{-f}\,\mathrm{d}\tilde{\lambda}^i} = e^{-n_i\lambda(A_i)}$$

$$[\mathbf{Z},\theta|\mathbf{O}]_+ \;\propto\; [\theta]\prod_j\left\{[\mathbf{Z}^{\mathbf{j}}_{\mathbf{1:r}}|\mathbf{O},\theta]\right\}\prod_i\left\{[\mathbf{Z}^{\mathbf{i}}|\theta]e^{-\sum_j f^i(\mathbf{Z}^i{}_j)}\right\},$$

with

$$\int [\mathbf{Z}^{\mathbf{i}}|\theta]e^{-\sum_j f^i(\mathbf{Z}^i{}_j)}\,\mathrm{d}\mathbf{Z}^i \;= e^{-n_i\lambda(A_i)}\,.$$