Multivariate extremes with censored data: how Dirichlet mixtures can help.

Anne Sabourin

Institut Mines-Télécom, Télécom ParisTech, CNRS-LTCI

Joint work with Philippe Naveau (LSCE, Saclay), Anne-Laure Fougères (ICJ, Lyon 1), Benjamin Renard (IRSTEA, Lyon).
Air quality

- Five air pollutants recorded in Leeds (UK), daily.
- Health issue: probability of a joint (simultaneous) excess of alert thresholds?
- Probability of regions far from the origin?
Censored Multivariate extremes: floods in the ‘Gardons’

joint work with Benjamin Renard

- Daily streamflow at 4 neighbouring sites
  (St Jean du Gard, Mialet, Anduze, Alès).
- Probability of simultaneous floods?
- Historical data (earliest: 1604) → censored data;
- Not enough clean data (6 days).

Gard river Neppel et al. (2010)
Multivariate extremes

- Random vectors $\mathbf{Y} = (Y_1, \ldots, Y_d)$; $Y_j \geq 0$

- Margins: $Y_j \sim F_j, 1 \leq j \leq d$ (any).

- **Standardization** (unit Fréchet margins): $X_j = -1/\log F_j(Y_j)$

- Joint extremes: $\mathbf{X}$’s distribution above large thresholds?

$$P(\mathbf{X} \in A \mid \mathbf{X} \in A_0)? \quad (A \subset A_0, \mathbf{0} \notin A_0), \ A_0 \ ‘far \ from \ the \ origin’.$$
Polar decomposition

- Polar coordinates: \( R = \sum_{j=1}^{d} X_j \) (\( L_1 \) norm) ; \( W = \frac{X}{R} \).

- \( W \in \text{simplex } S_d = \{ w : w_j \geq 0, \sum_j w_j = 1 \} \).

Characterize \( \mathbb{P}(X \in A \mid A \in A_0) \)

\[ \iff \]

Characterize \( \mathbb{P}(R > r, W \in B \mid R > r_0) \)
Limit law of excesses, fundamental result

- Radial homogeneity (under hypothesis of regular variation, i.e. convergence of normalized componentwise maxima)

\[ \mathbb{P}(R > rt, W \in B \mid R \geq t) \rightarrow_{t \to \infty} \frac{1}{r} H(B) \]

Model for excesses above large radial threshold \( r_0 \):

\[ \mathbb{P}(R > r, W \in B \mid R \geq r_0) \approx \frac{r_0}{r} H(B) \]

- Above large thresholds \( r_0 \), \( R \perp W \);
- \( H (+ \text{ margins}) \) rules the joint distribution
Angular distribution

- $H$ (+ margins) rules the joint distribution enditemize

- One condition only for genuine $H$: moments constraint,

$$\int w \, dH(w) = (\frac{1}{d}, \ldots, \frac{1}{d}). \quad \text{Center of mass } = \text{center of simplex.}$$

- Few constraints: non parametric family!
Estimating the angular measure

  
  **Issues**: asymptotic variance, Bayesian inference with $d > 2$, censored data

- **Compromise**: Mixture of countably many parametric models → Infinite-dimensional model

  **Dirichlet mixture model**
  
  (Boldi, Davison, 2007; S., Naveau, 2013)

  - How to deal with the **moments constraint** on $H$?
  - **MCMC** methods work in moderate dimension ($d = 5$)?
  - Does it still work with **censored data**?
Dirichlet distribution

\[ \forall \mathbf{w} \in \mathcal{S}_d, \quad \text{diri}(\mathbf{w} \mid \mu, \nu) = \frac{\Gamma(\nu)}{\prod_{i=1}^{d} \Gamma(\nu_i)} \prod_{i=1}^{d} w_i^{\nu_i\mu_i - 1}. \]

- \( \mu \in \mathcal{S}_d \): location parameter (point on the simplex) : ‘center’;
- \( \nu > 0 \): concentration parameter.
Dirichlet distribution

\[ \forall \mathbf{w} \in \mathbb{S}_d, \text{diri}(\mathbf{w} \mid \mu, \nu) = \frac{\Gamma(\nu)}{\prod_{i=1}^{d} \Gamma(\nu\mu_i)} \prod_{i=1}^{d} \mathbf{w}_i^{\nu\mu_i - 1}. \]

- \( \mu \in \mathbb{S}_d \): location parameter (point on the simplex): ‘center’;
- \( \nu > 0 \): concentration parameter.
Dirichlet mixture model

- $\mu = \mu_{1:k}$, $\nu = \nu_{1:k}$, $p = p_{1:k}$, $\psi = (\mu, p, \nu)$,

$$h_{\psi}(w) = \sum_{m=1}^{k} p_m \text{diri}(w | \mu_{m}, \nu_m)$$

- Moments constraint $\to$ on $(\mu, p)$:

$$\sum_{m=1}^{k} p_m \mu_{m} = \left( \frac{1}{d}, \ldots, \frac{1}{d} \right).$$

Weakly dense family ($k \in \mathbb{N}$) in the space of admissible angular measures.
Bayesian inference with non censored data

- Moments constraints
  - Prior construction?
  - Parameter generation for MCMC sampling?
  
  Difficult for dimension $\geq 2$.

- Re-parametrization S., Naveau (13) : work with unconstrained parameter in a product space
  
  - Weak posterior consistency
  
  - MCMC with reversible jumps manageable in moderate dimension ($\approx 5$).
Results in the re-parametrized version

- **Theoretically (Asymptotics):**
  - Posterior consistency: \( \forall U \) weakly open in \( \Theta \), containing \( \theta_0 \),
    \[ \pi_n(U) = \pi(U|\text{data}_{1:n}) \xrightarrow{n \to \infty} 1 \]
  - Markov chain’s ergodicity: \( \sum_{t=1}^{T} g(\theta_t) \xrightarrow{T \to \infty} \mathbb{E}_{\pi_n}(g) \)

- **Empirically:** convergence checks.
  Better mixing:

![Graphs showing empirical checks and better mixing](image-url)
Results in the re-parametrized version

- Theoretically (Asymptotics):
  - Posterior consistency: $\forall U$ weakly open in $\Theta$, containing $\theta_0$, $\pi_n(U) = \pi(U|\text{data}_{1:n}) \xrightarrow{n \to \infty} 1$.
  - Markov chain’s ergodicity: $\sum_{t=1}^{T} g(\theta_t) \xrightarrow{T \to \infty} \mathbb{E}_{\pi_n}(g)$

- Empirically: convergence checks.
  Better coverage of credible sets ($d=5$, bivariate margins, simulated data)
Results in the re-parametrized version

- Theoretically (Asymptotics):
  - Posterior consistency: $\forall U$ weakly open in $\Theta$, containing $\theta_0$, $\pi_n(U) = \pi(U|\text{data}_{1:n}) \xrightarrow{n \to \infty} 1$.
  - Markov chain’s ergodicity: $\sum_{t=1}^{T} g(\theta_t) \xrightarrow{T \to \infty} \mathbb{E}_{\pi_n}(g)$

- Empirically: convergence checks.
  for $d = 2$ performance $\sim$ bivariate non-parametric model


Solid line: DM. dotted: alternative non parametric model)
Inference with censored data

  - GEV models
  - Explicit expression for censored likelihood.
Issues

- Intervals overlapping threshold: extreme data or not?

- Censored likelihood: density $\frac{dr}{r^2} dH(w)$ integrated over boxes: no explicit expression.
Undetermined data (overlapping threshold)

Considering 'undetermined data' as missing $\Rightarrow$ biais!
Undetermined data (overlapping threshold)

Data in region $R \iff$ not in region $R^c$ . . .
Well defined likelihood in a Poisson model
Poisson model

\[ \left\{ \left( \frac{t}{n}, \frac{X_t}{n} \right) , 1 \leq t \leq n \right\} \sim \text{Poisson Process}(\text{Leb} \times \mu_*) \text{ on } [0, 1] \times R^c_{u,n} \]

\[
\frac{d\mu_*}{dr \times dw}(r, w) = \frac{d}{r^2} \times \text{dirichlet mixture}(w).
\]

Likelihood of overlapping data :

\[
P \left[ N \left\{ \left( \frac{t_2}{n} - \frac{t_1}{n} \right) \times \frac{1}{n} R^c_i \right\} = 0 \right] = \exp \left[ -(t_2 - t_1)\mu_*(R^c_i) \right]
\]
Numerical issues

Two types of terms in the likelihood without closed form:

- Censored regions overlapping threshold:
  \[ \exp \left\{ -\left( t_2 - t_1 \right) \mu_i (R_i^c) \right\}; \]

- Classical censoring above threshold:
  \[ \int_{\text{censored region}} \frac{d\lambda}{d\mathbf{x}}. \]
‘Censored’ likelihood : and data augmentation

- Data augmentation: Generate missing components under univariate conditional distributions.

  *One more Gibbs step, no more numerical integration.*

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

Dirichlet $\Rightarrow$ Explicit univariate conditionals

Exact sampling of censored data on censored interval
Censored regions overlapping threshold

\[ e^{-(t_2,i-t_1,i)\mu_*(R_i^c)} \Leftrightarrow \begin{cases} 	ext{augmentation Poisson process } N_i \text{ on } E_i \supset R_i^c. \\ + \\ \text{Functional } \varphi(N_i) = \varphi(\#\{\text{points in } R_i^c\}) \end{cases} \]

\[ \left[ z, \theta | \text{Obs} \right] \propto [N_i] \varphi(N_i) \]

density terms, prior, augmented missing components
Simulated data (Dirichlet, $d = 4$, $k = 3$ components), same censoring as real data

Pairwise plot and angular measure density (true/ posterior predictive)
Angular predictive density for Gardons data
Conclusion

- Bayesian Dirichlet model for multivariate large excesses:
  - ‘non’ parametric, suitable for moderate dimension, adaptable to censored data.
  - Two packages R:
    - `DiriXtremes`, MCMC algorithm for Dirichlet mixtures,
    - `DiriCens`, implementation with censored data.

- Towards high dimension (GCM grid, spatial fields)
  - Impose reasonable structure (sparse) on Dirichlet parameters?

- Possible application: Posterior sample → Simulation of regional extremes?
References

M.-O. Boldi and A. C. Davison.
A mixture model for multivariate extremes.

Gómez, G., Calle, M. L., and Oller, R.
Frequentist and bayesian approaches for interval-censored data.

Statistics for near independence in multivariate extreme values.

Flood frequency analysis using historical data: accounting for random and systematic errors.

Bayesian Dirichlet mixture model for multivariate extremes: a re-parametrization.
*Computation. Stat and Data Analysis*

Likelihood estimation for censored random vectors.

The art of data augmentation.