A Bayesian semi-parametric approach to extreme regime identification

Fernando Ferraz do Nascimento
CGBEST-UFPI (Federal University of Piaui, Brazil)
Dani Gamerman
DME-UFRJ (Federal University of Rio de Janeiro, Brazil)
Richard Davis
(Columbia University, USA)

1 Introduction

- Limiting tail behavior of distributions are known to follow one of three possible limiting distributions, depending on the domain of attraction of the observational model
- Fisher and Tippett (1928): Three different possibilities for the tail shape are named Gumbel ($\xi = 0$), Fréchet ($\xi > 0$) and Weibull ($\xi < 0$).

Pickands (1975) proved that if X is a random variable whose distribution function F, with endpoint x_F, is in the domain of attraction of a GEV distribution, then as $u \to x_F$, the conditional distribution function $F(x|u) = P(X \le u + x|X > u)$ is the distribution function of a generalized Pareto distribution (GPD), whose density is provided below.

$$g(x|\Psi) = \begin{cases} \frac{1}{\sigma}\left(1+\xi\frac{(x-u)}{\sigma}\right)^{-(1+\xi)/\xi}, & \text{if } \xi \neq 0 \\ \frac{1}{\sigma}\exp\{-(x-u)/\sigma\}, & \text{if } \xi = 0 \end{cases}, \quad (1)$$

where $x - u > 0$ for $\xi \ge 0$ and $0 \le x - u < -\sigma/\xi$ for $\xi < 0$. Thus, the GPD is always bounded from below by u, is bounded from above by $u - \sigma/\xi$ if $\xi < 0$ and unbounded from above if $\xi \ge 0$.

- Importance of a shape identification signal: Coles (2001), on pages 80 and 81, report numerical instabilities in the likelihood function when $\xi \approx 0$.
- Our approach: Evaluation of the probabilities of the three limiting regimes, afforded by a mixed distribution for the shape ξ, using non-informative priors
- Stepehnson and Tawn (2004): Similar approach in the study of maxima via the GEV. Only considered the possibilities $\xi = 0$ and $\xi \neq 0$, and the prior probabilities of the regimes are fixed

2 Model

Observational model of Nascimento et al. (2012)
Their mixture model with Gammas and GPD (MGPD, in short) has density given by

$$f(x|\theta,\mathbf{p},\Psi) = \begin{cases} \sum_{j=1}^{k} p_j f_G(x \mid \mu_j,\nu_j), & \text{if } x \le u \\ \left[1-\sum_{j=1}^{k} p_j F_G(x \mid \mu_j,\nu_j)\right] g(x|\Psi), & \text{if } x > u \end{cases}$$

where f_G and F_G are respectively the density and cumulative functions of a Gamma distributions, and g is the GPD density with parameters $\Psi = (\xi, \sigma, u)$.

2.1 Prior distribution

A mixed distribution is proposed for ξ, considering separately the probabilities associated with the three possible limiting tail behaviors: Gumbel ($\xi = 0$), Fréchet ($\xi > 0$) and Weibull ($\xi < 0$).
This mixed setting can be rephrased with the insertion of latent quantities (Z_ξ, Q_ξ). The three-dimensional quantity $Z_\xi = (Z_\xi^+, Z_\xi^0, Z_\xi^-)$, where $Z_\xi^+ + Z_\xi^0 + Z_\xi^- = 1$, indicates the signal of ξ, where $Z_\xi^+ = 1$ indicates that $\xi > 0$, $Z_\xi^0 = 1$ indicates that $\xi = 0$ and $Z_\xi^- = 1$ indicates that $\xi < 0$.
The quantity Q_ξ shows the probability of ξ be positive, negative or null, given by $Q_\xi = (q_\xi^+, q_\xi^0, q_\xi^-)$, where $q_\xi^+ + q_\xi^0 + q_\xi^- = 1$. The joint prior of these parameters is given by

$$\begin{aligned} p(\xi, Z_\xi, Q_\xi) &= p(\xi \mid Z_\xi, Q_\xi)p(Z_\xi \mid Q_\xi)p(Q_\xi) \\ &= p(\xi \mid Z_\xi)p(Z_\xi \mid Q_\xi)p(Q_\xi), \end{aligned}$$

The distribution of $\xi \mid Z_\xi$ is assigned in the following specification

$$\begin{aligned} \xi \mid Z_\xi^+ = 1 &\sim Gamma(a_\xi, b_\xi), \\ \xi \mid Z_\xi^0 = 1 &\sim \delta_{\xi=0} \\ \xi \mid Z_\xi^- = 1 &\sim U(-0.5, 0), \end{aligned}$$

The conditional distribution $Z_\xi \mid Q_\xi$: multinomial prior with parameters $(q_\xi^+, q_\xi^0, q_\xi^-)$
Q_ξ: Dirichlet distribution with parameter $\alpha_\xi = (\alpha_+, \alpha_0, \alpha_-)$. and 0, otherwise. In the lack of prior information, the vector α_ξ may be chosen to provide little information on Q_ξ.
Regimes probabilities are estimated from the data over the entire range of possibilities $[0, 1]$, unlike Stephenson and Tawn (2004) where only a discrete subset of these values are allowed.

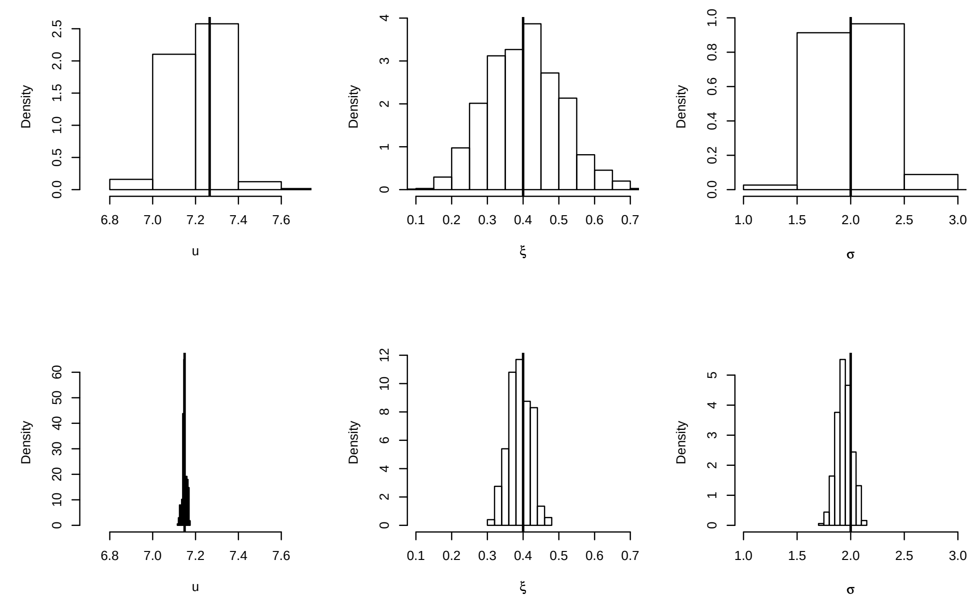
2.2 Posterior distribution

$$\pi(\theta,\mathbf{p},\Psi,Z_\xi,Q_\xi|\mathbf{x}) \propto \prod_{i:x_i\le u}\left(\sum_{j=1}^{k} p_j f_G(x_i|\mu_j,\eta_j)\right)$$
$$\prod_{i:x_i\ge u}\left[1-\sum_{j=1}^{k} p_j F_G(u|\mu_j,\eta_j)\right] g(x_i \mid \Psi)$$
$$\prod_{j=1}^{k}\left[p_j^{\gamma_j}\eta_j^{c_j-1}\exp\left(-\frac{d_j}{\eta_j}\right)\mu_j^{a_j-1}\exp\left(-\frac{b_j}{\mu_j}\right)\right]\frac{1}{2}\exp\left(-\frac{(u-\mu_u)}{2\sigma_u}\right)^2\frac{1}{\sigma}$$
$$p(\xi \mid Z_\xi)(q_\xi^+)^{z_\xi^+ + \alpha_+}(q_\xi^0)^{z_\xi^0+\alpha_0}(q_\xi^-)^{z_\xi^- + \alpha_-},$$

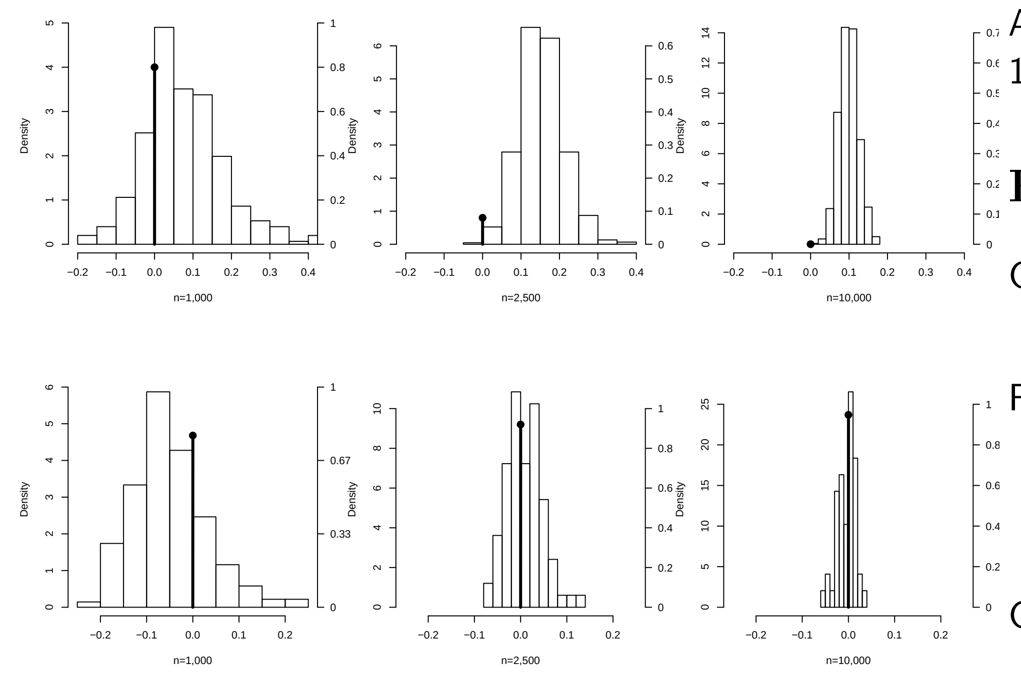
where the first and second line above comes from the likelihood, the third and fourth lines refers to the prior density of parameters, where $p(\xi \mid Z_\xi)$ is the density of (2), with a mixed distribution.
Estimation: MCMC methods: Metropolis-Hastings algorithm (Gamerman and Lopes, 2006)

3 Simulations

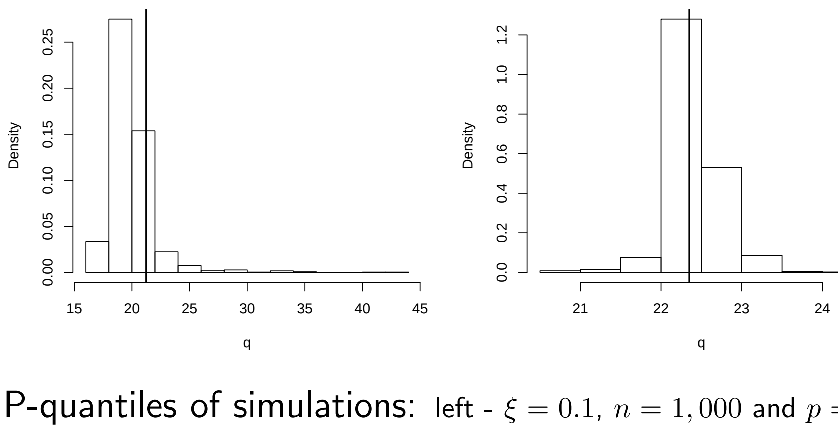
- Simulation studies were performed in different settings of parameter values
- Performed with samples of sizes 1,000, 2,500 and 10,000
- ξ was fixed with the values $\xi = (-0.4, -0.1, 0, 0.1, 0.4, 1)$



Posterior histogram for the GPD parameters with $\xi = 0.4$: top row - $n = 1,000$; bottom row - $n = 10,000$. Vertical lines: True value of parameter.



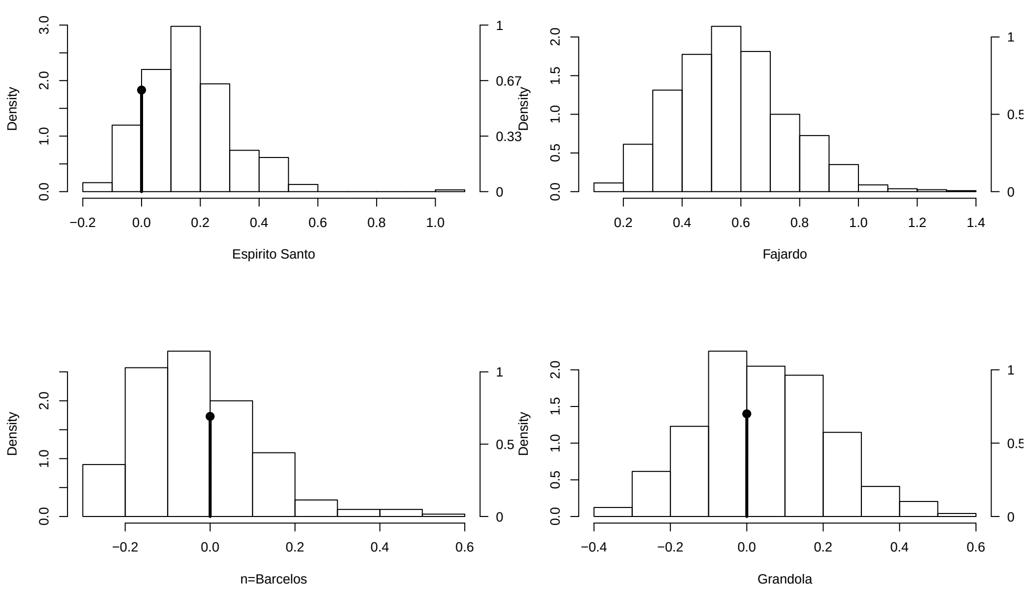
Posterior histogram for ξ in simulations: top row - $\xi = 0.1$; bottom row - $\xi = 0$. Vertical lines: probability lumps at $\xi = 0$. Their respective posterior probabilities are scaled according to the right y-axis



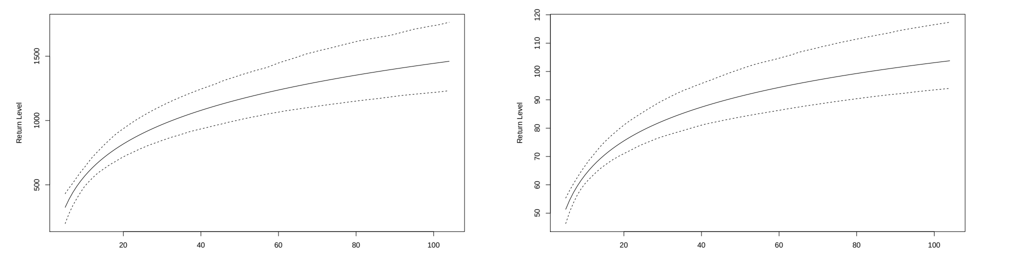
P-quantiles of simulations: left - $\xi = 0.1$, $n = 1,000$ and $p = 0.999$; right - $\xi = 0$, $n = 10,000$ and $p = 0.9999$. Vertical lines: true values of quantiles.

4 Applications

- Two environmental datasets
- Fajardo and Espiritu Santo rivers in Puerto Rico. The data were recorded daily from April 1967 to September 2002
- Monitoring stations in Portugal: Barcelos in the North, and Grandola in the South



Posterior distribution for ξ in the applications. The respective posterior mass probabilities at 0 are scaled according to the right y-axis.



Expected return levels for the applications. Left: Espiritu Santo river flow. Right: rainfall at Barcelos station. Full line: posterior mean; dashed lines: 95% credibility limits.

5 Conclusions

- New approach to GPD, where ξ has a mixed nature
- Simulations: efficiency in detecting the true ξ regime for n large
- Applications: Appoint some data with large probability of Gumbel regime
- Precise identification of large quantiles

5.1 Thanks

The research of the first author was supported by grants from CAPES, from Brazil. The research of the second author was supported by grants from CNPq and FAPERJ, from Brazil. The research of the third author was supported in part by ARO MURI grant W911NF-12-1-0385 and NSF grant DMS 1107031.

References

Coles, S. G. (2001). An Introduction to Statistical Modelling of Extreme Values. Springer.

Fisher, R. A. and Tippett, L. H. C. (1928). "On the estimation of the frequency distributions of the largest and smallest sumber of a sample." Proceedings of the Cambridge Philosophycal Society, 24, 180–190.

Gamerman, D. and Lopes, H. F. (2006). Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference (2nd edition). Baton Rouge: Chapman and Hall/CRC.

Nascimento, F. F., Gamerman, D., and Lopes, H. F. (2012). "A Semiparametric Bayesian approach to extreme value estimation." Statistics and Computing, 22, 661–675.

Pickands, J. (1975). "Statistical inference using extreme order statistics." Annals of Statistics, 3, 119–131.

Stepehnson, A. and Tawn, J. (2004). "Bayesian Inference for Extremes: Accounting for the Three Extremal Types." Extremes, 7, 291–307.