On Reparameterisations of the Poisson Process Model for Extremes in a Bayesian Framework

Théo Moins, Julyan Arbel, Anne Dutfoy, Stéphane Girard

On reparameterisations of the Poisson process model for extremes in a Bayesian framework

Théo Moins[1] & Julyan Arbel[1] & Anne Dutfoy[2] & Stéphane Girard[1]

[1] *Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France*
{theo.moins, julyan.arbel, stephane.girard}@inria.fr
[2] *EDF R&D dept. Périclès, 91120 Palaiseau, France*
anne.dutfoy@edf.fr

Résumé. Combiner l'analyse des valeurs extrêmes avec des méthodes bayésiennes a plusieurs avantages, comme la prise en compte d'information *a priori* ou encore la possibilité d'étudier des cas irréguliers en statistique fréquentiste. Nous nous attardons ici sur un modèle d'extrêmes par processus de Poisson, et proposons une approche alternative à une étude récente sur une reparamétrisation du modèle qui orthogonalise les paramètres pour améliorer l'échantillonnage *a posteriori* par méthode de Monte-Carlo par chaînes de Markov (MCMC).

Mots-clés. Théorie des valeurs extrêmes, Processus de Poisson, Inférence bayésienne.

Abstract. Combining extreme value analysis with Bayesian methods has several advantages, such as the consideration of prior information or the ability to study irregular cases for frequentist statistics. We focus here on a model of extremes by Poisson process, and propose an alternative of a recent study on a parameterisation of the model which orthogonalizes the parameters to improve posterior sampling by Markov chain Monte-Carlo method (MCMC).

Keywords. Extreme-Value theory, Poisson processes, Bayesian inference.

1 Introduction

Bayesian inference provides tools to estimate parameters of extreme value models, quantify their uncertainty (Arbel et al., 2019), and exploit prior information if available. In particular, Markov chain Monte Carlo (MCMC) methods have various advantages for extreme value parameter inference, which are summed up in Coles and Tawn (1996). However, the interdependence of the parameters can compromise their estimation, and in particular the convergence of the Markov chain. Among the extreme models, the Poisson process allows generalising the two most frequent models, namely block maxima and peak-over-threshold. After a presentation of the extreme value model based on a Poisson process, we present two reparametrisations to facilitate Bayesian inference. Both approaches are then compared from a theoretical and experimental point of view.

2 Poisson process characterisation of extremes

Let $(X_1, \ldots, X_n)$ be i.i.d. random variables with distribution function F, and $M_n = \max\{X_1, \ldots, X_n\}$, whose distribution is F^n. The domain of attraction of F is defined to be the set of distributions G such that there exist two sequences $a_n > 0$ and b_n such that

$$F^n(a_n x + b_n) \to G(x) \text{ as } n \to \infty.$$

The extreme-value theorem states that if F belongs to the maximum domain of attraction of a distribution G, then G is a generalised extreme value (GEV) distribution:

$$G(x) = G(x \mid \boldsymbol{\theta}) = \begin{cases} \exp\left(-\left\{1+\xi\left(\frac{x-\mu}{\sigma}\right)\right\}_+^{-\frac{1}{\xi}}\right) & \text{if } \xi \neq 0\,, \\ \exp(-\exp(-\frac{x-\mu}{\sigma})) & \text{if } \xi = 0, \end{cases}$$

where $\{x\}_+ = \max\{0, x\}$ and $\boldsymbol{\theta} = (\mu, \sigma, \xi) \in \mathbb{R} \times \mathbb{R}^+ \times \mathbb{R}$. Asymptotically, a_n and b_n can be absorbed in the parameters in such a way that it is sufficient to estimate $\boldsymbol{\theta}$.

An alternative is to consider values that exceed a high threshold u. A second extreme-value theorem, known as the Pickands theorem, states that if F belongs to the maximum domain of attraction of $G(\cdot \mid \mu, \sigma, \xi)$, then the distribution function of the exceedances $P(X - u \mid X > u)$ can be approximated by a Generalised Pareto Distribution (GPD):

$$P(X < y + u \mid X > u) \xrightarrow[u \to x^*]{} \begin{cases} 1 - \left\{1 + \xi \frac{y}{\tilde{\sigma}_u}\right\}_+^{-\frac{1}{\xi}} & \text{if } \xi \neq 0\,, \\ 1 - \exp\left(-\frac{y}{\tilde{\sigma}_u}\right) & \text{if } \xi = 0, \end{cases}$$

where x^* is the upper endpoint of X. The shape parameter ξ is the same as in the GEV model, and the relation $\tilde{\sigma}_u = \sigma + \xi(u - \mu)$ links the two scale parameters.

Summarised by Coles (2001), a third way to characterise extreme observations comes from the theory of point processes, and unifies the two previous models. From the extreme value theorem, one can show that the point process N_n related to the point sequence $\{X_1, \ldots, X_n\}$ is such that for intervals $I_u = [u, +\infty)$ with sufficiently large u, we have

$$N_n(I_u) \sim \mathcal{B}(n, p)\ , \text{ with } p \approx \begin{cases} \frac{1}{n}\left(1 + \xi\left(\frac{u-\mu}{\sigma}\right)\right)^{-\frac{1}{\xi}} & \text{if } \xi \neq 0\,, \\ \frac{1}{n}\exp(-\frac{u-\mu}{\sigma}) & \text{if } \xi = 0. \end{cases}$$

As $n \to +\infty$, the binomial distribution converges to a Poisson distribution $\mathcal{P}(\Lambda(I_u))$, with

$$\Lambda(I_u) = \begin{cases} \left(1 + \xi\left(\frac{u-\mu}{\sigma}\right)\right)^{-\frac{1}{\xi}} & \text{if } \xi \neq 0\,, \\ \exp(-\frac{u-\mu}{\sigma}) & \text{if } \xi = 0. \end{cases} \qquad (1)$$

This result combined with the independence property for distributions of N_n on non-overlapping sets is sufficient to say that N_n converges to a non-homogeneous Poisson process, with intensity measure for a fixed u given by (1):

$$N_n \xrightarrow{d} N, \quad \text{with } N(I_u) \sim \mathcal{P}(\Lambda(I_u)).$$

Clearly the GEV model is a special case of this limiting Poisson process, since:

$$P(M_n < z) = P(N_n(I_z) = 0) \xrightarrow[n \to +\infty]{} P(N(I_z) = 0) = \exp(-\Lambda(I_z)) = G(z \mid \boldsymbol{\theta}).$$

However, the parameters (μ, σ, ξ) are here related to the overall maximum of the dataset, and it is frequent to study maxima of m smaller blocks (such as annual maxima). To this end, the intensity measure $\Lambda(I_u)$ is multiplied by a scaling factor m equal to the number of blocks. In the same way, the threshold excess model is also a special case as it can be derived from the Poisson process model. One difference between estimation with GPD and Poisson process is the threshold dependence of $\tilde{\sigma}_u$ on the first case, contrary to σ which is here the same as in the GEV model, and does not depend on u.

Then, we are able to define the likelihood for n observations $(x_1, \ldots, x_n)$ above u:

$$L(\boldsymbol{\theta} \mid \mathbf{x}) = \exp\left(-m\Lambda(I_u \mid \boldsymbol{\theta})\right) \prod_{i=1}^{n} \lambda(x_i \mid \boldsymbol{\theta}),$$

$$\text{with} \quad \lambda(x \mid \boldsymbol{\theta}) = \begin{cases} \sigma^{-1}\left(1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right)^{-\frac{1+\xi}{\xi}} & \text{if } \xi \neq 0\,, \\ \sigma^{-1}\exp(-\frac{x-\mu}{\sigma}) & \text{if } \xi = 0. \end{cases}$$

The choice of the scaling factor m will have an impact on μ and σ, but ξ stays invariant with respect to m. More precisely, Wadsworth et al. (2010) show that if (μ_1, σ_1, ξ) are parameters associated with m_1 block maxima and (μ_2, σ_2, ξ) are parameters associated with m_2 block maxima, then

$$\mu_2 = \mu_1 - \frac{\sigma_1}{\xi}\left(1 - \left(\frac{m_2}{m_1}\right)^{-\xi}\right), \quad \text{and} \quad \sigma_2 = \sigma_1\left(\frac{m_2}{m_1}\right)^{-\xi}. \tag{2}$$

3 Two reparameterisations for Bayesian inference

Parameter orthogonality is defined as a property of a set of parameters $\boldsymbol{\theta}$ which leads to a diagonal Fisher information matrix $I(\boldsymbol{\theta})$. Sharkey and Tawn (2017) show empirically that orthogonality improves the convergence of the MCMC to the joint posterior distribution of the parameters. We successively describe their method to achieve near-orthogonality, and suggest another one proposed by Chavez-Demoulin and Davison (2005) for another purpose, that seems to be more effective for this objective.

3.1 Near-orthogonality by tuning m (Sharkey and Tawn, 2017)

In view of (2), it is easy to transform the parameters μ and σ by changing the value of the scaling factor. Sharkey and Tawn (2017) use this degree of freedom to find an m that minimises the asymptotic covariance between parameters, in order to change it before using the Metropolis–Hastings algorithm, and at the end, restore the parameters corresponding to the initial number of blocks. As the asymptotic covariance matrix can be found by inverting the Fisher information matrix, the authors describe their problem as finding a value of m that near-orthogonalizes the parameters (μ, σ, ξ).

The details of the derivation will not be given here, but one should note that they are only valid for $\xi > -1/2$. In the end, asymptotic covariances can be written as functions of $\left(x = -\frac{1}{\xi}\log\left\{1+\xi\left(\frac{u-\mu}{\sigma}\right)\right\}_+, \sigma, \xi\right)$ as:

$$\begin{aligned}
\mathrm{ACov}(\mu,\sigma) &= \frac{\sigma^2}{m\xi^2}e^x\left(\xi(1+\xi)^2x^2 - (1+3\xi)(1+\xi)x + \xi^3 + (1+\xi)(1+2\xi)\right.\\
&\quad\left. + e^{-\xi x}(1+\xi)(1+2\xi)(x-1)\right),\\
\mathrm{ACov}(\mu,\xi) &= \frac{\sigma}{m\xi^2}e^x(1+\xi)\left(\xi(1+\xi)x - (1+2\xi)(1-e^{-\xi x})\right),\\
\mathrm{ACov}(\sigma,\xi) &= \frac{\sigma}{m}e^x(1+\xi)\left((1+\xi)x - 1\right).
\end{aligned}$$

Denoting by ρ_{θ_1,θ_2} the asymptotic correlation between two of the three parameters, the authors noted that a range of values can also work for m between m_1 and m_2, where

$$m_1 = \underset{m}{\operatorname{argmin}}\{|\rho_{\mu,\sigma}| + |\rho_{\mu,\xi}|\} \text{ and } m_2 = \underset{m}{\operatorname{argmin}}\{|\rho_{\mu,\sigma}| + |\rho_{\sigma,\xi}|\}.$$

They also found on their experiments that m_1 cancels $\mathrm{ACov}(\mu,\sigma)$, and that m_2 cancels $\mathrm{ACov}(\sigma,\xi)$. A numerical method is used in Sharkey and Tawn (2017) to approximate m_1 and m_2 as functions of ξ, so their framework consists of estimating ξ (with maximum likelihood for example), to obtain $\hat{m}_1(\xi)$ and $\hat{m}_2(\xi)$ and choose a value in this interval to run the MCMC.

3.2 Orthogonal parameterisation (our proposal)

More directly, there exists a parameterisation of the Poisson process that leads directly to orthogonality. Suggested by Chavez-Demoulin and Davison (2005), it consists of the following change of variable:

$$(r,\nu,\xi) = \left(m\left(1+\xi\left(\frac{u-\mu}{\sigma}\right)\right)^{-1/\xi}, (1+\xi)(\sigma+\xi(u-\mu)), \xi\right).$$

Parameter r represents the intensity of the Poisson process, which is the expected number of exceedances, and the two others can be seen as an orthogonal parametrisation of the GPD distribution with scale $\tilde{\sigma}_u = \sigma + \xi(u-\mu)$ and shape ξ. The motivations of Chavez-Demoulin and Davison (2005) for this transformation are different[1], but this parameterisation can be used here to ensure full orthogonality, which means that $I(r,\nu,\xi)$ is diagonal.

[1] To avoid computations difficulties when using a generalised linear model for extremes.

4 Comparison of the two approaches

The approach of Sharkey and Tawn (2017) implies to study the x roots of $\mathrm{ACov}(\mu, \sigma)$ and $\mathrm{ACov}(\sigma, \xi)$ to respectively deduce $\hat{m}_1(\xi)$ and $\hat{m}_2(\xi)$. The corresponding value of m can be found by observing that $x = \log(\frac{r}{m})$, with $r = m\left(1 + \xi\left(\frac{u-\mu}{\sigma}\right)\right)^{-1/\xi}$ the expected number of observations. Although r is unknown before estimation, it can be easily estimated by the actual number of observations n, allowing us to deduce the value of m from a given x. The authors provide an approximation method to numerically compute $\hat{m}_1(\xi)$ and $\hat{m}_2(\xi)$. In contrast here, we investigate the existence, uniqueness and position of the roots from a theoretical point of view.

For $\mathrm{ACov}(\sigma, \xi)$, we directly have $x_1 = \frac{1}{1+\xi}$ as the unique root. Moreover, as $\xi > -\frac{1}{2}$, we have $x_1 > 0$, which motivates us to study the sign of the root x_2 for $\mathrm{ACov}(\mu, \sigma)$. Indeed, if x_2 is unique and $x_2 < 0$, then the choice $x = 0$ which cancels the third asymptotic covariance $\mathrm{ACov}(\mu, \xi)$ will always be reasonable as it will stay in the targeted interval, between the two other roots. In addition, $x = 0$ corresponds to the choice $m = r$ (which in practice translates into $m = n$), and is a simple choice as it does not require any estimation of ξ. The interest of the choice $m = n$ has already been mentioned in Wadsworth et al. (2010) to improve the mixing property of the chain. Unfortunately, a study of function for $\mathrm{ACov}(\mu, \sigma)$ shows that the properties of uniqueness and positivity are only valid in the case where $\xi > 0$. In that case, studies of Wadsworth et al. (2010) and Sharkey and Tawn (2017) corroborate the choice of $m = n$. However, it is not the case when $-\frac{1}{2} < \xi < 0$. The study shows that x_2 is not negative here, and worse, may not be unique (Fig. 1).

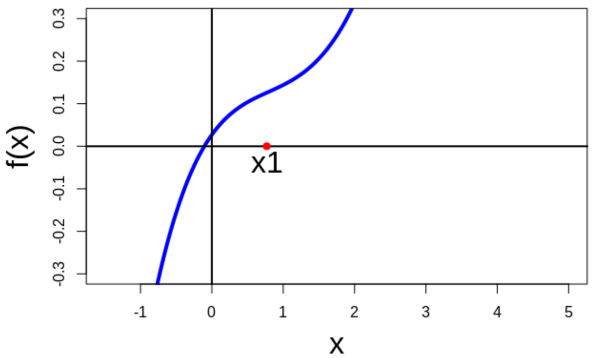
Thus, instead of tuning m, using the orthogonal parameterisation of Chavez-Demoulin and Davison (2005) is more adapted, as has the following three advantages:

1. It exactly orthogonalises the three parameters.
2. It is more accurate in the sense that there is no need to estimate r (as r is one of the parameters).
3. It is valid for all $\xi > -1/2$.

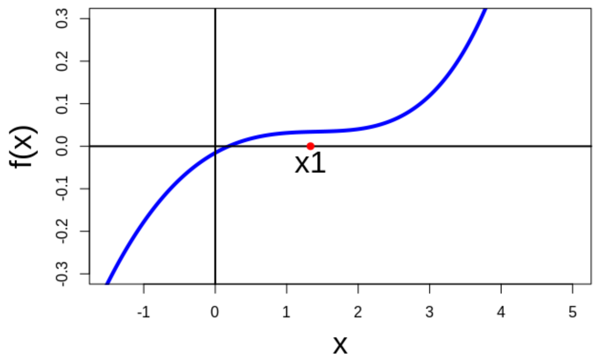
Moreover, by plugging the variables (r, ν) in (2), we can show that the invariance property with respect to m holds for the three parameters, and so the parametrisation is totally independent from the choice of m. In the communication, we shall also present the experimental part, comparing the convergence and mixing properties of the Markov chains corresponding to different parameterisations on various datasets.

References

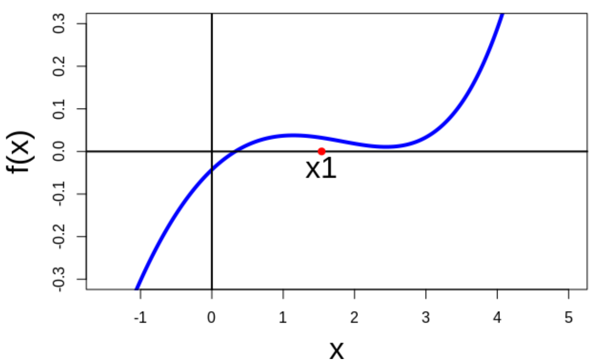
Arbel, J., Crispino, M., and Girard, S. (2019). "Dependence properties and Bayesian inference for asymmetric multivariate copulas." *Journal of Multivariate Analysis*, 174.



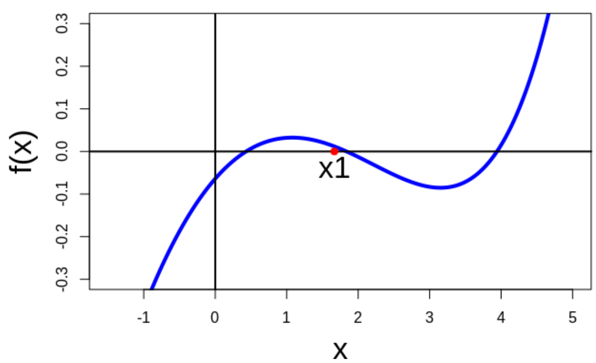
(a) $\xi = 0.3$



(b) $\xi = -0.25$



(c) $\xi = -0.35$



(d) $\xi = -0.4$

Figure 1: Graphs of ACov(μ, σ) as a function of x for different values of ξ. When $\xi > 0$ (a), 0 effectively belongs to the segment $[x_1, x_2]$, but it is not the case anymore when $\xi < 0$, with different scenarios. The covariance function can stay monotonic (b), or may be locally decreasing (c), up to having multiple roots (d).

Chavez-Demoulin, V. and Davison, A. C. (2005). "Generalized additive modelling of sample extremes." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1), 207–222.

Coles, S. (2001). *An introduction to statistical modeling of extreme values*. Springer Series in Statistics. London: Springer-Verlag.

Coles, S. G. and Tawn, J. A. (1996). "A Bayesian analysis of extreme rainfall data." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 45(4), 463–478.

Sharkey, P. and Tawn, J. A. (2017). "A Poisson process reparameterisation for Bayesian inference for extremes." *Extremes*, 20(2), 239–263.

Wadsworth, J., Tawn, J. A., and Jonathan, P. (2010). "Accounting for choice of measurement scale in extreme value modeling." *The Annals of Applied Statistics*, 4(3), 1558–1578.