Projet de recherche : analyse bayésienne robuste de lois d'extrêmes, application en hydrologie belge et en météorologie corse

M2 2024

1 Contexte

La construction de lois *a priori* sur les paramètres des distributions de valeurs extrêmes est une tâche difficile. Les experts n'ont en général aucune intuition sur la signification des paramètres et préfèrent faire des évaluations sur des quantités observables. Leurs jugements sont fondés sur l'expérience passée, où des événements intéressants se sont rarement produits, ce qui entraÃ®ne une incertitude importante dans les évaluations. La transformation de ces expertises en lois *a priori* peut être fortement affectée par l'arbitraire introduit par le statisticien, par exemple dans le choix de leurs formes fonctionnelles.

La robustesse bayésienne (ou analyse de sensibilité bayésienne : [2, 3]) est une approche visant á lutter contre l'impossibilité pratique de spécifier exactement les distributions a priori, les modèles statistiques et les fonctions de perte, c'est-á-dire les trois ingrédients de l'approche bayésienne. En particulier, le choix de la distribution a priori est l'aspect le plus critique de cette approche. Son application aux lois issues de la théorie des valeurs extrêmes, dans des cadres appliquées où celle-ci est usuellement utilisée (par exemple pour modéliser le comportement de variables météorologiques extrêmes, telle la température, la pluviométrie, l'humidité relative, etc.), n'a á ce jour pas encore été menée.

Dans la pratique, l'approche bayésienne robuste se fonde sur la méthodologie suivante : on considère une classe de distribution *a priori*, et une gamme de valeur couverte par la quantité d'intérêt (ici, principalement les quantiles *a posteriori* et des périodes de retour) quand la loi *a priori* (prior) varie dans cette classe. Si l'intervalle est petit, alors tout prior dans la classe peut être choisi puisque le choix n'affecte pas l'estimation de la quantité d'intérêt. Si la fourchette est large, les experts doivent fournir des informations complémentaires afin de réduire la taille de la classe du prior. La procédure doit être répétée jusqu'á ce qu'une petite fourchette soit obtenue ou qu'aucun affinement ne soit possible. Dans ce dernier cas, l'analyse doit être effectuée en

utilisant un seul prior (peut-être optimal au regard de certains critères) mais en indiquant la fourchette de la quantité d'intérêt et en reconnaissant comment l'estimation est affectée par ce choix particulier.

Ce projet vise donc á mettre en place une telle analyse.

2 Applications

On considère deux situations faisant intervenir les lois de maxima d'un phénomène naturel :

— **un jeu de données réelles de maxima journaliers annuels de débits de la Meuse** en une station de mesure située près de la ville de Liège (Belgique), téléchargeable á l'adresse

`https://perso.lpsm.paris/~bousquet/projets/max-meuse.txt`

Les mesures sont données en m^3/s. On s'intéresse ici á l'estimation des niveaux de retour á 4 ans, puis aux niveaux correspondant á des probabilités de dépassement d'au plus 0.1 et 0.001. Une information *a priori* est donnée dans la table 1. Celle-ci est issue d'une expertise produite á partir de modèles de simulation qui tentent de prendre en compte la variation prédictive du débit dans des conditions de changement climatique au cours du 21ème siècle.

Percentile order	Discharge (m^3/s)
5%	1250 ($\pm$200)
50%	2000 ($\pm$100)
75%	2100 ($\pm$100)

TABLE 1 – Prior predictive information on daily maxima discharge per year, extrapolated by numerical analysis of physically-based climate models.

— **un jeu de données réelles de maxima journaliers annuels de la pluviométrie** á Penta-di-Casinca (Haute Corse), téléchargeable á l'adresse

`https://perso.lpsm.paris/~bousquet/projets/pluviometry-corsica.csv`

Les mesures sont données en mm. On s'intéresse ici á l'estimation des niveaux de retour á 50 puis 100 ans. Une information historique *a priori*, issue de l'in-

terrogation d'un expert de Météo-France, est donnée dans la table 2.

Percentile order	Pluviometry P (mm)
25%	75 ($\pm$20)
50%	100 ($\pm$20)
75%	150 ($\pm$20)

TABLE 2 – Prior predictive information on daily maxima pluviometry per year, extrapolated by an expert from daily maxima measured at a nearby station.

3 Formalisation

3.1 Principe général

On considère pour une grandeur X d'intérêt la loi des valeurs extrêmes généralisées (GEV) de fonction de répartition

$$F(x;\theta) = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-1/\xi}\right\},$$

et de densité

$$f(x;\theta) = \frac{1}{\sigma}\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-1/\xi-1}\exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-1/\xi}\right\},$$

avec $\mu \in \mathbb{R}$, $\xi \in \mathbb{R}$ et $\sigma \in \mathbb{R}^+$. On note $\theta = (\mu, \sigma, \xi) \in \Theta$. Au regard des cas d'étude, on considère donc disposer pour chaque cas, outre un échantillon $x_1, \ldots, x_n$, de spécifications *a priori* de la forme $F(x_{q_i}) = q_i, i = 1, \ldots, n$. En introduisant une distribution *a priori* de densité $\pi(\theta)$, on a alors :

$$F(x_{q_i}) = \int_\Theta F(x_{q_i};\theta)\pi(\theta)d\theta = q_i, i = 1, \ldots, n.$$

Nous souhaitons évaluer comment les données expérimentales et les hypothèses sur la distribution priore pourraient affecter les quantiles et donc les niveaux de retour. De façon générale, les données observées $\mathbf{x} = (X_1, \ldots, X_m)$ conduisent á la vraisemblance $l_{\mathbf{x}}(\theta) = \prod_{j=1}^{m} f(X_j;\theta)$. L'intérêt réside dans l'obtention a posteriori d'une certaine quantité, disons $g(\theta)$, qui peut être estimée par

$$G(\pi) \quad = \quad \frac{\int_\Theta g(\theta) l_{\mathbf{x}}(\theta)\pi(d\theta)}{\int_\Theta l_{\mathbf{x}}(\theta)\pi(d\theta)}. \qquad (1)$$

3.2 Approche robuste via une classe de moments contraints généralisés

Nous considérons une variable aléatoire X avec une distribution GEV. Nous supposons que l'expert est capable de spécifier uniquement des quantiles sur la quantité observable X (sans condition sur le paramètre θ), c'est-á-dire qu'il fournit uniquement des déclarations comme $F(x_{q_i}) = q_i, i = 1, \ldots, n$. En introduisant une distribution préalable $\pi(\theta)$, les instructions deviennent alors

$$F(x_{q_i}) = \int_\Theta F(x_{q_i}; \theta)\pi(\theta)d\theta = q_i, i = 1, \ldots, n.$$

Nous souhaitons évaluer comment les données expérimentales et les hypothèses sur la distribution a priori pourraient affecter ces quantiles (et même d'autres).

Une distribution a priori correspondant á ces quantiles pourrait être trouvée numériquement, au moins dans une approximation raisonnable, mais un tel choix serait sans doute arbitraire, fondé sur la commodité du statisticien plutÃ´t que sur une évaluation efficace par l'expert. C'est pourquoi une approche bayésienne robuste est adoptée, en considérant toutes les distributions a priori compatibles avec les quantiles évalués et en étudiant ensuite l'influence d'un tel choix sur les quantités d'intérêt, á savoir les quantiles et les rendements.

La classe des priors admissibles est un cas particulier de la classe généralisée des moments contraints présentée dans [1], et donnée par

$$\Gamma = \{\pi : \int_\Theta H_i(\theta)\pi(\theta)d\theta \leq \alpha_i, \ ; i = 1, ..., n\}$$

où H_i sont des fonctions intégrées π et $\alpha_i, i = 1, \ldots, n$, sont des nombres réels fixes. Si nous prenons $H_i(\theta) = F(x_{q_i})$, $\alpha_i = q_i$ et, par souci de simplicité, l'égalité au lieu des bornes d'inégalité, alors la classe Γ est celle de tous les priors menant á ces quantiles, en supposant un modèle GEV.

L'approche bayésienne robuste s'intéresse á la mesure de l'effet d'une classe de priors sur la quantité d'intérêt (1). La mesure la plus courante est fournie par la gamme

$$\sup_{\pi\in\Gamma} G(\pi) - \inf_{\pi\in\Gamma} G(\pi).$$

La robustesse est atteinte lorsque cette gamme est *petite* (selon un jugement subjectif d'un décideur).

Nous nous concentrerons ici uniquement sur la manière de calculer $\sup_{\pi\in\Gamma} G(\pi)$, soit l'équivalent de $\inf_{\pi\in\Gamma} G(\pi)$.

Le théorème 3 de [1] montre que

$$\sup_{\pi\in\Gamma} G(\pi) = \sup_{(\theta,\mathbf{p})\in T} \frac{\sum_{j=1}^{n+1} g(\theta_j) l_{\mathbf{x}}(\theta_j) p_j}{\sum_{j=1}^{n+1} l_{\mathbf{x}}(\theta_j) p_j},$$

où $\theta = (\theta_1, \ldots, \theta_{n+1})'$, $\mathbf{p} = (p_1, \ldots, p_{n+1})'$ et l'ensemble $T \subset \Theta^{n+1} \times [0,1]^{n+1}$ est défini par les conditions suivantes :

- $\sum_{j=1}^{n+1} F(x_{q_i}; \theta_j) p_j = q_i, \ ; i = 1, \ldots, n$
- $\sum_{j=1}^{n+1} p_j = 1.$

Par conséquent, $\sup_{\pi \in \Gamma} G(\pi)$ est recherché dans le sous-ensemble des distributions extrêmes $\sum_{j=1}^{n+1} p_j \delta_{\theta_j}$, avec $\delta.$ la mesure de Dirac, satisfaisant les conditions ci-dessus.

Les quantités d'intérêt sont des quantiles á des probabilités données (et des temps de retour conséquents), mais elles ne peuvent être obtenues á partir d'une fonction $G(\pi)$ pour un choix adéquat de la fonction $g(\theta)$. Par conséquent, nous calculerons les limites supérieures et inférieures de $F(x|\mathbf{x})$, $x > 0$, et nous obtiendrons les limites des quantiles par une fonction inverse. Le processus est exigeant en termes de calculs (et conduit á des solutions approximatives), car de nombreux problèmes d'optimisation doivent être résolus pour obtenir les limites supérieures et inférieures de $F(x|\mathbf{x})$ sur une grille suffisamment fine, puis une fonction inverse doit être calculée numériquement pour obtenir les limites des quantiles.

Par conséquent, nous calculons d'abord $\sup$ et $\inf$ de

$$\frac{\sum_{j=1}^{n+1} F(x; \theta_j) l_{\mathbf{x}}(\theta_j) p_j}{\sum_{j=1}^{n+1} l_{\mathbf{x}}(\theta_j) p_j}$$

sur une grille de valeurs x et tracer les deux courbes $\inf_{\pi \in \Gamma} F(x|\mathbf{x})$ et $\sup_{\pi \in \Gamma} F(x|\mathbf{x})$. Supposons que l'intérêt se situe dans la plage a posteriori du quantile x_α d'ordre α. Nous traÃ§ons une ligne horizontale en correspondance de α et nous l'intersectons avec les courbes ci-dessus : les points d'intersection donnent les limites inférieure et supérieure de x_α.

4 Quelques indications pour guider ce travail

Ce travail de recherche vise á produire des estimateurs des quantités d'intérêt proposées plus haut, pour les deux cas d'étude, en mettant en oeuvre la démarche robuste proposée ci-dessus. Il sera intéressant de comparer les résultats avec une démarche classique où une loi *a priori* relativement arbitraire est choisie.

Il est donc conseillé de commencer par formaliser le problème, notamment en rappelant les principau résultats et en détaillant les formules plus haut, puis de choisir un cas simple de loi *a priori* auquel comparer les résultats. Il semble aussi important de se munir d'une procédure de simulation de données, afin de reproduire l'expérimentation et de tester la "robustesse" générale de l'approche (elle-même dite robuste).

On attend de ce travail, outre une formalisation et une mise en oeuvre, une rédaction de code Python ou R bien documentée.

Références

[1] B. Betrò, M. Męczarski, and F. Ruggeri. Robust bayesian analysis under generalized moments conditions. *Journal of Statistical Planning and Inference*, 41 :257–266, 1994.

[2] D. Ríos Insua, F. Ruggeri, and B. Vidakovic. Some results on posterior regret γ–minimax estimation. *Statistics and Decision*, 13 :315–331, 1995.

[3] D. Ríos Insua and F. (eds) Ruggeri. *Robust Bayesian Analysis*. Lecture Notes in Statistics, Springer :New York, 2000.