Bayesian computational methods for estimating extreme quantiles from environmental data

Théo Moins

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : MSTII - Mathématiques, Sciences et technologies de l'information, Informatique
Spécialité : Mathématiques Appliquées
Unité de recherche : Laboratoire Jean Kuntzmann

Méthodes bayésiennes computationnelles pour l'estimation de quantiles extrêmes à partir de données environnementales

Bayesian computational methods for estimating extreme quantiles from environmental data

Présentée par :

Théo MOINS

Direction de thèse :

Stéphane GIRARD DIRECTEUR DE RECHERCHE, INRIA CENTRE GRENOBLE-RHONE-ALPES	Directeur de thèse
Julyan ARBEL CHARGE DE RECHERCHE HDR, INRIA CENTRE GRENOBLE-RHONE-ALPES	Co-directeur de thèse
Anne DUTFOY INGENIEUR DE RECHERCHE, EDF R&D	Co-encadrant de thèse

Rapporteurs :

CLEMENT DOMBRY
PROFESSEURE DES UNIVERSITES, UNIVERSITE DE FRANCHE-COMTE
ROBIN RYDER
MAITRE DE CONFERENCES HDR, UNIVERSITE PARIS 9 - DAUPHINE

Thèse soutenue publiquement le **19 septembre 2023**, devant le jury composé de :

STEPHANE GIRARD DIRECTEUR DE RECHERCHE, INRIA CENTRE GRENOBLE-RHONE-ALPES	Directeur de thèse
CLEMENT DOMBRY PROFESSEURE DES UNIVERSITES, UNIVERSITE DE FRANCHE-COMTE	Rapporteur
ROBIN RYDER MAITRE DE CONFERENCES HDR, UNIVERSITE PARIS 9 - DAUPHINE	Rapporteur
ANTONIO CANALE ASSOCIATE PROFESSOR, UNIVERSITA DEGLI STUDI DI PADOVA	Examinateur
ANNE-CATHERINE FAVRE PROFESSEURE DES UNIVERSITES, GRENOBLE INP	Présidente
JULYAN ARBEL CHARGE DE RECHERCHE HDR, INRIA CENTRE GRENOBLE-RHONE-ALPES	Co-directeur de thèse

Invités :

ANNE DUTFOY
INGENIEUR DE RECHERCHE, EDF R&D

Remerciements

Malgré un timing serré (comme toujours) pour rédiger ces remerciements, je vais tâcher d'écrire quelque chose que ChatGPT n'aurait pas été capable de générer[1]. Ces remerciements sont d'autant plus importants car cette période a probablement été la plus épanouissante de ma vie, à la fois sur le plan professionnel et personnel. Sans avoir le sentiment d'avoir produit une thèse hors du commun, je suis extrêmement heureux de l'avoir accomplie sans encombres pendant trois ans, malgré un contexte parfois compliqué (notamment avec les confinements).

Je voudrais commencer par souligner à quel point ce travail n'est *pas* un accomplissement personnel ! Je reconnais avoir eu du flair pour saisir et exploiter une opportunité qui allait me convenir. Pour le reste, c'est l'environnement dans lequel j'ai été, et particulièrement ces trois dernières années, qui m'ont permis d'accomplir ces travaux. La chance que j'ai eue dès le début a été de faire un stage de deux mois dans une certaine équipe Mistis à l'INRIA, un peu par hasard. Cela m'a permis de découvrir ce que c'était de travailler dans cet environnement, ce qui m'a ensuite incité à revenir en thèse sans aucune hésitation, et prenant la décision du doctorat en toute connaissance de cause.

Je tiens à remercier tout d'abord (et très chaleureusement) ceux qui étaient au front avec moi, mes directeurs de thèse, Julyan et Stéphane. Je ne pouvais pas espérer mieux, car les deux possèdent de nombreuses qualités qui, en prime, se complètent parfaitement, ce qui décuple la qualité de l'encadrement. J'ai eu la chance énorme d'avoir été très bien suivi et d'avoir eu deux superviseurs très disponibles, sans pour autant avoir été mis sous pression au travail. Cet équilibre subtil et précieux a grandement contribué au bon déroulement de cette thèse. Je ne vais pas dresser une liste exhaustive de leurs qualités (en vrac : la pédagogie, la patience, l'empathie, l'humour !), mais je reste très reconnaissant envers eux.

Je souhaite également remercier l'ensemble de l'équipe Statify que j'ai pu côtoyer pendant ces trois ans, car c'est grâce à eux que cet environnement de travail est aussi familial et particulièrement agréable au quotidien. Un remerciement tout particulier à mes grandes soeurs du laboratoire, Masha et Lulu. J'ai l'impression d'avoir traversé toute la thèse avec vous, et vous avez toujours été là pour moi, ce dont je suis vraiment reconnaissant. Dédicace aussi à Geoffroy et Jacopo, on ne fera peut être jamais de start-up deep learning pour optimiser les tickets de caisses à la cantine mais finalement c'est peut être pas plus mal.

Je souhaite aussi remercier EDF qui a financé ma thèse, en particulier Anne Dutfoy qui m'a accompagné tout au long de ces trois ans et a complété mon encadrement de manière extrêmement pertinente. Elle m'a également permis de présenter mes travaux à plusieurs reprises au sein d'EDF R&D, et m'a toujours apporté un soutien précieux dans

[1] Bien que ce dernier corrigera quand même mon français, ne crachons pas dans la soupe.

mes démarches avec EDF. Je tiens à remercier tous les membres de la R&D que j'ai pu rencontrer, en particulier Merlin Keller et Nicolas Bousquet pour leur intérêt pour mes travaux. Je remercie également les rapporteurs de cette thèse, Robin Ryder et Clément Dombry, ainsi que les autres membres du jury, Anne-Catherine Favre et Antonio Canale, pour leur participation et leur intérêt pour mon travail.

Enfin, je tiens à exprimer ma gratitude envers tous mes proches, qui m'ont tout apporté et qui continuent de le faire. Je ne veux pas trop m'étaler parce que je n'arrive pas à exprimer à quel point je suis reconnaissant. Comme dirait Cijee, "j'ai pas les mots", et c'est vrai, je n'ai vraiment pas les mots ! Mais tout de même, merci à ma grande famille de coeur : Pitou, Lambert, Pérot, Clément (x3!), Sten, Juju, William, Marie... Je m'arrête ici dans cette liste non exhaustive, j'espère que les personnes se reconnaîtront, et j'espère aussi pouvoir leur dire en personne, notamment pendant les festivités qui accompagnent la thèse. Spécial hommage à Edmond pour ce modèle LaTeX et pour les photos Github, deux contributions plus que significatives. Et 1000 mercis à ma famille de sang, le sang de la veine, celle qui a toujours été là pour moi, depuis toujours. Bien sûr, mes parents qui ont toujours été présents, la fratrie Marie, Lucie et Tom, ma marraine préférée, ma grand-mère (car je suis et serai toujours son chouchou), et les cousines Léna et Élise. Et pour conclure, merci à Youyou, la plus belle rencontre que j'ai pu faire dans ma modeste vie.

Ils ont certainement dû me prendre pour une petite merde
Un accroc de la coke ou un pauvre mec qui n'a plus de nerfs
Cette étrange manière d'écrire doit sans doute déplaire
Mais les vieilles habitudes c'est dur de les perdre
À part ça quoi de neuf ? Rien de béné, toujours tendance à être peiner,
Nullement l'intention d'écrire un truc khéné
Sans aucune prétention je donne tout pour que ça puisse sonner vrai
Ils sont contre moi à mon triste regret
Je serais contre eux car à part mes gars je trouvais personne quand j'avais le ventre creux
Faut bien passer aux choses sérieuses après toutes ces nombreuses périodes creuses
Je sais pas trop quelle voie j'emprunte je n'ai encore rien vu de la vie
Je dois laisser mes empreintes en donnant mon avis
J'attends pas de flatterie sur mon image mets le son à fond et rend fou tout le voisinage
Je suis prêt pour le grand voyage, grimper ou tomber de haut
Spécial hommage à tous les mecs qui parlent mal dans mon dos
Hommage aux groupes qui sont rentré réglo à tous ceux qui ont le monopole pour le moment
C'est pour tous ceux qui m'attendent au tournant ; traitres et médias
Ceux qui me font passer pour mort et tous les trous de balles
C'est pour ceux qui m'ont donné du shit, donné du fric quand j'avais rien
C'est pour tous ceux qui apporté amour, force et soutient
Tous tous tous tous les gens de mon coin libre ou en détention
Une pensée pour eux toujours avec bonne intention
Hey les gars tenez le coup, on se verra bientôt
Que Dieu vous aide à garder la forme et le moral chaque jour
C'est pour ma famille surtout les 2 êtres qui m'ont donné naissance
Ma vie est là leurs c'est pour FF mon groupe et le reuf ma click
C'est le moment, niquons tout, y'a pas de bleuf sur ma zik

Le Rat Luciano - De un

Résumé

Cette thèse se situe à l'intersection de deux domaines de recherche : la statistique des valeurs extrêmes et la statistique bayésienne. L'objectif principal est d'utiliser des méthodes bayésiennes pour l'estimation de quantiles extrêmes de données environnementales. L'utilisation du point de vue bayésien est motivée par différentes problématiques liées à l'estimation des quantiles extrêmes. Tout d'abord, cela permet de directement prendre en compte différentes sources d'incertitudes dans un estimateur ponctuel, par exemple en utilisant des lois dites prédictives. Ensuite, cela permet d'accéder à des intervalles de crédibilité pour quantifier la marge d'erreur autour de l'estimation. Enfin, un dernier objectif est de fournir des éléments de réponse quant à la quantification des limites de crédibilité d'extrapolation, c'est-à-dire de déterminer jusqu'où il est raisonnable d'extrapoler la queue de distribution pour l'estimation de quantiles par exemple.

La première contribution de cette thèse porte sur l'amélioration de méthodes bayésiennes computationnelles par la reparamétrisation de modèles d'extrêmes. En particulier, l'étude met en évidence deux avantages à l'utilisation d'une paramétrisation dite orthogonale. D'abord, elle améliore significativement la convergence d'algorithmes MCMC. Ensuite, elle facilite le calcul de la loi *a priori* de Jeffreys pour le modèle d'extrêmes caractérisé par un processus de Poisson, et permet de démontrer la propreté de la loi *a posteriori* associée. Cette analyse est complétée par l'utilisation d'un *a priori* semi-informatif appelé PC prior, qui est également calculé à partir de la vraisemblance du processus de Poisson.

La deuxième contribution concerne l'amélioration du diagnostic de Gelman–Rubin noté $\hat{R}$ pour la convergence des algorithmes MCMC. Une nouvelle version est proposée, basée sur une version localisée qui permet d'identifier un problème de convergence sur un quantile donné de la loi cible. Sa construction repose sur une étude théorique qui permet, entre autre, d'associer un seuil à partir duquel on estime que les chaînes MCMC n'ont pas convergé à un niveau de confiance fixé. Le cas multivarié est également traité, et des simulations sur des modèles bayésiens viennent compléter la proposition.

La troisième contribution de la thèse consiste en des résultats préliminaires sur le comportement de différents estimateurs bayésiens à taille d'échantillon fini. L'objectif est de comprendre comment les estimateurs se comportent dans la queue, en prenant en compte l'incertitude associée à l'estimation des paramètres. Les résultats portent sur le domaine d'attraction des lois prédictives (*a priori* et *a posteriori*), ainsi que sur un équivalent asymptotique de deux méthodes pour estimer un niveau de retour extrême, dans le cas d'une loi *a priori* uniforme sur le paramètre de forme.

Enfin, la dernière contribution de cette thèse est l'application du modèle et de tout les résultats précédents à des séries de données environnementales. Cela permet une estimation de niveaux de retour centennaux, millénaux et décamillénaux de débits de rivières et de vitesses de vents, ainsi que d'apporter des éléments de réponse sur les limites d'extrapolation dans la queue de distribution.

Abstract

This thesis lies at the intersection of two research domains: extreme value statistics and Bayesian statistics. The main objective here is to use Bayesian methods for the estimation of extreme quantiles, and in particular the return levels of environmental datasets. The adoption of a Bayesian paradigm is motivated by various challenges associated with the estimation of extreme quantiles. Firstly, it allows for the direct consideration of different sources of uncertainty in a point estimator, for example by using the so-called predictive distributions. Secondly, it enables access to credible intervals to quantify the estimation error. Lastly, one aim is to provide insights into quantifying the limits of extrapolation, in other words, determining how far it is reasonable to extrapolate the tail of the distribution with a reasonable error for quantile estimation.

The first contribution of this thesis focuses on enhancing computational Bayesian methods through the reparameterization of extreme value models. In particular, the study highlights two advantages of employing an orthogonal parametrization. This first leads to a significant improvement in the convergence of MCMC algorithms. Second, it facilitates the derivation of the Jeffreys prior for the Poisson process characterization of extremes, thereby demonstrating posterior propriety. This investigation of the prior is further complemented by the use of a semi-informative prior called the PC prior, which is also calculated for this Poisson process likelihood.

The second contribution concerns the improvement of a convergence diagnostic for MCMC algorithms known as Gelman–Rubin diagnostic and denoted by $\hat{R}$. A new version denoted $\hat{R}(x)$ is proposed, based on a localized approach that diagnoses convergence issues on a specific quantile of the target distribution. Its construction relies on a theoretical study that enables, among other things, the association of a confidence level with a threshold indicating the lack of convergence of the MCMC chains. The multivariate case is addressed, and simulations on Bayesian models are conducted and support the proposal.

The third contribution of the thesis consists of preliminary results regarding the tail behavior of different Bayesian estimators for finite sample sizes. The aim is to understand how these estimators behave in the tail, when taking into account the uncertainty associated with parameter estimation. The results cover the domain of attraction of predictive distributions (prior and posterior) and provide an asymptotic equivalence for two estimation methods of extreme return levels, under a uniform prior on the shape parameter.

Lastly, the final contribution of this thesis entails the application of the model and all the previous results to a series of environmental datasets. This allows for the estimation of centennial, millennial, and decamillennial return levels for different datasets of river flows and wind speeds, while also providing insights into the extrapolation limits in the tail of the distribution.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Contents

Résumé

Ce chapitre présente les fondements théoriques et les concepts qui vont ensuite être employés dans le reste du manuscrit. Nous commençons par introduire la théorie des valeurs extrêmes univariées en Section 1.1 et discutons de trois méthodes statistiques couramment utilisées pour estimer les événements extrêmes : l'estimation par dépassement de seuil, par maxima de blocs et la caractérisation par un processus de Poisson non homogène, qui généralise les deux premiers modèles. Ensuite, nous nous intéressons au paradigme bayésien en Section 1.2, en mettant l'accent sur les aspects computationnels qui seront explorés dans les chapitres suivants. Après l'introduction à ces deux branches de la statistique, nous passons en revue la littérature existante à leur intersection en Section 1.3 et présentons différents domaines de recherche concernant les modèles bayésiens de valeurs extrêmes. Enfin, nous concluons dans la Section 1.4 en résumant les motivations d'Électricité de France (EDF) qui co-finance la thèse, et en donnant un aperçu des contributions apportées.

Abstract

This chapter presents the theoretical foundations and concepts that form the basis of this thesis. We begin by introducing univariate extreme value theory in Section 1.1 and discuss three statistical models commonly used for estimating extreme events: peaks-over-threshold, block maxima, and point process characterization, which integrates the first two methods. Next, we delve into the Bayesian paradigm in Section 1.2, focusing on the computational aspects that will be explored in the subsequent chapters. By introducing these two fields, we review the existing literature at their intersection in Section 1.3, and outline various research areas concerning Bayesian extreme value modelling. Finally, we conclude in Section 1.4 by summarizing the motivations of *Électricité de France* (EDF, co-funding this thesis) and providing an overview of the contributions made here.

1.1 Introduction to univariate extreme value theory

While traditional statistical methods often focus on characterizing the central tendencies and variability of data, extreme value theory (EVT) is specifically concerned with the tail behavior of probability distributions. It is rooted in a theory developed by Fisher and Tippett (1928) and Gnedenko (1943) on the convergence in law of the maximum value of a sequence of independent and identically distributed random variables, and has then been extended with the results of Pickands (1975), Balkema and De Haan (1974) on the convergence in law of excesses above a threshold. This theory provides valuable insights into the probabilities of events that fall outside the range of typical occurrences, allowing us to better understand and quantify the risks associated with extreme phenomena. We refer to Coles (2001), Beirlant et al. (2006) for great introductions to the field, and to Haan and Ferreira (2006), Resnick (2008) for a more theoretical in-depth analysis. In this study, we focus exclusively on the univariate case, considering scalar random variables that are independent and identically distributed (i.i.d.). Usually, one of two following problems is addressed:

1. **Small probability estimation:** this problem involves determining the probability associated with a given extreme quantile, which is typically larger than the largest observed value.
2. **Extreme quantile estimation:** here, the goal is to determine the quantile associated with a low probability, which tends to approach zero as the number of observations increases.

In the following, our primary objective is to address the problem of estimating an extreme quantile which is called return level in environmental studies (see Definition 3). We start by introducing in Section 1.1.1 and Section 1.1.2 the two fundamental theorems on which all extreme models rely on and their subsequent applications. Then, in Section 1.1.3, we introduce a unified perspective that combines both approaches. Finally, in Section 1.1.4, we review the application to environmental data.

1.1.1 Analyzing extremes through maxima modelling

1.1.1.1 Asymptotic behavior of the maxima

Let $(X_1, \dots, X_n)$ be $n \in \mathbb{N}$ i.i.d. random variables with cumulative distribution function (cdf) F and survival function $\bar{F} := 1 - F$. The first result relates to the distribution of the maxima $M_n := \max\{X_1, \dots, X_n\}$. In the i.i.d. case, the cdf of M_n can be denoted as F^n. Similarly to how a sample mean needs to be standardized to converge to a non-degenerate distribution (which is the standard normal distribution according to the central limit theorem), we are interested in the limit behavior of the standardized sample maxima. We introduce the concept of maximum domain of attraction:

Definition 1. *A cdf F is said to belong to a maximum domain of attraction if and only if there exist two sequences $a_n > 0$, b_n and a non degenerate cdf G such that for all $x \in \mathbb{R}$*

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

In the following, this property is denoted $F \in \mathrm{MDA}(G)$. Equivalently, one could write that $F \in \mathrm{MDA}(G)$ if and only if $(M_n - b_n)/a_n \xrightarrow{d} Y$, where $Y \sim G$. Here, the normalizing sequence b_n plays the role of a location parameter, shifting the distribution, while the sequence a_n acts as a scale parameter, controlling the spread of the distribution. A question arises: what is the limiting distribution that replaces the Gaussian distribution in the convergence of the empirical mean? The answer is provided by the extreme value theorem, also known as the Fisher-Tippett-Gnedenko theorem:

Theorem 1 (Fisher–Tippett–Gnedenko)**.** *If $F \in \mathrm{MDA}(G)$, then there exist $\xi \in \mathbb{R}$ and normalizing sequences such that G can be written*

$$G_\xi(x) := \begin{cases} \exp\left(-\{1+\xi x\}_+^{-\frac{1}{\xi}}\right) & \text{if } \xi \neq 0\,, \\ \exp(-\exp(-x)) & \text{if } \xi = 0, \end{cases} \tag{1.1}$$

where $\{x\}_+ = \max\{0, x\}$.

This distribution is known as the Generalized Extreme Value (GEV) distribution. Unlike regular distributions, its support depends on the parameters involved:

$$\mathrm{supp}(G_\xi) = \{x \in \mathbb{R} \quad \text{s.t. } 1 + \xi x > 0\}\,. \tag{1.2}$$

When $\xi \neq 0$, the use of $\{.\}_+$ in Equation (1.1) can be omitted by considering Equation (1.2) as the support of the distribution. Note also that the case $\xi = 0$ is a continuous extension of $\xi \neq 0$.

Theorem 1 introduces a parameter ξ, known as the extreme-value index or tail index, which acts as a shape parameter. In the estimation process, the normalizing sequences a_n and b_n are two other parameters that will be estimated, hereafter referred to as σ and μ respectively. As the name suggests, the GEV distribution encompasses three distinct behaviors determined by the sign of ξ, resulting in three different domains of attraction:

- **If $\xi = 0$**: F is said to belong to the Gumbel maximum domain of attraction (Gumbel, 1958). This domain includes distributions such as normal, exponential, gamma, and lognormal, among others. It exhibits a light tail behavior, meaning that its survival function decreases as an exponential function.
- **If $\xi > 0$**: F is said to belong to the Fréchet maximum domain of attraction (Fréchet, 1927). This domain includes heavy-tailed distributions, where the survival function decreases as a power function. Typically, these distributions have finite moments of order smaller than $1/\xi$ only, indicating that larger values of ξ correspond to heavier tails. Examples of heavy-tailed distributions include Cauchy, Pareto, and Student distributions.
- **If $\xi < 0$**: F is said to belong to the Weibull domain of attraction (Weibull, 1951). By examining Equation (1.2), it is clear that $\xi < 0$ imposes an upper bound condition on the support of G_ξ. Therefore, this domain comprises short-tailed distributions with a finite endpoint, such as uniform or beta distributions.

It is worth noting that the condition of belonging to a maximum domain of attraction in Theorem 1 is satisfied by a vast majority of known cdf F (see p.145 of Embrechts et al., 2013, for additional examples). Figure 1.1 shows examples of survival functions $\bar{G}_\xi = 1 - G_\xi$ and the corresponding probability density functions (pdf) for different values of ξ.

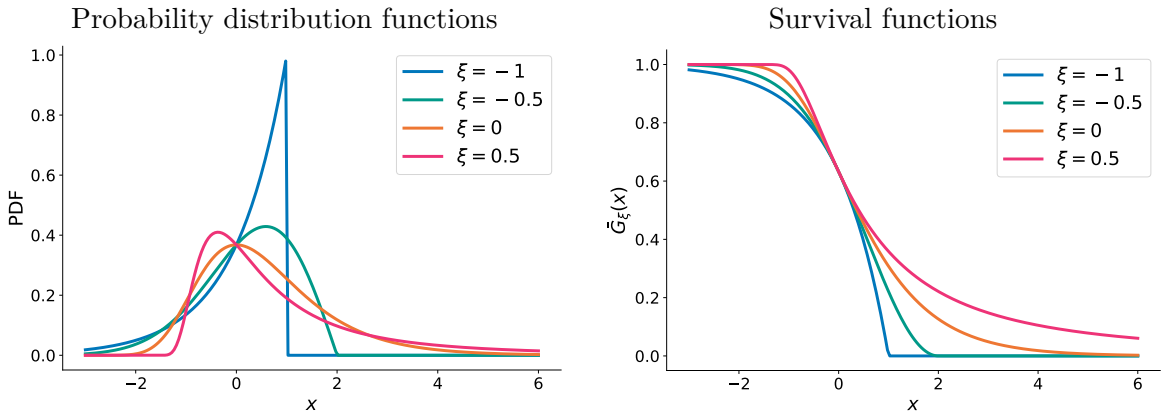


Figure 1.1: Probability density functions (left plot) and survival functions (right plot) of GEV distributions with $\xi \in \{-1, -1/2, 0, 1/2\}$.

1.1.1.2 Block maxima approach

How to use this result in practice? Let us first note that one can derive an approximation in the tail of the survival function and its inverse from Theorem 1: as $F \in \mathrm{MDA}(G_\xi)$, one has

$$\log(1 - \bar{F}(x)) \simeq \frac{1}{n} \log(G_\xi((x-\mu)/\sigma)). \tag{1.3}$$

So for large x, a Taylor expansion yields

$$\bar{F}(x) \simeq -\frac{1}{n} \log(G_\xi((x-\mu)/\sigma) = \begin{cases} \frac{1}{n}\left(1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right)^{-\frac{1}{\xi}} & \text{if } \xi \neq 0\,, \\ \frac{1}{n}\exp\left(-\frac{x-\mu}{\sigma}\right) & \text{if } \xi = 0. \end{cases} \tag{1.4}$$

Similarly, inverting Equation (1.4) gives an approximation of the quantile associated with a small probability p:

$$\bar{F}^{-1}(p) \simeq \mu + \sigma G_\xi^{-1}(\exp(-np)) = \begin{cases} \mu + \sigma\frac{(np)^{-\xi}-1}{\xi} & \text{if } \xi \neq 0\,, \\ \mu - \sigma\log(np) & \text{if } \xi = 0. \end{cases} \tag{1.5}$$

Then, given a finite number of observations n, a common approach to estimating extreme quantiles is to assume that the maximum M_n follows a GEV distribution, and substitute the values of the parameters (μ, σ, ξ) with their estimators $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ in Equation (1.5). However, obtaining reliable parameter estimates requires having a sample of maxima observations, which is not directly available. To overcome this challenge, one approach proposed by Gumbel (1958) involves dividing the dataset into non-overlapping blocks and extracting the maximum value from each of them. These block maxima are then used to estimate the parameters (μ, σ, ξ). Estimators commonly employed include maximum likelihood estimation (Prescott and Walden, 1983) and probability weighted moments (Hosking et al., 1985).

The block maxima approach offers a straightforward and intuitive way to estimate extreme events by focusing on the maximum values within each block. However, it should be noted that this method may discard a significant amount of data. Furthermore, the choice of the block size is challenging and can be seen as a bias-variance tradeoff (see Figure 1.2):

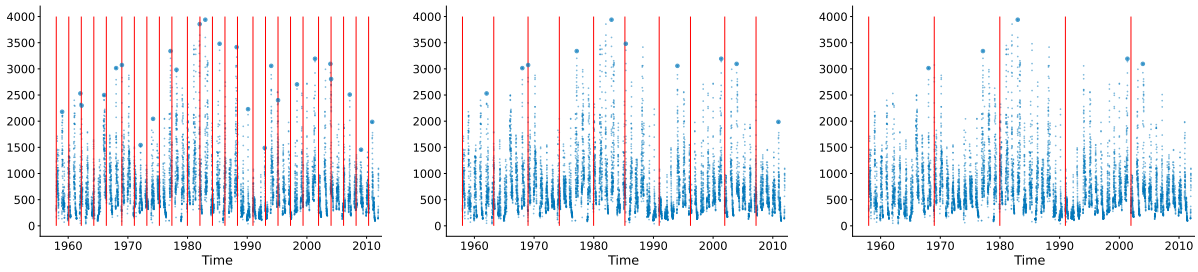


Figure 1.2: Examples of three block size selections for a dataset of river flow at Tours (see Chapter 5).

- A small block size provides more extreme events for estimation, which reduces the variability of parameter estimates but introduces bias. This bias arises because the maxima of a small number of samples may deviate from the GEV distribution which is only asymptotic.
- On the other hand, a larg block size reduces bias but increases variance due to the limited number of samples available for parameter estimation.

Thus, selecting an appropriate block size requires careful consideration, and involves finding a balance that minimizes the overall error.

1.1.2 Analyzing extremes through excesses modelling

1.1.2.1 Asymptotic behavior of excesses over a threshold

An alternative approach to estimating extreme quantiles involves considering the excesses above a threshold instead of the maximum values of the distribution. Let X be a random variable with the same cdf F as $X_1, \ldots, X_n$, and let x_F denote its endpoint (which can be finite or infinite). The distribution function of the excesses of X over a threshold $u < x_F$ can be expressed as follows for $y \geq 0$:

$$\mathbb{P}(X < y + u \mid X > u) = \frac{F(u+y) - F(u)}{1 - F(u)}, \tag{1.6}$$

or equivalently, the survival function is

$$\mathbb{P}(X > y + u \mid X > u) = \frac{\bar{F}(u+y)}{\bar{F}(u)}. \tag{1.7}$$

The second extreme-value theorem, known as the Pickands theorem (Pickands, 1975), provides an approximation when the threshold u is asymptotically high.

Theorem 2 (Pickands, 1975)**.** *$F \in \mathrm{MDA}(G_\xi)$ if and only if there exist $\sigma_u > 0$ and $\xi \in \mathbb{R}$ such that the law of excesses can be uniformly approximated by a Generalized Pareto Distribution (*GPD*):*

$$\sup_{y \in (0, x_F - u)} \left| \frac{\bar{F}(u+y)}{\bar{F}(u)} - \bar{H}_{\sigma_u, \xi}(y) \right| \xrightarrow[u \to x_F]{} 0, \tag{1.8}$$

with

$$\bar{H}_{\sigma_u,\xi}(y) = \begin{cases} \left\{1 + \xi \frac{y}{\sigma_u}\right\}_+^{-\frac{1}{\xi}} & \text{if } \xi \neq 0\,, \\ \exp\left(-\frac{y}{\sigma_u}\right) & \text{if } \xi = 0. \end{cases} \tag{1.9}$$

This theorem establishes that the law of excesses converges uniformly to a GPD with two parameters: σ_u, which may depend on u, and ξ. In particular, $F \in \mathrm{MDA}(G_\xi)$ implies the convergence for all $y \in \mathrm{supp}(G_\xi)$ of

$$\mathbb{P}(X > y + u \mid X > u) \xrightarrow[u \to x_F]{} \bar{H}_{\sigma_u,\xi}(y), \tag{1.10}$$

Equation (1.4) provides an outline of the proof. Assuming $\xi \neq 0$ [1], we obtain:

$$\begin{aligned} \mathbb{P}(X > y + u \mid X > u) &= \frac{\bar{F}(u+y)}{\bar{F}(u)} \\ &\simeq \frac{\left(1 + \xi\left(\frac{y+u-\mu}{\sigma}\right)\right)^{-\frac{1}{\xi}}}{\left(1 + \xi\left(\frac{u-\mu}{\sigma}\right)\right)^{-\frac{1}{\xi}}} \\ &= \left(1 + \xi\frac{y}{\sigma_u}\right)^{-\frac{1}{\xi}} = \bar{H}_{\sigma_u,\xi}(y), \end{aligned}$$

where $\sigma_u = \sigma + \xi(u - \mu)$. This demonstrates the intrinsic connection between the GPD parameters and the GEV ones, for the approximation of the maxima. The shape parameter ξ is shared by both models and plays the same role, while the scale parameter σ_u can be expressed as a function of (μ, σ, ξ) and linearly depends on u. Note that for the GPD, the case where $\xi = 0$ corresponds to an exponential distribution, and $\xi = -1$ corresponds to a uniform distribution on $[0, \sigma_u]$.

1.1.2.2 Peaks-over-threshold approach

Similarly to the result obtained for the GEV distribution, one can derive an approximation of the survival function by considering the change of variable $x = y + u$:

$$\bar{F}(x) \simeq \bar{F}(u)\bar{H}_{\sigma_u,\xi}(x-u) = \begin{cases} \bar{F}(u)\left(1 + \xi\left(\frac{x-u}{\sigma_u}\right)\right)^{-\frac{1}{\xi}} & \text{if } \xi \neq 0\,, \\ \bar{F}(u)\exp\left(-\frac{x-u}{\sigma_u}\right) & \text{if } \xi = 0. \end{cases} \tag{1.11}$$

By inverting the expression, one can approximate the quantile function for a small probability p as

$$\bar{F}^{-1}(p) \simeq u + \sigma_u H^{-1}_{\sigma_u,\xi}\left(1 - \frac{p}{\bar{F}(u)}\right) \simeq \begin{cases} u + \sigma_u \frac{\left(\frac{p}{\bar{F}(u)}\right)^{-\xi} - 1}{\xi} & \text{if } \xi \neq 0\,, \\ u - \sigma_u \log\left(\frac{p}{\bar{F}(u)}\right) & \text{if } \xi = 0. \end{cases} \tag{1.12}$$

Expressions (1.11) and (1.12) are analogous to expressions (1.4) and (1.5) in the GEV case. Maximum likelihood estimators (Davison and Smith, 1990, Zhou, 2010) and probability

[1] All computations can be extended to the case where $\xi = 0$.

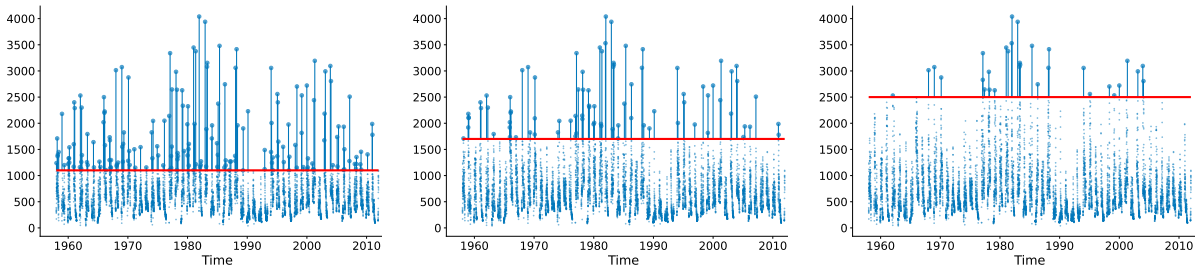


Figure 1.3: Examples of three threshold selections for a dataset of river flow at Tours (see Chapter 5).

weighted moments (Hosking and Wallis, 1987) are commonly used estimation methods in this context. In practice, a value is chosen for $\bar{F}(u)$, and the associated u is a location parameter that needs to be estimated, similarly to μ in the GEV case. However, just as samples of maxima are not directly available for parameter estimation in the GEV case, the excesses are not observed initially in the peaks-over-threshold method. The common practice, known as the peaks-over-threshold method, involves selecting a number n_u of excesses, and so choosing $\bar{F}(u) = \frac{n_u}{n}$. If the order statistics are denoted by $x_{(1)} \leq \ldots \leq x_{(n)}$, the associated threshold u is then estimated as $x_{(n-n_u)}$, and the exceedances are defined as $y_i := x_{(n-n_u+i)} - x_{(n-n_u)}$, for $i = 1, \ldots, n_u$.

The discussion on the choice of u is similar to the one regarding the block size: selecting a small value of n_u leads to a large variance in the estimations due to the limited number of observations, while choosing a large value of n_u introduces a significant bias because the asymptotic tail approximation is no longer accurate (see Figure 1.3). One method of threshold elicitation, described in Coles (2001, Chapter 4), involves choosing the lowest threshold that provides a reasonable asymptotic approximation, based on a graphical method. Typically, if X follows a GPD and $\xi < 1$, then

$$\mathbb{E}[X - u \mid X > u] = \frac{\sigma_u}{1 - \xi},$$

and for all $v > u$,

$$\mathbb{E}[X - v \mid X > v] = \frac{\sigma_v}{1 - \xi} = \frac{\sigma_u + \xi v}{1 - \xi}.$$

Thus, the mean of the excesses is a linear function of u when the observations are GPD. Consequently, the threshold can be selected by plotting a mean residual life plot (*i.e.*, the sample mean excesses) and identifying the smallest threshold that exhibits linearity in u. However, it is important to note that this method, as discussed in Coles (2001, Chapter 4) and other works, does not always lead to a clear conclusion, see Section 5.2 for a further discussion. The question of choosing an appropriate threshold in the peaks-over-threshold method, as well as the selection of the block size in the block maxima approach, remains therefore an open area of research.

1.1.3 A unifying model: the Poisson process characterization

A third way to characterize the extreme value behavior comes from the theory of point processes, and has the main advantage of unifying the GEV and GPD models presented before. We present here an intuitive way for obtaining this model similarly to Coles (2001,

Chapter 7), and refer to Leadbetter et al. (1983, Chapter 5) for more theoretical details on the construction.

1.1.3.1 Statistical perspective of Poisson process

The starting point here shifts from considering the distribution of single observations to focusing on the underlying point process associated with these observations. Let us briefly review the definition of a point process.

Definition 2. *A point process N on a measurable space $\mathcal{A}$ is a counting measure used to describe the random occurrence of events in a given subset $A \subset \mathcal{A}$.*

In other words, for any measurable subset $A \subset \mathcal{A}$, $N(A)$ represents the number of observations within A. The characteristics of the point process are determined by consistently specifying the probability distribution of the random variable $N(A)$ for all measurable subsets A of the underlying space. For the sake of clarity, we consider the case where $\mathcal{A}$ is one-dimensional, but the concept can be generalized to multidimensional spaces. For any $A \subset \mathcal{A}$, we call intensity measure $\Lambda(A)$ the expected number of points in A, so that $\Lambda(A) = \mathbb{E}(N(A))$, and $\lambda(t)$ the associated intensity function, such that:

$$\Lambda(A) = \int_A \lambda(t)\mathrm{d}t.$$

The simplest example of a point process is a homogeneous Poisson process, defined on $\mathcal{A} = \mathbb{R}^+$ by two properties:

- For all intervals $[t_1, t_2]$ with $0 \leq t_1 \leq t_2$, $N([0, t_2]) - N([0, t_1]) \sim \mathcal{P}(\lambda(t_2 - t_1))$ with $\lambda > 0$.
- For two non-overlapping intervals $[t_1, t_2]$ and $[t_3, t_4]$ with $t_1 < t_2 < t_3 < t_4$, $N([0, t_2]) - N([0, t_1])$ and $N([0, t_4]) - N([0, t_3])$ are independent random variables.

Here, the intensity function is constant ($\lambda(t) = \lambda$), and the intensity measure (which is the mean of the Poisson distribution) is proportional to the interval size: $\Lambda([t_1, t_2]) = \lambda(t_2 - t_1)$. To generalize this process while maintaining the independence property for non-overlapping intervals, a varying intensity measure based on the location can be introduced:

$$N([0, t_2]) - N([0, t_1]) \sim \mathcal{P}(\Lambda([t_1, t_2])), \text{ with } \Lambda([t_1, t_2]) = \int_{t_1}^{t_2} \lambda(t)\mathrm{d}t.$$

From a statistical perspective, when estimating the point process, we consider the non-homogeneous Poisson case and assume that the intensity function can be parameterized by $\boldsymbol{\theta}$, so that $\lambda(t) = \lambda(t; \boldsymbol{\theta})$. The aim is to estimate the parameters of the intensity based on a set of observed points $t_1 < \ldots < t_n$. The information contained in the observed points includes the occurrence of events as well as the fact that there are n points in the observation period $[0, t]$. Thus, the likelihood can be expressed as

$$L(n, t_1, \ldots, t_n; \boldsymbol{\theta}) = L(t_1, \ldots, t_n \mid N([0, t]) = n; \boldsymbol{\theta})\mathbb{P}(N([0, t]) = n; \boldsymbol{\theta}).$$

The following result can be found in Chapter 2 of Ross (1996):

Proposition 1. *Let $(T_1, \dots, T_n)$ be the arrival time of N observations. The joint density f of $(T_1, \dots, T_n)$ given that $N([0,t]) = n$ can be written*

$$f(t_1, \dots, t_n \mid N(t) = n) = n! \frac{\prod_{i=1}^{n} \lambda(t_i)}{\Lambda([0,t])^n} \mathbb{I}\{0 < t_1 < \dots < t_n < t\}.$$

Therefore, since $N([0,t]) \sim \mathcal{P}(\Lambda([0,t]))$, we obtain that the likelihood associated with a non-homogeneous Poisson process is

$$L(n, t_1, \dots, t_n; \boldsymbol{\theta}) = \exp\left(-\Lambda([0,t]; \boldsymbol{\theta})\right) \prod_{i=1}^{n} \lambda(t_i; \boldsymbol{\theta}). \tag{1.13}$$

1.1.3.2 Poisson process for extremes

Recall from Equation (1.4) that for each variables $X_1, \dots, X_n$ and $u \in \text{supp}(G_\xi)$, we have

$$\mathbb{P}\left(X_i > u\right) \approx \begin{cases} \frac{1}{n}\left(1 + \xi\left(\frac{u-\mu}{\sigma}\right)\right)^{-\frac{1}{\xi}} & \text{if } \xi \neq 0\,, \\ \frac{1}{n}\exp(-\frac{u-\mu}{\sigma}) & \text{if } \xi = 0. \end{cases} \tag{1.14}$$

This value can be interpreted as the probability of each point belonging to the interval $I_u = [u, +\infty)$. Given the i.i.d observations assumption, we can deduce that the point process N_n associated with the point sequence $\{X_i\ ,\ i = 1, \dots, n\}$ is such that

$$N_n(I_u) \sim \mathcal{B}(n,p)\ \text{, with } p = \begin{cases} \frac{1}{n}\left(1 + \xi\left(\frac{u-\mu}{\sigma}\right)\right)^{-\frac{1}{\xi}} & \text{if } \xi \neq 0\,, \\ \frac{1}{n}\exp(-\frac{u-\mu}{\sigma}) & \text{if } \xi = 0. \end{cases} \tag{1.15}$$

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

$$\Lambda(I_u) = \begin{cases} \left(1 + \xi\left(\frac{u-\mu}{\sigma}\right)\right)^{-\frac{1}{\xi}} & \text{if } \xi \neq 0\,, \\ \exp(-\frac{u-\mu}{\sigma}) & \text{if } \xi = 0. \end{cases} \tag{1.16}$$

The property holds for all I_u, and the independence property for the distributions of N_n on non-overlapping sets is sufficient to conclude that N_n converges to a non-homogeneous Poisson process (NHPP) N with an intensity measure for a given u given by Equation (1.16):

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

By combining this result with Equation (1.13) and denoting $\boldsymbol{\theta} := (\mu, \sigma, \xi)$, we finally obtain

$$\begin{aligned} L(n_u, x_1, \dots, x_{n_u}; \boldsymbol{\theta}) &= \exp\left(-\Lambda(I_u; \boldsymbol{\theta})\right) \prod_{i=1}^{n_u} \lambda(x_i; \boldsymbol{\theta}), \\ &= \exp\left(-\left(1 + \xi\left(\frac{u-\mu}{\sigma}\right)\right)^{-\frac{1}{\xi}}\right) \sigma^{-n_u} \prod_{i=1}^{n_u} \left(1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right)^{-\frac{1+\xi}{\xi}}. \end{aligned}$$

1.1.3.3 Connection with GEV and GPD case

By construction, we directly have for $x > 0$

$$\mathbb{P}(M_n < x) = \mathbb{P}(N_n(I_x) = 0) \xrightarrow[n\to+\infty]{} \mathbb{P}(N(I_x) = 0) = \exp(-\Lambda(I_x)) = G_\xi((x-\mu)/\sigma),$$

so the distribution of M_n effectively converges to a GEV with the Poisson process model.

However, estimating (μ, σ, ξ) with this model is typically related to the overall maximum of the dataset. In many cases, it is more common to study the maxima of smaller blocks, such as annual maxima. To account for this, the intensity measure $\Lambda(I_u; \boldsymbol{\theta})$ needs to be multiplied by m. One interpretation of this rescaling is by introducing a time dimension in the point process. If the sequence of points is now represented as $\left(\frac{i}{n+1}, X_i\right)$, then the intensity measure for a given $A = [t_1, t_2] \times [u, +\infty)$ becomes

$$\Lambda(I_u; \boldsymbol{\theta}) = \begin{cases} (t_2 - t_1)\left(1 + \xi\left(\frac{u-\mu}{\sigma}\right)\right)^{-\frac{1}{\xi}} & \text{if } \xi \neq 0, \\ (t_2 - t_1)\exp(-\frac{u-\mu}{\sigma}) & \text{if } \xi = 0. \end{cases} \tag{1.17}$$

Therefore, within each period equal to one block (e.g., one year for annual maxima), the maximum follows a GEV distribution. The distribution of the global maximum can now be seen as the maximum of m smaller blocks, and its distribution is now denoted by G_ξ^m. It is worth noting that raising a GEV distribution to a power m results in another GEV distribution, but with different parameters (max-stability property). According to Wadsworth et al. (2010), if $(\mu_{n_i}, \sigma_{n_i}, \xi)$ $(i \in \{1, 2\})$ are parameters for n_i GEV observations, then the following relationship holds

$$\mu_{n_2} = \mu_{n_1} - \frac{\sigma_{n_1}}{\xi}\left(1 - \left(\frac{n_2}{n_1}\right)^{-\xi}\right), \quad \sigma_{n_2} = \sigma_{n_1}\left(\frac{n_2}{n_1}\right)^{-\xi}. \tag{1.18}$$

Note that ξ is invariant to the choice of m.

The GPD case can also be derived from the point process representation. Indeed, it can be shown that

$$\mathbb{P}(X_i > y + u \mid X_i > u) = \frac{\Lambda(I_{u+y}; \boldsymbol{\theta})}{\Lambda(I_u; \boldsymbol{\theta})} = \left(1 + \xi\frac{y}{\sigma_u}\right)^{-\frac{1}{\xi}}, \tag{1.19}$$

with the same definition of $\sigma_u = \sigma + \xi(u - \mu)$ as in the GPD case. Thus, the Poisson process point of view allows for estimating the shape parameter associated with the GEV distribution, which is threshold-invariant. The advantages of this approach have been discussed by Smith (1989) and Coles (2001), Chapter 7, in handling non-i.i.d or non-stationary cases.

1.1.4 Application to environmental studies

1.1.4.1 Review of applications

Extreme value theory has broad applications in various domains. For instance, it is used in finance to determine worst-case scenarios for market crashes or extreme fluctuations (e.g., Finkenstadt and Rootzén, 2003, Embrechts et al., 2013). In engineering, it assists in evaluating the structural integrity of bridges, dams, or platforms against extreme

loads (e.g., Castillo, 1988). In insurance, extreme value theory enables insurers to assess potential losses arising from catastrophic events (e.g., Smith, 2003). For environmental studies, which is the focus of this thesis, it allows researchers and practitioners to better understand, predict, and manage the risks associated with natural hazards. In particular, various applications exist:

- Extreme weather events, such as hurricanes (e.g. Casson and Coles, 2000), heavy rainfall (e.g. Coles and Tawn, 1996), or droughts (e.g. Engeland et al., 2004). It helps in understanding the distribution and occurrence of these events, and assessing their impacts on ecosystems and human populations.
- Hydrology, for example to estimate the probability of extreme floods (see Pan et al. (2022) for a recent overview on extreme value for flood frequency analysis), storm surges (Butler et al., 2007), or wave height (Wadsworth et al., 2010).
- Extreme air pollution, which focuses on concentrations of pollutants, for example CO concentration (Sharma et al., 1999) or NO_2 levels (Castro-Camilo et al., 2021).

See also the book of Bousquet and Bernardara (2021) for various environmental examples, and Chapter 5 for different case studies on river flow and wind speed time series.

1.1.4.2 Return level estimation

All environmental quantities mentioned above correspond in practice to times series, *i.e.* measurements taken at regular intervals. Consequently, practitioners often use concepts like annual exceedance probability, which correspond to a probability that a given value is exceeded in a one-year period. This concept can be extended to the notions of return periods and return values.

Definition 3. *A return level ℓ_T associated with a return period T is defined as the quantile of order $1-\frac{1}{T}$ of the maximum values observed over a given period. For instance, if there are n_y observations per year, a T-year return level ℓ_T satisfies the equation*

$$\mathbb{P}(M_{n_y} < \ell_T) = 1 - \frac{1}{T}, \tag{1.20}$$

with $M_{n_y} = \max\{X_1, \dots, X_{n_y}\}$.

The return level ℓ_T can be interpreted as the value that is expected to be exceeded once every T years on average. Similarly to the duality of probability and quantile estimation, one can be interested in estimating the T-year return level associated with a given return period of T years, or conversely, estimating the return period T associated with a given return level ℓ_T. For return level estimation, Theorem 1 provides an approximation of Equation (1.20) in the case of sufficiently large n_y. This approximation can be inverted to obtain an estimate of the return level ℓ_T:

$$\ell_T \simeq \mu + \sigma G_\xi^{-1}(1-\frac{1}{T}) = \begin{cases} \mu + \frac{\sigma}{\xi}\left((-\log(1-1/T))^{-\xi} - 1\right) & \text{if } \xi \neq 0\,, \\ \mu - \sigma\log(-\log(1-1/T)) & \text{if } \xi = 0. \end{cases} \tag{1.21}$$

Note also that

$$\mathbb{P}(M_{n_y} < \ell_T) = (1-\bar{F}(\ell_T))^{n_y} \simeq 1 - n_y\bar{F}(\ell_T),$$

so another approach which uses the peaks-over-threshold method involves considering the quantile of order $1/(n_y T)$ and using the approximation in Equation (1.11) to obtain

$$\ell_T \simeq u + H^{-1}_{\sigma_u,\xi}\left(1 - \frac{1}{n_y T \bar{F}(u)}\right) \simeq \begin{cases} u + \sigma_u \frac{\left(n_y T \bar{F}(u)\right)^{\xi} - 1}{\xi} & \text{if } \xi \neq 0\,, \\ u - \sigma_u \log\left(\frac{1}{n_y T \bar{F}(u)}\right) & \text{if } \xi = 0. \end{cases} \tag{1.22}$$

It is important to note that all these results are derived under the assumption of i.i.d. data, which is often unrealistic for environmental time series due to factors such as non-zero dependencies between consecutive observations, seasonality, time trends, etc. Some commonly used preprocessing steps of the data are described in Chapter 5.

1.2 Introduction to Bayesian statistics

The Bayesian viewpoint of statistics, named after Thomas Bayes who proposed Bayes' rule in 1763, was further developed by Laplace in the 18th century. Laplace contributed to the theoretical but also computational aspects of Bayesian statistics, introducing the Laplace approximation (Tierney and Kadane, 1986). In the 20th century, significant advancements were made in Bayesian computation, driven by the computational power of machines to simulate random draws (Metropolis et al., 1953, Hastings, 1970), see Martin et al. (2020) for historical insights. Foundational work in Bayesian statistics was also established during this time (Jeffreys, 1939, Savage, 1954, Berger, 1988). Comprehensive introductions of the Bayesian viewpoint include the books of Robert (2007) and Gelman et al. (2013).

1.2.1 The Bayesian paradigm

The fundamental idea of Bayesian statistics is to provide a full probability model for both observable quantities, which are the data, and unobservable quantities, which are the parameters of the model. This allows the available information about any quantity of interest to be summarized using a probability distribution. We consider the general case of n i.i.d. realisations $\boldsymbol{x}^{(n)} = (x_1, \ldots, x_n)$ of $X_1, \ldots, X_n$, that are distributed according to a likelihood $p(\boldsymbol{x}^{(n)} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} p(x_i \mid \boldsymbol{\theta})$, given a vector of parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. It is important to make the distinction between the observables, which are quantities that can be directly measured (such as a dataset $\boldsymbol{x}^{(n)}$ or a future observation x), and the parameters, which are unobservable and must be inferred using the observable quantities.

Given a model $p(\boldsymbol{x}^{(n)} \mid \boldsymbol{\theta})$, frequentist statisticians treat the unobservable $\boldsymbol{\theta}$ as deterministic and aim to find estimators that are functions of the random observations. In contrast, Bayesian statisticians consider the unknown parameters as random variables and assign them a prior distribution $p(\boldsymbol{\theta})$. This prior combined with $p(\boldsymbol{x}^{(n)} \mid \boldsymbol{\theta})$ yield the joint distribution of $\boldsymbol{\theta}$ and $\boldsymbol{x}^{(n)}$, denoted as $p(\boldsymbol{\theta}, \boldsymbol{x}^{(n)})$. By applying Bayes' rule, we obtain the posterior distribution of $\boldsymbol{\theta}$ given the observed data $\boldsymbol{x}^{(n)}$.

Definition 4. *The posterior distribution of $\boldsymbol{\theta}$ is a conditional density that updates the prior information on the parameters $p(\boldsymbol{\theta})$ based on the observed data $\boldsymbol{x}^{(n)}$ using Bayes' rule:*

$$p(\boldsymbol{\theta} \mid \boldsymbol{x}^{(n)}) = \frac{p(\boldsymbol{x}^{(n)} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\boldsymbol{x}^{(n)})} = \frac{p(\boldsymbol{x}^{(n)} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int_{\Theta} p(\boldsymbol{x}^{(n)} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}}. \tag{1.23}$$

The denominator $p(\boldsymbol{x}^{(n)})$, known as the evidence or the marginal distribution of $\boldsymbol{x}^{(n)}$, can be seen as a normalizing constant. Usually, its explicit expression is not known, except for certain special cases called conjugate distributions:

Definition 5. *A family of prior distributions is said to be conjugate for a given likelihood if it leads to a posterior distribution that belongs to the same family of distributions.*

Examples of likelihoods with a known conjugate prior include simple cases such as Gaussian (with Gaussian conjugate prior for the mean), binomial (with a beta conjugate prior), or Pareto distributions (with a gamma conjugate prior). In most cases, the calculation of the integral for the evidence is not tractable, and we only have access to the unnormalized posterior density:

$$p(\boldsymbol{\theta} \mid \boldsymbol{x}^{(n)}) \propto p(\boldsymbol{\theta}) \prod_{i=1}^{n} p(x_i \mid \boldsymbol{\theta}). \tag{1.24}$$

Despite the challenges in computing the exact distribution, this posterior represents all the available information about the parameters after observing $\boldsymbol{x}^{(n)}$. From this posterior distribution, point estimation problems can be reformulated as scalar summaries of the posterior. This includes obtaining point estimates of the parameters, making predictions, quantifying the uncertainty around estimates, etc. A global formulation is the derivation of the posterior mean of a given function $f(\boldsymbol{\theta})$ of the parameters:

$$\mathbb{E}_{p(\cdot|\boldsymbol{x}^{(n)})}[f(\boldsymbol{\theta})] = \int_{\Theta} f(\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{x}^{(n)}) \mathrm{d}\boldsymbol{\theta}. \tag{1.25}$$

Different choices of the function f can cover posterior quantities such as posterior momentsn as well as other observable quantities such as the posterior predictive distribution:

Definition 6. *A predictive distribution is the probability distribution of a new observation x where all the unobservable are marginalized and all the observable are conditioned. If there is no observations, we obtain the prior predictive distribution:*

$$p(x) = \int p(x \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}. \tag{1.26}$$

Otherwise, the probability can be conditioned by the observation of $\boldsymbol{x}^{(n)}$ to obtain the posterior predictive:

$$p(x \mid \boldsymbol{x}^{(n)}) = \int p(x \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{x}^{(n)}) \mathrm{d}\boldsymbol{\theta}. \tag{1.27}$$

Note that choosing $f(\boldsymbol{\theta}) = p(x \mid \boldsymbol{\theta})$ allows to include the posterior predictive distribution into the formalism of Equation (1.25).

1.2.2 Modelling prior information

In order to carry out a Bayesian analysis, it is necessary to have a prior distribution $p(\boldsymbol{\theta})$ in addition to the likelihood $p(\boldsymbol{x}^{(n)} \mid \boldsymbol{\theta})$. Although the assumption that data follows a given distribution is widely accepted by anyone practicing parametric statistics, the assumption that parameters themselves are random is often cited as a drawback of Bayesian statistics, which are perceived as subjective. Historically, in the debates between frequentists and

Bayesians, this subjectivist view has not only been an argument to be countered for Bayesian statisticians, but has also served as the foundation for a new epistemological interpretation of probabilities (De Finetti, 1937). However, it should be noted that nothing in the theory indicates a systematic choice of a prior distribution, and so the choice of prior is a question that Bayesians have been trying to address since the inception of Bayesian statistics.

We provide a brief overview of different approaches to selecting a prior distribution, which can be categorized into two main groups. The first group consists of informative and weakly-informative priors. These priors are chosen with the intention of influencing the posterior distribution based on external knowledge or beliefs. It can incorporate additional information obtained from experts or impose constraints or regularization on the model. The second group includes uninformative priors, also known as objective priors. These priors are employed when no external information is available or when the goal is to maintain objectivity in the inference process. Uninformative priors aim to provide a prior distribution that is as non-informative as possible, allowing the data to have a stronger influence on the resulting posterior distribution.

1.2.2.1 Informative priors

Even when additional expert knowledge is available, expressing it in terms of a prior distribution for the parameters can be challenging: there may be multiple compatible prior choices, the information may not directly relate to the parameters, and quantifying the significance of the information in posterior inference is difficult (see an attempt in Jones et al., 2022). We briefly review several strategies for choosing informative priors and refer to O'Hagan (2019), Mikkola et al. (2023) for complete and recent overviews.

- **Conjugate priors**: When available, conjugate priors (see Definition 5) are often the most natural choice for modelling. They are informative in the sense that they require specifying the parameters of the prior distribution, but their explicit formulation simplifies subsequent analysis. However, even when conjugate priors are available, the choice among them can be a topic of discussion (see Robert, 2007, Chapter 3).

- **Hierarchical priors**: In cases where a prior is elicited, it usually comes with parameters $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$ to fix. To formalize the uncertainty around the parameters of the prior, usually called *hyperparameters*, one can choose to consider them as random too with distribution $p(\boldsymbol{\lambda})$, giving:

  $$p(\boldsymbol{\theta}) = \int_{\boldsymbol{\Lambda}} p(\boldsymbol{\theta} \mid \boldsymbol{\lambda}) p(\boldsymbol{\lambda}) \mathrm{d}\boldsymbol{\Lambda}.$$

  This kind of models with an additional layer are called hierarchical, see Chapter 10 in Robert (2007) and Chapter 5 in Gelman et al. (2013) for discussions. Hierarchical models offer several advantages: they are flexible and allow for more realistic modelling of real-world phenomena, for example by explicitly accounting for the heterogeneity that exists between different groups or clusters. However, they may also come with computational challenges and identifiability issues.

- **Predictive priors**: Applied statisticians often have insights at the observable level rather than directly on the parameters. In some cases, prior information in terms of

a prior predictive distribution $p(x)$ can be elicited. However, deducing the prior $p(\boldsymbol{\theta})$ from the prior predictive distribution involves solving the integral Equation (1.26) which requires additional assumptions, such as Tikhonov regularization (Gribok et al., 2004). Alternatively, the likelihood can be reparameterized as a function of observable quantities of interest, such as probabilities or quantiles, and a prior can be specified for them based on expert information (Coles and Tawn, 1996). Recent developments suggest using an expert's information on the probabilities of a partition of the observable space, modeled by a Dirichlet distribution (Hartmann et al., 2020). See also Perepolkin et al. (2021) for an extension on quantiles instead of probabilities elicitation.

- **Empirical Bayes priors**: Another approach for incorporating information into the prior is to estimate the prior distribution using the data. This approach, known as empirical Bayes, typically uses frequentist estimations like maximum likelihood to estimate the hyperparameters in the prior. It is important to note that empirical Bayes violates the strict principles of Bayesian inference, since it uses the data to inform the prior, which is supposed to represent prior beliefs. While it can provide efficient data-driven estimates, it can also lead to issues such as overfitting the prior to the data. However, empirical Bayes has shown interesting asymptotic properties, see for instance Chapter 4 in Berger (1988) or Rousseau and Szabo (2017).

1.2.2.2 Uninformative priors

When no information is initially available, modelling a prior can be challenging. A first approach is to consider a uniform prior distribution (also known as flat prior), assuming that it does not favor any specific values. However, the choice of parameterization is crucial, as a uniform prior may not remain uniform after a change of coordinates. A change of variable can therefore reveal hidden information in a flat prior. To address this issue, one can consider the parameter's invariances. For example, if μ represents a location parameter for the density of a random variable X such that $p(x \mid \mu) = g(x - \mu)$, then a random variable $Y = X + a$ has a distribution expressed as $g(x - (\mu + a))$. Thus, an invariant prior for μ should treat Y and X equally, meaning that $g(\mu) = g(\mu + a)$ for all $a \in \mathbb{R}$, and leading to a uniform prior on the real line : $g(\mu) \propto 1$. Unless used on a bounded support, this prior is improper as it integrates to ∞, but can be used if it yields a proper posterior distribution. Similarly, one can show that an invariant prior for a scale parameter $\sigma > 0$ is proportional to $\frac{1}{\sigma}$, which is equivalent to a uniform prior for $\log(\sigma)$.

In the general case, we briefly review different families of uninformative priors, and refer to Chapter 3 in Robert (2007) for more details on each of them.

- **Jeffreys prior**: A solution to construct a distribution that is invariant under reparameterization in the general case is proposed by Jeffreys (1939), and is still one of the most popular uninformative prior.

  Definition 7. *Jeffreys prior for the parameters $\boldsymbol{\theta}$ is defined as the square root of the determinant of Fisher information matrix:*

  $$p_J(\boldsymbol{\theta}) \propto \sqrt{\det \mathcal{I}(\boldsymbol{\theta})}, \tag{1.28}$$

  with

  $$\mathcal{I}(\boldsymbol{\theta}) = \mathbb{E}\left[-\frac{\partial^2}{\partial \boldsymbol{\theta}^2} \log p(x \mid \boldsymbol{\theta}) \mid \boldsymbol{\theta} \right]. \tag{1.29}$$

Under this definition, a reparameterization $\boldsymbol{\phi} = h(\boldsymbol{\theta})$ with h being a continuously differentiable function results in $p_J(\boldsymbol{\phi}) \propto \sqrt{\det \mathcal{I}(\boldsymbol{\phi})}$, making the expression equivalent to if we had initially considered the parameters $\boldsymbol{\phi}$. It is worth noting that Jeffreys prior, besides being often improper, does not adhere to the likelihood principle, which states that the evidence provided by the observations should only be contained in the likelihood (Berger and Wolpert, 1988). Furthermore, there is no uniqueness when it comes to priors invariant to reparameterization. For instance, the product of univariate Jeffreys priors is another candidate that may possess better optimality properties in specific cases (Kass and Wasserman, 1996). Additional issues associated with Jeffreys' prior are discussed in Robert (2007), Chapter 3, but according to the author, this prior remains "*the best 'automated' technique to derive noninformative prior distributions*".

- **Reference prior**: An extension of Jeffreys prior is known as reference priors (Bernardo, 1979, Berger et al., 2009). It corresponds to a distribution that minimizes its divergence from the posterior distribution, typically the Kullback–Leibler divergence. This definition refines the concept of non-informativity by aiming to influence the posterior as minimally as possible. The mathematical formulation of this problem involves maximizing a mutual information, which can be analytically intractable. In one dimension, this prior coincides with Jeffreys rule and, under certain regularity conditions, holds for higher dimensions as well (Clarke and Barron, 1994).

- **Matching prior**: Finally, another family of uninformative priors is matching priors (Datta and Sweeting, 2005), which is an uninformative prior that aim at producing a posterior probability at a given region asymptotically close to the corresponding frequentist coverage probability. This typically results in a posterior confidence set asymptotically close to a frequentist confidence interval, or to posterior predictive quantiles close to the real one. More formally, if we denote $Q_\alpha(\boldsymbol{x}^{(n)}, \pi)$ the quantile of order $(1-\alpha)$ of the Bayesian posterior predictive distribution for a given prior π, then π is a matching prior for predictive if for a given random new random variable X_{n+1}, we have
$$\mathbb{P}(X_{n+1} > Q_\alpha(\boldsymbol{x}^{(n)}, \pi) \mid \boldsymbol{\theta}) = \alpha + O(n^{-1}). \tag{1.30}$$

For all these priors presented here, the term "uninformative" can be misleading and has faced criticism (Gelman et al., 2017, Lemoine, 2019). Although they are employed when no prior information is available, the priors themselves always contain some form of information, and understanding how this information influences posterior inference can be challenging. Uninformative priors can, in fact, be strongly informative, in the sense that they have a significant influence on posterior distributions (Lemoine, 2019). It is sometimes suggested to use the so-called weakly informative priors instead, which are nearly flat priors with very high variance, to mitigate computational issues associated with uninformative priors. This is typically recommended in the Bayesian workflow of Gelman et al. (2020).

1.2.3 Computational methods

In the general case of Equation (1.25), the expression of the posterior mean of a given functional $f(\boldsymbol{\theta})$ cannot be derived analytically and requires the use of an approximation method. However, its computation is usually challenging for two main reasons:

- **Non-explicit form**: As mentioned before, it involves the posterior distribution, which is known only up to a constant in most of the cases.
- **High dimensionality**: The dimension of the integral corresponds to the dimension of the parameter space, which can quickly become too large for simple algorithms. Moreover, the geometry of high dimension distributions poses challenges for many computational methods (see Betancourt, 2019).

We introduce Markov chain Monte Carlo (MCMC) algorithms which aim at addressing these challenges. These methods are widely employed in Bayesian statistics and offer the advantage of being applicable to a wide range of models. In recent years, the development of probabilistic programming languages has also facilitated MCMC implementation. Notable examples include PyMC3 (Salvatier et al., 2016), Nimble (de Valpine et al., 2017), or Stan (Carpenter et al., 2017). These tools have made it easier to apply MCMC techniques to various statistical problems. For detailed information on MCMC methods, we recommend books such as Gilks et al. (1995), Robert and Casella (2004), Brooks et al. (2011).

1.2.3.1 Metropolis–Hastings algorithms

MCMC methods combine two fundamental concepts:

- **Monte Carlo approximation**: This involves approximating the integral in Equation (1.25) by a finite sum of random samples $f(\boldsymbol{\theta}^{(1)}), \dots, f(\boldsymbol{\theta}^{(N)})$. If $\boldsymbol{\theta}^{(i)}$ is distributed according to the posterior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{x}^{(n)})$ for $i \in \{1, \dots, N\}$, then the law of large number states that

$$\frac{1}{N} \sum_{i=1}^{N} f(\boldsymbol{\theta}^{(i)}) \xrightarrow[N \to +\infty]{} \int_{\Theta} f(\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{x}^{(n)}) d\boldsymbol{\theta} = \mathbb{E}_{p(\cdot \mid \boldsymbol{x}^{(n)})}[f(\boldsymbol{\theta})]. \tag{1.31}$$

  In practice, this approach involves considering a finite sample size N and computing the empirical mean to approximate the theoretical one.

- **Markov chain generation**: The Monte Carlo approximation requires generating values from the posterior distribution, which is often impossible as it is only known up to a constant. Instead, the values $\boldsymbol{\theta}^{(i)}$ are generated from a Markov chain, meaning that they are sampled according to a Markov kernel $K(\cdot, \cdot)$ that depends on the previous value:

$$\boldsymbol{\theta}^{(i+1)} \mid \boldsymbol{\theta}^{(i)} \sim K(\boldsymbol{\theta}^{(i)}, \cdot). \tag{1.32}$$

  With an appropriate choice of $K(\cdot, \cdot)$, it is possible to show that the limiting distribution of $\boldsymbol{\theta}^{(i)}$ is the target distribution (in this case, the posterior distribution), and that the result in Equation (1.31) holds true with the right choice of $K(\cdot, \cdot)$. See Chapter 6 in Robert and Casella (2004) for more details on the convergence properties of Markov chains and MCMC algorithms.

An ubiquitous MCMC method when $\boldsymbol{\theta} = (\theta_1, \dots, \theta_d)$ with $d \geq 2$ is the Gibbs sampler (Geman and Geman, 1984), where the idea is to draw values of θ conditionally on all other components (denoted full conditional distributions) and iterate for all components $i \in 1, \dots, d$. The general procedure for a fixed number of iterations N is outlined in Algorithm 1.

Algorithm 1: Gibbs sampler

Initialize $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \ldots, \theta_d^{(0)})$;
for $i \leftarrow 1$ **to** N **do**
  Sample $\theta_1^{(i)} \sim p(\theta_1 \mid \boldsymbol{x}^{(n)}, \theta_2^{(i-1)}, \theta_3^{(i-1)}, \ldots, \theta_d^{(i-1)})$;
  Sample $\theta_2^{(i)} \sim p(\theta_2 \mid \boldsymbol{x}^{(n)}, \theta_1^{(i)}, \theta_3^{(i-1)}, \ldots, \theta_d^{(i-1)})$;
  $\vdots$
  Sample $\theta_d^{(i)} \sim p(\theta_d \mid \boldsymbol{x}^{(n)}, \theta_1^{(i)}, \theta_2^{(i)}, \ldots, \theta_{d-1}^{(i)})$;
end

By using it, we can generate samples $\boldsymbol{\theta}^{(i)}$ that depend on the previous samples solely through $\boldsymbol{\theta}^{(i-1)}$. Remarkably, these samples possess the property of converging to the target distribution given simple conditions (Roberts and Polson, 1994). However, in situations where the full conditional distributions are not readily available, a more general approach is to employ the Metropolis–Hastings algorithm (Metropolis et al., 1953, Hastings, 1970). This technique involves sampling a value $\boldsymbol{\theta}^*$ from a known distribution called the proposal distribution $q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(i)})$, and accept this new value or not according to an acceptance ratio $\alpha(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(i)})$. Algorithm 2 outlines the steps of the Metropolis–Hastings algorithm.

Algorithm 2: Metropolis–Hastings algorithm

Initialize $\boldsymbol{\theta}^{(0)}$;
for $i \leftarrow 1$ **to** N **do**
  Sample $\boldsymbol{\theta}^* \sim q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(i-1)})$;
  Sample $u \sim \mathcal{U}[0, 1]$;
  Compute $\alpha(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(i-1)}) = \min\left(1, \frac{p(\boldsymbol{\theta}^* \mid \boldsymbol{x}^{(n)})}{p(\boldsymbol{\theta}^{(i-1)} \mid \boldsymbol{x}^{(n)})} \frac{q(\boldsymbol{\theta}^{(i-1)} \mid \boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(i-1)})}\right)$;
  if $u < \alpha(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(i-1)})$ **then**
    $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^*$;
  else
    $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^{(i-1)}$;
  end
end

Note that the Gibbs sampler (Algorithm 1) is a special case of Metropolis–Hastings with the proposal $q(\theta_k^* \mid \boldsymbol{\theta}^{(i)}) = p(\theta_k^* \mid \theta_1^{(i)}, \cdots, \theta_d^{(i)}, \boldsymbol{x}^{(n)})$. Routine calculations show that this choice of proposal results in $\alpha(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(i)}) = 1$, which mean that every candidate generated by the proposal is accepted. In the general case, the acceptance probability depends on two ratios:

- The posterior ratio $\frac{p(\boldsymbol{\theta}^* \mid \boldsymbol{x}^{(n)})}{p(\boldsymbol{\theta}^{(i-1)} \mid \boldsymbol{x}^{(n)})}$, where the normalizing constant simplifies and favors samples with higher posterior density.
- The proposal ratio $\frac{q(\boldsymbol{\theta}^{(i-1)} \mid \boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(i-1)})}$ which acts as a correction term. This ratio simplifies to 1 in the case of symmetric proposals, where $q(\boldsymbol{\theta}^{(i-1)} \mid \boldsymbol{\theta}^*) = q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(i-1)})$.

The Metropolis–Hastings algorithm allows for control over the dependence between $\boldsymbol{\theta}^{(i)}$

and $\boldsymbol{\theta}^{(i-1)}$ through the choice of the proposal distribution. For example, if $q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(i)}) = \mathcal{N}(\boldsymbol{\theta}^{(i)}, \delta^2 \boldsymbol{I})$, the parameter δ can be adjusted to tune the autocorrelation. However, this tuning is usually hard, espicially in high dimension (Betancourt, 2017): in this example of proposal, a small value of δ leads to successive values of $\boldsymbol{\theta}$ being too close to each other, while a large value can cause the algorithm to get stuck on a value for multiple steps. Achieving a balance between exploration of the parameter space and avoiding excessive duplication is crucial. Typically, a good proposal distribution results in an acceptance rate ranging from 15% to 45% (Roberts and Rosenthal, 2001).

It is also worth noting that every Metropolis–Hastings algorithm requires an initialization value $\boldsymbol{\theta}^{(0)}$, which can significantly impact the algorithm's efficiency for a finite number of iterations N. To mitigate this influence, it is recommended to define a burn-in period where the initial iterations are discarded for inference, reducing the impact of the starting value. Another approach to reduce the influence of the starting value is to run multiple chains in parallel (Gelman and Rubin, 1992). Although this approach is debated and some prefer using all available iterations for a long chain (Geyer, 1992), advancements in technology, particularly parallel computing, have made running multiple chains a standard practice in most probabilistic programming languages. Also, it helps for diagnosing convergence issues as it enables the chains distribution to be compared with each other to identify defect in some of them (see Section 1.2.4).

1.2.3.2 Gradient-based Metropolis algorithms

Several advancements in MCMC algorithms have been proposed to enhance their efficiency by incorporating the gradient of the target distribution into the proposal step. One example of such improvement is the Metropolis-adjusted Langevin algorithm (Roberts and Tweedie, 1996) where the proposal is based on Langevin dynamics and the gradient information is utilized to simulate a diffusion process.

Another notable development in this direction is Hamiltonian Monte Carlo (HMC, Neal, 1996). HMC takes inspiration from the behavior of a mass in a gravitational field and solves the Hamiltonian equations to propose new points in the state space. The key idea is therefore to introduce an auxiliary momentum variable $\boldsymbol{\xi}^*$ and perform a conservative exploration of the log-posterior. In this context, the Hamiltonian $H(\boldsymbol{\theta}, \boldsymbol{\xi})$ is

$$H(\boldsymbol{\theta}, \boldsymbol{\xi}) = -\log p(\boldsymbol{\xi}, \boldsymbol{\theta} \mid \boldsymbol{x}^{(n)}) = -\log p(\boldsymbol{\xi} \mid \boldsymbol{\theta}, \boldsymbol{x}^{(n)}) - \log p(\boldsymbol{\theta} \mid \boldsymbol{x}^{(n)}), \tag{1.33}$$

where the two terms can be respectively seen as kinetic and potential energies. The corresponding equations of motion are

$$\frac{d\theta_k}{dt} = \frac{\partial H}{\partial \xi_k}, \qquad \frac{d\xi_k}{dt} = \frac{\partial H}{\partial \theta_k}. \tag{1.34}$$

Solving them until a given time allows to obtain a candidate that will be accepted or not depending on the acceptance ratio. A basic version of the method is shown in Algorithm 3, with a discrete leapfrog integrator that corresponds to the three-step update of $\boldsymbol{\xi}_l$ and $\boldsymbol{\theta}_l$ at each time step $l \in \{1, \dots, L\}$, with L a given number of discrete steps. This time discretization allows the integrator to avoid error propagation. See Betancourt (2017) for more details and extensions, and Hoffman and Gelman (2014) for an improvement of HMC named No-U-Turn sampler.

Algorithm 3: Hamiltonian Monte Carlo

Initialize $\boldsymbol{\theta}^{(0)}$;
for $i \leftarrow 1$ **to** N **do**
  Sample $\boldsymbol{\xi}^* \sim \mathcal{N}(0, M)$;
  Let $\boldsymbol{\theta}_0 = \boldsymbol{\theta}^{(i-1)}$ and $\boldsymbol{\xi}_0 = \boldsymbol{\xi}^*$;
  for $l \leftarrow 1$ **to** L **do**
    $\boldsymbol{\xi}_{l-1/2} = \boldsymbol{\xi}_{l-1} + \epsilon \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}_{l-1} \mid \boldsymbol{x}^{(n)})/2$;
    $\boldsymbol{\theta}_l = \boldsymbol{\theta}_{l-1} + \epsilon \boldsymbol{\xi}_{l-1/2}$;
    $\boldsymbol{\xi}_l = \boldsymbol{\xi}_{l-1/2} + \epsilon \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}_l \mid \boldsymbol{x}^{(n)})/2$;
  end
  Sample $u \sim \mathcal{U}[0, 1]$;
  Compute $\alpha(\boldsymbol{\theta}_L, -\boldsymbol{\xi}_L \mid \boldsymbol{\theta}_0, \boldsymbol{\xi}_0) = \min\left(1, \frac{p(\boldsymbol{\theta}_L, -\boldsymbol{\xi}_L \mid \boldsymbol{x}^{(n)})}{p(\boldsymbol{\theta}_0, \boldsymbol{\xi}_0 \mid \boldsymbol{x}^{(n)})}\right)$;
  if $u < \alpha(\boldsymbol{\theta}_L, -\boldsymbol{\xi}_L \mid \boldsymbol{\theta}_0, \boldsymbol{\xi}_0)$ **then**
    $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}_L$;
  else
    $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}_0$;
  end
end

1.2.3.3 Other computational methods

Although they will not be utilized in the subsequent sections of the thesis, it is worth mentioning that there exist alternative computational methods for Bayesian inference. Some of these methods include other Monte Carlo techniques such as importance sampling or sequential Monte Carlo (Chopin and Papaspiliopoulos, 2020). Another kind of method is integrated nested Laplace algorithms (Rue et al., 2009) based on the Laplace approximation (Tierney and Kadane, 1986). In cases where the models are highly complex, such as Bayesian neural networks, variational inference (Blei et al., 2017) offers another alternative, by approximating the posterior distribution with a simpler one and aiming to find the parameters of this approximation by maximizing the Kullback–Leibler (KL) divergence between the true posterior and the approximation. Finally, approximate Bayesian computation (ABC, Sisson et al., 2018) is a computational framework used in Bayesian statistics to approximate posterior distributions when direct computation of the likelihood function is not feasible.

1.2.4 MCMC convergence diagnostics

In order to ensure that an MCMC method provides a good approximation of the target distribution, it is crucial to assess the convergence of the Markov chain(s) within a fixed number of iterations, since MCMC guarantees are only asymptotic. In particular, two fundamental properties need to be verified:

- **Stationarity:** If the chains are still in the early stages of exploration and have not adequately traversed the parameter space, the generated distributions will not accurately represent the target distribution.

- **Mixing:** It is possible for the chains to appear to have reached stationarity, but they may have done so by exploring only a subspace of the parameter space. This subset depends on the initial values, and the distribution will not have fully converged to the target distribution.

Various diagnostic indicators have been developed to assess the convergence of MCMC algorithms. We will describe univariate versions of these indicators for a specific component θ_i, with $i \in 1, \dots, d$, simply denoted by θ from now on. Generalizing these diagnostics to higher dimensions in $\boldsymbol{\theta}$ is challenging, as discussed in Chapter 3.

1.2.4.1 Autocorrelation at lag-t

Unlike independent Monte Carlo samplers, the elements of Markov chains are not independent, as each sample at time t is used to generate the next one at time $t+1$. This dependency impacts the quality of the approximation, as it reduces the effective information contained in each element of the chain. The simplest way to estimate autocorrelation for a given chain is to use the sample autocorrelation function at lag-t, which measures the correlation between elements of the sequence that are distant from each other by t steps. For a given parameter θ with corresponding MCMC samples $(\theta^{(1)}, \dots, \theta^{(N)})$, the autocorrelation at lag-t, denoted by $\text{acf}_t(\theta)$, is calculated as follows:

$$\text{acf}_t(\theta) = \frac{\frac{1}{N-t}\sum_{k=1}^{N-t}(\theta^{(k)} - \bar{\theta})(\theta^{(k+t)} - \bar{\theta})}{\frac{1}{N-1}\sum_{k=1}^{N}(\theta^{(k)} - \bar{\theta})^2},$$

with $\bar{\theta} = \frac{1}{N}\sum_{k=1}^{N}\theta^{(k)}$. Other methods for autocorrelation estimation exist and are used in practice, see Chapter 13 in Gelman et al. (2013).

1.2.4.2 Effective Sample Size

The Effective Sample Size (ESS) represents the number of i.i.d. samples that would yield the same variance as the correlated samples obtained from the MCMC algorithm. For a given θ, let us assume that M chains $(\theta^{(1,l)}, \dots, \theta^{(N,l)})$ are simulated, where $l \in 1, \dots, M$. Denote $\bar{\theta}^{(.,l)} = \frac{1}{N}\sum_{k=1}^{N}\theta^{(k,l)}$ and $\bar{\theta}^{(.,.)} = \frac{1}{N}\sum_{l=1}^{M}\bar{\theta}^{(.,l)}$. In the case of independent samples, we have $\text{Var}(\bar{\theta}^{(.,.)}) = \frac{1}{MN}\text{Var}(\theta \mid \boldsymbol{x}^{(n)})$. Otherwise, in the presence of correlation, we have the following asymptotic result:

$$MN \times \text{Var}(\bar{\theta}^{(.,.)}) \xrightarrow[N\to+\infty]{} \left(1 + 2\sum_{t=1}^{\infty}\text{acf}_t(\theta)\right)\text{Var}(\theta \mid \boldsymbol{x}^{(n)}). \tag{1.35}$$

From this result, we can define the ESS as

$$\text{ESS} := \frac{MN}{1 + 2\sum_{t=1}^{\infty}\text{acf}_t(\theta)}. \tag{1.36}$$

Thus, the ESS is estimated by truncating the sum in Equation (1.36) and estimating the autocorrelations.

1.2.4.3 Potential scale reduction factor $\hat{R}$

The potential scale reduction factor, denoted by $\hat{R}$, introduced by Gelman and Rubin (1992), is another diagnostic indicator used to assess the convergence of MCMC chains when multiple chains are employed. It compares the distributions of the chains with each other by examining the between-chain variance $\hat{B}$ and the within-chain variance $\hat{W}$. It is calculated as:

$$\hat{R} := \sqrt{\frac{\hat{W} + \hat{B}}{\hat{W}}}, \tag{1.37}$$

with

$$\hat{B} = \frac{1}{M-1}\sum_{l=1}^{M}\left(\bar{\theta}^{(.,l)} - \bar{\theta}^{(.,.)}\right)^2, \text{ and } \hat{W} = \frac{1}{M}\sum_{l=1}^{M} s_l^2, \text{ where } s_l^2 = \frac{1}{N-1}\sum_{k=1}^{n}\left(\theta^{(k,l)} - \bar{\theta}^{(.,l)}\right)^2. \tag{1.38}$$

The quantities $\hat{W} + \hat{B}$ and $\hat{B}$ both converge to the posterior variance as $n \to \infty$, but one by overestimating and the other by underestimating. More details are given in Chapter 3 which aims at improving this MCMC convergence diagnostic.

To sum up Gelman et al. (2013) recommends to use $\hat{R}$ and ESS with the following rule of thumb:

$$\hat{R} \leq 1.01 \implies \text{“Chains are mixing well”.}$$
$$\text{ESS} \geq 400 \implies \text{“Enough data for estimation”.}$$

1.3 Bayesian methods for univariate extreme value modelling

We are interested in applying Bayesian statistics (introduced in Section 1.2) to extreme value models (introduced in Section 1.1). Similarly to the frequentist approach, this involves assuming that observations are exactly distributed according to a GEV, GPD, or non homogeneous Poisson process (NHPP) distribution, and estimating the parameters $\boldsymbol{\theta} := (\mu, \sigma, \xi)$ (or $\boldsymbol{\theta} := (\sigma_u, \xi)$ for GPD) within the Bayesian framework. For extreme value analysis, there are several advantages to using Bayesian methods. Coles and Powell (1996) mentions the ability to incorporate prior information, the flexibility to handle any value of the shape parameter ξ (unlike frequentist estimates), and the access to the posterior predictive distribution (Definition 6), which allows for natural prediction models that consider parameter uncertainty.

Several articles and book chapters have been written to introduce Bayesian methods for extreme value analysis. A literature review can be found in Coles and Tawn (1996), and introductions are provided in Chapter 9 of Coles (2001), Smith (2003), and Coles (2003), which all cover similar content. Beirlant et al. (2006) adds a semi-parametric component in Chapter 11, connecting Bayesian methods with Hill and Weissman estimators in the frequentist case. See Ameraoui et al. (2016), Beirlant et al. (2018), Li et al. (2019) for extensions of Bayesian extremes estimate in the semi-parametric setting. The book by Dey and Yan (2016) includes two chapters dedicated to Bayesian methods. A first chapter of Stephenson (2016) updates the previous chapters, while a second one of Erhardt and Sisson (2016) focuses on methods using approximate Bayesian computation (ABC) specifically

for multivariate extremes. Finally, Bousquet (2021) provides a comprehensive overview of Bayesian extreme value methods, and also covers model selection and calibration compared to previous works. For a review of Bayesian implementations in extreme value analysis, the Bayesian section in Belzile et al. (2022) can be consulted.

We focus here on the development of Bayesian methods for univariate extreme value analysis. However, it is worth noting that many Bayesian models for multivariate extremes can also be found in the literature (e.g. Sabourin et al., 2013, Dombry et al., 2017, Jóhannesson et al., 2022). We organize our review into three research areas: prior modelling, uncertainty quantification, and hyperparameter elicitation.

1.3.1 Prior modelling

Various strategies exist for modelling prior distributions, both in informative and non-informative cases (see Section 1.2.2). In the context of prior modelling, Table 1.1 summarizes the current research on univariate extremes by providing information on each proposed prior, including whether it is informative or uninformative, the type of prior it corresponds to, and the likelihood it is used with (GEV, GPD, or Non-Homogeneous Poisson Process (NHPP)). It is important to note that Table 1.1 focuses specifically on articles that originally proposed new priors. In practice, in other Bayesian extremes papers, vague priors are commonly used, or alternatively, since the demonstration of posterior propriety by Northrop and Attalides (2016), uniform or Jeffreys have gained popularity and have been employed in various studies (e.g., Sharkey and Tawn, 2017, Beranger et al., 2021).

1.3.1.1 Informative prior

All the categories of informative priors mentioned in Section 1.2.2 have been developed in the literature on extreme value models. While no conjugate prior exists for these likelihoods, some quasi-conjugate priors, that exhibit conjugacy for certain parameters conditional on others, have been suggested. Examples include the works by Parent and Bernier (2003), Bousquet and Keller (2017) which use historical data, as well as Diebolt et al. (2005) which exploits the mixture of gamma distributions property of the GPD.

The perspective of predictive inference has also played a significant role in extreme value modelling. Given the challenges in interpreting the three parameters (μ, σ, ξ) for non-experts, it is more intuitive for applied statisticians to have information about the scale of observations. Coles and Tawn (1996) propose eliciting a prior on three quantiles (q_1, q_2, q_3) instead of the parameters. To respect the ordering constraint $q_1 < q_2 < q_3$, a positive prior is placed on the quantile difference, chosen as a gamma prior. An expression in terms of (μ, σ, ξ) can then be obtained with a change of variable. It should be noted that the choice of setting independent priors on quantile differences induces a dependence structure that can be interesting to discuss. Building on this work, Stephenson and Tawn (2004) modifies the approach to assign a non-zero posterior probability for $\xi = 0$. Gaioni et al. (2010) generalizes the framework above to handle more than three quantiles and proposes to use normal priors.

Hierarchical priors have also been employed for extreme value problems, for modelling extreme mixtures (Walshaw, 2000, Bottolo et al., 2003), or for a relaxed version of the GEV likelihood (Zorzetto et al., 2020). An empirical Bayes method has been suggested to achieve asymptotic results of the posterior distribution, in particular contraction rates

Reference	(Un)informative	Category	Likelihood	Comments
Coles and Tawn (1996)	Informative	Predictive level	GEV	Gamma on the quantile difference
Walshaw (2000)	Informative	Hierarchical	GEV	Multivariate normal
Bottolo et al. (2003)	Informative	Hierarchical	GEV	Mixture of GEV
de Zea Bermudez and Turkman (2003)	Uninformative	Vague	GPD	Pareto if $\xi > 0$, gamma otherwise
Smith (2003)	Uninformative	Vague	GPD	Normal with high variance
Parent and Bernier (2003)	Informative	Quasi-conjugate	GPD	Uses historical data
Stephenson and Tawn (2004)	Informative	Predictive level	GEV	Non-zero probability of $\xi = 0$
Diebolt et al. (2005)	Informative	Quasi-conjugate	GPD	Restriction to $\xi > 0$
Castellanos and Cabras (2007)	Uninformative	Jeffreys	GPD	Posterior propriety
Gaioni et al. (2010)	Informative	Predictive level	GEV	Multivariate normal
Ho (2010)	Uninformative	Matching prior	GPD	Posterior propriety but $\xi > 0$
Cabras (2013)	Uninformative	Jeffreys	NHPP	Approximate conditional likelihood
Northrop and Attalides (2016)	Uninformative	Uniform Jeffreys MDI	GPD GEV	Posterior propriety
Bousquet and Keller (2017)	Informative	Quasi-conjugate	GEV	Distinction of the 3 MDA
Opitz et al. (2018)	Informative	PC	GPD	See Simpson et al. (2017)
Zorzetto et al. (2020)	Informative	Hierarchical	GEV	Relaxed version of the asymptotic model
Castro-Camilo et al. (2021)	Informative	PC	GPD	Bounded version of PC prior
Padoan and Rizzelli (2022)	Informative	Empirical	GEV	Asymptotic results
Moins et al. (2023)	Uninformative Informative	Jeffreys PC	NHPP	Posterior propriety

Table 1.1: Original prior distribution propositions for univariate Bayesian extreme models.

(Padoan and Rizzelli, 2022). Lastly, a principled approach called PC, proposed by Simpson et al. (2017), has been applied to construct priors in the GPD case (Opitz et al., 2018). Castro-Camilo et al. (2021) has further modified the PC approach, constraining the prior to be bounded between 0 and 1/2.

1.3.1.2 Uninformative prior

Common choices for modelling uninformative priors in extreme value models include vague priors (*i.e.* priors with high variance), uniform priors (which may be improper), and Jeffreys' prior. At first, Smith (2003) and de Zea Bermudez and Turkman (2003) employ vague priors to ensure posterior propriety. In the GPD case, de Zea Bermudez and Turkman (2003) specifically chooses a Pareto distribution for ξ when $\xi > 0$ and a gamma distribution otherwise. Castellanos and Cabras (2007) demonstrates the posterior propriety of Jeffreys' prior for the GPD when $\xi > -1/2$, while Northrop and Attalides (2016) extends this study to uniform and maximal data information (MDI) priors for both the GEV and GPD cases (see Chapter 2 for more details). Ho (2010) proposes a matching prior for the case when $\xi > 0$, which yields a proper posterior despite the prior being improper.

1.3.2 Uncertainty quantification using posterior predictive

In the context of extreme value modelling, a Bayesian analysis offers the advantage of accessing to the entire posterior distribution, allowing for straightforward quantification of uncertainty. The posterior predictive distribution is particularly useful when estimating quantities in the observable space such as extreme quantiles, as it provides a way to quantify uncertainty. Numerous studies have examined the properties and applications of posterior predictive distributions in extreme value models. Davison (1986) and Smith (1999) explore the properties of posterior predictive distributions and provide an extreme value example. Engelund and Rackwitz (1992) investigates the use of predictive distributions for the three domain of attraction with an uninformative prior, and reveal using simulations that this approach can lead to unreasonable decisions. de Zea Bermudez et al. (2001) advocates for the posterior predictive distribution in a Poisson-GPD model. Fawcett and Walshaw (2016) extends the use of posterior predictive distributions to spatially dependent and non-stationary extreme value models, specifically estimating posterior predictive return levels. Additionally, Fawcett and Green (2018) and Jonathan et al. (2021) examine different Bayesian estimators of extreme quantiles to gain insights into the best approach for incorporating parameter uncertainty in the estimation process. More detailed information on this topic can be found in Chapter 4.

1.3.3 Bayesian elicitation of hyperparameters

As mentioned in Section 1.1, the block maxima and peaks-over-threshold approaches require to specify a hyperparameter that is crucial for obtaining efficient estimators: the block size in the GEV model, and the threshold in the GPD one. Various works have focused on threshold elicitation, considering the threshold as a parameter that delineates the boundary between the bulk and the tail of the distribution to be modelled. These works incorporate all observations, not just the excesses, and estimate the threshold within the Bayesian framework along with a model for the bulk and a model for the tail. Frigessi et al.

(2002) proposes a dynamically weighted mixture model with a GPD and a light-tailed distribution, eliminating the need for estimating a threshold explicitly. The threshold is then explicit in Behrens et al. (2004) and Tancredi et al. (2006), where a mixture of uniform for observations below the threshold is used in the latter. Other methods, such as Hundecha et al. (2009), MacDonald et al. (2011), Solari and Losada (2012), do Nascimento et al. (2012) utilize parametric or non-parametric models for the bulk. Recently, Martín et al. (2022) fits all the observations to a stable distribution which allows the author to deduce the GPD parameters. Some approaches do not require fitting the data below the threshold, such as the likelihood ratio test proposed by Wadsworth and Tawn (2012) and the use of posterior predictive checks by Lee et al. (2015) for threshold selection. Additionally, Northrop et al. (2017) performs Bayesian model averaging by combining estimations from multiple thresholds instead of choosing a single one. Other hyperparameters in the Bayesian framework have been studied as well, including the measurement scale (Wadsworth et al., 2010) and the scaling factor for the NHPP (Sharkey and Tawn, 2017). Further details on the latter can be found in Chapter 2.

1.4 Thesis outline

1.4.1 Context

Extreme weather events, although by definition very rare, can cause considerable human and material damage. Fires, floods, droughts, and cold waves have multiplied as a result of climate change, in addition to earthquakes, torrential rains, extreme winds, and so on. The more extreme these events are, the more dramatic their consequences can be. For the French electrical company *Électricité de France* (EDF), it is essential to quantify the risk associated with such events in order to ensure the proper design of infrastructures. EDF's R&D has been engaged in the modeling and analysis of extreme values in both univariate and multivariate contexts for over a decade as part of the MADONE project (*Méthodes pour les Agressions d'Origine Naturelle Externe*). The specific objective is the statistical study of meteorological variables such as temperature, flow rate, or wind speed at different sites to justify the sizing of structures to organizations like the Nuclear Safety Agency (ASN).

In particular, the goal is to determine return levels (Definition 3) associated with centennial, millennial, or even deca-millennial return periods. However, such estimations require significant extrapolation, which therefore come with significant uncertainty. These sources of uncertainty are multiple: those related to data, model, estimation, etc. Quantifying them allows for verifying the reliability of the estimates given by a model, particularly to answer the crucial question

How far is it reasonable to extrapolate the tail of the distribution?

This thesis is co-funded by EDF R&D, and follows previous works that have been done in the same environment and with the same problematics. First, the thesis of Clément Albert (Albert, 2018) focused on estimating the deterministic extrapolation error for the use of an extreme model. Then an internship, done by Valentin Chevalier, focussed on linking the extrapolation error to the mean squared error (MSE). Finally, Tony Zheng began exploring Bayesian methods in an internship for the estimation error (in contrast

to the deterministic error studied in Albert (2018)). This thesis is thus in continuation of this internship and examines the Bayesian approach to estimate extreme events.

1.4.2 Contributions

We provide an overview of the following chapters of the manuscript and emphasize its key contributions:

- In Chapter 2, we propose a reparameterization of the Poisson process characterization of univariate extremes for Bayesian inference, which leads to two benefits: an improvement of MCMC convergence, and the calculation of Jeffreys and PC prior for the NHPP. The framework is then applied on an environmental dataset for return level estimation of Garonne flow data (France).

  This work was presented in 4 conferences (ISBA 2021; JDS 2021; AppliBUGS day; EVA 2023), and led to a communication paper (Moins et al., 2021b) and a journal article accepted at the Computational Statistics and Data Analysis (CSDA) journal (Moins et al., 2023).

- In Chapter 3, we focus on a purely Bayesian computational problem and aim at better understanding the behavior of $\hat{R}$ diagnostic introduced in Section 1.2. This leads us to propose a localized version that focuses on quantiles of the target distribution. We obtain key theoretical properties of the associated population value, along with experimental guarantees of robustness for various experiments.

  This work was presented in 5 conferences (ISBA 2022; BAYSM 2022; One World YoungStatS webinar; Energy Forecasting Innovation Conference; CMStats 2022), 2 posters (BayesComp-ISBA; Bayes@CIRM), and led to a communication paper (Moins et al., 2022b), the discussion of Vehtari et al. (2021) Bayesian Analysis article (Moins et al., 2021a) that has been extended as a paper accepted at the Bayesian Analysis journal (Moins et al., 2023).

- In Chapter 4, our aim is to investigate the characteristics of different Bayesian quantities in the context of a finite number of observations in the GPD case. Our primary focus lies in examining the behavior of prior and posterior predictive distributions, as well as Bayesian estimators of the return level. By analyzing the prior predictive, we gain insights into how the choice of prior impacts extreme observations. Furthermore, studying the posterior quantities allows us to explore the boundaries of extrapolation when working with a limited amount of data.

- In Chapter 5, we illustrate the findings of the previous chapters on various environmental datasets provided by EDF: three river flows and three wind speed datasets jointly observed in three French cities: Tours, Reims and Orange. The aims are to check the behavior on different real-world datasets, to provide a Bayesian estimate of extreme return levels, to compare the results with previous internal EDF studies, and to answer to the main question on the limits of extrapolation in practical examples.

- Finally, we conclude the manuscript with a summary of our work and discuss several research perspectives for each contributions.

Chapter 2

Reparameterization of extreme value framework for improved Bayesian workflow

Contents

Résumé

Nous nous intéressons dans ce chapitre aux bénéfices d'une reparamétrisation orthogonale du processus de Poisson pour les extrêmes introduit en Partie 1.1 dans un cadre bayésien. Selon la définition de Jeffreys (1939), des paramètres sont dits orthogonaux si la matrice d'information de Fisher associée est diagonale. Dans le cadre bayésien, nous mettons en évidence deux avantages d'un tel changement de variable pour le processus de Poisson pour les extrêmes.

Tout d'abord, la convergence des méthodes de Monte Carlo par chaînes de Markov (MCMC) est améliorée lorsqu'elle est appliquée à des paramètres orthogonaux. Cette proposition repose sur des diagnostics de convergence comme l'autocorrelation des chaînes de Markov, la taille d'échantillon effective, ou une version améliorée de $\hat{R}$ présentée au Chapitre 3.

Ensuite, un second avantage de l'orthogonalisation est qu'elle simplifie le calcul de certaines lois *a priori* qui dépendent de l'information de Fisher. En particulier, la loi *a priori* non informative de Jeffreys ainsi que la loi *a priori* semi-informative nommée PC prior (Simpson et al., 2017) sont calculées pour le processus de Poisson pour les extrêmes. Les résultats montrent que les distributions sont impropres mais conduisent à une loi *a posteriori* propre, c'est-à-dire intégrable.

Ces améliorations de l'inférence bayésienne sont ensuite appliquées à l'estimation de niveaux de retour de données de débit de la Garonne (France). Les résultats reflètent comment le PC prior permet d'ajouter de l'information *a priori* sur l'indice de queue pour réduire l'incertitude de l'estimation du niveau de retour, en particulier la taille des intervalles de crédibilité.

Les résultats de ce chapitre sont présentés sous la forme d'un article accepté pour publication à CSDA (Moins et al., 2023). La Partie 2.1 rappelle le cadre bayésien des extrêmes dans lequel nous nous plaçons, puis présente les travaux déjà existants sur la reparamétrisation, du point de vue bayésien d'une part et sur la paramétrisation orthogonale d'autre part. La Partie 2.2 présente la reparamétrisation orthogonale et les intuitions qui permettent de supposer qu'elle va aider à la convergence des chaînes MCMC, tandis que la Partie 2.3 s'attarde sur le calcul des lois *a priori* de Jeffreys et PC. Les résultats sont ensuite illustrés sur les données de débit de la Garonne en Partie 2.4, avant de conclure en Partie 2.5. Les annexes contiennent une remarque additionnelle sur la quasi-orthogonalité proposée par Sharkey and Tawn (2017) (Annexe 2.A), des preuves détaillées (Annexe 2.B) ainsi que des expériences additionnelles (Annexe 2.C).

Abstract

In this chapter, we focus on the benefits of an orthogonal reparameterization of the Poisson process for extremes introduced in Section 1.1 within a Bayesian framework. According to the definition by Jeffreys (1939), parameters are said to be orthogonal if the associated Fisher information matrix is diagonal. In the Bayesian context, we highlight two advantages of such a variable change for the Poisson process for extremes.

Firstly, the convergence of Markov chain Monte Carlo (MCMC) methods is improved when applied to orthogonal parameters. This proposition relies on convergence diagnostics such as the autocorrelation of Markov chains, effective sample size, and an improved version of $\hat{R}$ presented in Chapter 3.

A second advantage of orthogonalization is that it simplifies the computation of certain prior distributions that depend on the Fisher information. In particular, the uninformative Jeffreys prior and the semi-informative PC (penalized complexity) prior (Simpson et al., 2017) are calculated for the Poisson process for extremes. The results show that the distributions are improper but lead to a proper posterior distribution.

These improvements in Bayesian inference are then applied to the estimation of return levels of Garonne flow data (France). The results demonstrate how the PC prior allows adding prior information on the tail index to reduce the uncertainty in the estimation of return levels, particularly the size of credibility intervals.

The results of this chapter are presented in the form of an article accepted in CSDA (Moins et al., 2023). Section 2.1 recalls the Bayesian extreme framework we adopt and presents existing work on reparameterization from both a Bayesian perspective and an orthogonal parameterization point of view. Section 2.2 introduces the orthogonal reparameterization and the insights that suggest it will aid in the convergence of MCMC chains, while Section 2.3 focuses on the computation of Jeffreys and PC prior distributions. The results are then illustrated using Garonne flow data in Section 2.4, followed by a conclusion in Section 2.5. The appendices contain additional remarks on the quasi-orthogonality proposed by Sharkey and Tawn (2017) (Appendix 2.A, details of all proofs (Appendix 2.B), and additional experiments (Appendix 2.C).

2.1 Introduction

Studying the long-term behavior of environmental variables is necessary to understand the risks of hazardous meteorological events such as floods, storms, or droughts. To this end, models from extreme value theory allow us to extrapolate data in the distribution tails, in order to estimate extreme quantiles that may not have been observed (see Coles, 2001, for an introduction). In particular, a key quantity to estimate is the return level ℓ_T associated with a given period of T years, the level that is exceeded on average once every T years. Assessing the resistance of facilities to natural disasters such as dams to floods that occur on average once every 100 years or 1 000 years is critical for companies such as *Électricité de France* (EDF). Moreover, characterizing the uncertainty on the estimation of this return level is also of interest, which encourages the choice of the Bayesian paradigm. However, performing Bayesian inference requires multiple steps that must be managed by the user, from the choice of the model to the evaluation and validation of computations. This has been recently formalized by Gelman et al. (2020) in the form of a Bayesian workflow. After introducing models stemming from extreme value theory in Section 2.1.1, we briefly review in Section 2.1.2 one particular step of the workflow, reparameterization, and more specifically the choice of an orthogonal parameterization.

2.1.1 Extreme-value models

Three different frameworks exist to model extreme events, leading to different likelihoods: one by block maxima, one by peaks-over-threshold, and one that unifies both through a Poisson process characterization.

Block maxima model Let M_n be the maximum of n i.i.d random variables with cumulative distribution function (cdf) F. We assume that F belongs to the maximum domain of attraction of a non-degenerate cdf G, meaning that there exist two sequences $a_n > 0$ and b_n such that $(M_n - b_n)/a_n$ converges in distribution to the cdf G. The extreme value theorem (e.g., Haan and Ferreira, 2006, Chapter 1) states that G is necessarily a generalized extreme-value (GEV) distribution, with cdf:

$$G(x) = \begin{cases} \exp\left(-\{1+\xi x\}_+^{-1/\xi}\right) & \text{if } \xi \neq 0\,, \\ \exp(-\exp(-x)) & \text{if } \xi = 0, \end{cases} \tag{2.1}$$

where $\{x\}_+ = \max\{0, x\}$. Consequently, for a finite value of n, one can consider the approximation $\mathbb{P}(M_n \leq x) \approx G((x-b_n)/a_n) =: G(x \mid b_n, a_n, \xi)$, and focus on the estimation of the three parameters of the GEV distribution. Here, as the dataset is fixed, the dependence in n for the location and scale parameters will be omitted. To obtain a sample of maxima, one can divide the dataset into m blocks of size n/m and extract the maximum from each of them.

Peaks-over-threshold model Alternatively, one can consider observations that exceed a high threshold u. Let X be a random variable with cdf F. Pickands theorem (Pickands, 1975) states that, if F belongs to the maximum domain of attraction of G with $\mathbb{P}(M_n \leq x) \approx G(x \mid \mu, \sigma, \xi)$, then the distribution of the exceedances $X - u \mid X > u$ is, as u

converges to the upper endpoint of F, a generalized Pareto distribution (GPD), with cdf

$$H(y \mid \tilde{\sigma}, \xi) = \begin{cases} 1 - \{1 + \xi \frac{y}{\tilde{\sigma}}\}_+^{-1/\xi} & \text{if } \xi \neq 0\,, \\ 1 - \exp\left(-\frac{y}{\tilde{\sigma}}\right) & \text{if } \xi = 0, \end{cases} \tag{2.2}$$

where the shape parameter ξ is the same as in (2.1) and the GPD and GEV scales are linked by $\tilde{\sigma} = \sigma + \xi(u - \mu)$. To obtain a sample of n_u excesses, the peaks-over-threshold method focusses on the n_u largest values of the dataset. It thus requires the estimation of the quantile of order $1 - n_u/n$, which can be seen as the third parameter to estimate, in addition to $\tilde{\sigma}$ and ξ. The most classical choice is to estimate this intermediate quantile by the $(n - n_u)$th order statistic.

Poisson process characterization of extremes Finally, these two approaches can be generalized by a third one, using a non-homogeneous Poisson process. We present here an intuitive way for obtaining this model similarly to Coles (2001, Chapter 7), and refer to (Leadbetter et al., 1983, Chapter 5) for theoretical details. We start by observing that, for large n, $F^n(x) \approx G(x \mid \mu, \sigma, \xi)$, for x in the support of G denoted by $\text{supp}(G(\cdot \mid \mu, \sigma, \xi)) = \left\{x \in \mathbb{R} \quad \text{s.t. } 1 + \xi\left(\frac{x-\mu}{\sigma}\right) > 0\right\}$. Hence, considering a large threshold $u \in \text{supp}(G(\cdot \mid \mu, \sigma, \xi))$, a Taylor expansion yields

$$n \log F(u) \simeq -n(1 - F(u)) \simeq \log G(u \mid \mu, \sigma, \xi),$$

or, equivalently,

$$\mathbb{P}\left(X > u\right) \simeq -\frac{1}{n} \log G(u \mid \mu, \sigma, \xi). \tag{2.3}$$

Equation (2.3) can be seen as the probability of X to belong to $I_u := [u, +\infty)$. In the case of n i.i.d random variables, one can deduce that the associated point process N_n is such that $N_n(I_u) \sim \mathcal{B}(n, p_n)$ with p_n given by Equation (2.3). As $n \to \infty$, the binomial distribution $\mathcal{B}(n, p_n)$ converges to the Poisson distribution $\mathcal{P}(\Lambda(I_u))$, with $\Lambda(I_u) = -\log G(u \mid \mu, \sigma, \xi)$. This property being valid for all I_u together with the independence property on non-overlapping sets imply that N_n converges to a non-homogeneous Poisson process, with intensity measure $\Lambda(I_u)$: $N_n \xrightarrow{d} N$, with $N(I_u) \sim \mathcal{P}(\Lambda(I_u))$. This model generalizes the block maxima one since

$$\mathbb{P}(M_n < x) = \mathbb{P}(N_n(I_x) = 0) \to \mathbb{P}(N(I_x) = 0) = \exp(-\Lambda(I_x)) = G(x \mid \mu, \sigma, \xi),$$

as $n \to \infty$. However, an estimation of the parameters (μ, σ, ξ) with this model is related to the overall maximum M_n of the dataset, and it is frequent to study maxima of m smaller blocks $M_{n/m}$, where m is typically the number of years in the observations and so $M_{n/m}$ corresponds to annual maxima. To do so, the intensity measure is multiplied by m, which modifies the parameterization and in particular the value of μ and σ: Wadsworth et al. (2010) shows that, if $(\mu_{k_i}, \sigma_{k_i}, \xi)$ $(i \in \{1, 2\})$, are parameters for k_i GEV observations, then

$$\mu_{k_2} = \mu_{k_1} - \frac{\sigma_{k_1}}{\xi}\left(1 - \left(\frac{k_2}{k_1}\right)^{-\xi}\right), \quad \sigma_{k_2} = \sigma_{k_1}\left(\frac{k_2}{k_1}\right)^{-\xi}. \tag{2.4}$$

The threshold excess model can also be derived from the point process representation, since $\mathbb{P}(X > y + u \mid X > u) \simeq 1 - H(y \mid \tilde{\sigma}, \xi)$, with $\tilde{\sigma} = \sigma + \xi(u - \mu)$. Moreover, in contrast to the peaks-over-threshold model where an intermediate quantile needs to be estimated, the Poisson model directly includes a third location parameter μ.

In the following, we will focus mainly on this latter model, and treat the peaks-over-threshold method as a special case in Section 2.4.1.

Bayesian inference Using the Bayesian paradigm in extreme value models is advantageous in comparison to the frequentist approach, see Coles and Powell (1996) for a general review, and Stephenson (2016) or Bousquet (2021) for more recent overviews. For the Poisson process characterization of extremes, Bayesian inference consists in fixing a scaling factor m and a threshold u to get a number of $n_u \geq 1$ observations exceeding u denoted by $\boldsymbol{x} = (x_1, \ldots, x_{n_u})$. The likelihood of these observations can be written as

$$L(\boldsymbol{x}, n_u \mid \mu, \sigma, \xi) = \mathrm{e}^{-m\left(1+\xi\left(\frac{u-\mu}{\sigma}\right)\right)^{-1/\xi}} \sigma^{-n_u} \prod_{i=1}^{n_u} \left(1 + \xi\left(\frac{x_i - \mu}{\sigma}\right)\right)^{-1-1/\xi}. \tag{2.5}$$

A complete Bayesian model requires also the specification of a prior $p(\mu, \sigma, \xi)$, to obtain the posterior $p(\mu, \sigma, \xi \mid \boldsymbol{x}, n_u)$ using Bayes' theorem, $p(\mu, \sigma, \xi \mid \boldsymbol{x}, n_u) \propto p(\mu, \sigma, \xi) L(\boldsymbol{x}, n_u \mid \mu, \sigma, \xi)$. This posterior summarizes the information on the parameters after observations, and can be used to extract point estimators, build credible intervals, or write the probability of a new observation $\tilde{x}$ given data $\boldsymbol{x}$ using the posterior predictive:

$$p(\tilde{x} \mid \boldsymbol{x}, n_u) = \int p(\tilde{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{x}, n_u) \mathrm{d}\boldsymbol{\theta}, \qquad \boldsymbol{\theta} = (\mu, \sigma, \xi). \tag{2.6}$$

These quantities of interest are rarely explicit, and are often derived by sampling approaches. A recent survey of extreme value softwares (Belzile et al., 2022) contains a Bayesian section, and a comparison with frequentist methods. In the general Bayesian case, an overview of the Bayesian workflow is given in Gelman et al. (2020), and we focus here on the particular step of reparameterization for the likelihood $L(\boldsymbol{x}, n_u \mid \mu, \sigma, \xi)$ in the case where Markov chain Monte Carlo (MCMC) methods are used to approximate the posterior distribution.

2.1.2 Reparameterization

Although the choice of parameterization of a statistical model does not alter the model *per se*, it does reshape its geometry, which in turn may impact computational aspects of sampling algorithms such as efficiency or accuracy. For these methods, a crucial complication for chain convergence is parameter correlation. This notion of correlation between parameters can be associated with a notion of asymptotic orthogonality, leading to independence of posterior components.

Parameterization and Bayesian inference It has been known for several decades that parameterization is crucial for good mixing of MCMC chains, especially when the correlation between the coordinates is large. See Gilks et al. (1995, Chapter 6) for a great introduction for Gibbs sampling and Metropolis–Hastings algorithm. More general computations are conducted by Roberts and Sahu (1997) in the normal case, but this convergence rate is less explicit in the general case, see for example Roberts and Polson (1994). For Metropolis–Hastings, if the structure of the kernel is not similar to the one of the target density (which is a typical case if there is a complex dependence between parameters), then too many candidates generated by the kernel are rejected and the same problem as for Gibbs sampling occurs. For more recent MCMC algorithms such as Hamiltonian Monte Carlo (HMC, Neal, 1996) and its variant NUTS (Hoffman and Gelman, 2014), Betancourt and Girolami (2015) gives an example of the benefit of reparameterization for hierarchical models. More generally, Betancourt (2019) studies reparameterization

from a geometric perspective, in order to show its equivalence with adapted versions of HMC on Riemannian manifolds.

Due to the difficulty of obtaining general results on reparameterization and MCMC convergence, a significant part of the research focuses on specific models, such as hierarchical models (Papaspiliopoulos et al., 2003, Browne et al., 2009), linear regression (Gilks et al., 1995), or mixed models (Gelfand et al., 1995, 1996).

For extreme value models, Diebolt et al. (2005) uses a continuous mixture of exponential distributions in the GPD case. Opitz et al. (2018) also suggests to use the median instead of the usual scale parameter to reduce correlation for Integrated Nested Laplace approximation (INLA). An alternative Monte Carlo algorithm, the ratio-of-uniforms method, is also implemented for extreme value models in the `revdbayes` package (Northrop, 2022a). The influence of parameterization is also considered in this framework as the acceptance rate can be altered because of correlated parameters (see Appendix 2.C.4). Parameter transformations are also studied in order to make likelihood-based inference suitable in the high-dimensional case in Jóhannesson et al. (2022). Finally, Belzile et al. (2022) proposes a reparameterization trick that can be used to obtain a suitable initial value for optimization routines.

Orthogonal parameterization As seen before, reducing dependence between coordinates is desirable for MCMC methods. Dependence can be characterized using asymptotic covariance and the notion of orthogonality according to Jeffreys (1939): parameters are said to be orthogonal when the Fisher information is diagonal. From this definition, having orthogonal parameters leads to asymptotic posterior independence when a Bernstein–von Mises theorem holds (e.g., Van der Vaart, 2000, Chapter 10). However, the problem of finding an orthogonal parameterization is seldom feasible when there are more than three parameters, since the number of equations is then greater than the number of unknown variables. In the case of three parameters, there are as many equations as there are unknowns, but the non linear system does not necessarily lead to a solution (Huzurbazar, 1950).

The main use of orthogonal parameterization is to make parameters of interest independent of nuisance parameters (Cox and Reid, 1987). Other definitions of orthogonality are also proposed to be more adapted to the inferential context (Tibshirani and Wasserman, 1994) or to ensure consistency of the parameter of interest (Woutersen, 2011). For Bayesian inference, Tibshirani and Wasserman (1994) compares different definitions and suggests a strong assumption of normality for the posterior. In the following, we keep the most popular definition of orthogonality due to Jeffreys (1939), as we are not interested in properties associated with the estimation of a given parameter of interest, but rather on the dependence structure between parameters. However, up to our knowledge, there is no clear evidence in the literature of a direct link between parameter orthogonality and mixing properties of the corresponding MCMC chains, such as a better convergence rate. In Section 2.4, we bring some empirical evidence on the interest of orthogonality in extreme value models.

2.1.3 Contributions and outline

In this paper, we study the benefits of reparameterization for the Poisson process characterization of extremes in a Bayesian context. In particular, it is shown that the orthogonal

parameterization is useful for several reasons: we argue in Section 2.2 that it improves the performance of MCMC algorithms in terms of convergence, and we show in Section 2.3 that it also facilitates the derivation of priors such as Jeffreys and an informative variant on the shape parameter using penalized complexity (PC) priors (Simpson et al., 2017). These results are then illustrated by experiments in Section 2.4, first on simulations to compare the different parameterizations, and second on a real dataset of the Garonne river flow. Proofs as well as additional experiments are provided in the Appendix, and the code corresponding to the experiments is available online.[1]

2.2 Reaching orthogonality for extreme Poisson process

An attempt to reparametrize the Poisson process for extremes in order to improve MCMC convergence already exists in the literature (Sharkey and Tawn, 2017), but has several limitations that are detailed here. Instead, we suggest to use the fully orthogonal parameterization of Chavez-Demoulin and Davison (2005).

Near-orthogonality with hyperparameter tuning Based on the relationship between parameters given in Equation (2.4), Sharkey and Tawn (2017) suggests to change the scaling factor m before using Metropolis–Hastings algorithm in order to optimize MCMC convergence. To this aim, they minimize the non-diagonal elements of the inverse Fisher information matrix corresponding to asymptotic covariances and then retrieved the parameters corresponding to the initial number of blocks from Equation (2.4). As the calculations cannot be achieved explicitly, the authors found empirically that the values m_1 and m_2 that cancel respectively the asymptotic covariances $\mathrm{ACov}(\mu, \sigma)$ and $\mathrm{ACov}(\sigma, \xi)$ are such that any $m \in [m_1, m_2]$ improves the MCMC convergence. Approximations of m_1 and m_2 are then given as functions of ξ, and therefore a preliminary estimation of ξ (typically obtained using maximum likelihood estimation) is required to obtain $\hat{m}_1(\xi)$ and $\hat{m}_2(\xi)$, and to choose a value in this interval before running an MCMC with the right choice of m. Despite leading to significant improvement of the convergence of Markov Chains, this method suffers from several limitations. First, preliminary estimation of the shape parameter ξ is required, to compute $\hat{m}_1(\xi)$ and $\hat{m}_2(\xi)$ and choose a value in the corresponding interval, which adds complexity and computational burden on the overall framework. Moreover, it also affects the accuracy of orthogonalization, as the expressions of m_1 and m_2 are found empirically, then are approximated by $\hat{m}_1(\xi)$ and $\hat{m}_2(\xi)$, and finally computed at $\hat{\xi}$ which adds a new source of uncertainty. One way to lighten the method would be to suggest a simpler choice of m, for example $m = n_u$, which leads to a satisfactory behaviour as noticed by Wadsworth et al. (2010). However, it is shown in Appendix 2.A that this choice presents some flaws and does not bring any general guarantee of orthogonality.

Orthogonal parameterization Alternatively, there exists a parameterization of the Poisson process that leads to orthogonality. Suggested by Chavez-Demoulin and Davison (2005), it consists of the change of variable

$$(r, \nu, \xi) = \left(m \left(1 + \xi \left(\frac{u - \mu}{\sigma} \right) \right)^{-1/\xi}, (1 + \xi)(\sigma + \xi(u - \mu)), \xi \right), \tag{2.7}$$

[1] https://github.com/TheoMoins/ExtremesPyMC

while the inverse transformation is

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

With this parameterization, the likelihood is

$$L(\boldsymbol{x}, n_u \mid r, \nu, \xi) = e^{-r} \left(\frac{r}{m} \right)^{n_u} \left(\frac{\nu}{1+\xi} \right)^{-n_u} \prod_{i=1}^{n_u} \left(1 + \frac{\xi(1+\xi)}{\nu} (x_i - u) \right)^{-1-1/\xi}. \tag{2.8}$$

Under this form, we can directly see that r is orthogonal to ν and ξ, as the likelihood factorizes with respect to r and (ν, ξ). Parameter $r \geq 0$ represents the intensity of the Poisson process, which is the expected number of exceedances, while the two other ones can be seen as an orthogonal parameterization of the GPD with scale $\tilde{\sigma}_u = \sigma + \xi(u - \mu)$ and shape ξ. Under this parameterization and provided $\xi > -1/2$, the Fisher information matrix $\mathcal{I}(r, \nu, \xi)$ is

$$\mathcal{I}(r, \nu, \xi) = \operatorname{diag}\left(\frac{1}{r}, \frac{r}{\nu^2(1+2\xi)}, \frac{r}{(1+\xi)^2} \right), \tag{2.9}$$

where $\operatorname{diag}(\boldsymbol{u})$ denotes the diagonal matrix with diagonal equal to vector $\boldsymbol{u}$. Calculations are provided in Appendix 2.B. Therefore, the orthogonal parameterization of Chavez-Demoulin and Davison (2005) is more adapted than the tuning of m since it directly yields the optimal solution sought by Sharkey and Tawn (2017). Moreover, it is obtained without recourse to any optimization procedure or approximation. Finally, by plugging (r, ν) into Equation (2.4), one can show that the invariance property with respect to m holds for the three parameters, and so the parameterization is independent of the choice of m.

Generalisation to covariates The inclusion of covariates in a model holds both theoretical and practical significance, as it enables the incorporation of factors such as temporal trends. A notable advantage of this approach is that if the parameters are orthogonal and each of them depends on distinct parameters, then these parameters will also be orthogonal to one another. To elaborate further, let $\boldsymbol{C}_i$ represent the set of covariates associated with the observation x_i. In the most comprehensive scenario, where all relevant covariates are considered, the model can be expressed as follows:

$$\begin{cases} r_i = f_r(\boldsymbol{\theta}_r, \boldsymbol{C}_i), \\ \nu_i = f_\nu(\boldsymbol{\theta}_\nu, \boldsymbol{C}_i), \\ \xi_i = f_\xi(\boldsymbol{\theta}_\xi, \boldsymbol{C}_i), \end{cases}$$

where $\boldsymbol{\theta}_r$, $\boldsymbol{\theta}_\nu$ and $\boldsymbol{\theta}_\xi$ are three vectors of parameters. The log-likelihood can be written as

$$\sum_{i=1}^{n} \ell_i(r_i, \nu_i, \xi_i) = \sum_{i=1}^{n} \ell_i(f_r(\boldsymbol{\theta}_r, \boldsymbol{C}_i), f_\nu(\boldsymbol{\theta}_\nu, \boldsymbol{C}_i), f_\xi(\boldsymbol{\theta}_\xi, \boldsymbol{C}_i)).$$

By leveraging the property that no parameters are shared, along with the chain rule, one can derive the following expression for the Fisher information associated with the jth coordinate $\theta_{\nu,j}$ of $\boldsymbol{\theta}_\nu$ and the kth coordinate $\theta_{\xi,k}$ of $\boldsymbol{\theta}_\xi$:

$$\mathbb{E}\left(-\frac{\partial^2 \ell_i}{\partial \theta_{\nu,j} \partial \theta_{\xi,k}} \right) = \frac{\partial f_\nu}{\partial \theta_{\nu,j}} \frac{\partial f_\xi}{\partial \theta_{\xi,k}} \mathbb{E}\left(-\frac{\partial^2 \ell_i}{\partial \nu_i \partial \xi_i} \right) = 0.$$

As a result, every parameter in $\boldsymbol{\theta}_\nu$ is orthogonal to all parameters in $\boldsymbol{\theta}_\xi$. Similar calculations reveal that $\boldsymbol{\theta}_r$ is orthogonal to both $\boldsymbol{\theta}_\nu$ and $\boldsymbol{\theta}_\xi$. Hence, when covariates are defined based on orthogonal parameters, it leads to block-wise orthogonality, with the associated Fisher information matrix comprising three blocks in this case.

2.3 Priors invariant to reparameterization

In the case where no external information is available about the parameters, the choice of the prior distribution should be made with caution. Typically, the term "uninformative prior" or "objective prior" can be misleading, as it refers to priors used when one does not have preliminary information, but the prior itself does contain information. As an example, a flat prior over the range of possible values is only flat for a given parameterization. After a change of variable, a uniform prior does not necessarily remain uniform (e.g., Robert, 2007, Chapter 3). This problem is all the more serious as our study which deals with reparametrization: here, we derive two priors that enjoy the property of being invariant with respect to reparameterization.

2.3.1 Jeffreys prior

Jeffreys prior (Jeffreys, 1946) is built with the aim of invariance: if $\mathcal{I}(\boldsymbol{\theta})$ denotes the Fisher information matrix associated with parameters $\boldsymbol{\theta}$, it is defined as

$$p_{\text{J}}(\boldsymbol{\theta}) \propto \sqrt{\det \mathcal{I}(\boldsymbol{\theta})}. \tag{2.10}$$

Under this prior, one can show that a reparameterization $\boldsymbol{\phi} = h(\boldsymbol{\theta})$ with h a continuously differentiable function yields $p_{\text{J}}(\boldsymbol{\phi}) \propto \sqrt{\det \mathcal{I}(\boldsymbol{\phi})}$. This prior is computed for the GPD by Castellanos and Cabras (2007) and for the GEV under a modified version where $p_{\text{J}}(\mu, \sigma, \xi) \propto \sqrt{\det \mathcal{I}(\sigma, \xi)}$ by Kotz and Nadarajah (2000). Up to our knowledge, Jeffreys prior has never been computed for the Poisson process characterization of extremes. The orthogonalization done in Equation (2.7) directly provides Jeffreys prior with respect to (r, ν, ξ):

Proposition 1. *Jeffreys prior associated with a Poisson process for extremes with parameters (r, ν, ξ) from Equation (2.7) exists provided $\xi > -1/2$, and is given by*

$$p_{\text{J}}(r, \nu, \xi) \propto \frac{r^{1/2}}{\nu(1+\xi)(1+2\xi)^{1/2}}. \tag{2.11}$$

Moreover, the invariance to reparameterization property directly provides the expression of Jeffreys prior on (μ, σ, ξ).

Corollary 1. *Jeffreys prior associated with a Poisson process for extremes with original parameters (μ, σ, ξ) exists provided $\xi > -1/2$, and can be written as*

$$p_J(\mu, \sigma, \xi) \propto \frac{\left(1 + \xi\left(\frac{u-\mu}{\sigma}\right)\right)^{-\frac{3}{2\xi}-1}}{\sigma^2(1+\xi)(1+2\xi)^{1/2}}. \tag{2.12}$$

This prior cannot be defined for $\xi \leq -1/2$, as it corresponds to a case where the Fisher information matrix is infinite. However, this assumption is not too restrictive as

the great majority of models of interest belong to a maximum domain of attraction with $\xi \in (-1/2, 1/2)$. Note that this prior, similarly to the uniform one, is improper in the sense that the integral over the range of parameters is infinite. Consequently, it is necessary to check whether the posterior is proper or not to be able to use it. Castellanos and Cabras (2007) shows that the posterior is proper when using Jeffreys prior in the GPD case, while Northrop and Attalides (2016) shows that it is never the case with GEV likelihood. For the Poisson process, we show the following result:

Proposition 2. *Jeffreys prior for a Poisson process for extremes yields a proper posterior distribution, as long as* $\xi > -1/2$.

A proof is provided in Appendix 2.B.

2.3.2 Penalized complexity prior for the shape parameter

The shape parameter ξ plays a crucial role in the inference, as it tunes the heaviness of the tail distribution: it is heavy if $\xi > 0$, light if $\xi = 0$ and finite (*i.e.* with a finite right end-point) if $\xi < 0$. The case $\xi = 0$ can be seen as a simpler model with an exponential decrease of the survival function, where the GPD cdf in Equation (2.2) simplifies to an exponential distribution. This concentration of an entire maximum domain of attraction at a single value of ξ complicates the study, since it is for example difficult to distinguish heavy tails with low ξ from light tails (Stephenson and Tawn, 2004). However, this change of regime can have significant consequences when it comes to extrapolation. It should also be noted that a vast majority of datasets have distribution with $|\xi| \leq 1/2$. It is therefore natural, even in a non-informative framework, to penalize high values of $|\xi|$. One way to do this is to use penalized complexity (PC) priors (Simpson et al., 2017): the idea is to consider a prior that penalizes exponentially the distance between a model $p_\xi := p(\cdot \mid \xi)$ with a given ξ and the baseline p_0 with $\xi = 0$. The general formula is

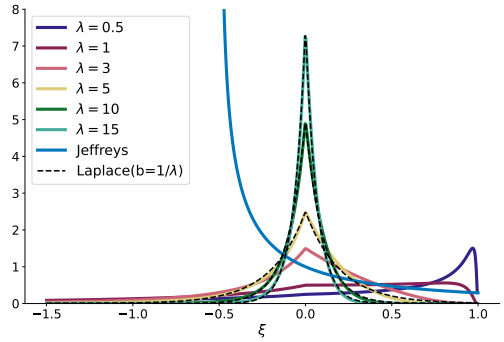
$$p_{\mathrm{PC}}(\xi \mid \lambda) = \lambda \exp(-\lambda d(\xi)) \left| \frac{\partial d(\xi)}{\partial \xi} \right|,$$

with $\lambda > 0$, $d(\xi) = \sqrt{2\mathrm{KL}\left(p_\xi || p_0\right)}$ and $\mathrm{KL}(p_\xi || p_0)$ the Kullback–Leibler divergence between p_ξ and p_0: $\mathrm{KL}(p_\xi || p_0) = \int p_\xi(x) \log\left(p_\xi(x)/p_0(x)\right) \mathrm{d}x$. Parameter λ acts as a scaling parameter and controls the range of acceptable values for ξ. This prior has the advantage of being proper and invariant to reparameterization on ξ. The computation with GPD has already been done by Opitz et al. (2018) for the case $\xi \geq 0$: the authors prove that $d(\xi)$ is finite only if $\xi < 1$, and is $d(\xi) = \sqrt{2}\xi/\sqrt{1-\xi}$ for $0 \leq \xi < 1$. Then, they show that it can be approximated by an exponential distribution on ξ in the case $\xi \to 0$, when λ can be taken sufficiently large and to favor $\xi = 0$. A first observation is that routine calculations extend this definition both to negative values of ξ, and to the Poisson process characterization where the density of observation is also GPD.

Proposition 3. *The PC prior associated with a Poisson process for extremes exists for any* $\xi < 1$ *and is*

$$p_{\mathrm{PC}}(\xi \mid \lambda) = \frac{\lambda}{2} \left(\frac{1 - \xi/2}{(1-\xi)^{3/2}} \right) \exp\left(-\lambda \frac{|\xi|}{\sqrt{1-\xi}} \right). \tag{2.13}$$

This prior is plotted for several values of λ in Figure 2.1. As observed by Opitz et al. (2018), the PC prior is very similar to a Laplace$(0, 1/\lambda)$ when λ is large enough so that



λ	$\mathrm{I}_{\mathrm{PC}}(0.95 \mid \lambda)$	$\mathrm{I}_{\mathrm{Lap}}(0.95 \mid \lambda)$
0.5	$[-36.8, 0.97]$	$[-6.00, 6.00]$
1	$[-9.88, 0.90]$	$[-2.99, 2.99]$
3	$[-1.61, 0.61]$	$[-0.99, 0.99]$
5	$[-0.80, 0.44]$	$[-0.59, 0.59]$
10	$[-0.34, 0.25]$	$[-0.29, 0.29]$
15	$[-0.22, 0.18]$	$[-0.19, 0.19]$

Figure 2.1: Left panel: examples of PC priors $p_{\mathrm{PC}}(\cdot \mid \lambda)$ with λ ranging from 0.5 to 15, and Jeffreys prior (blue curve) represented for fixed values of $(\mu, \sigma) = (0, 1)$. The black dashed lines represent Laplace distributions with scale parameter equal to $1/\lambda$, for $\lambda \in \{5, 10, 15\}$. Note that Laplace distributions $p_{\mathcal{L}}(\cdot \mid 1/\lambda)$ approximate well $p_{\mathrm{PC}}(\cdot \mid \lambda)$ when $\lambda \geq 10$. Right panel: credible intervals at 95% for PC and Laplace priors, resp. $\mathrm{I}_{\mathrm{PC}}(0.95 \mid \lambda)$ and $\mathrm{I}_{\mathrm{Lap}}(0.95 \mid \lambda)$.

the peaks at 0 dominates over the endpoint at 1. In the case where λ is small (typically $\lambda \leq 1$), an asymptote appears at the upper bound $\xi = 1$ which could have an undesirable influence in the posterior distribution. Thus, for the least informative case, a value of $\lambda = 1$ is sufficiently small as it does not favor values close to 0 nor those close to 1. In the case when 0 is favoured with a high λ and the true value of ξ differs from 0, the estimation may be altered compared to the uninformative case: see Appendix 2.C.5 for an analysis on simulated data. For the two other parameters, one can consider Jeffreys' rule on (r, ν) in order to obtain a non-informative prior for (r, ν) while keeping invariance to reparameterization property. ξ is therefore considered *a priori* independent of (r, ν). In view of the Fisher information matrix in Equation (2.9), we obtain $p_{\mathrm{J}}(r, \nu) \propto 1/\nu$. Similarly to Jeffreys prior in Section 2.3.1, the resulting prior is improper but the following proposition can be shown:

Proposition 4. *The prior defined as* $p(r, \nu, \xi) \propto p_{\mathrm{PC}}(\xi) p_{\mathrm{J}}(r, \nu) \propto p_{\mathrm{PC}}(\xi)/\nu$ *for the Poisson process for extremes yields a proper posterior distribution.*

The proof, detailed in Appendix 2.B, relies on a result of Northrop and Attalides (2016). Note that this result still holds if $p_{\mathrm{PC}}(\xi)$ is replaced by its approximation through the Laplace distribution.

2.4 Experiments

The benefits of the orthogonal reparameterization in the Poisson process model are illustrated here on simulations and a real environmental dataset. Appendix 2.C contains additional experiments, notably using Hamiltonian Monte Carlo (HMC) instead of MCMC (Appendix 2.C.1), under various maximum domains of attraction (Appendix 2.C.2), in other models than the Poisson process model, namely the GPD and GEV ones (Appendix 2.C.3), using the ratio-of-uniforms algorithm instead of MCMC (Appendix 2.C.4) and finally with replications and comparison with maximum likelihood (Appendix 2.C.5). All experiments are done using PyMC3 library (Salvatier et al., 2016), and the corresponding code is available online[1].

2.4.1 Simulations with the Poisson process model

Data generation We start by comparing the different parameterizations on exceedances generated with the Poisson process model described in Section 2.1.1. For a given value of (μ, σ, ξ) and hyperparameters (u, m), the data generation proceeds in two steps: first, a number of events n_u is simulated using a Poisson distribution with parameter $\Lambda(I_u)$ as defined in Section 2.1.1. Then, for each point $i \in \{1, \dots, n_u\}$, the position x_i knowing that $x_i \in I_u$ is sampled from a GPD with parameters $(u, \tilde{\sigma}, \xi)$, with $\tilde{\sigma} = \sigma + \xi(u - \mu)$. An example with $(m, u, \mu, \sigma, \xi) = (40, 30, 50, 15, -0.25)$ is detailed here, leading to an expected number of observations $\Lambda(I_u) \approx 126$.

Experimental setup For MCMC hyper-parameters such as number of chains, burn-in period per chain or initialization, we keep the default values suggested in the PyMC3 library: four chains (which corresponds to our number of cores) with 1 000 iterations each, and a burn-in period of 1 000 iterations. In addition to these choices, this library offers the possibility to choose among different sampling methods, such as the traditional Metropolis–Hastings algorithm, but also more modern MCMC algorithms like Hamiltonian Monte Carlo (HMC, Neal, 1996), or the No-U-Turn sampler (NUTS, Hoffman and Gelman, 2014) which is the default choice in PyMC3. We choose to compare the different reparameterizations on Metropolis–Hastings draws (after burn-in), and the behaviour on NUTS is also investigated (Appendix 2.C.1). We show that 1 000 iterations are sufficient for the chains to converge when the parameterization is well chosen. However, note that the algorithm only takes a few seconds to run, so this number of iterations can easily be increased for real data applications, as done in Section 2.4.2. Finally, Jeffreys prior (computed in Section 2.3.1) is choosen for all configurations, but experiments have shown similar results with the PC prior of Section 2.3.2.

Convergence diagnostic Our aim is to discriminate the different parameterizations according to the rate of convergence of the MCMC chains to their target. Different indicators exist to assess the quality of MCMC approximation. First, given a finite number of samples, autocorrelation plots as functions of lag measure how good the posterior approximation is, as the dependence between the chain elements reduces the effective information available for inference. To measure this, a common practice relies on the effective sample size, defined as $\text{ESS} = MN(1 + 2\sum_{t=1}^{\infty} \rho_t)^{-1}$, with M the number of chains of size N, and ρ_t the autocorrelation at lag t. The ESS corresponds to an equivalent number of independent draws, and so quantifies the amount of effective data for estimation (e.g., Gelman et al., 2013, Section 11.5). Here, the evolution of ESS with the number of draws is reported for each configuration. To complete the diagnostic, the potential scale reduction factor (commonly denoted by $\hat{R}$) also aims at bringing an indication about the state of convergence by computing the ratio of two estimators of the posterior variance. Generally $\hat{R} \geq 1$, and if it is greater than a given threshold, a convergence issue is raised. We use here a refinement of $\hat{R}$ named $\hat{R}_\infty$ (Moins et al., 2023), based on a local version $\hat{R}(x)$ which aims at ensuring the convergence at a given quantile x of the distribution. Then, $\hat{R}_\infty$ is defined as the supremum of the $\hat{R}(x)$ values: $\hat{R}_\infty := \sup_{x \in \mathbb{R}} \hat{R}(x)$. This scalar summary amounts to considering the value of $\hat{R}(x)$ associated with the worse quantile approximation by the MCMC chains.

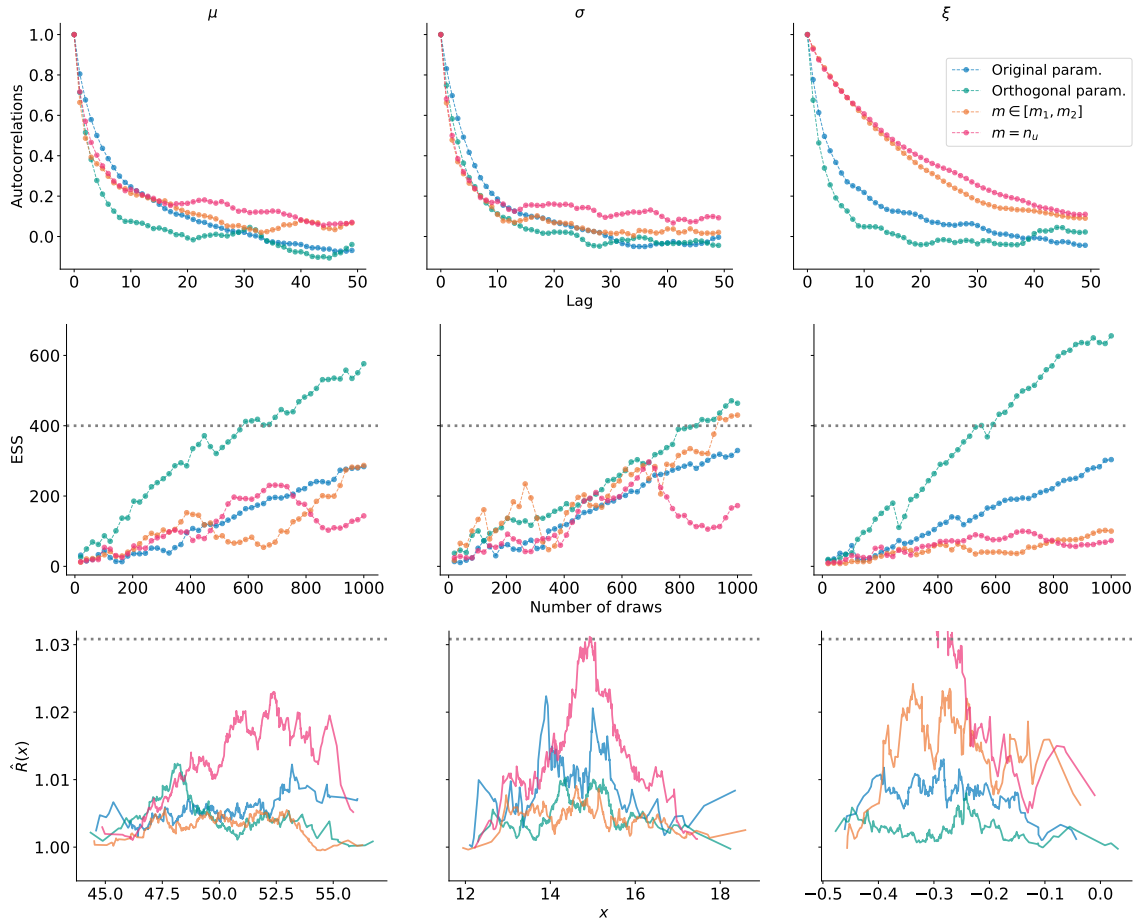


Figure 2.2: Convergence diagnostic plots for Poisson parameters (μ, σ, ξ) with $\xi < 0$, after 1 000 Metropolis–Hastings draws and a burn-in of 1 000, for four different parameterizations: the original one (in blue), the Sharkey and Tawn (2017) update with $m \in [\hat{m}_1, \hat{m}_2]$ (in orange), the Wadsworth et al. (2010) update with $m = n_u$ (in magenta), and the orthogonal parameterization (in green). Top row: autocorrelations as functions of the lag. Second row: evolution of ESS with the number of draws (the gray line corresponds to value of 400 recommended in Gelman et al. (2013)). Bottom row: $\hat{R}(x)$ as a function of the quantile x, with the adapted threshold of 1.03 (Moins et al., 2023). Some curves are truncated for visibility purposes, as they are taking much larger values than the threshold.

Results Results are reported in Figure 2.2, with four parameterizations that are compared for MCMC efficiency. (i) The orthogonal parameterization (r, ν, ξ) of Equation (2.7), and three triplets (μ, σ, ξ) associated with the following choices of m: (ii) the original m (same as the one used for generation), (iii) $m = n_u$ which is the choice of Wadsworth et al. (2010) and the package `revdbayes` (Northrop, 2022a), and (iv) $m \in [m_1, m_2]$ as suggested by Sharkey and Tawn (2017) (see Section 2.2).

In order to compare the same quantities, all convergence diagnostics are computed with the original parameterization (μ, σ, ξ) and m, consequently after a transformation of the chains for the other parameterizations. Figure 2.2 confirms that the orthogonal parameterization behaves best in the case $\xi < 0$: the Markov chains have the lowest autocorrelations, the lowest value of $\hat{R}(x)$ for almost all x, and this parameterization is the only one that satisfies the recommendation of ESS lower than 400 for estimation (Gelman et al., 2013). Conversely, the two parameterizations that suggest a change for m seem to suffer from a lack of convergence, even more than the original parameterization. For the cases $\xi > 0$ and $\xi = 0$ detailed in Appendix 2.C.2, the orthogonal parameterization is still best, but the behaviour of the three other parameterizations is reversed: the one with no change for m is the one with the largest convergence issues. Some intuitions about the behaviour of parameterizations that rely on changing m, in particular in the case $\xi < 0$, can be found in Appendix 2.A. We also refer to Appendix 2.C.3 for a study of the GPD and GEV cases. As a conclusion, the orthogonal parameterization is effective in the three maximum domains of attraction, for both Poisson process and GPD models.

2.4.2 Case study on river flow data

We apply our framework to 36 160 daily measurements of the Garonne river flow (France), from 1915 to 2013.

Preprocessing Before selecting a threshold and running a MCMC algorithm, some common preprocessing steps on daily environmental data are required: first because of seasonality, only the rainy season from December to May is considered, which reduces the number of observations to 18 043. The observations are also not independent; an autocorrelation plot suggests a three-day correlation in measurements. Therefore, clusters of exceedances of parameter $r = 3$ days are considered here, which means that two exceedances that occurred in less than three days are merged as one observation (the largest one in the cluster). Previous EDF studies (see for instance Albert, 2018) agree with traditional threshold elicitation methods (e.g., Coles, 2001, Chapter 4) to consider a threshold of $u = 2\,000$ m^3/s for estimation. In the end, $n_u = 182$ clusters of exceedances are obtained and represented in Figure 2.3. A temporal trend could be suspected there. A possible way to model such a phenomenon would be to include covariates in the orthogonal parameters, see the last paragraph of Section 2.2.

Return level estimation We are interested in estimating the T-year return level ℓ_T, which is exceeded on average once every T years. This is obtained by solving the equation $G(\ell_T \mid \mu, \sigma, \xi) = 1 - 1/T$, with G the GEV cdf defined in Equation (2.1):

$$\ell_T = \mu - \frac{\sigma}{\xi}\left(1 - (-\log(1 - 1/T))^{-\xi}\right). \tag{2.14}$$

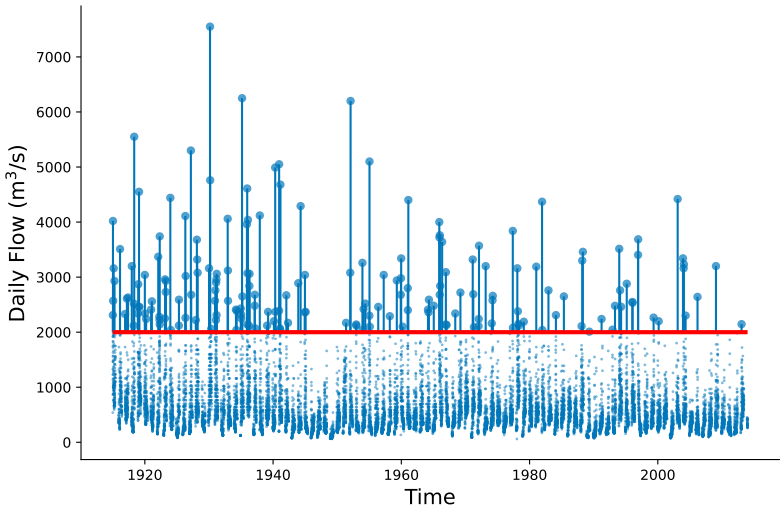


Figure 2.3: Plot of $n_u = 182$ exceedances of the Garonne river flow between 1915 and 2013 above the threshold $u = 2\,000$ (represented in red).

	Post. Mean	Post. SD	95%-CI	ESS	$\hat{R}_\infty$
μ	2 560.8	84.1	[2 409.8, 2 724.1]	3 473	≈ 1.0
σ	919.6	73.2	[787.2, 1 063.3]	2 709	≈ 1.0
ξ	0.015	0.077	[−0.120, 0.164]	2 702	≈ 1.0

Table 2.1: Posterior summaries (mean, standard deviation (SD), credible interval (CI) at 95%) and convergence diagnostics (ESS and $\hat{R}_\infty$) for (μ, σ, ξ) associated with annual maxima ($m = 99$).

Here, as the data span 99 years, we fix $m = 99$ in order to obtain parameters associated with annual maxima. The same setup as in Section 2.4.1 is then run with 5 000 draws from Metropolis–Hastings algorithm with the orthogonal parameterization. Convergence diagnostic values are reported in Figure 2.4 and show no evidence of lack of convergence, along with a very satisfactory effective sample size for estimation (final values can be found in Table 2.1 along with $\hat{R}_\infty$ for each parameter). Results of posterior summaries for (μ, σ, ξ) are reported in Table 2.1: looking at the posterior for ξ, the three maximum domains of attraction cannot be excluded, although the 95% credible interval (CI) is tight around zero. This may suggest that $\xi = 0$ and an exponential decrease of the survival function. Return levels for annual maxima are displayed in the left panel of Figure 2.5, and show that the model seems to fit the data correctly. These curves are obtained by computing the mean and 2.5%/97.5% quantiles on the posterior distribution of ℓ_T for any given return period T. This is more accurate than the version where pointwise posterior quantities of (μ, σ, ξ) are plugged in Equation (2.14) (see Jonathan et al., 2021, for a comparison). The obtained posterior mean of ℓ_T, is 6 949 m^3/s for the 100-year level and 9 266 m^3/s for the 1 000-year one. These results corroborate a study conducted in Albert et al. (2020), where the estimated value of 10 000 m^3/s for the 1 000-year return level belongs to the credible interval in Figure 2.5.

Prior influence on the return level estimation uncertainty Looking at the posterior distribution for ξ, one can reasonably make the assumption that $\xi = 0$ and therefore

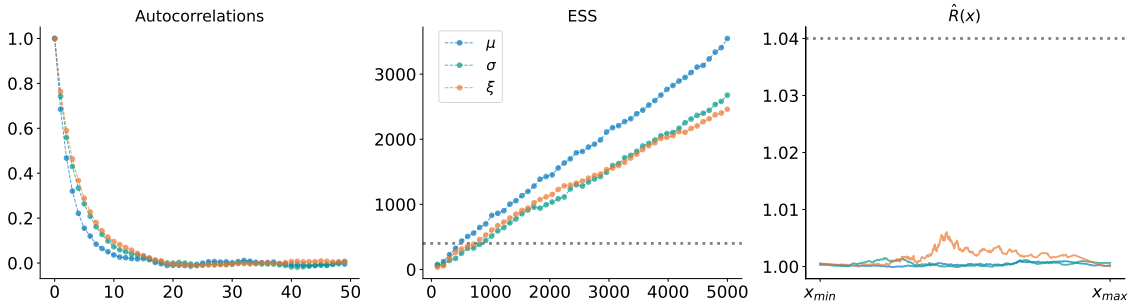


Figure 2.4: Convergence diagnostic plots for Garonne river flow data, after 5 000 Metropolis–Hastings draws and a burn-in of 1 000. Left: autocorrelations as functions of the lag. Middle: evolution of ESS with the number of draws (the gray line corresponds to value of 400 recommended in Gelman et al. (2013)). Right: $\hat{R}(x)$ as a function of the quantile x, with the adapted threshold of 1.04 (see Moins et al., 2023).

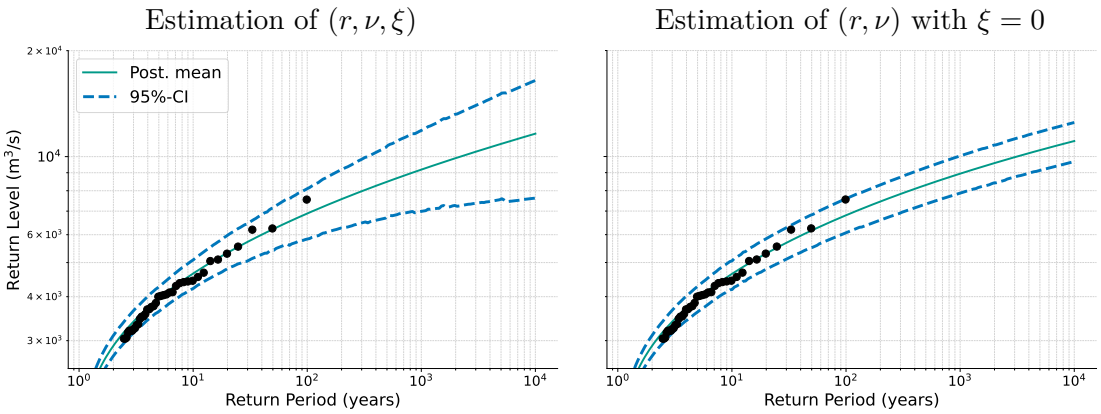


Figure 2.5: Return levels for annual maxima of Garonne flow data. Full green curves correspond to return levels obtained with posterior mean return level, and the dashed ones to the bounds of the 95% credible interval (CI). On the left, all three parameters (r, ν, ξ) are estimated, while on the right, only (r, ν) are estimated with the assumption that $\xi = 0$. Black points represent the observed annual maxima.

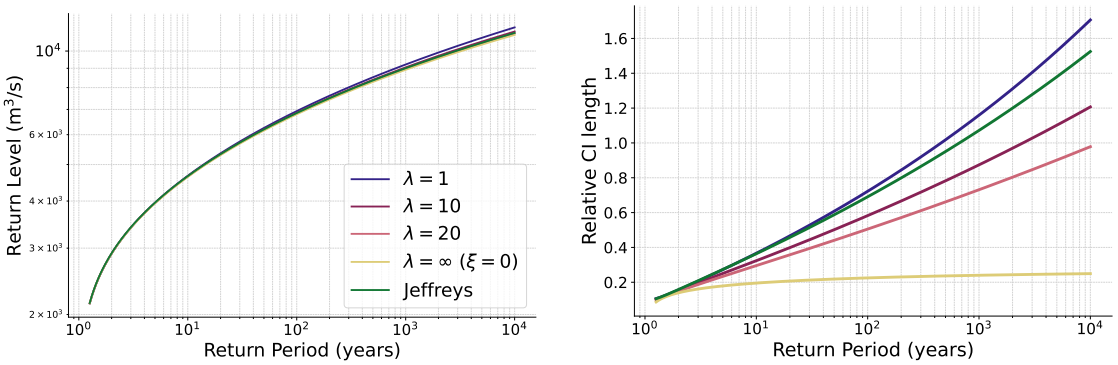


Figure 2.6: Comparison of return levels with different priors as functions of return period (log scale). On the left: return levels with posterior mean parameters. On the right: return level credible interval (CI) length relative to the point estimate (in %).

assume an exponential decrease for the survival function of the river flow. In this case, the remaining location parameter μ and scale parameter σ can be estimated with fixed $\xi = 0$. The resulting posterior summaries are very close to the ones of Table 2.1. As a result, the return level curves with posterior mean parameters (see Figure 2.5) are very similar in both cases. However, as the uncertainty on the shape parameter is excluded when fixing $\xi = 0$, the return levels credible intervals change drastically and become very concentrated around means, as shown in the right panel of Figure 2.5. In fact, this reflects that most of the uncertainty on the estimated return level is due to the estimation of the shape parameter, and so knowing its value greatly facilitates the extrapolation. PC priors allow us to navigate between these two extreme cases thanks to the hyperparameter λ. Looking at the left panel of Figure 2.6, it appears that the return level curves associated with posterior means are not affected by those differences of priors. However, the larger λ, the more information is added about the closeness of ξ to zero, and the smaller the length of the credible interval (note however that this does not give any guarantee on the estimation bias). This behaviour is illustrated on the right panel of Figure 2.6: denoting by $\ell_T^{(\mathrm{m})}$, $\ell_T^{(2.5\%)}$, and $\ell_T^{(97.5\%)}$ respectively the posterior mean, and the posterior quantiles at 2.5% and 97.5% of the return level, then the right plot in Figure 2.6 displays the length of the credible interval for the return level estimation, relatively to the estimator $\ell_T^{(\mathrm{m})}$: $(\ell_T^{(97.5\%)} - \ell_T^{(2.5\%)})/\ell_T^{(\mathrm{m})}$. This ratio is expected to grow with T, as the uncertainty increases in the tail. When $\lambda = 1$, this growth is similar to the one associated with Jeffreys prior, which can be seen as a noninformative case. For example, one can see that the size of the credible interval is already greater than the posterior estimation for the 1 000-year return level (ratio greater than one). Using $\lambda = 10$ corresponds to a confidence of 95% of having ξ between -0.3 and 0.3 with the version approximated by a Laplace distribution (see the table in Figure 2.1), reduces by approximately 20% the size of the credible interval for $T = 1\,000$. The length when ξ is fixed at zero is drastically lower than in the other cases, even those concerning PC priors with large λ values.

2.5 Conclusion

In this paper we demonstrate the benefits of using an orthogonal parameterization in the sense of Jeffreys (1939) for Bayesian inference of extreme value models. First, orthogonal

parameters facilitate the convergence of MCMC algorithms such as Metropolis–Hastings or NUTS (Section 2.2 and Appendix 2.A). This improvement is "free" in the sense that it is obtained at no extra computational cost, except a simple change of variable if one interest lies in the original parameters (μ, σ, ξ). This conclusion is confirmed by convergence diagnostics such as autocorrelation, effective sample size, and local $\hat{R}$, on simulations in the three maximum domains of attraction (Section 2.4.1 and Appendix 2.C).

Secondly, the orthogonal parameterization also facilitates the computation of Jeffreys prior (Section 2.3.1): we show that this uninformative prior is defined for $\xi > -1/2$ and is improper, but leads to a proper posterior. Posterior propriety is a necessary condition for using this prior in practice when no external information is available. However, this uninformative case is actually far from the reality of most of the applications: even without any expert information, a shape parameter in the range $(-1, 1)$ already includes a vast majority of the distributions arising in natural phenomena. Therefore as an alternative, a PC prior on ξ can be used instead and allows users to control the prior knowledge they want to include on ξ (Section 2.3.2). In particular, it penalizes the values of ξ that move away from 0, and navigate between the uninformative case and the deterministic one where $\xi = 0$. In addition to its flexibility, this prior enjoys the same advantages as Jeffreys prior: invariance to reparameterization and posterior propriety. Additionally, it can be defined without any restriction for ξ if one uses the approximation by a Laplace distribution (otherwise, $\xi < 1$). This prior information on ξ impacts the posterior uncertainty around the return level estimation. By applying our framework on river flow data (Section 2.4.2), we showed that the length of the credible interval for the return level can be significantly reduced by adding prior information of ξ. However, the uncertainty around the return level can be quantified differently, by using the quantiles of the posterior predictive distribution defined in (2.6), see Fawcett and Green (2018) for a comparison. In future work, it would be interesting to also investigate the influence of the prior on the posterior predictive return levels.

Acknowledgement

We would like to thank the anonymous reviewers and an Editor for their careful reading and for providing us with valuable comments that helped us improving the manuscript. S. Girard acknowledges the support of the Chair Stress Test, Risk Management and Financial Steering, led by the École polytechnique and its Foundation and sponsored by BNP Paribas. J. Arbel acknowledges the support of the French National Research Agency (ANR-21-JSTM-0001).

2.A Approaching orthogonality by choosing $m = n_u$

Sharkey and Tawn (2017) suggests to take a value of the scaling factor m that minimises the off-diagonal terms of the asymptotic covariance matrix (that is the inverse Fisher information matrix), denoted by $\text{ACov} := \mathcal{I}^{-1}(\mu, \sigma, \xi)$. Those terms exist only if $\xi > -1/2$ (see Proposition 2 and its proof in Appendix 2.B) and can be written as functions of

$x = -\frac{1}{\xi}\log\left\{1+\xi\left(\frac{u-\mu}{\sigma}\right)\right\}_+$, σ, and ξ as:

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

Denote by $\rho_{\cdot,\cdot}$ the asymptotic correlation between two out of the three parameters, the authors note that a range of values may also work for m between m_1 and m_2, where

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

They also find on their experiments that m_1 cancels $\rho_{\mu,\sigma}$, and that m_2 cancels $\rho_{\sigma,\xi}$. A numerical method is used in Appendix Sharkey and Tawn (2017) to approximate m_1 and m_2 as functions of ξ. Therefore, this approach requires to study the roots x_1 of $\mathrm{ACov}_{\sigma,\xi}$ and x_2 of $\mathrm{ACov}_{\mu,\sigma}$ to respectively derive $\hat{m}_1(\xi)$ and $\hat{m}_2(\xi)$. Without any approximation, we directly have $x_1 = 1/(1+\xi)$ as the unique root for $\mathrm{ACov}_{\sigma,\xi}$. Moreover, as $\xi > -1/2$, 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 $m = n_u$), and is a simple choice as it does not require any estimation of ξ. The interest of the choice $m = n_u$ has already been mentioned in Wadsworth et al. (2010) to improve the mixing property of the chain. Unfortunately, a study of function $x \mapsto \mathrm{ACov}_{\mu,\sigma}(x)$ shows that the properties of uniqueness and positivity of x_2 are only valid in the case where $\xi > 0$. In that case, the works of Wadsworth et al. (2010) and Sharkey and Tawn (2017) corroborate the choice of $m = n_u$. However, it is not the case anymore when $-1/2 < \xi < 0$. It can be shown that x_2 is not negative here, and worse, may not be unique. This can be seen as a counter-indication for frameworks that aim at reducing the three asymptotic covariances at the same time by tuning the scaling factor m.

2.B Proofs

Proof of Proposition 1 The log-likelihood l using the (r,ν,ξ) parameterization of Equation (2.7) can be written as:

$$\begin{aligned}
l(r,\nu,\xi \mid \boldsymbol{x}, n_u) = -r + n_u\log\left(\frac{r}{m}\right) - n_u\log(\nu) + n_u\log(1+\xi)\\
-\left(1+\frac{1}{\xi}\right)\sum_{i=1}^{n_u}\log\left\{1+\frac{\xi(1+\xi)}{\nu}(x_i-u)\right\}_+.
\end{aligned}$$

Under this form, one can directly see that r is orthogonal to ν and ξ. The second derivatives of l are

$$\frac{\partial^2 l}{\partial r^2} = -\frac{n_u}{r^2}, \qquad \frac{\partial^2 l}{\partial r \partial \nu} = 0, \qquad \frac{\partial^2 l}{\partial r \partial \xi} = 0,$$

$$\frac{\partial^2 l}{\partial \nu^2} = \frac{n_u}{\nu^2} + \frac{\xi(1+\xi)^3}{\nu^4} \sum_{i=1}^{n_u} \frac{(x_i - u)^2}{\left\{1 + \frac{\xi(1+\xi)}{\nu}(x_i - u)\right\}_+^2} - \frac{2(1+\xi)^2}{\nu^3} \sum_{i=1}^{n_u} \frac{(x_i - u)}{\left\{1 + \frac{\xi(1+\xi)}{\nu}(x_i - u)\right\}_+},$$

$$\frac{\partial^2 l}{\partial \nu \partial \xi} = -\frac{(1+2\xi)(1+\xi)^2}{\nu^3} \sum_{i=1}^{n_u} \frac{(x_i - u)^2}{\left\{1 + \frac{\xi(1+\xi)}{\nu}(x_i - u)\right\}_+^2} + \frac{2(1+\xi)}{\nu^2} \sum_{i=1}^{n_u} \frac{(x_i - u)}{\left\{1 + \frac{\xi(1+\xi)}{\nu}(x_i - u)\right\}_+},$$

$$\begin{aligned}\frac{\partial^2 l}{\partial \xi^2} = &-\frac{n_u}{(1+\xi)^2} + \frac{(1+2\xi)^2(1+\xi)}{\xi \nu^2} \sum_{i=1}^{n_u} \frac{(x_i - u)^2}{\left\{1 + \frac{\xi(1+\xi)}{\nu}(x_i - u)\right\}_+^2} \\ &+ \frac{2(1+\xi-\xi^2)}{\xi^2 \nu} \sum_{i=1}^{n_u} \frac{(x_i - u)}{\left\{1 + \frac{\xi(1+\xi)}{\nu}(x_i - u)\right\}_+} - \frac{2}{\xi^3} \sum_{i=1}^{n_u} \log\left\{1 + \frac{\xi(1+\xi)}{\nu}(x_i - u)\right\}_+ .\end{aligned}$$

Focussing on the expectations, as we observe a Poisson process, the information is contained in the number n_u of observed points (we write N_u the corresponding random variable) and the position of jumping events x_i (we write X_i the corresponding random variable, with the same distribution as X). Here, N_u is distributed according to a Poisson distribution with parameter r, and $X - u$ is a GPD random variable with parameters $(\frac{\nu}{1+\xi}, \xi)$. For example, deriving the following expectations is the cornerstone to obtain the Fisher information matrix:

$$\begin{aligned}\mathbb{E}_{N_u,X}\left[\sum_{i=1}^{N_u} \frac{(X_i - u)^2}{\left\{1 + \frac{\xi(1+\xi)}{\nu}(X_i - u)\right\}_+^2}\right] &= \mathbb{E}_{N_u}\left[\mathbb{E}_{X|N_u}\left[\sum_{i=1}^{N_u} \frac{(X_i - u)^2}{\left\{1 + \frac{\xi(1+\xi)}{\nu}(X_i - u)\right\}_+^2}\right]\right] \\ &= \mathbb{E}_{N_u}[N_u]\, \mathbb{E}_{X|N_u}\left[\frac{(X-u)^2}{\left\{1 + \frac{\xi(1+\xi)}{\nu}(X - u)\right\}_+^2}\right] \\ &= r\frac{1+\xi}{\nu} \int_u^{+\infty} (x-u)^2 \left\{1 + \frac{\xi(1+\xi)}{\nu}(x-u)\right\}_+^{-\frac{1}{\xi}-3} \mathrm{d}x.\end{aligned}$$

The above integral exists provided $\xi > -1/2$ and we obtain

$$\mathbb{E}_{N_u,X}\left[\sum_{i=1}^{N_u} \frac{(X_i - u)^2}{\left\{1 + \frac{\xi(1+\xi)}{\nu}(X_i - u)\right\}_+^2}\right] = \frac{2r\nu^2}{(1+\xi)^3(1+2\xi)}.$$

Similarly, the remaining expected values can be written as

$$\mathbb{E}_{N_u,X}\left[\sum_{i=1}^{N_u} \frac{(X_i - u)}{\left(1 + \frac{\xi(1+\xi)}{\nu}(X_i - u)\right)}\right] = \frac{r\nu}{(1+\xi)^2},$$

$$\mathbb{E}_{N_u,X}\left[\sum_{i=1}^{N_u} \log\left(1 + \frac{\xi(1+\xi)}{\nu}(X_i - u)\right)\right] = r\xi.$$

Plugging these values into the Fisher coefficients yields the result:

$$I(r, \nu, \xi) = \mathrm{diag}\left(\frac{1}{r}, \frac{r}{\nu^2(1+2\xi)}, \frac{r}{(1+\xi)^2}\right).$$

Proof of Proposition 2 Let us show that the following integral exists for any $n_u \geq 1$:

$$C_{n_u} = \int_{\mathcal{S}} \frac{r^{1/2}e^{-r}}{\nu(1+\xi)(1+2\xi)^{1/2}} \left(\frac{r(1+\xi)}{m\nu}\right)^{n_u} \prod_{i=1}^{n_u} \left(1 + \frac{\xi(1+\xi)}{\nu}(x_i - u)\right)^{-1-\frac{1}{\xi}} \mathrm{d}r\mathrm{d}\nu\mathrm{d}\xi,$$

where $\mathcal{S}$ is the integration domain:

$$\mathcal{S} = \left\{(r, \nu, \xi) \in \mathbb{R}^3 \text{ s.t. } \xi > -\frac{1}{2},\, r > 0,\, \nu \geq \{-\xi(1+\xi)((\max_i x_i) - u)\}_+\right\}.$$

Let us consider the case of one observation ($n_u = 1$):

$$\begin{aligned}
&\int_{-\frac{1}{2}}^{+\infty} (1+2\xi)^{-\frac{1}{2}} \int_0^{+\infty} r^{\frac{3}{2}} e^{-r} \int_{\{-\xi(1+\xi)(x-u)\}_+}^{+\infty} \nu^{-2} \left(1 + \frac{\xi(1+\xi)}{\nu}(x-u)\right)^{-\frac{1}{\xi}-1} \mathrm{d}\nu\mathrm{d}r\mathrm{d}\xi \\
&= \int_{-\frac{1}{2}}^{0} (1+2\xi)^{-\frac{1}{2}} \int_0^{+\infty} r^{\frac{3}{2}} e^{-r} \int_{-\xi(1+\xi)(x-u)}^{+\infty} \nu^{-2} \left(1 + \frac{\xi(1+\xi)}{\nu}(x-u)\right)^{-\frac{1}{\xi}-1} \mathrm{d}\nu\mathrm{d}r\mathrm{d}\xi \\
&+ \int_0^{+\infty} (1+2\xi)^{-\frac{1}{2}} \int_0^{+\infty} r^{\frac{3}{2}} e^{-r} \int_0^{+\infty} \nu^{-2} \left(1 + \frac{\xi(1+\xi)}{\nu}(x-u)\right)^{-\frac{1}{\xi}-1} \mathrm{d}\nu\mathrm{d}r\mathrm{d}\xi \\
&= \int_{-\frac{1}{2}}^{0} (1+2\xi)^{-\frac{1}{2}} \int_0^{+\infty} r^{\frac{3}{2}} e^{-r} \left[\frac{1}{(1+\xi)(x-u)} \left(1 + \frac{\xi(1+\xi)}{\nu}(x-u)\right)^{-\frac{1}{\xi}}\right]_{-\xi(x-u)\left(\frac{r}{m}\right)^{\xi}}^{+\infty} \mathrm{d}r\mathrm{d}\xi \\
&+ \int_0^{+\infty} (1+2\xi)^{-\frac{1}{2}} \int_0^{+\infty} r^{\frac{3}{2}} e^{-r} \left[\frac{1}{(1+\xi)(x-u)} \left(1 + \frac{\xi(1+\xi)}{\nu}(x-u)\right)^{-\frac{1}{\xi}}\right]_0^{+\infty} \mathrm{d}r\mathrm{d}\xi \\
&= \frac{1}{(x-u)} \int_{-\frac{1}{2}}^{+\infty} (1+\xi)^{-1}(1+2\xi)^{-\frac{1}{2}} \int_0^{+\infty} r^{\frac{3}{2}} e^{-r} \mathrm{d}r\mathrm{d}\xi \\
&= \frac{3\pi^{\frac{3}{2}}}{4(x-u)} < \infty.
\end{aligned}$$

Therefore, the posterior is proper for $n_u = 1$. It is well-known that it stays so for $n_u > 1$ as can be seen by induction. For instance for $n_u = 2$, the posterior writes

$$p(\boldsymbol{\theta} \mid x_1, x_2) \propto p(x_1, x_2 \mid \boldsymbol{\theta})p(\boldsymbol{\theta}) = p(x_2 \mid \boldsymbol{\theta})p(x_1 \mid \boldsymbol{\theta})p(\boldsymbol{\theta}) \propto p(x_2 \mid \boldsymbol{\theta})p(\boldsymbol{\theta} \mid x_1) \leq p(\boldsymbol{\theta} \mid x_1)$$

which is integrable.

Proof of Proposition 4 Similarly to the proof of Proposition 2, the aim is to show the existence of the following integral for any n_u:

$$C_{n_u} = \int_{\mathcal{S}} \frac{p_{\mathrm{PC}}(\xi \mid \lambda)}{\nu} e^{-r} \left(\frac{r}{m}\right)^{n_u} \left(\frac{\nu}{1+\xi}\right)^{-n_u} \prod_{i=1}^{n_u} \left(1 + \frac{\xi(1+\xi)}{\nu}(x_i - u)\right)^{-1-\frac{1}{\xi}} \mathrm{d}r\mathrm{d}\nu\mathrm{d}\xi,$$

with $p_{\mathrm{PC}}(\xi \mid \lambda)$ defined in Equation (2.13), and $\mathcal{S}$ the following integration domain:

$$\mathcal{S} = \left\{(r, \nu, \xi) \in \mathbb{R}^3 \text{ s.t. } \xi < 1,\, r > 0,\, \nu \geq \{-\xi(1+\xi)((\max_i x_i) - u)\}_+\right\}.$$

In the general case for n_u, we have

$$C_{n_u} = \frac{\Gamma(n_u+1)}{m^{n_u}} \int_{-\infty}^{1} \int_{\{-\xi(1+\xi)(x-u)\}_+}^{+\infty} \frac{p_{\text{PC}}(\xi \mid \lambda)}{\nu} \left(\frac{\nu}{1+\xi}\right)^{-n_u} \prod_{i=1}^{n_u} \left(1 + \frac{\xi(1+\xi)}{\nu}(x_i - u)\right)^{-1-\frac{1}{\xi}} \mathrm{d}\nu \mathrm{d}\xi$$
$$= \frac{\Gamma(n_u+1)}{m^{n_u}} \int_{-\infty}^{1} \int_{\{-\xi(x-u)\}_+}^{+\infty} \frac{p_{\text{PC}}(\xi \mid \lambda)}{\sigma} \sigma^{-n_u} \prod_{i=1}^{n_u} \left(1 + \xi\left(\frac{x_i - u}{\sigma}\right)\right)^{-1-\frac{1}{\xi}} \mathrm{d}\sigma \mathrm{d}\xi.$$

The remaining integral corresponds to the normalizing constant of the posterior distribution of a GPD model with a prior of the form $p(\sigma, \xi) \propto p(\xi)/\sigma$. Since $p(\xi)$ is a proper density, Theorem 1 in Northrop and Attalides (2016) allows us to conclude that C_{n_u} is finite for any $n_u \geq 1$. Note that this result remains true with $p_{\text{PC}}(\xi \mid \lambda)$ replaced by a Laplace distribution as suggested in Section 2.3.2, since the prior on ξ remains proper.

2.C Additional experiments

2.C.1 Simulations using an Hamiltonian Monte Carlo algorithm

Hamiltonian Monte Carlo (HMC, Neal, 1996) and its variants such as NUTS (Hoffman and Gelman, 2014) are MCMC methods with a Markov kernel based on trajectories of particles computed using Hamiltonian dynamics. As a consequence, the performance of these methods is also sensitive to the choice of the parameterization (see Betancourt, 2019 for a formalization of the problem). We performed the same experiments as those in Section 2.4.1 and Appendix 2.C.2, using 500 NUTS iterations instead of 1 000 Metropolis–Hastings draws. The results obtained here are similar, and show that the orthogonal parameterization improves the efficiency of NUTS sampling. The case $\xi > 0$ is illustrated in Figure 2.7 with the same configuration as the one described in the first paragraph of Appendix 2.C.2. We observe similar trends as those in Figure 2.8: changing the value of m improves convergence, and using the orthogonal parameterization is even better. Moreover, NUTS seems to be more efficient on the three cases than with Metropolis–Hastings, as the chains seem to be less correlated compared to their equivalent in Figure 2.8, and the ESS can even be greater than the number of draws.

2.C.2 Simulations in other maximum domains of attraction

We study the influence of parameterizations for MCMC convergence in cases where $\xi > 0$ and $\xi = 0$.

Example with $\xi > 0$ Here, we set $(m, u, \mu, \sigma, \xi) = (5, 10, 30, 15, 0.7)$, which leads to an expected number of observations $r \approx 239$. Looking at autocorrelations, ESS and $\hat{R}(x)$ curves in Figure 2.8, we can first confirm the result of Sharkey and Tawn (2017) about the inefficiency of Metropolis–Hastings with the original parameterization: high autocorrelations, high $\hat{R}(x)$ (around 1.7 for the highest) and almost zero ESS even after 1 000 iterations indicate a severe convergence issue. Changing the value of m before the MCMC algorithm as suggested by Sharkey and Tawn (2017) or by Wadsworth et al.

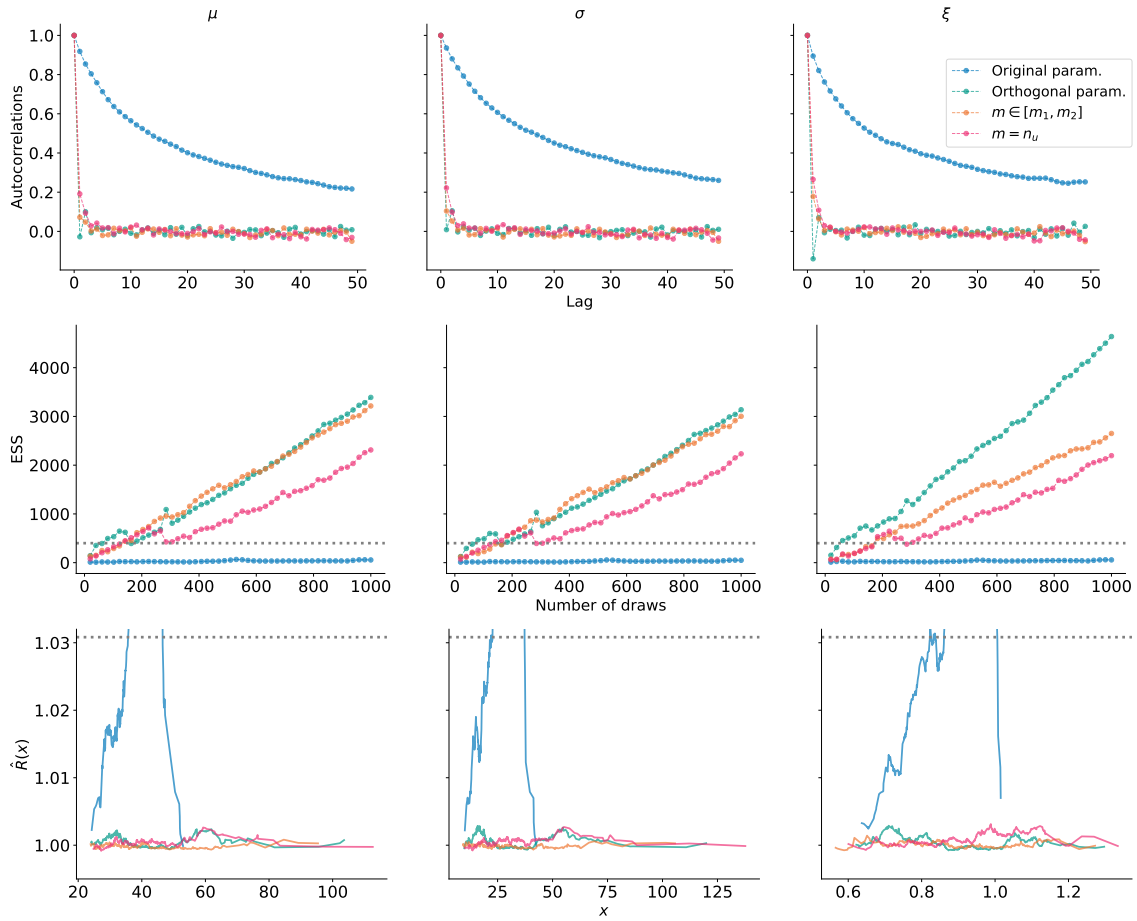


Figure 2.7: Convergence diagnostic plots for Poisson parameters (μ, σ, ξ) with $\xi > 0$, after 500 NUTS draws and a burn-in of 1 000, for four different parameterizations: the original one (in blue), the Sharkey and Tawn (2017) update with $m \in [\hat{m}_1, \hat{m}_2]$ (in orange), the Wadsworth et al. (2010) update with $m = n_u$ (in magenta), and the orthogonal parameterization (in green). Top row: autocorrelations as functions of the lag. Second row: evolution of ESS with the number of draws (the gray line corresponds to value of 400 recommended in Gelman et al. (2013)). Bottom row: $\hat{R}(x)$ as a function of the quantile x, with the adapted threshold of 1.03 (see Moins et al., 2023). The red curve is truncated for visibility purposes, as it is taking much larger values than the threshold.

(2010) improves inference significantly. Still, our orthogonal parameterization is even more efficient, especially for the estimation of the tail parameter ξ: the autocorrelation reduces even more rapidly with the lag, and the ESS increases faster with the number of draws. With the recommendations of ESS ≥ 400 for estimation (Gelman et al., 2013), our experimental setup is satisfactory only in the orthogonal case because of ξ. In contrast, more iterations are required to fulfill this condition for the parameterization recommended by Sharkey and Tawn (2017).

Example with $\xi = 0$ Finally when $\xi = 0$, the GPD and therefore the intensity $\Lambda(I_u)$ of the Poisson process defined in Section 2.1.1 reduce to an exponential model with location and scale parameters. Figure 2.9 shows an example in this case with $(m, u, \mu, \sigma, \xi) = (20, 20, 25, 5, 0)$, leading to $r \approx 54$ expected observations. Similarly to

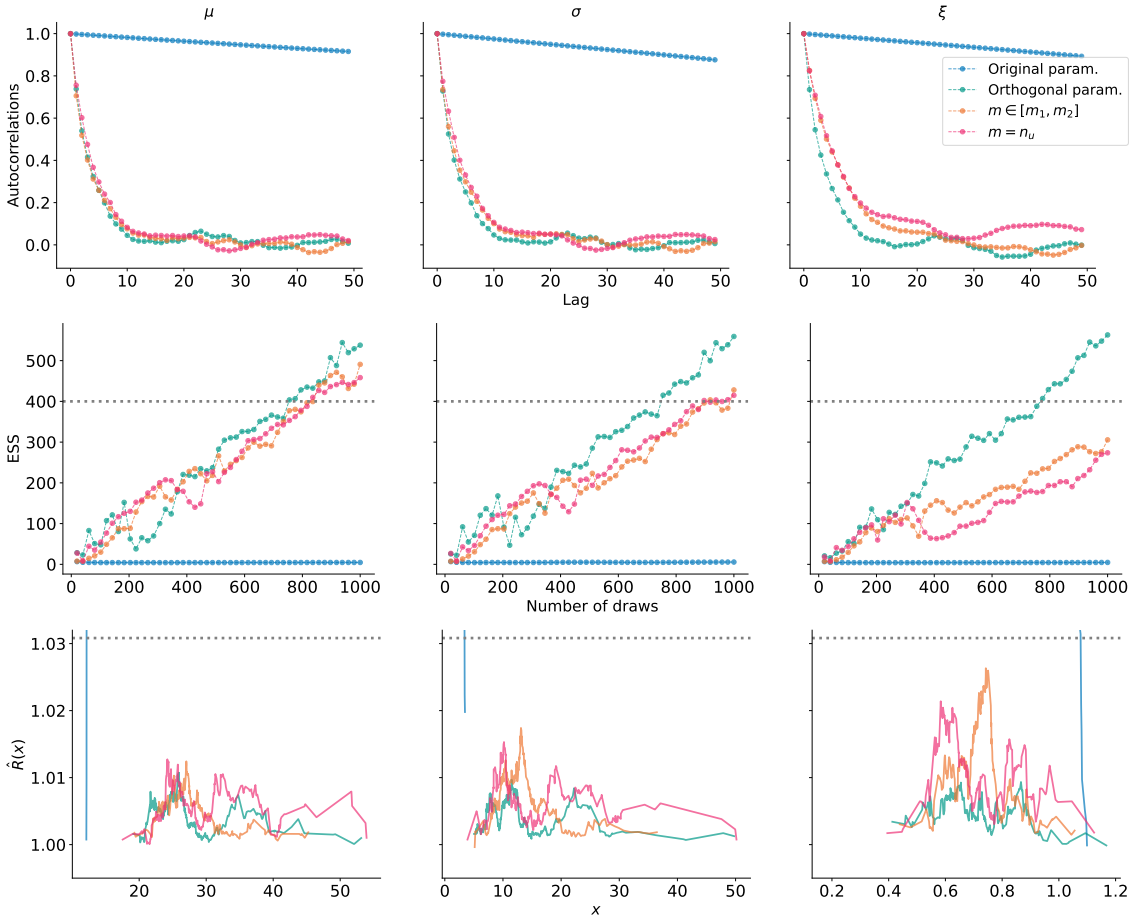


Figure 2.8: Convergence diagnostic plots for Poisson parameters (μ, σ, ξ) with $\xi > 0$, after 1 000 Metropolis–Hastings draws and a burn-in of 1 000, for four different parameterizations: the original one (in blue), the Sharkey and Tawn (2017) update with $m \in [\hat{m}_1, \hat{m}_2]$ (in orange), the Wadsworth et al. (2010) update with $m = n_u$ (in magenta), and the orthogonal parameterization (in green). Top row: autocorrelations as functions of the lag. Second row: evolution of ESS with the number of draws (the gray line corresponds to value of 400 recommended in Gelman et al. (2013)). Bottom row: $\hat{R}(x)$ as a function of the quantile x, with the adapted threshold of 1.03 (see Moins et al., 2023). The red curve is truncated for visibility purposes, as it is taking much higher values than the threshold.

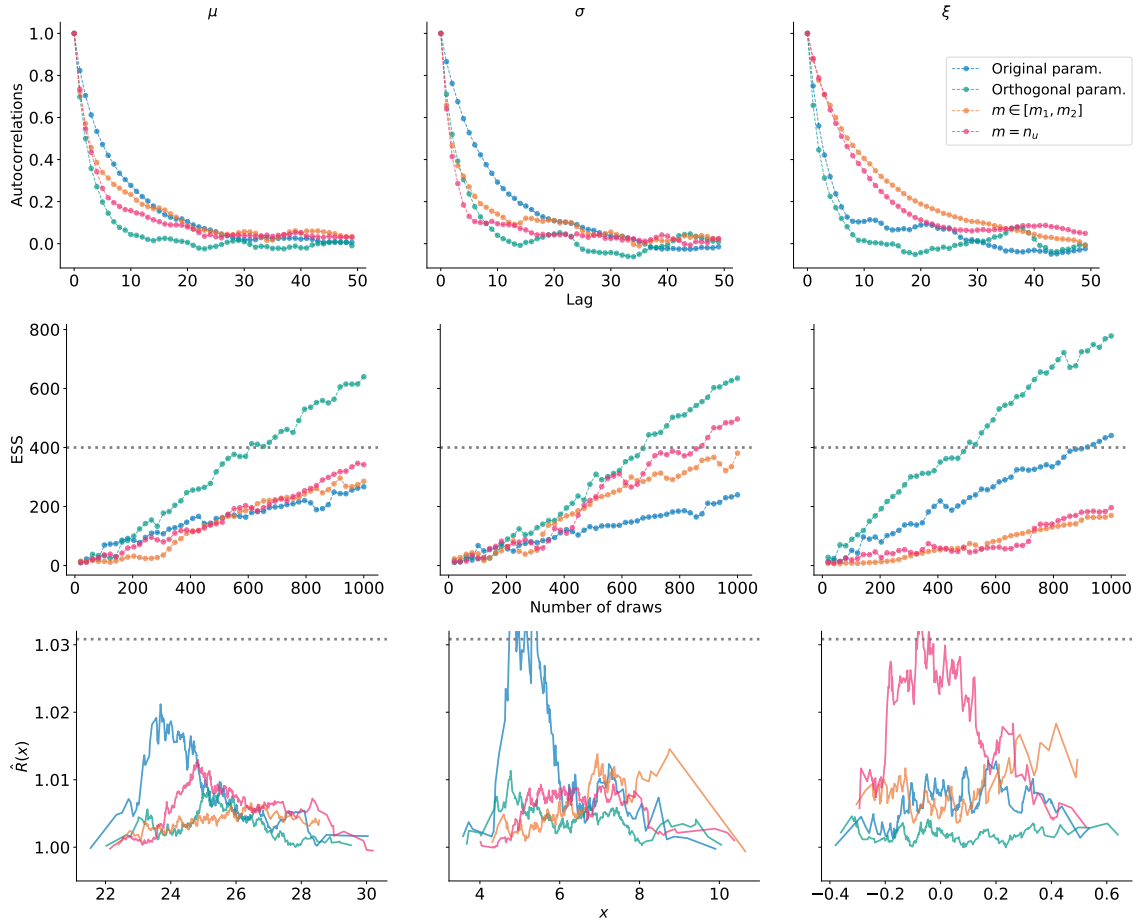


Figure 2.9: Convergence diagnostic plots for Poisson parameters (μ, σ, ξ) with $\xi = 0$, after 1 000 Metropolis–Hastings draws and a burn-in of 1 000, for four different parameterizations: the original one (in blue), the Sharkey and Tawn (2017) update with $m \in [\hat{m}_1, \hat{m}_2]$ (in orange), the Wadsworth et al. (2010) update with $m = n_u$ (in magenta), and the orthogonal parameterization (in green). Top row: autocorrelations as functions of the lag. Second row: evolution of ESS with the number of draws (the gray line corresponds to value of 400 recommended in Gelman et al. (2013)). Bottom row: $\hat{R}(x)$ as a function of the quantile x, with the adapted threshold of 1.03 (see Moins et al., 2023).

the case $\xi > 0$ in Section 2.4.1, this example illustrates that updating m like Sharkey and Tawn (2017) or Wadsworth et al. (2010) is beneficial for MCMC convergence, but less than using orthogonal parameterization. In the same way as in the two other maximum domains of attraction, this parameterization is the most efficient one for the convergence of Metropolis–Hastings algorithm.

2.C.3 GPD and GEV case

In the particular case of GPD (defined in Equation (2.2)) that arises in the traditional peaks over threshold model, the same observation can be made about the benefits of an orthogonal parameterization for (σ, ξ). More precisely, the transformation $(\nu, \xi) = (\sigma(1 + \xi), \xi)$ leads to an orthogonal Fisher information matrix for GPD (Chavez-Demoulin and Davison, 2005), and improves MCMC convergence as shown in Figure 2.10. The

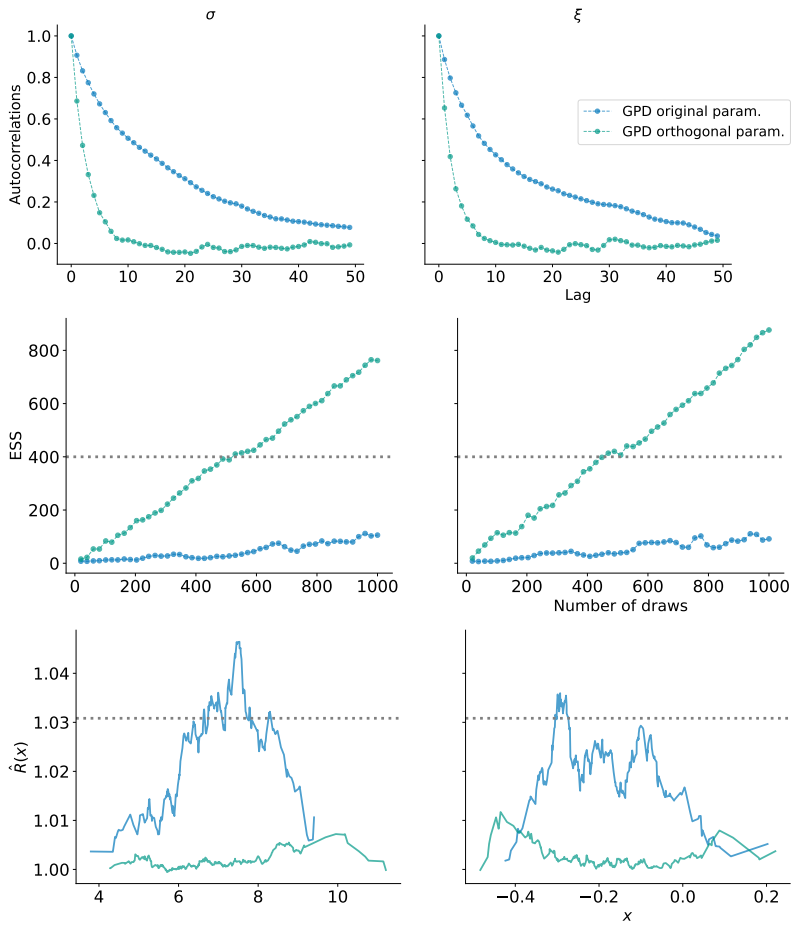


Figure 2.10: Convergence diagnostic plots for GPD parameters (σ, ξ) with $\xi < 0$, after 1 000 Metropolis–Hastings draws and a burn-in of 1 000, for two parameterizations, the original (in blue) and the orthogonal one (in green). Top row: autocorrelations as functions of the lag. Second row: evolution of ESS with the number of draws (the gray line corresponds to value of 400 recommended in Gelman et al. (2013)). Bottom row: $\hat{R}(x)$ as a function of the quantile x, with the adapted threshold of 1.03 (see Moins et al., 2023).

same experimental setup as in the Poisson process case is used here, with a choice of $(\sigma, \xi) = (5, -0.1)$ and $u = 25$. Again, all plots in Figure 2.10 show that the chains mixing is satisfactory only in the case of an orthogonal parameterization, while the original parameterization requires more iterations to be effective for inference. Up to our knowledge, there is no orthogonal parameterization for the GEV likelihood known in the literature. However, it should be noted that the parameters of the Poisson process model (μ, σ, ξ) correspond to those of the block maxima framework with m blocks (see Section 2.1.1). Consequently, we should expect a similar convergence issue for parameters (μ, σ, ξ) with GEV likelihood, and therefore an improvement in the MCMC convergence with the use of the orthogonal parameterization (r, ν, ξ) of the Poisson model.

2.C.4 Ratio-of-uniforms

The benefits of reparameterization for Bayesian inference can be extended to other sampling methods than MCMC. Typically, the efficiency of acceptance-rejection algorithms can be altered if the geometry of the acceptance region is too complex, and this can be due to correlation between parameters. The `rust` package (Northrop, 2022b) implements such an acceptance-rejection algorithm dedicated to extreme value models, the so-called generalized ratio-of-uniforms method. It consists in simulating uniformly values in a region that encloses an acceptance region, where the ratio of the obtained samples is distributed according to the target distribution (see Gilks et al. (1995, Chapter 5) for more details). As explained in the `revdbayes` documentation (Northrop, 2022a) which is built upon `rust`, the efficiency of this method highly depends on the probability of acceptance p_a. `revdbayes` already includes the possibility to use the reparameterization suggested by Wadsworth et al. (2010) with $m = n_u$ for the Poisson process, along with a rotate option to reduce dependence. We add the orthogonal parameterization in the comparison, and show the results in Figure 2.11. We set $(m, u, \mu, \sigma, \xi) = (100, 0, 1, 1, -0.1)$, and draw 10 000 samples for three configurations. As expected, the orthogonal parameterization slightly improves the probability of acceptance compared to the case where $m = n_u = 110$, which is already significantly better than the case where m is not changed ($m = 100$). Note that this package operates a transformation of variable before sampling to improve normality, which in view of the bottom row in Figure 2.11, may not be necessary for the orthogonal parameters.

2.C.5 Replications and comparison with maximum likelihood

Despite the fact that the Bayesian paradigm comes with several benefits (briefly described in Section 2.1.1), one can be interested in the comparison with frequentist estimator such as maximum likelihood estimation (MLE). From a frequentist point of view, this involves extracting a pointwise estimator from the posterior distribution, such as the posterior mean, and replicate the experiment to estimate the mean squared error (MSE). The two steps of the Bayesian workflow we study here are expected to impact the performance of these estimators. A parameterization which leads to poor convergence of the MCMC chains will affect the accuracy of estimation, and the prior can add a bias that may or may not be advantageous to the estimation.

For different values of ξ_0 between -0.5 and 1, we replicate 100 times the following experiment (this range includes a large number of models and allows us to have both Jeffreys and PC priors well defined): for $i = 1, \ldots, 100$, we generate samples $\boldsymbol{x}_i$ according

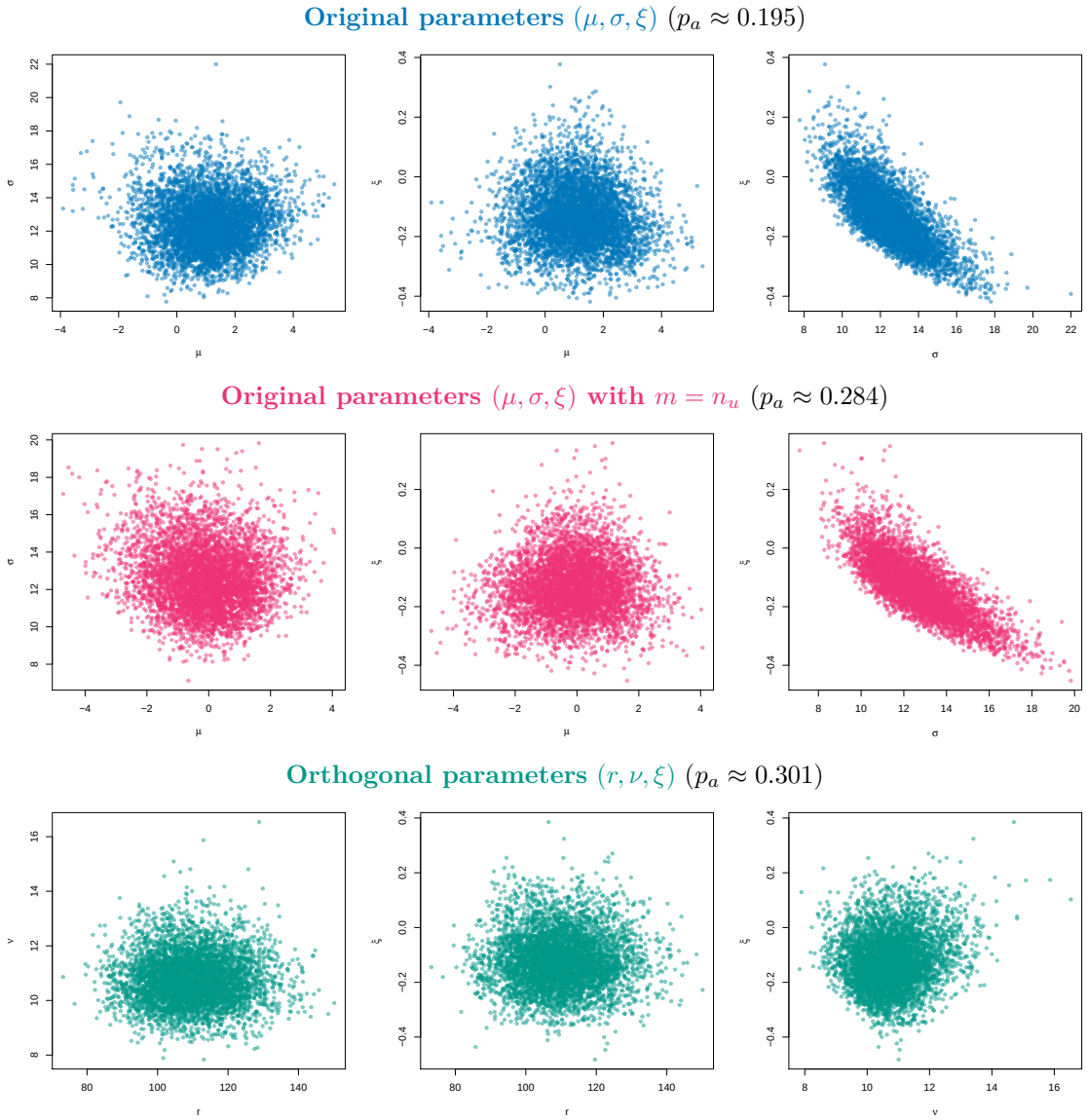


Figure 2.11: Pairwise plots of parameter values simulated using the ratio of uniform method, for three parameterizations: the original one (in blue), the Wadsworth et al. (2010) update with $m = n_u$ (in magenta), and the orthogonal parameterization (in green). The probability of acceptance p_a reflects the efficiency of the sampling method.

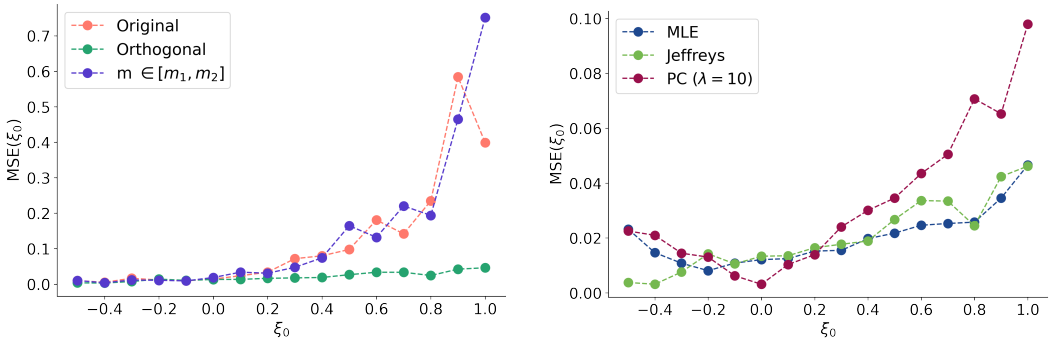


Figure 2.12: Mean squared error (MSE) on the estimation of ξ for a true value $\xi_0 \in [-1/2, 1]$. The computation is done on 100 replications for each value of ξ_0. Left panel: different parameterizations under Jeffreys prior. Right panel: Jeffreys and PC priors under orthogonal parameterization, along with MLE.

to a Poisson process distribution with parameters $(m, u, \sigma, \xi) = (1, 10, 15, \xi_0)$ and μ in a way such that the expected number of points is equal to $r = 100$:

$$\mu = u - \frac{\sigma}{\xi_0}(100^{-\xi_0} - 1).$$

Then, we run MCMC chains with the same configuration as in Section 2.4 and compute the posterior mean $\hat{\xi}_i = \mathbb{E}[\xi \mid \boldsymbol{x}_i]$. We these 100 experiments, we compute the MSE:

$$\text{MSE}(\xi_0) = \frac{1}{100}\sum_{i=1}^{100}(\hat{\xi}_i - \xi_0)^2.$$

First, we compare the different parameterizations for the Poisson process with the same Jeffreys prior. Results are displayed in the left panel of Figure 2.12, and illustrate the inaccuracy of the frameworks without reparameterization and with the update of Sharkey and Tawn (2017), due to lack of convergence of MCMC. This issue is getting worse as ξ_0 increases, and a bias/variance decomposition of the MSE shows that it is mostly due to the variance term. Then, for the same orthogonal parameterization, we compare Jeffreys prior, PC prior with a choice of $\lambda = 10$, and the MLE for the Poisson process, implemented using the extRemes package (Gilleland and Katz, 2016). Results in the right panel of Figure 2.12 show that the performance of the posterior mean estimation with Jeffreys prior is approximately the same as the MLE, except when ξ_0 is close to $-1/2$ where the asymptote behaviour of Jeffreys favours the estimation. This shows that, despite the uninformative construction, this prior can favour negative values of ξ_0 close to $-1/2$. The behaviour of PC prior is, as expected, penalizing the values of ξ far from $\xi_0 = 0$. When ξ_0 is around zero, this prior outperforms Jeffreys' one and MLE, but assuming a value near zero when $|\xi_0|$ is large can add a large bias.

Chapter 3

On the use of a local $\hat{R}$ to improve MCMC convergence diagnostic

Contents

Résumé

Diagnostiquer la convergence de méthodes de Monte Carlo par chaînes de Markov (MCMC) est crucial et reste un problème ouvert. Parmi les méthodes les plus populaires, le diagnostic de Gelman–Rubin (Gelman and Rubin, 1992), communément noté $\hat{R}$, est un indicateur qui vérifie la convergence des chaînes vers la distribution cible, en comparant les variances inter et intra-chaînes.

Plusieurs améliorations ont été suggérées depuis son introduction dans les années 90. Dans ce travail, nous cherchons à mieux comprendre le comportement de $\hat{R}$ en proposant une version localisée $\hat{R}(x)$ qui se concentre sur un quantile x de la distribution cible. Cette nouvelle version est d'abord définie à l'échelle de la population, c'est-à-dire que l'on définit une valeur théorique $R(x)$ qui sera estimée par $\hat{R}(x)$. Cela nous permet de déduire plusieurs propriétés, et notamment d'associer le choix du seuil de décision sur la convergence à un niveau de confiance α, qui lui est interprétable. Cette version localisée du diagnostic conduit naturellement à proposer un nouvel indicateur $\hat{R}_\infty$ sur l'ensemble de la distribution, qui permet à la fois de localiser la convergence dans différents quantiles de la distribution cible, tout en résolvant certains problèmes de convergence qui ne sont pas détectés par les autres versions de $\hat{R}$.

Une extension multivariée est ensuite proposée pour répondre à la question encore ouverte d'un diagnostic de convergence qui prendrait en considération la structure de dépendance. Notre analyse illustre la difficulté de la tâche par une différence de sensibilité par rapport à la direction de dépendance, ce qui implique un coût de calcul exponentiel en la dimension si l'on souhaite symétriser sur l'ensemble des directions. Or, par conséquent, ceci est peu raisonnable en grande dimension. Des simulations sont ensuite effectuées pour vérifier la sensibilité de ce diagnostic dans des cas classiques, ainsi que dans des cas construits mettant à défaut les autres versions dans la littérature.

La suite du chapitre est présentée sous la forme d'un article accepté pour publication (Moins et al., 2023). La Partie 3.1 dresse un état des lieux des différentes versions de $\hat{R}$ et de leurs limites, puis la Partie 3.2 définit la version univariée, et la Partie 3.3 l'extension multivariée. Enfin, les expériences en Partie 3.4 ainsi qu'une discussion en Partie 3.5 viennent achever le document principal. Plusieurs annexes viennent compléter ces travaux : des exemples additionnels prouvant la robustesse de $\hat{R}_\infty$ en Annexe 3.A, le détail des preuves dans en Annexe 3.B, l'estimation du seuil associé à un niveau de confiance donné en Annexe 3.C, et enfin des exemples de calculs explicites de la valeur théorique en Annexe 3.D.

Abstract

Diagnosing the convergence of Markov chain Monte Carlo (MCMC) is crucial and remains essentially an unsolved problem. Among the most popular methods, the Gelman–Rubin diagnostic (Gelman and Rubin, 1992), commonly referred to as $\hat{R}$, is an indicator that checks the convergence of chains to the target distribution by comparing the between-chain and within-chain variances.

Several improvements have been suggested since its introduction in the 90s. In this study, we aim to better understand the behavior of $\hat{R}$ by proposing a localized version $\hat{R}(x)$ that focuses on a quantile x of the target distribution. This new version is first defined at the population level, meaning we define a theoretical value $R(x)$ that will be estimated by $\hat{R}(x)$. This allows us to derive several theoretical properties, including associating the choice of threshold for the decision on convergence with a more interpretable confidence level α. This localized version of the diagnostic naturally leads to proposing a new indicator $\hat{R}_\infty$ over the entire distribution, which allows for both localizing the convergence in different quantiles of the target distribution and addressing some convergence issues not detected by other versions of $\hat{R}$.

A multivariate extension is then proposed to address the open question of convergence diagnostics that consider the dependency structure. Our analysis illustrates the difficulty of the task due to a difference in sensitivity with respect to the direction of dependence, which implies an exponential computational cost if one wishes to symmetrize and thus is not feasible in high dimensions. Simulations are then conducted to verify the sensitivity of this diagnostic in classical cases, as well as in toy cases that challenge other versions in the literature.

The rest of the chapter is presented as a manuscript accepted for publication (Moins et al., 2023). Section 3.1 provides an overview of the different versions of $\hat{R}$ and their limitations, followed by Section 3.2 defining the univariate version and Section 3.3 presenting the multivariate extension. Finally, the experiments in Section 3.4 and the discussion in Section 3.5 conclude the main document. Several appendices complement these works: additional examples demonstrating the robustness of $\hat{R}_\infty$ in Appendix 3.A, the detailed proofs in Appendix 3.B, the estimation of the threshold associated with a given confidence level in Appendix 3.C, and finally, explicit calculations of the theoretical value in Appendix 3.D.

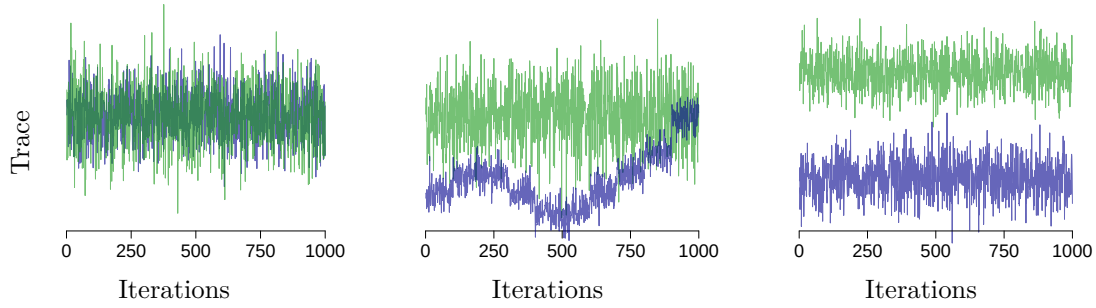


Figure 3.1: Traceplots illustrating convergence and two types of non-convergence of MCMC. Left: nothing indicates a convergence issue, as the two chains seem to have the same stationary distribution. Middle: the blue chain is still in an exploration phase and therefore is not stationary. Right: example where having multiple chains helps detecting a mixing issue despite a stationarity appearance of each.

3.1 Introduction

Markov chain Monte Carlo (MCMC) algorithms have strongly contributed to the popularity of Bayesian models to sample from posterior distributions, especially in high-dimensional or high computational settings. This success results in a variety of softwares increasingly used for a wide range of applications: Stan (Carpenter et al., 2017), PyMC3 (Salvatier et al., 2016), NIMBLE (de Valpine et al., 2017), or Pyro (Bingham et al., 2019), to cite a few. The fundamental idea behind these algorithms is the convergence of the sampling distribution to the target (typically the posterior) when the number of samples goes to infinity. A major challenge is therefore to know if the behaviour for a finite number of draws is satisfactory or not. This allows for a handle on the number of iterations to be drawn, which is all the more crucial in complex models with costly sampling schemes. See Roy (2020) for a recent literature review on convergence diagnostics.

3.1.1 Diagnosing MCMC convergence

Two frequently used properties to verify chains convergence are stationarity and mixing (see Vats and Flegal, 2021, for a discussion). *Stationarity* is related to the invariance property of the target distribution F for standard MCMC algorithms like Metropolis–Hastings or Gibbs sampling (Robert and Casella, 2004): if $\theta^{(i)}$ is the ith element of an MCMC chain, then $\theta^{(i)} \sim F$ implies $\theta^{(i+1)} \sim F$, so that as soon as an element of the chain is distributed according to F, all the following ones will be too. Thus, a chain whose distribution changes drastically during iterations is still in the exploration phase and is therefore not stationary (see middle panel in Figure 3.1). *Mixing* refers in practice to the exploration of the support of F: slow mixing chains correspond to chains that only explore a subset of the parameter space, which can lead to strong bias in the distribution (see Robert, 1995, for a more rigorous definition). A common way to limitate mixing issues is to run several chains in parallel with different starting points, which also allows comparing the chains together. Stationarity and mixing are two properties that can be treated separately: in principle, being stationary implies convergence to the target distribution and thus necessarily also mixing, but in practice there are examples of chains that seem

to have reached stationarity but are not mixing (see right panel in Figure 3.1), hence the need for comparing multiple chains.

We place ourselves in the case of several chains: consider m chains of size n, with $\theta^{(i,j)}$ denoting the ith draw from chain j. We focus here on the Gelman–Rubin diagnostic (Gelman and Rubin, 1992), named potential reduction scale factor and commonly denoted by $\hat{R}$. It is by far one of the most popular methods to assess MCMC convergence, used in particular in Stan, PyMC3, or NIMBLE. The original heuristic for $\hat{R}$ construction is the comparison between two estimators that converge to the target variance $\text{Var}[\theta]$, based on $\hat{W}$ and $\hat{B}$, respectively the estimated within- and between-variances. This diagnostic has the advantage of being scalar even in the case of a huge number of chains and comes with a rule of thumb that makes it very easy to use: generally $\hat{R} \geq 1$, and if it is greater than a given threshold (for example 1.01), then a convergence issue is raised. This was originally constructed to diagnose mixing issues only, but Gelman et al. (2013, Section 11.4) suggest splitting the chains in two before computing $\hat{R}$ to check for stationarity at the same time. We will also always consider this split version of $\hat{R}$ throughout this paper, thus focusing only on the problem of mixing diagnostic.

3.1.2 Different $\hat{R}$ versions and their limitations

The original $\hat{R}$ of Gelman and Rubin (1992) has some limitations that are listed here with associated improvements suggested in the literature.

L1. It must be compared to an arbitrary chosen threshold. To use $\hat{R}$, a threshold must be set to determine a convergence issue. Originally set to 1.1, Vats and Knudson (2021) note that this choice is arbitrary and usually too optimistic. Thus, the authors propose a threshold according to a confidence level based on a relationship made with effective sample size (ESS). This observation was then shared by Vehtari et al. (2021) who suggest dropping the threshold to 1.01. Driven by practical arguments, this choice remains unprincipled nor theoretically justified, which is related to the next limitation.

L2. It suffers from a lack of interpretability. How to interpret a given value of $\hat{R}$? By construction, $\hat{R}$ is a ratio of two quantities that must estimate the posterior variance. Therefore, having a value close to one can be seen as having two correct estimations of the same quantity, which is an indication of convergence. However to our knowledge, no study investigates the theoretical or population value R that $\hat{R}$ aims at estimating, which would shed light on what is actually diagnosed. Typically, chains such that $\hat{R} \approx 1$ do not necessarily correspond to mixing chains: Vehtari et al. (2021) exhibit some counter-examples in order to motivate a more robust version called rank-$\hat{R}$. Still, the different versions of $\hat{R}$ only allow to draw conclusions when they are significantly greater than 1, and the common properties of chains producing $\hat{R} \approx 1$ are not well known as they are constructed at the estimator level.

L3. It is not robust to certain types of non-convergence. Traditional $\hat{R}$ can be fooled, in the sense that $\hat{R} \approx 1$ without convergence. This motivates the construction of rank-$\hat{R}$ (Vehtari et al., 2021), based on two cases where the original $\hat{R}$ is not robust:

(i) When the mean of the target distribution is infinite: in that case $\hat{W}$ and $\hat{B}$ are ill-defined and $\hat{R} \approx 1$ even though the chains follow different distributions. One solution is to apply rank transformation on the chains before computing $\hat{R}$ (this version is named bulk-$\hat{R}$ by Vehtari et al., 2021).

(ii) When the means of the chains are equal: in that case, the variance of means $\hat{B}$ is zero, and so $\hat{R} \approx 1$ even if the variances of chains are different. Here in addition to the rank-transformation, transforming the chains to get the deviation from their median allows to overcome this problem (this version is named tail-$\hat{R}$ by Vehtari et al., 2021).

Defining rank-$\hat{R} = \max\{\text{bulk-}\hat{R}, \text{tail-}\hat{R}\}$ overcomes the two issues at the same time. However, this robustness can be seen as very specific and can easily be fooled by simple examples. One way is to consider chains with different distributions, but with (i) same mean (to fool bulk-$\hat{R}$), and (ii) same mean over the median (to fool tail-$\hat{R}$). For example, uniform $\mathcal{U}(\mu - 2\sigma, \mu + 2\sigma)$, normal $\mathcal{N}(\mu, \frac{\pi}{2}\sigma^2)$, or Laplace $\mathcal{L}(\mu, \sigma)$ distributions share the same mean (equal to μ) and same mean over the median (equal to σ), and thus mixing them yields rank-$\hat{R} \approx 1$. We provide an example and a more general framework to construct such cases in Appendix 3.A of the supplementary material. One illustration can also be found in the right column of Figure 3.3. Although these counter-examples may never appear in practice, they do show some fairly counter-intuitive results that the additional layer of computation carried by rank-$\hat{R}$ makes even more difficult to analyse.

L4. It does not target a specific quantity of interest. Another point raised by Vehtari et al. (2021) is that the convergence diagnostic does not depend on inferential features of interest. It might be more precise to speak of convergence for a given posterior quantity, typically a mean, higher order moment, or quantile. Typically, practitioners apply $\hat{R}$ on quantities of interest such as the log-likelihood, the posterior density, or quantiles. On their side, Vehtari et al. (2021) suggest a local transformation on ESS to obtain a tail-ESS associated with 5% and 95% quantiles.

L5. It is associated with a univariate parameter. Although the vast majority of Bayesian models have multivariate parameters, $\hat{R}$ focuses on univariate convergence (i.e. convergence of margins). Some multivariate extensions exist, like Brooks and Gelman (1998) or Vats et al. (2019), but do not seem to be universally accepted: for example Stan or PyMC3 use instead a table containing univariate $\hat{R}$ with one value per parameter. However, assessing convergence on margins misses the point of dependence among parameter components, and does not guarantee the convergence of the joint distribution. Another version of $\hat{R}$ called R^* is suggested by Lambert and Vehtari (2022) and can deal with multivariate parameters: the idea is to use a classification algorithm which, in the case of converging chains, would not be able to identify to which chain a sample belongs. To avoid a result depending on the seed of the experiment, the authors suggest to draw several samples from the simplex obtained with the classification algorithm. In addition to the interpretability issues mentioned previously, this method has the constraint of not being able to study only a scalar value but a histogram, to check to what extent it contains or not the value 1.

We take a step forward in addressing all these limitations with a localized version of $\hat{R}$ briefly introduced in Moins et al. (2021a) and developed here: we analyze $\hat{R}(x)$, a local version of $\hat{R}$ associated with a given quantile x, and the corresponding population value $R(x)$. This study leads us to propose a new indicator $\hat{R}_\infty$. In addition to being more interpretable, $\hat{R}_\infty$ shows better results than $\hat{R}$ in terms of MCMC convergence diagnostic, both on simulated experiments and on Bayesian models. As with all other versions of $\hat{R}$, this one can be applied to any MCMC algorithm: Metropolis–Hastings, HMC (Neal, 1996), NUTS (Hoffman and Gelman, 2014), etc.

The rest of the paper is organized as follows: we introduce in Section 3.2 the population

version $R(x)$ and the corresponding sample version $\hat{R}(x)$, as well as their scalar counterparts R_∞ and $\hat{R}_\infty$. Since this proposed version depends on a quantile x and is constructed at a population level, it is both targeting a specific quantity of interest and interpretable, addressing respectively limitations L4 and L2. We also establish several properties on the behaviour of $R(x)$ function and on the convergence of the estimator $\hat{R}(x)$, helping in establishing a threshold and addressing limitation L1. Our proposed approach to deal with the multivariate case of limitation L5 is described in Section 3.3. Some empirical results are given in Section 3.4, showing that our proposed solution helps overcoming many of convergence issues identified in limitation L3. We conclude in Section 3.5. All proofs and details of the calculations are provided in the supplementary material, and experiments are available online[1] as well as the R package `localrhat` (Moins et al., 2022a) containing our diagnostic implementation.

3.2 Local version of $\hat{R}$

Since the original version of Gelman and Rubin (1992), the heuristic for the construction of $\hat{R}$ was based on an analysis of variance. It consists in comparing two estimators of the posterior variance $\mathrm{Var}[\theta]$. The first one is the within-variance $\hat{W}$, which underestimates $\mathrm{Var}[\theta]$ as the bias of the estimator is (most of the time) strictly negative if the elements of the chains are not i.i.d, see Vats and Knudson (2021). The second one adds the between-variance $\hat{B}$ as a bias correction. This typically overestimates $\mathrm{Var}[\theta]$ if the initial values are chosen over-dispersed. As pointed out by Vats and Knudson (2021), following this heuristic does not exclude the use of other estimators of the bias than $\hat{B}$. Moreover, defining $\hat{R}$ at the sample level hinders a theoretical study of a population version to be conducted. Another justification can start with the law of total variance: assume that a univariate θ is sampled using m chains, and let $Z \in \{1, \dots, m\}$ be the corresponding index of the chain. Then,

$$\mathrm{Var}[\theta] = \mathbb{E}_Z[\mathrm{Var}_{\theta|Z}[\theta \mid Z]] + \mathrm{Var}_Z[\mathbb{E}_{\theta|Z}[\theta \mid Z]]. \tag{3.1}$$

The two terms in the right-hand side correspond respectively to the population versions of the within-variance W and the between-variance B. Replacing them by their estimated versions yields the original $\hat{R}$ formula of Gelman and Rubin (1992). In the following, we use (3.1) on a chains transformation which allows to localise convergence at a given quantile. For the theoretical study, we suppose stationarity of the chains to focus only on chain mixing issues. Thus, samples within a chain $j \in \{1, \dots, m\}$ may be correlated but are all distributed according to the same distribution F_j which may vary with j.

3.2.1 Population version

For all $x \in \mathbb{R}$, introduce the Bernoulli random variable $I_x = \mathbb{I}\{\theta \leq x\}$, where $\mathbb{I}\{\cdot\}$ denotes the indicator function. Similarly to the Raftery–Lewis diagnostic (Raftery and Lewis, 1992), the idea of our local convergence estimate is decidedly simple: we use I_x in place of θ in the original Gelman–Rubin construction. The population within-chain and between-chain variances at point x are then defined respectively as $W(x) = \mathbb{E}\left[\mathrm{Var}[I_x \mid Z]\right]$ and

[1] https://theomoins.github.io/localrhat/Simulations.html

$B(x) = \mathrm{Var}[\mathbb{E}[I_x \mid Z]]$. Note that both quantities exist whatever the tail heaviness of θ distribution thanks to introduction of the indicator function, thus relaxing moment conditions of the original $\hat{R}$. We define the associated population $R(x)$ as

$$R(x) = \sqrt{\frac{W(x) + B(x)}{W(x)}}.$$

It turns out that under the assumption of stationarity for each chain, $R(x)$ can be expressed in closed-form with respect to the chains' distribution.

Proposition 1. *Suppose that, for any $j \in \{1, \dots, m\}$, $\mathbb{P}(Z = j) = 1/m$ and θ given $Z = j$ has cumulative distribution function (cdf) F_j. Then, one has for any $x \in \mathbb{R}$:*

$$R(x) = \sqrt{1 + \frac{\sum_{j=1}^{m} \sum_{k=j+1}^{m} \left(F_j(x) - F_k(x)\right)^2}{m \sum_{j=1}^{m} F_j(x)(1 - F_j(x))}}. \tag{3.2}$$

Thus, using I_x instead of θ defines a local convergence estimate at any point x which quantifies a distance between the F_j's. This allows for diagnosing convergence relatively to a quantile one wants to estimate (for a posterior credible interval for example). The following proposition states straightforward properties of $R(x)$ emanating from (3.2):

Proposition 2. *The population $R(x)$ satisfies the following properties:*

(i) $R(x) \geq 1$ *for all* $x \in \mathbb{R}$.

(ii) $R(x) = 1$ *for all* $x \in \mathbb{R}$ *if and only if* $F_1 = \cdots = F_m$.

(iii) $R(x) \to 1$ *as* $|x| \to \infty$.

(iv) $R(x)$ *inherits continuity property of* $F_1, \dots, F_m$ *if the support of the* F_j*'s are overlapping.*

Based on these results and in order to summarize this continuous index into a scalar one, we may also consider its supremum over $\mathbb{R}$:

$$R_\infty = \sup_{x \in \mathbb{R}} R(x). \tag{3.3}$$

Note that, in view of Proposition 2(iv), R_∞ is finite simply as soon as the F_j's are continuous with overlapping supports. Considering R_∞ amounts to considering the local version $R(x)$ corresponding to the quantile x with the poorest convergence when no information is given on the posterior interval used for inference.

3.2.2 Sample version

Population version $R(x)$ can be estimated by replacing the $F_j(x)$'s in (3.2) by their empirical counterparts $\hat{F}_j(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{\theta^{(i,j)} \leq x\}$. This is equivalent to computing the original version of $\hat{R}$ on indicator variables $I_x^{(i,j)} = \mathbb{I}\{\theta^{(i,j)} \leq x\}$ instead of $\theta^{(i,j)}$. This connects with the Raftery–Lewis diagnostic (Raftery and Lewis, 1992) and more recently with Vehtari et al. (2021) who suggest this transformation for effective sample size (ESS) to construct graphical diagnostics or "tail-versions" of this diagnostic. Moreover, a rank-normalization

step is added in Vehtari et al. (2021)'s to prevent from infinite moments, although using $I_x^{(i,j)}$ ensures the index existence whatever the $\theta^{(i,j)}$ distribution is. Skipping this step for $\hat{R}$ yields an explicit expression of what is estimated in the stationary case with (3.2). This makes the diagnostic more interpretable and allows us to obtain key theoretical results for the associated theoretical R and R_∞.

Note that for a given number of chains m and chain length n, $\hat{R}(x)$ can only take $m(n+1)$ different values, as the computation is based on nm indicator variables. Thus, the best accuracy we can obtain for $\hat{R}_\infty$ for a given n and m consists in evaluating $\hat{R}(x)$ at all the $\theta^{(i,j)}$'s. This can be accelerated by subsampling, often with limited decrease in accuracy.

3.2.3 Convergence properties

Let us assume that all m chains are mutually independent and have converged to a common distribution so that $F_1 = \cdots = F_m =: F$. Assume, moreover, that a Markov chain central limit theorem holds (see for instance Robert and Casella, 2004, Theorem 6.65), so that we can write

$$\sqrt{n}(\hat{F}_j(x) - F(x)) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2(x)\right), \tag{3.4}$$

as $n \to \infty$, for all $j \in \{1, \dots, m\}$ and where $\sigma^2(x)$ is some asymptotic variance. In particular in the i.i.d. setting, $\sigma^2(x) = F(x)(1-F(x))$. Letting $\hat{F}(x) = \frac{1}{nm}\sum_{j=1}^m \sum_{i=1}^n \mathbb{I}\{\theta^{(i,j)} \leq x\} = \frac{1}{m}\sum_{j=1}^m \hat{F}_j(x)$ and taking into account of the independence between chains yield

$$\sqrt{nm}(\hat{F}(x) - F(x)) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2(x)\right), \tag{3.5}$$

as $n \to \infty$, and $\sigma(x)/\sqrt{nm}$ can be interpreted as the Monte Carlo standard error (MCSE) associated with $\hat{F}(x)$. Following the definition of the ESS used in Gong and Flegal (2016) or Vats et al. (2019), we can define a local-ESS as the ratio of the target variance to the squared MCSE:

$$\mathrm{ESS}(x) = nm\frac{F(x)(1-F(x))}{\sigma^2(x)}. \tag{3.6}$$

This quantity is in line with the definition of ESS for quantile of Vehtari et al. (2021), and has already been studied by Raftery and Lewis (1992) who focus on this indicator transformation and approximate the resulting process as a two-state Markov chain. This yields an explicit expression of the stationary distribution F, which can be used to obtain an expression of $\mathrm{ESS}(x)$ as a function of the transition probabilities. Several limitations of this two-state Markov chain approximation are raised by Brooks and Roberts (1999), Doss et al. (2014), for example. A more general way to estimate $\mathrm{ESS}(x)$ is to apply the same idea as in the definition of the local $\hat{R}(x)$: use any estimator of ESS (Robert and Casella, 2004, Gelman et al., 2013) on indicator variables $I_x^{(i,j)}$ instead of $\theta^{(i,j)}$.

Combining the asymptotic result (3.5) with expression (3.6) yields the following large n limiting distribution result on $\hat{R}(x)$ (χ^2_{m-1} denotes the chi-square distribution with $m-1$ degrees of freedom).

Proposition 3. *Assume that all m chains are mutually independent and have converged to a common distribution $F \coloneqq F_1 = \cdots = F_m$. Then:*

(i) *The distribution of $\hat{R}_\infty$ does not depend on the underlying distribution F.*

m	$R_{\lim,\alpha}(x)$	ESS(x)	α
4	1.01	50	0.80
		100	0.57
		200	0.26
		400	0.04
		800	$< 10^{-3}$
		1500	$< 10^{-6}$

m	$R_{\lim,\alpha}(x)$	ESS(x)	α
2	1.005	400	0.05
4	1.010		
8	1.017		
15	1.029		
50	1.080		
100	1.144		

Table 3.1: Left: Type I error α as a function of ESS(x) when $R_{\lim,\alpha}(x) = 1.01$ and $m = 4$. Right: $R_{\lim,\alpha}(x)$ as a function of m when ESS$(x) = 400$ and $\alpha = 0.05$.

(ii) *For any* $x \in \mathbb{R}$, $\text{ESS}(x)(\hat{R}^2(x) - 1) \xrightarrow{d} \chi^2_{m-1}$ *as* $n \to \infty$.

Note that casting the problem of convergence monitoring in terms of analysing components of variance from multiple sequences dates back to Gelman and Rubin (1992), Section 2.2, and earlier works by Fosdick (1959), Gelfand and Smith (1990). Let us highlight that the assumption $F_1(x) = \cdots = F_m(x)$ is equivalent to the ANOVA hypothesis $\mathbb{E}(I_x^{(\cdot,1)}) = \cdots = \mathbb{E}(I_x^{(\cdot,m)})$ and that the statistics studied in Proposition 3(ii) can similarly be rewritten in terms of the ANOVA test statistics: $\hat{R}^2(x) - 1 = \hat{B}(x)/\hat{W}(x)$, where $\hat{B}(x)$ and $\hat{W}(x)$ are the respective empirical counterparts of $B(x)$ and $W(x)$. These interpretations can then be used to derive a statistical test on the convergence of the chains. To this end, note also that the limit in distribution of Proposition 3(ii) still holds when ESS(x) is replaced by a consistent estimator $\widehat{\text{ESS}}(x)$. This result allows computing the type I error associated with the null hypothesis that $\hat{R}(x) = 1$, in other terms that all the chains have converged to a common distribution at x. Let $z_{m-1,1-\alpha}$ be the quantile of level $1-\alpha$ of the χ^2_{m-1} distribution, and introduce the associated threshold

$$R_{\lim,\alpha}(x) \coloneqq \sqrt{1 + \frac{z_{m-1,1-\alpha}}{\text{ESS}(x)}}. \tag{3.7}$$

The type I error is then given by $\mathbb{P}(\hat{R}(x) \geq R_{\lim,\alpha}(x)) \simeq \alpha$. As an illustration, some values of α are reported for the threshold $R_{\lim,\alpha}(x) = 1.01$, $m = 4$ chains and different values of ESS(x) in the left panel of Table 3.1. For example, it appears that the probability of having $\hat{R}(x) > 1.01$ and ESS$(x) = 400$ when convergence is reached is 0.04, and decreases quickly for larger values of ESS(x).

3.2.4 Threshold elicitation

Threshold for the local $\hat{R}(x)$. Proposition 3(ii) allows us to associate a threshold for $\hat{R}(x)$ to a type I error α, using the definition of $R_{\lim,\alpha}(x)$ in (3.7). Some values are displayed in the right panel of Table 3.1 for a fixed ESS$(x) = 400$ and $\alpha = 0.05$. It appears that the value of 1.01, the recent recommendation of Vehtari et al. (2021), seems to be coherent for $\hat{R}(x)$ and a moderate number of chains, typically the default configuration in Stan ($m = 4$), JAGS ($m = 3$) or PyMC3 ($m = \max\{n_c, 2\}$ with n_c the number of cores). However, the value of m must be doubled if a split version is used, and when m increases the threshold becomes more severe and it may be appropriate to consider a higher (i.e. less stringent) one: for example, a threshold of 1.1 can be enough provided the number

	$R_{\infty,\text{lim}}$			
m \ α	0.005	0.01	0.05	0.1
2	1.018	1.016	1.012	1.010
3	1.023	1.022	1.016	1.014
4	1.027	1.025	1.020	1.018
8	1.038	1.037	1.031	1.028
10	1.043	1.041	1.036	1.033
20	1.080	1.076	1.062	1.056

Table 3.2: Empirical quantiles $R_{\infty,\text{lim}}$ of the $\hat{R}_\infty$ distribution under the null hypothesis that all chains follow the same distribution for a target ESS of 400, based on 2000 replications.

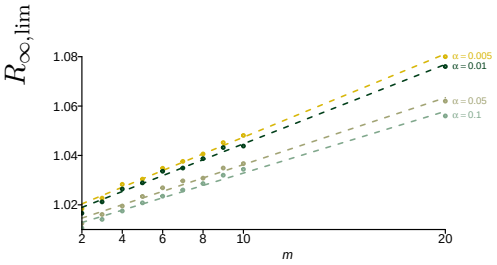


Figure 3.2: Illustration of the values in Table 3.2 and of the linearity with m for a fixed α.

of chains m is larger than 100. The case of a large number of chains has been recently studied by Margossian et al. (2022) who suggest a new version of $\hat{R}$ for this configuration. Note that a similar observation about the stringency of the threshold can be made with rank-$\hat{R}$, see Appendix 3.C for more details.

Therefore, we recommend to keep the threshold of 1.01 as a general rule of thumb for $\hat{R}(x)$, except if the number of chains is too large or if one wants to have a more precise threshold. In such a case it only requires to provide α, m and a target value $\text{ESS}(x)$ to compute $R_{\text{lim},\alpha}(x)$ using (3.7).

Threshold for the supremum $\hat{R}_\infty$. Proposition 3 does not induce any threshold $R_{\infty,\text{lim}}$ for $\hat{R}_\infty$, since Proposition 3(ii) only establishes the pointwise convergence of the empirical process $\hat{R}(\cdot)$. However, Proposition 3(i) shows that under the null hypothesis where all chains follow a common distribution F, the latter F is irrelevant to the $\hat{R}_\infty$ statistic. Such an independence to the underlying distribution F makes it possible the use of a quantile of $\hat{R}_\infty$ as a threshold associated with a given probability α and number of chains m. Table 3.2 provides estimations of $R_{\infty,\text{lim}}$ using replications for several values of α and m and a fixed number of effective samples of 400, as recommended by Vehtari et al. (2021) (more details are provided in Appendix 3.C). Here, we can see that a fixed rule of thumb for a range of m would be too imprecise, as the quantile values increase rapidly with m. Nevertheless, Table 3.2 illustrates a linear relationship between m and the appropriate threshold for a given α.

In the simulations in Section 3.2.5 and in the experiments in Section 3.4, we mostly consider $m = 4$ and therefore choose a threshold of 1.02, which is a little more accurate than 1.01 by looking at Table 3.2. Note that if $m = 8$ or if a split version of $\hat{R}_\infty$ is used with $m = 4$, then a threshold of 1.03 should be preferred. In the `localrhat` R package (Moins et al., 2022a), the computation of $\hat{R}_\infty$ comes with the associated threshold at 5% based on the calculations in Table 3.2, as well as a p-value associated with the obtained $\hat{R}_\infty$.

3.2.5 Illustrative examples

In this section, we consider toy distributions for the chains, where the computation of R_∞ can be done explicitly. In particular, we first focus on two cases raised by Vehtari et al. (2021) of deficient behaviour of the traditional $\hat{R}$. Then, we exhibit a failure situation for rank-$\hat{R}$. All these theoretical behaviours are illustrated on a simulation study. Further applications to Bayesian inference are provided in Section 3.4, and other examples where $\hat{R}$ and rank-$\hat{R}$ fail in Appendix 3.D.

Example 1: Chains with same mean and different variances. To tackle the first situation of poor behaviour of the traditional $\hat{R}$, we consider m chains following centered uniform distributions with different variances. More specifically, assume that the $m-1$ first chains have the cdf $F_1 = \cdots = F_{m-1}$ of the uniform distribution $\mathcal{U}(-\sigma, \sigma)$ while the last chain has the cdf F_m of the uniform distribution $\mathcal{U}(-\sigma_m, \sigma_m)$ with $0 < \sigma \leq \sigma_m$. In such a case, the between-variance is zero and it is thus expected that $\hat{R} \approx 1$. In contrast, Lemma 3 in Appendix 3.D provides an explicit expression for $R(x)$ as well as

$$R_\infty = \sqrt{1 + \frac{m-1}{m}\left(1 - \frac{2}{1 + \sigma_m/\sigma}\right)}.$$

It appears that R_∞ is an increasing function of σ_m/σ starting from $R_\infty = 1$ when $\sigma_m/\sigma = 1$, and upper-bounded by $\sqrt{2 - 1/m}$ when $\sigma_m/\sigma \to \infty$. Results are illustrated in the left column of Figure 3.3. In the bottom row, the histograms of replications confirm that $\hat{R}_\infty$ is able to spot the same convergence issue as the one Vehtari et al. (2021) suggests.

Example 2: Chains with heavy-tails and different locations. As a second example of poor behaviour of $\hat{R}$, we consider chains following Pareto(α, η) distributions, with cdf

$$F(x \mid \alpha, \eta) = 1 - (x/\eta)^{-\alpha}, \quad \forall x \in [\eta, +\infty),$$

shape parameter $\alpha > 0$ and lower bound $\eta > 0$. Let us recall that such a distribution is heavy-tailed (Embrechts et al., 2013, Table 3.4.2) and has an infinite first moment when $\alpha \leq 1$. We focus on the case where one chain is shifted from the other ones: $F_1(x) = \cdots = F_{m-1}(x) = F(x \mid \alpha, \eta)$ and $F_m(x) = F(x \mid \alpha, \eta_m)$ with $0 < \eta \leq \eta_m$ and $\alpha \leq 1$. Here, the within- and between-variances do not exist and it is expected in practice that $\hat{R} \approx 1$. In contrast, R_∞ can be written as

$$R_\infty = \sqrt{1 + \frac{1}{m}\left(\left(\frac{\eta_m}{\eta}\right)^\alpha - 1\right)},$$

see Lemma 4 in the supplementary material. Clearly, R_∞ is an increasing function of η_m/η starting from $R_\infty = 1$ when $\eta_m = \eta$ and such that $R_\infty \to \infty$ as $\eta_m/\eta \to \infty$. Results are shown in the middle column of Figure 3.3. This experiment corresponds to the second example of convergence issue raised by Vehtari et al. (2021). The same observations as for Example 1 can be made here: $\hat{R}_\infty$ is prone to indicating a convergence issue than rank-$\hat{R}$.

Example 3: Chains with same mean and mean over the median. Finally, we come back to the example described in Section 3.1.2 where both $\hat{R}$ and rank-$\hat{R}$ fail to detect non-convergence. Following the method described in Appendix 3.A, we consider $m-1$

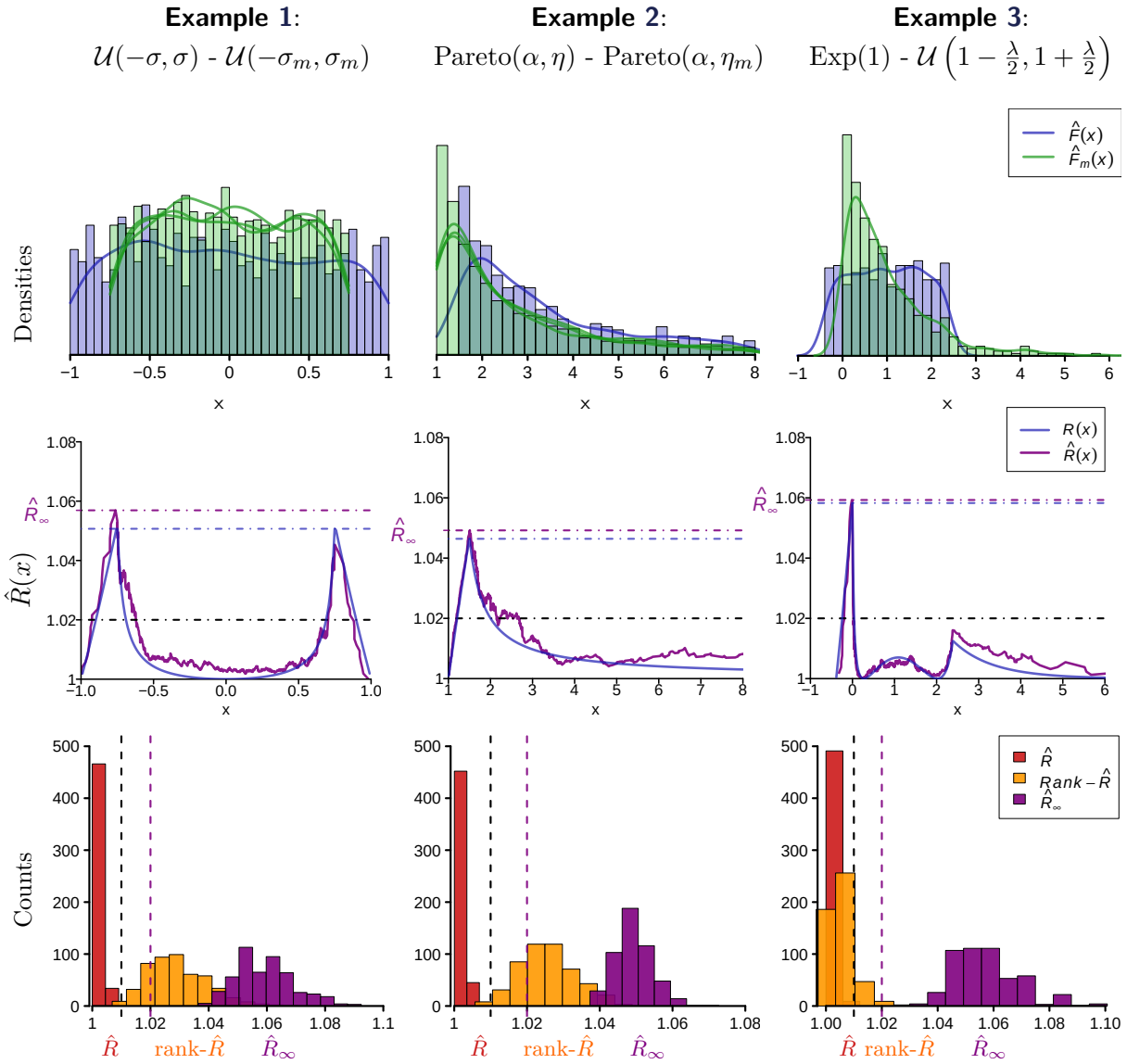


Figure 3.3: Illustrations with $m = 4$ chains, $n = 200$ independent iterations each. Top row: Simulation of $F_1 = \cdots = F_{m-1}$ in green distinct from F_m in blue. For the uniform example (left), $\sigma = 3/4$ and $\sigma_m = 1$, for the Pareto (middle) $\eta = 1$ and $\eta_m = 1.5$, and for the uniform (right) $\lambda = 4\log(2)$. Second row: The corresponding population version $R(x)$ and empirical version $\hat{R}(x)$ as functions of x for one replication. Bottom row: Histograms of 500 replications of $\hat{R}$, rank-$\hat{R}$ and $\hat{R}_\infty$. Dashed lines correspond to the threshold of 1.01 for $\hat{R}$ and rank-$\hat{R}$ and 1.02 for $\hat{R}_\infty$ (see Section 3.2.3).

exponential chains Exp(1) and one uniform $\mathcal{U}(1-2\log 2, 1+2\log 2)$. This results in chains with same mean and mean over the median. Results are illustrated in the right panel of Figure 3.3: the histograms of replications confirm that $\hat{R}_\infty$ is able to detect the convergence issue that neither $\hat{R}$ nor rank-$\hat{R}$ are able to detect. Here, the explicit calculation of R_∞ is not feasible, but Lemma 5 in the supplementary material provides another example where the computation can be done, with uniform and Laplace distributions.

3.3 Multivariate extension

3.3.1 Population version and algorithm for multivariate diagnosis

Our $R(x)$ can naively be adapted to the multivariate case: assume now that the parameter is multivariate and write $\boldsymbol{\theta} = (\theta_1, \dots, \theta_d) \in \mathbb{R}^d$ with $d \geq 2$, and denote by $\theta_p^{(j)}$ the coordinate $p \in \{1, \dots, d\}$ from chain $j \in \{1, \dots, m\}$. Similarly to the univariate case, $\hat{R}$ can be computed on the indicator variables $I_{\boldsymbol{x}}^{(j)} = \mathbb{I}\{\theta_1^{(j)} \leq x_1, \dots, \theta_d^{(j)} \leq x_d\}$ for any $\boldsymbol{x} = (x_1, \dots, x_d) \in \mathbb{R}^d$. Under the assumptions of Proposition 1, all calculations remain valid in dimension d and therefore the expression of $R(\boldsymbol{x})$ is formally the same as in (3.2):

$$R(\boldsymbol{x}) = \sqrt{\frac{W(\boldsymbol{x}) + B(\boldsymbol{x})}{W(\boldsymbol{x})}} = \sqrt{1 + \frac{\sum_{j=1}^m \sum_{k=j+1}^m \left(F_j(\boldsymbol{x}) - F_k(\boldsymbol{x})\right)^2}{m \sum_{j=1}^m F_j(\boldsymbol{x})(1 - F_j(\boldsymbol{x}))}}. \tag{3.8}$$

The properties listed in Proposition 2 in the univariate case remain true as well. The associated R_∞ is defined as $R_\infty(F_1, \dots, F_m) = \sup_{\boldsymbol{x} \in \mathbb{R}^d} R(\boldsymbol{x})$, while $\hat{R}(\boldsymbol{x})$ is computed by replacing the cumulative distribution functions in (3.8) by their empirical counterparts. Note also that all values computed in Table 3.1 and Table 3.2 remain identical in this multivariate extension. However, those results are not giving information about the sensitivity to convergence issues, which in the multivariate case can come from margins but also from the dependence structure.

It is easily seen that, if the marginal distributions of $F_1, \dots, F_m$ coincide, then R_∞ is the same as the one associated with uniform margins (see Lemma 1 in the supplementary material). In other words, we have $R_\infty(F_1, \dots, F_m) = R_\infty(C_1, \dots, C_m)$ where C_j is the copula defined in $[0,1]^d$ associated with F_j, $j \in \{1, \dots, m\}$. This suggests that a multivariate diagnosis can be conducted in two steps as follows:

1. Compute the univariate $\hat{R}_{\infty,p}$ separately on each of the coordinates $p \in \{1, \dots, d\}$. If $\hat{R}_{\infty,p} < R_{\infty,\lim}^{(M)}$ for all $p \in \{1, \dots, d\}$, with $R_{\infty,\lim}^{(M)}$ a choice of margins threshold, then all of them are deemed to have converged and to be identically distributed.

2. Compute the multivariate $\hat{R}_\infty$ to check the dependence structure convergence. If $\hat{R}_\infty < R_{\infty,\lim}^{(C)}$, with $R_{\infty,\lim}^{(C)}$ a copula threshold, then the dependence structure is also deemed to have converged, and so has the multivariate distribution.

The test for convergence is now separated in two parts: 1. convergence of the margins, and 2. convergence of the copula knowing that the margins have converged. It can easily be shown that, up to a first order approximation, one way to obtain a type I error α for the global two-step test is to consider a level $\alpha/2$ for each of the two components. The first step corresponds to d univariate tests, so for $R_{\infty,\lim}^{(M)}$ one can use the univariate threshold

$R_{\infty,\lim}$ defined in Section 3.2.4 with a level $\alpha/2d$, corresponding to a Bonferroni correction for the error level $\alpha/2$. In the following subsections, we focus on the second step of the algorithm: the theoretical properties of the multivariate $\hat{R}_\infty$ in the case of convergence on the margins, which will provide insights for choosing $R^{(C)}_{\infty,\lim}$. Values of $R^{(M)}_{\infty,\lim}$ and $R^{(C)}_{\infty,\lim}$ are then given as functions of (α, d, m) in Table 3.3. As a general rule, one can reasonably use for $\alpha = 0.05$ the values $(R^{(M)}_{\infty,\lim}, R^{(C)}_{\infty,\lim}) = (1.03, 1.03)$ in the case of $m = 4$ chains, and $(R^{(M)}_{\infty,\lim}, R^{(C)}_{\infty,\lim}) = (1.04, 1.05)$ if $m = 8$ or if a split version is used with $m = 4$, with limited variations around these values for varying dimension d.

3.3.2 Upper bounds

Let us first consider the case of $m = 2$ chains with uniform margins and associated copulas C_1 and C_2. For all $\boldsymbol{u} = (u_1, \ldots, u_d) \in [0,1]^d$, one has

$$R(\boldsymbol{u}) = \sqrt{1 + \frac{(C_1(\boldsymbol{u}) - C_2(\boldsymbol{u}))^2}{2\left(C_1(\boldsymbol{u})(1 - C_1(\boldsymbol{u})) + C_2(\boldsymbol{u})(1 - C_2(\boldsymbol{u}))\right)}}. \tag{3.9}$$

In addition to having the usual lower bound of 1, the next lemma allows establishing an upper bound on $R_\infty(C_1, C_2)$.

Lemma 1. *Let C_1, C_2, C_- and C_+ be copulas such that:*

$$\text{for all } \boldsymbol{u} \in [0,1]^d, \begin{cases} C_-(\boldsymbol{u}) \le C_1(\boldsymbol{u}) \le C_+(\boldsymbol{u}), \\ C_-(\boldsymbol{u}) \le C_2(\boldsymbol{u}) \le C_+(\boldsymbol{u}). \end{cases} \tag{3.10}$$

Then, $R_\infty(C_1, C_2) \le R_\infty(C_-, C_+)$.

Let W_d and M_d the lower and upper Fréchet–Hoeffding bounds in dimension d (see Nelsen, 2006, Theorem 2.10.12):

$$W_d(\boldsymbol{u}) \coloneqq \max\left\{1 - d + \sum_{i=1}^{d} u_i, 0\right\} \quad \text{and} \quad M_d(\boldsymbol{u}) \coloneqq \min\left\{u_1, \ldots, u_d\right\}.$$

Any copula is bounded from below and from above by W_d and M_d respectively, in the sense of (3.10). Thus, applying Lemma 1 with $(C_-, C_+) = (W_d, M_d)$ yields:

Proposition 4. *For any d-variate copulas C_1 and C_2,*

$$R_\infty(C_1, C_2) \le \sqrt{\frac{d+1}{2}},$$

Unlike the univariate version (see for instance Example 2 in Section 3.2.5), the value of R_∞ associated with the convergence of the dependence structure is upper-bounded, with a bound that grows with the dimension. This difference of behaviour could be used for example to tune the threshold for the multivariate case. However this bound, although it is the "best possible" (Nelsen, 2006, Theorem 2.10.13), is tight only in the case $d = 2$ since W_d is no more a copula when $d > 2$. It may also be too loose since it compares the extreme case of one chain with comonotonic dependence and another one with anti-comonotonic dependence. Some refinements are proposed in Section 3.3.3.

In the case of $m > 2$ chains, the previous bounding technique does not apply anymore, and we propose the following result based on bounding pairwise R_∞'s:

Corollary 1. *For any $m \geq 2$ and d-variate copulas $(C_1, \ldots, C_m)$,*

$$R_\infty(C_1, \ldots, C_m) \leq \sqrt{1 + \frac{m-1}{2}(d-1)}.$$

Although this limit is not tight in the general case, it coincides with the upper bound of Proposition 4 when $m = 2$. Let us also note that, for any fixed $m \geq 2$, the upper bound of $R_\infty(C_1, \ldots, C_m)$ diverges at a fixed $\sqrt{d}$ rate as the dimension increases.

3.3.3 Influence of the dependence direction on the sensitivity of $\hat{R}_\infty$

When $m = 2$, one way to refine the upper bound established in Proposition 4 is to assume that both copulas are modelling either positive of negative dependence. More specifically, let us recall the notions of positive lower orthant dependence (PLOD) and negative lower orthant dependence (NLOD) (see Nelsen, 2006, Section 5.7). The random vector $(\theta_1, \ldots, \theta_d)$ is

- PLOD if $\forall \boldsymbol{x} \in \mathbb{R}^d, \quad \mathbb{P}(\theta_1 \leq x_1, \ldots, \theta_d \leq x_d) \geq \prod_{i=1}^d \mathbb{P}(\theta_i \leq x_i)$,
- NLOD if $\forall \boldsymbol{x} \in \mathbb{R}^d, \quad \mathbb{P}(\theta_1 \leq x_1, \ldots, \theta_d \leq x_d) \leq \prod_{i=1}^d \mathbb{P}(\theta_i \leq x_i)$.

Both properties can be characterized in terms of the associated copula. The PLOD (resp. NLOD) property holds if and only if $C(\boldsymbol{u}) \geq \Pi_d(\boldsymbol{u})$ (resp. $C(\boldsymbol{u}) \leq \Pi_d(\boldsymbol{u})$) for all $\boldsymbol{u} \in [0,1]^d$ where Π_d is the independent copula defined by $\Pi_d(\boldsymbol{u}) \coloneqq \prod_{i=1}^d u_i$. Note that this does not define a total order on copulas since some copulas are neither PLOD nor NLOD. Nevertheless, it allows us to derive refined bounds for R_∞ in the NLOD and PLOD cases.

For PLOD, the upper bound is in not closed-form for any dimension d, but simple bounds can be derived in the two extreme cases $d = 2$ and $d \to \infty$.

Corollary 2. *Let $m = 2$. For any two PLOD d-variate copulas C_1 and C_2, $R_\infty(C_1, C_2) \leq R_\infty(\Pi_d, M_d)$ with*

$$\begin{cases} R_\infty(\Pi_2, M_2) = \sqrt{\frac{1}{2} + \frac{1}{\sqrt{3}}} \approx 1.038 & \text{if } d = 2, \\ \sqrt{\frac{d}{2\log d}}(1 + o(1)) \leq R_\infty(\Pi_d, M_d) \leq \sqrt{\frac{d+1}{2}} & \text{as } d \to \infty. \end{cases}$$

Conversely, the upper bound can be computed explicitly in the NLOD case.

Corollary 3. *Let $m = 2$. For any two NLOD d-variate copulas C_1 and C_2, $R_\infty(C_1, C_2) \leq R_\infty(\Pi_d, W_d)$ with*

$$R_\infty(\Pi_d, W_d) = \sqrt{1 + \frac{1}{2}\frac{1}{\left(1 - \frac{1}{d}\right)^{-d} - 1}}.$$

Let us stress that positive and negative dependence are handled differently by R_∞. When $d = 2$, the PLOD and NLOD bounds (respectively equal to 1.04 and 1.08) are significantly lower than the value $\sqrt{3/2} \approx 1.22$ corresponding to the global bound, with a value higher in the NLOD case than in the PLOD one. However, this observation is quickly inverted when d increases: for NLOD, $R_\infty(\Pi_d, W_d)$ is bounded and converges to $\sqrt{1 + \frac{1}{2(e-1)}} \approx 1.136$ as $d \to \infty$, which strongly constrains the range of values that can

be obtained whatever the dimension. In contrast, the upper bound $R_\infty(\Pi_d, M_d)$ in the PLOD case diverges with the dimension, at the same rate (up to a logarithmic factor) as in the general case, see Proposition 4. Thus, the sensitivity of R_∞ strongly depends on the sign of dependence and asymptotically favours PLOD dependence when d increases.

This difference can be explained by the construction of $R(x)$ itself (and thus R_∞), which favours a dependence direction in $\mathbb{R}^d$ due to the computation of $\mathbb{I}\{\theta_1^{(\cdot)} \leq x_1, \dots, \theta_d^{(\cdot)} \leq x_d\}$. One way to overcome this issue in the bivariate case is to compute two versions of R_∞, denoted respectively by R_∞^+ and R_∞^-, based respectively on $\mathbb{I}\{\theta_1^{(\cdot)} \leq x_1, \theta_2^{(\cdot)} \leq x_2\}$ and $\mathbb{I}\{\theta_1^{(\cdot)} \leq x_1, \theta_2^{(\cdot)} \geq x_2\}$. Note that R_∞^+ coincides with the construction proposed in Section 3.3.1.

Corollary 4. *Let $m = 2$. Then, $R_\infty^+(\Pi_2, M_2) = R_\infty^-(W_2, \Pi_2)$ and $R_\infty^+(W_2, \Pi_2) = R_\infty^-(\Pi_2, M_2)$.*

It appears that PLOD and NLOD upper bounds are exchanged by computing R_∞^- instead of R_∞^+, which makes R_∞^- more sensitive to negative dependence than positive dependence (in the bivariate case). One way to consider symmetrically both dependencies would be to consider $\hat{R}_\infty^{(\max)} = \max(R_\infty^+, R_\infty^-)$. However, in dimension d, considering all directions would imply the computation of 2^{d-1} different R_∞, which would be too expensive for large d. Similar curse of dimensionality occurs in the multivariate extension of the Kolmogorov–Smirnov test, see for example Lopes et al. (2007) for improvements of the naive multidimensional version of the test. Computing $\hat{R}_\infty^{(\max)}$ is still feasible for small values of d: typically for $d \leq 6$ we were able to replicate values in our experiments. Therefore, we provide in Table 3.3 (Appendix 3.C) the estimated threshold $R_{\infty,\lim}^{(C)}$ associated with the maximum of $\hat{R}_\infty$ in all possible directions when $d \leq 6$.

One alternative in the high-dimensional case could be to apply $\hat{R}$ on an indicator function associated with a univariate function of the parameters, to return to the case described in Section 3.2. Typically in a Bayesian model, one could use the log-likelihood $l_{\boldsymbol{\theta}} = \log p(y \mid \boldsymbol{\theta})$ when it is available, and compute $\hat{R}_\infty$ with $\mathbb{I}\{l_{\boldsymbol{\theta}} \leq x\}$. Similarly, the log posterior as implemented in Stan can also be used, as suggested in the Stan reference manual (Carpenter et al., 2017). Ensuring convergence for all x on the log posterior may be satisfying for multivariate diagnosis, as it is illustrated in Example 3.9.

3.3.4 Multivariate illustrative examples

Similarly to Section 3.2.5, we illustrate our theoretical study in the multivariate case with simulations based on toy distributions for the chains. Especially, we consider multivariate normal distributions, and focus on the case where all the margins are the same (typically distributed according to a standard normal distribution). This leads to

$$\boldsymbol{\theta}^{(i,j)} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_j),$$

$i \in \{1, \dots, n\}$ and $j \in \{1, \dots, m\}$, where $\boldsymbol{\Sigma}_j$ is the covariance matrix of the chain j, with diagonal elements equal to one to keep standard Gaussian margins.

Example 4: Bivariate normal distributions with different correlation terms. In the bivariate case, the dependence structure is driven by only one value, which is the off-diagonal element $\rho_j \in (-1, 1)$ of $\boldsymbol{\Sigma}_j$. Similarly to other examples, we suppose that we

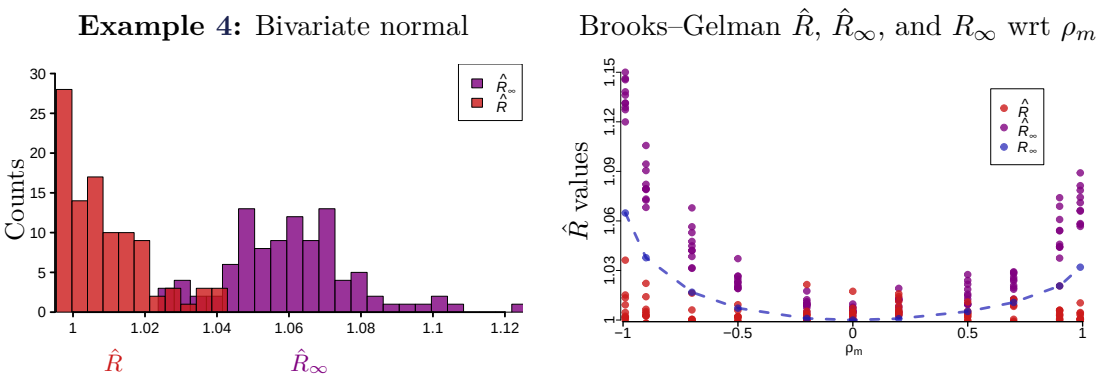


Figure 3.4: Behaviour of Brooks–Gelman $\hat{R}$ (in orange) and multivariate $\hat{R}_\infty$ (in violet) in the case of chains with bivariate normal distributions, with different off-diagonal elements in the covariance matrix. On the left: Histograms with 100 replications with one standard normal chain and one with $\rho_m = 0.9$. On the right: The same experiment with 10 replications for different values of ρ_m, plotted as a function of ρ_m, and the corresponding population R_∞ in blue.

have $m-1$ converging chains with identity covariance matrix ($\rho_1 = \cdots = \rho_{m-1} = 0$) while $\rho_m \in (-1, 1)$ for the last one.

Results are shown in Figure 3.4, with a comparison of $\hat{R}_\infty$ with the multivariate $\hat{R}$ of Brooks and Gelman (1998). The histogram on the left represents the values of the two diagnostics for 100 replications with $m = 2$, $n = 200$ and $\rho_m = 0.9$. Despite a large difference on the covariance term between the chains, we can see that Brooks–Gelman $\hat{R}$ fails to correctly diagnose this difference, as most of the values are between 1 and 1.01, contrary to $\hat{R}_\infty$. Due to the i.i.d nature of the example, the recent proposal of Vats and Knudson (2021) for a multivariate $\hat{R}$ does not detect any convergence issue as the diagnostic is not based on a comparison between chains. This difference of behaviour is confirmed on the right panel of Figure 3.4, which illustrates 10 replications of both diagnostics as a function of ρ_m. For instance, if $\rho_m = 0$ then the four chains are identically distributed and no convergence issue should be raised. Conversely, the value of $\hat{R}$ should increase when $|\rho_m| \to 1$, as the difference between the last chain and the other ones increases. For the Brooks–Gelman version, we can see that the value of $\hat{R}$ is almost constant and thus insensitive to ρ_m, which is not satisfactory, contrary to $\hat{R}_\infty$ which has a parabolic shape.

As discussed in Section 3.3.3, the behaviour of $\hat{R}_\infty$ is not symmetric when $\rho_m \to -1$ and $\rho_m \to 1$: the upper bound corresponding to positive dependence diverges with the dimension (Corollary 2 for PLOD copulas) whereas the one for negative dependence is bounded by approximately 1.14 (Corollary 3 for NLOD copulas). This leads to the intuition that the convergence diagnostic is more sensitive in the PLOD case than in the NLOD, but this observation is asymptotic and when $d = 2$, the two bounds are respectively equal to 1.08 and 1.04, so the statement is reversed. This asymmetry is illustrated in Figure 3.4 on theoretical R_∞ (in blue) and estimations $\hat{R}_\infty$ (in purple).

Example 5: Evolution of the behaviour when the dimension increases. In the general case of dimensionality $d > 2$, we still compare $m-1$ chains that follow a multivariate standard normal distribution with one that has a given covariance matrix $\boldsymbol{\Sigma}_m$. To obtain $\boldsymbol{\Sigma}_m$, we generate a matrix $\boldsymbol{S}$ according to Wishart distribution with d degrees of freedom, and we transform $\boldsymbol{S}$ in order to have one on the diagonal to keep the

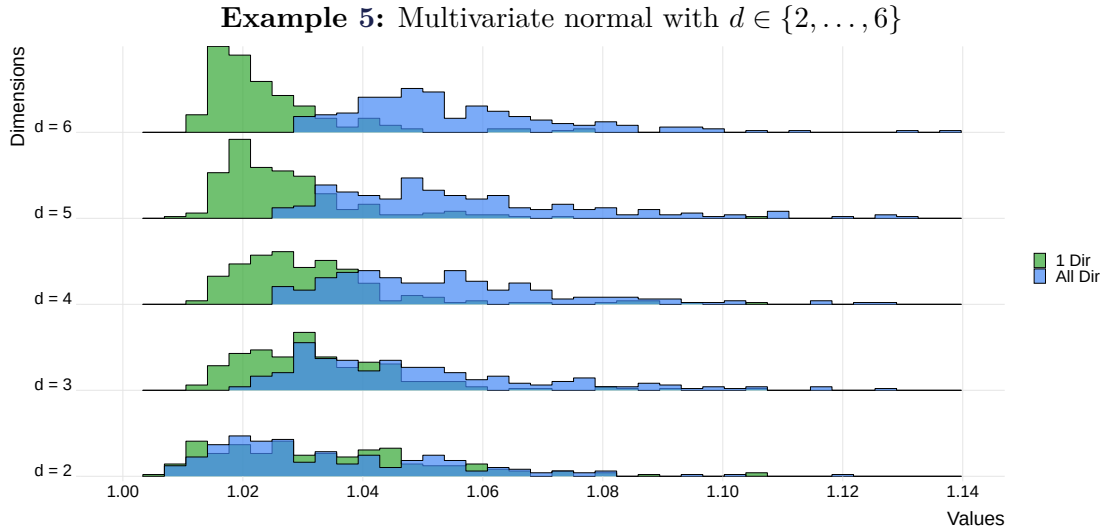


Figure 3.5: Comparison between $\hat{R}_\infty$ computed on one direction (in green), and $\hat{R}_\infty^{(\max)}$, the maximum of $\hat{R}_\infty$ computed on all possible indicator functions (in blue). For each $d \in \{2, \dots, 6\}$, 200 replications are done where a new covariance matrix is generated for the normal distribution, which leads to different directions of dependence among the replications.

same margins for all chains (while remaining semi definite positive):

$$\boldsymbol{\Sigma}_m = \boldsymbol{D}^{-1/2}\boldsymbol{S}\boldsymbol{D}^{-1/2}, \quad \text{with} \quad \boldsymbol{D} = \operatorname{diag}\left(s_{1,1}, \dots, s_{d,d}\right).$$

To illustrate the influence of the dependence direction (Section 3.3.3), a new matrix $\boldsymbol{\Sigma}_m$ is generated for each simulation, in order to have varying directions across replications. Then, we compare $\hat{R}_\infty$ with $\hat{R}_\infty^{(\max)}$, the maximum of $\hat{R}_\infty$ over all 2^{d-1} possible directions for the indicator functions.

Results are shown in Figure 3.5, where 200 replications are shown for $\hat{R}_\infty$ and $\hat{R}_\infty^{(\max)}$ for $d \in \{2, \dots, 6\}$. As $\hat{R}_\infty^{(\max)}$ requires the computation of 2^{d-1} different $\hat{R}_\infty$, obtaining these histograms quickly becomes infeasible for larger dimensions. When $d = 2$, we can see that there is no significant difference between $\hat{R}_\infty$ and $\hat{R}_\infty^{(\max)}$, but as the dimension increases the values of $\hat{R}_\infty$ become more concentrated and closer to one. Indeed, as the number of possible directions increases exponentially, it is more and more rare to obtain the one to which $\hat{R}_\infty$ is sensitive. On the contrary, $\hat{R}_\infty^{(\max)}$ seems to stay robust with respect to this curse of dimensionality in terms of sensitivity, as the histograms look invariant when d increases.

3.4 Empirical results

In Section 3.2.5 and Section 3.3.4, we considered toy examples where the distribution of the chains is known in order to control the value of the population R_∞ and illustrate the robustness when other versions of $\hat{R}$ fail. Here we extend to other models in a more practical case for Bayesian inference. We adopt a baseline similar to the one used by Lambert and Vehtari (2022) to illustrate the behaviour of $\hat{R}_\infty$ on Bayesian models, and add a multivariate example studied in Vats et al. (2019). For all examples in this section, we choose 4 chains and therefore a threshold $R_{\infty,\text{lim}} = 1.02$ in the univariate case (according

Example 6: Autoregressive model

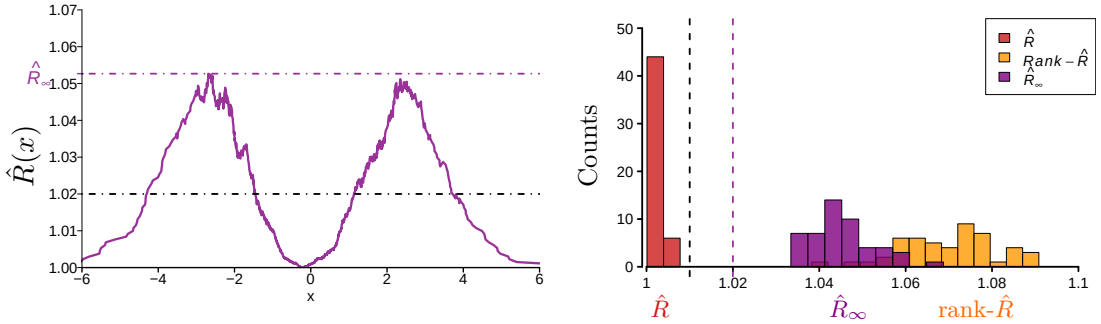


Figure 3.6: Behaviour of $\hat{R}_\infty$ on the autoregressive example described in Section 3.4, with $m = 4$ chains of size $n = 500$ and $(\sigma, \sigma_m, \rho) = (1, 2, 1/2)$. On the left: $\hat{R}(x)$ as a function of x for one replication. On the right: Histograms of 50 replications of $\hat{R}$, rank-$\hat{R}$ and $\hat{R}_\infty$. The dashed lines correspond to thresholds of 1.01 and 1.02.

to Section 3.2.3), and $(R^{(C)}_{\infty,\text{lim}}, R^{(M)}_{\infty,\text{lim}}) = (1.03, 1.03)$ in the multivariate one (according to Section 3.3.1). For each univariate study, we plot an example of $\hat{R}(x)$ as a function of x, and we recommend this illustration to users who want to analyse more carefully a given value of $\hat{R}_\infty$. Together with this figure, we also show histograms of replications to check the behaviour of the different $\hat{R}$ more rigorously. All experiments are done on R using `rstan` library (Stan Development Team, 2021) and the package `localrhat` that we propose with this paper (Moins et al., 2022a). Additional experiments have also been conducted on Python using OpenTURNS (Baudin et al., 2017). All the code concerning these experiments and the additional ones are available in the online appendix (link in the Introduction).

Example 6: Autoregressive model with different variances. The first example is a basic autoregressive model to study the behavior of $\hat{R}_\infty$ in the case of Markov chains with different variances: we consider m chains of size n such that for $i \in \{1, \dots, n-1\}$ and $j \in \{1, \dots, m\}$,

$$\theta^{(i+1,j)} = \rho\theta^{(i,j)} + \epsilon_{i,j}, \quad \text{with} \quad \epsilon_{i,j} \sim \mathcal{N}(0, \sigma_j^2),$$

where $\rho \in (0, 1)$ and $\sigma_j > 0$. In particular, assume that the first $m-1$ chains are generated using the same process: $\sigma_1 = \cdots = \sigma_{m-1} = \sigma$, while for the last chain $\sigma_m \neq \sigma$.

Results are illustrated in Figure 3.6 with $m = 4$, $\sigma = 1$, $\sigma_m = 2$ and $\rho = 1/2$ on 50 replications, and an example of $\hat{R}(x)$ as a function of x on the left panel. Similarly to the rank-$\hat{R}$ replications, the $\hat{R}_\infty$ values remain far from the threshold of 1.02 which confirms the sensitivity to this convergence defect. This corroborates in a more practical case the results of Example 1 in Section 3.2.5, on the sensitivity of $\hat{R}_\infty$ on chains with same mean and different variances. Note that the value $R(0) = 1$ is due to the fact that all the chains share the same median equal to zero.

Example 7: HMC on Cauchy distribution. As an extension of Example 2 in Section 3.2.5, we analyze the behaviour of $\hat{R}_\infty$ in the case of heavy-tailed distributions. We run Hamiltonian Monte Carlo (HMC) (Neal, 1996) using Stan on Cauchy distributions for 50 variables. We consider the one with the most important mixing issue diagnosed with

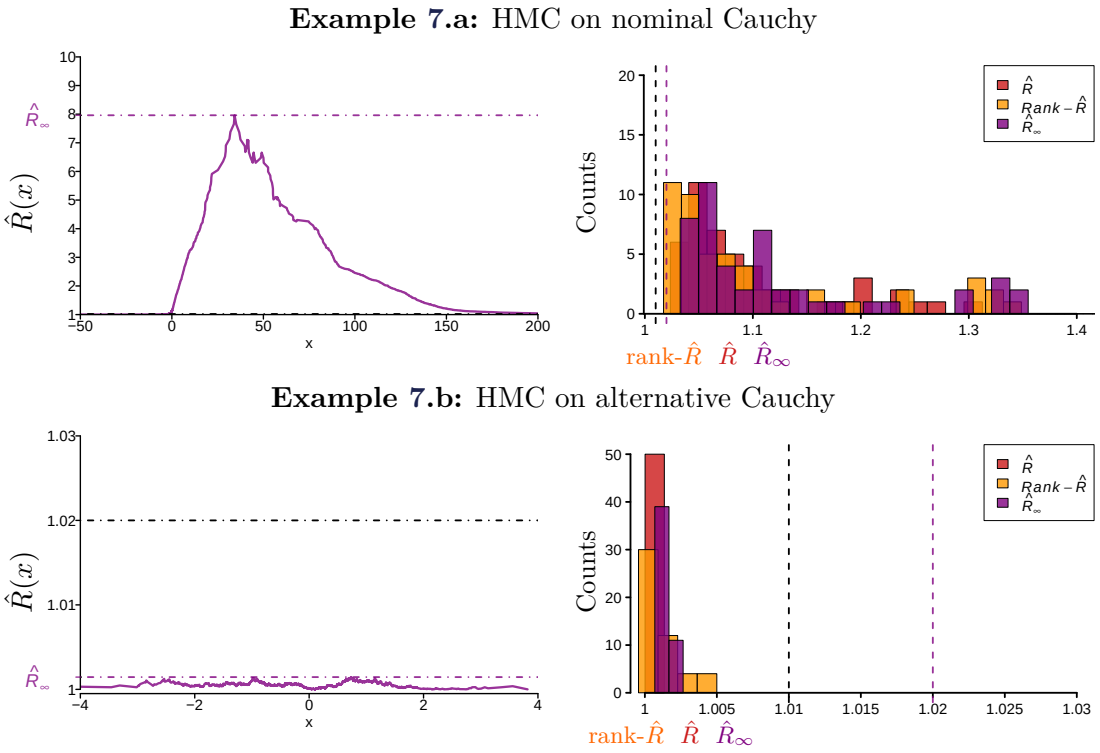


Figure 3.7: Behaviour of $\hat{R}_\infty$ on the Cauchy example described in Section 3.4 for the two parameterisations. On the left: $\hat{R}(x)$ as a function of x for one replication. On the right: Histograms of 50 replications of $\hat{R}$, rank-$\hat{R}$ and $\hat{R}_\infty$. The dashed lines correspond to thresholds of 1.01 and 1.02.

$\hat{R}_\infty$. Due to the tail heaviness of Cauchy distributions, the HMC iterations on a given chain can get trapped in a tail, which causes mixing issues. One solution to avoid this is to use an alternative parameterisation (Moins et al., 2023) that avoids sampling from a heavy-tailed distribution:

Example 7.a: Nominal parameterisation

$$x_j \sim \text{Cauchy}(0, 1), \quad j \in \{1, \dots, 50\}.$$

Example 7.b: Alternative parameterisation

$$x_j = a_j / \sqrt{b_j}, \quad a_j \sim \mathcal{N}(0, 1), \quad b_j \sim \chi^2_1.$$

One would expect convergence issues with the nominal parameterisation and not with the alternative one. For both, the process of selecting the worst parameters among the 50 ones is iterated for the generation of replications, and results are shown in Figure 3.7. Histograms on the top right confirm the risk of diverging chains with the nominal parameterisation, as all the values are above 1.02 for all the versions of $\hat{R}$. This means that it is very likely to have at least one chain out of the 50 with a convergence issue in this experiment. This divergence can be really extreme, as it is shown on the top left panel where the value of $\hat{R}_\infty$ is over seven, due to a mixing issue in the right tail of the distribution. The opposite occurs with the other parameterisation, as all the convergence diagnostics indicate no mixing issues (see bottom row of Figure 3.7), which means no counter-indications that the chains for the 50 variables have converged. Looking at $\hat{R}(x)$ function on one replication in the bottom left panel, the curve seems to be very noisy and close to 1 compared to 1.02 (even sometimes less than 1) so the difference with 1 seems only due to Monte Carlo noise.

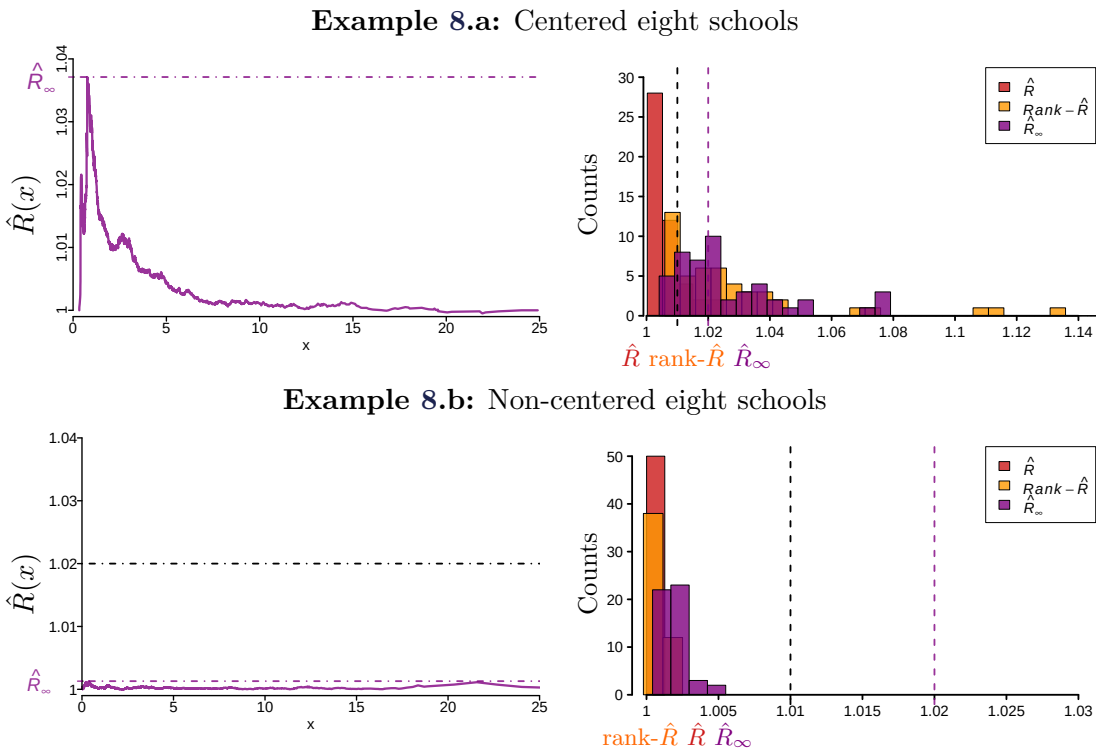


Figure 3.8: Behaviour of $\hat{R}_\infty$ on the hierarchical example for τ described in Section 3.4 for the centered and non-centered version. On the left: $\hat{R}(x)$ as a function of x for one replication. On the right: Histograms of 50 replications of $\hat{R}$, rank-$\hat{R}$ and $\hat{R}_\infty$. The dashed lines correspond to thresholds of 1.01 and 1.02.

Example 8: Hierarchical Bayesian model on two parameterisations. As a classical Bayesian example, we consider using HMC on a hierarchical Bayesian model and in particular the eight-school (Gelman et al., 2013, Section 5.5), where two parameterisations are possible to model the problem:

Example 8.a: Centered parameterisation (CP)
$$\theta_j \sim \mathcal{N}(\mu, \tau), \quad y_j \sim \mathcal{N}(\theta_j, \sigma_j^2).$$

Example 8.b: Non-centered parameterisation (NCP)
$$\bar{\theta}_j \sim \mathcal{N}(0, 1), \quad \theta_j = \mu + \tau\bar{\theta}_j, \quad y_j \sim \mathcal{N}(\theta_j, \sigma_j^2).$$

In the CP parameterisation, a prior dependence is between (μ, τ) and the population parameters θ_j, whereas in the other case (NCP), $\bar{\theta}_j$ is a priori independent of (μ, τ), and θ_j is just a function of $\bar{\theta}_j$ and (μ, τ) (see for example Papaspiliopoulos et al., 2003). Vehtari et al. (2021) argue in favour of the NCP for the eight-school example, by analysing the convergence of the chains associated with the parameter τ.

We also focus on computing $\hat{R}_\infty$ for τ: results and comparison with other versions of $\hat{R}$ are shown in Figure 3.8. In the first row, we can see that the $\hat{R}_\infty$ diagnostic confirms the one of rank-$\hat{R}$, as the two corresponding histograms are similar in the top right panel and conclude for a lack of convergence in most of the cases. However, for both diagnostics, a significant number of cases are also below 1.02 (respectively 1.01 for rank-$\hat{R}$), which is represented on the top left panel. In spite of this, the bottom row of Figure 3.8 shows a clear difference and NCP seems to help for chain convergence.

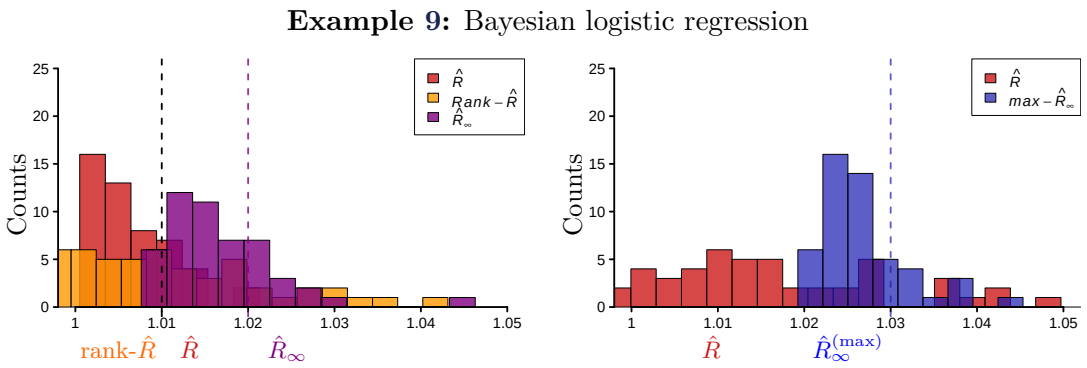


Figure 3.9: Behaviour of multivariate and univariate $\hat{R}_\infty$ on the Bayesian logistic regression example, with $m = 4$ chains of size $n = 200$. On the left: Histograms of 50 replications of $\hat{R}$, rank-$\hat{R}$ and univariate $\hat{R}_\infty$ all applied on the log-posterior. On the right: Histograms of 50 replications of Brooks–Gelman $\hat{R}$ and $\hat{R}_\infty^{(\max)}$. The dashed line corresponds to different thresholds: on the left, 1.01 in black for $\hat{R}$ and rank-$\hat{R}$, 1.02 in violet for $\hat{R}_\infty$, and on the right 1.03 in blue for $\hat{R}_\infty^{(\max)}$.

Example 9: Bayesian logistic regression. This example is related to the extension of $\hat{R}_\infty$ in the multivariate case as proposed in Section 3.3. As a multivariate Bayesian example, we run Stan on a basic hierarchical logistic model using the dataset `logit` available in the R package `mcmc`:

$$\boldsymbol{\beta} \sim \mathcal{N}(0, 0.35^2 \boldsymbol{I}_4), \quad y_j \sim \text{Bernoulli}\left(\frac{1}{1 + e^{-\boldsymbol{x}_j^\top \boldsymbol{\beta}}}\right).$$

Here the posterior is intractable and Vats et al. (2019) showed that the posterior coefficients $\boldsymbol{\beta}$ could be significantly correlated, encouraging a multivariate diagnostic to check the convergence of the dependence structure. We run $m = 4$ chains each of size $n = 200$ after a burn-in of 100. In this configuration, despite a low number of iterations, all the different univariate $\hat{R}_\infty$ are mostly below 1.02 when replicated, and the rank-$\hat{R}$ are below 1.01.

When applied to the log posterior, the diagnostic is less clear and results are shown in the left panel of Figure 3.9: a significant part of the histogram for $\hat{R}_\infty$ is below the threshold, meaning that the number of iterations is almost sufficient but is not yet. Looking at the right plot of Figure 3.9, we notice in this example that the sensitivity of $\hat{R}_\infty^{(\max)}$ is approximately the same as the univariate version on the left, as the proportion of values over the threshold is similar (the choice of $R_{\infty,\text{lim}}^{(C)} = 1.03$ is made according to Table 3.3). Although the computation of $\hat{R}_\infty^{(\max)}$ is possible here as the number of dimensions is small, computing a univariate $\hat{R}_\infty$ on the log posterior instead seems satisfactory here.

3.5 Discussion

In this paper we propose a new version of the Gelman–Rubin diagnostic called $\hat{R}_\infty$, which improves MCMC convergence diagnostics on several aspects. Firstly, it uses a localized version $\hat{R}(x)$ which assesses convergence at a given quantile x of the target distribution. Moreover, it is also based on a theoretical study of what $\hat{R}(x)$ is actually estimating: assuming stationarity to focus only on the mixing property, the population version can be seen as a distance measure between the distributions of the chains. This allows us to obtain convergence properties of $\hat{R}(x)$ and to tune the usual threshold of 1.01 (Section 3.2.3)

based on a given confidence level and on the number of chains. We show theoretically (Section 3.2.5) and using experiments (Section 3.4) that our version is efficient to diagnose convergence. Finally, we suggest a two-step algorithm for a multivariate diagnosis (Section 3.3.1), and reinforce the second step to consider all the directions of the space, as we show that the natural extension cannot be used directly (Section 3.3.3). Therefore, in the high-dimensional case where this computation is likely to be too expensive, we suggest to replace it by a univariate calculation on the log-likelihood or the log-posterior. Diagnosing convergence in the multivariate case remains an open problem, and this is our hope that the local approach advocated here will trigger more research in this direction in the future.

Acknowledgement

We would like to thank a Reviewer and an Editor for providing us with valuable comments that helped us improving the manuscript. Specifically, comments from an Editor allowed us to deal with multi-stage testing in an appealing and satisfactory way. S. Girard acknowledges the support of the Chair Stress Test, Risk Management and Financial Steering, led by the École polytechnique and its Foundation and sponsored by BNP Paribas. J. Arbel acknowledges the support of the French National Research Agency (ANR-21-JSTM-0001).

3.A Construction of rank-$\hat{R}$ false negatives

In this section we detail different ways to construct false negative distributions on the chains that fool rank-$\hat{R}$, in the sense that they return rank-$\hat{R} \approx 1$ with non-identical distributions.

3.A.1 Two-parameter distributions

As mentioned in Section 3.1.2, one way to fool rank-$\hat{R}$ is to fix two constraints on the distributions, which are identical mean and mean over the median on all chains. Many two-parameter distributions can be tuned to respect these constraints. Consider for example a uniform chain $\mathcal{U}(-2\sigma, 2\sigma)$ and another one from a Laplace distribution $\mathcal{L}(0, \sigma)$, where $\sigma > 0$. The resulting R_∞ does not depend on the scale parameter σ and is given by

$$R_\infty = \sqrt{1 + \frac{1}{2(2e^2 - 1)}} \approx 1.018,$$

see Lemma 5. Note that computations are done in the case of $m = 2$ chains, and so the traditional threshold of 1.01 holds for $\hat{R}_\infty$ (see Table 3.2). Therefore we expect $\hat{R}_\infty$ to diagnose convergence, and not $\hat{R}$ nor rank-$\hat{R}$. Results are provided in Figure 3.10 and confirm this hypothesis.

Example 3.10: Laplace $\mathcal{L}(0,\sigma)$ and $\mathcal{U}(-2\sigma, 2\sigma)$

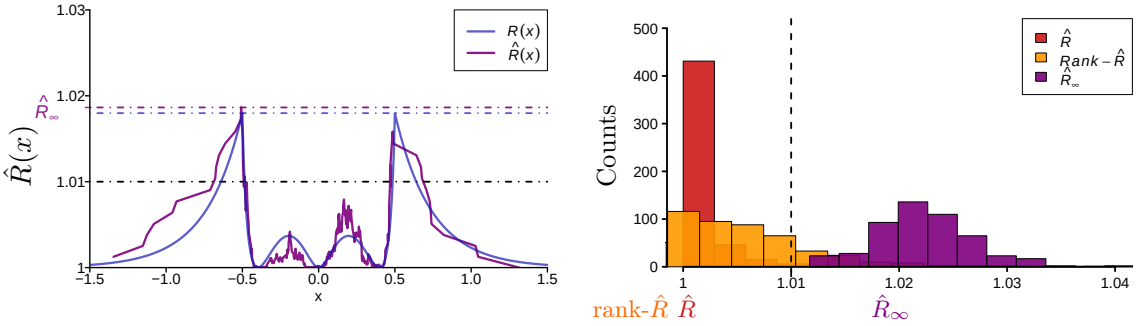


Figure 3.10: Behaviour of $\hat{R}_\infty$ on the case of $m = 2$ chains, $n = 500$ each where F_1 is a Laplace distribution $\mathcal{L}(0, 1/4)$ and F_2 is $\mathcal{U}(-1/2; 1/2)$. On the left: $\hat{R}(x)$ as a function of x for one simulation. On the right: Histograms of 500 replications of $\hat{R}$, rank-$\hat{R}$ of and $\hat{R}_\infty$. The dashed line corresponds to the threshold of 1.01 which holds here for all versions of $\hat{R}$.

3.A.2 General framework with Generalized Pareto Distribution

To find a general way to construct counter-examples, let us consider the Generalised Pareto Distribution (GPD), parametrised by (μ, σ, ξ):

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

with $\{.\}_+ = \max\{0, .\}$. The support of this distribution depends on the parameters: $[\mu; +\infty)$ if $\xi \geq 0$, $[\mu; \mu - \sigma/\xi)$ otherwise. The expectation exists only if $\xi < 1$ and is equal to $\mu + \frac{\sigma}{1-\xi}$, and the median is given by $x_{\text{med}} = \mu + \sigma\frac{2^\xi - 1}{\xi}$.

An interesting property here is that conditioned on exceeding a value, the distribution is still a GPD distribution: If $X \sim \text{GPD}(\mu, \sigma, \xi)$ then $X \mid X > u \sim \text{GPD}(u, \tilde{\sigma}, \xi)$ with $\tilde{\sigma} = \sigma + \xi(u - \mu)$. Therefore, $X \mid X > x_{\text{med}} \sim \text{GPD}(u, \tilde{\sigma}, \xi)$, and

$$\begin{aligned} \mathbb{E}(X \mid X > x_{\text{med}}) &= x_{\text{med}} + \frac{\sigma + \xi(x_{\text{med}} - \mu)}{1 - \xi}, \\ &= \mu + \frac{\sigma}{1-\xi}\left(1 + \frac{2^\xi - 1}{\xi}\right), \\ &= \mathbb{E}(X) + \sigma\frac{(2^\xi - 1)}{\xi(1-\xi)}. \end{aligned}$$

Then, by considering $(m-1)$ chains that follow a $\text{GPD}(\mu_1, \sigma_1, \xi_1)$, and one that follows a $\text{GPD}(\mu_2, \sigma_2, \xi_2)$, we can solve the system of two equations that links the two means and the two means over the medians, to obtain a range of possible parameters:

$$\begin{cases} \mu_1 + \frac{\sigma_1}{1-\xi_1} = \mu_2 + \frac{\sigma_2}{1-\xi_2}, \\ \sigma_1\frac{(2^{\xi_1}-1)}{\xi_1(1-\xi_1)} = \sigma_2\frac{(2^{\xi_2}-1)}{\xi_2(1-\xi_2)}. \end{cases} \tag{3.11}$$

In order to obtain different values of parameters, we should choose $\xi_1 \neq \xi_2$ and $\sigma_1 \neq \sigma_2$. One way to characterize the set of solutions is as follows:

1. Fix ξ_1 and ξ_2 such that $\xi_1 \neq \xi_2$, and define $\lambda = \frac{f(\xi_1)}{f(\xi_2)}$ with $f(\xi) = \frac{(2^\xi - 1)}{\xi(1-\xi)}$.

2. Fix σ_1, and using the second equation of (3.11) define $\sigma_2 = \lambda\sigma_1$.
3. Finally, choose μ_1 and μ_2 such that $\mu_1 - \mu_2 = \sigma_1(\frac{\lambda}{1-\xi_2} - \frac{1}{1-\xi_1})$.

An example of a solution is as follows:

1. Choose $(\mu_1, \sigma_1, \xi_1) = (0, 1, 0)$, the standard exponential distribution Exp(1), and $\xi_2 = -1$, a uniform distribution.
2. Following the method described before, we obtain $\lambda = \frac{f(\xi_1)}{f(\xi_2)} = 4\log(2)$, so $\sigma_2 = \lambda\sigma_1 = 4\log(2)$.
3. Following the last point, we set $\mu_2 = \mu_1 - \sigma_1(\frac{\lambda}{1-\xi_2} - \frac{1}{1-\xi_1}) = 1 - 2\log(2)$.

To conclude, if some chains follow an Exp(1) distribution and the other ones a $\mathcal{U}(1 - 2\log(2); 1+2\log(2))$ distribution, the difference between the chains should not be detected by the rank-$\hat{R}$, which is illustrated in the last column of Figure 3.3. Similarly to the other examples in Section 3.2.5, $\hat{R}_\infty$ manages to diagnose the convergence issue contrary to other versions.

3.B Proofs

3.B.1 Proofs in the univariate case

Proof of Proposition 1. Let$x \in \mathbb{R}$. The within- and between-variances are given by:

$$\begin{aligned}
W(x) &= \mathbb{E}\left[\mathrm{Var}[I_x \mid Z]\right] \\
&= \frac{1}{m}\sum_{j=1}^{m}\left(\mathbb{E}\left[I_x^2 \mid Z = j\right] - \mathbb{E}^2\left[I_x \mid Z = j\right]\right) \qquad (3.12)\\
&= \frac{1}{m}\sum_{j=1}^{m}\left(\mathbb{P}\left(\theta \le x \mid Z = j\right) - \mathbb{P}^2\left(\theta \le x \mid Z = j\right)\right) \\
&= \frac{1}{m}\sum_{j=1}^{m}\left(F_j(x) - F_j^2(x)\right), \qquad (3.13)\\
B(x) &= \mathrm{Var}\left[\mathbb{E}[I_x \mid Z]\right] = \mathbb{E}\left[F_Z(x)^2\right] - \mathbb{E}[F_Z(x)]^2 \\
&= \frac{1}{m}\sum_{j=1}^{m}F_j^2(x) - \left(\frac{1}{m}\sum_{j=1}^{m}F_j(x)\right)^2 \qquad (3.14)\\
&= \frac{m-1}{m^2}\sum_{j=1}^{m}F_j^2(x) - \frac{2}{m^2}\sum_{j<k}F_j(x)F_k(x) \\
&= \frac{1}{m^2}\sum_{j<k}(F_j^2(x) + F_k^2(x) - 2F_j(x)F_k(x)) \\
&= \frac{1}{m^2}\sum_{j<k}\left(F_j(x) - F_k(x)\right)^2, \qquad (3.15)
\end{aligned}$$

where (3.12) and (3.14) are a consequence of $\mathbb{P}(Z = j) = 1/m$ for all $j \in \{1, \cdots, m\}$. The conclusion follows.

Proof of Proposition 2. Proofs of (i) and (ii) are straightforward, let us focus on (iii) and (iv).

(iii): Without loss of generality, we assume that all F_j's are defined on $\mathbb{R}$ and we denote $\bar{F}_j = 1 - F_j$ the associated survival function. From Proposition 1, we can write, for any $x \in \mathbb{R}$:

$$R^2(x) = \frac{\left(\frac{1}{m}\sum_{j=1}^m F_j(x)\right)\left(\frac{1}{m}\sum_{j=1}^m \bar{F}_j(x)\right)}{\frac{1}{m}\sum_{j=1}^m \bar{F}_j(x)F_j(x)}.$$

Denote by $f(x) \simeq g(x)$ two functions f and g that are asymptotically equivalent when $x \to a$, $a \in \mathbb{R} \cup \{-\infty, +\infty\}$. Consider first the case $x \to -\infty$. Clearly, for all $j \in \{1, \dots, m\}$, $F_j(x) \to 0$ and $\bar{F}_j(x) \to 1$ as $x \to -\infty$, so that

$$\left(\frac{1}{m}\sum_{j=1}^m F_j(x)\right)\left(\frac{1}{m}\sum_{j=1}^m \bar{F}_j(x)\right) \simeq \frac{1}{m}\sum_{j=1}^m F_j(x),$$
$$\text{and} \quad \frac{1}{m}\sum_{j=1}^m \bar{F}_j(x)F_j(x) \simeq \frac{1}{m}\sum_{j=1}^m F_j(x),$$

leading to $R(x) \to 1$ as $x \to -\infty$. Second, remarking that $R(x)$ is symmetric with respect to $(F_j(x), \bar{F}_j(x))$, we similarly have $R(x) \to 1$ as $x \to \infty$.

(iv): It is clear that $B(x)$ and $W(x)$ are continuous since the F_j's are continuous, so the only thing to prove is that the denominator $W(x)$ never vanishes, except if extending by continuity is possible. Remarking that for all x, we have $F_j(x)(1-F_j(x)) \geq 0$, $W(x) = 0$ if and only if $F_j(x) = 0$ or $F_j(x) = 1$ for $j \in \{1, \dots, m\}$. Almost all combinations are avoided by the assumption that the supports must overlap, except $F_1(x) = \cdots = F_m(x) = 0$ and $F_1(x) = \cdots = F_m(x) = 1$. In these two latter cases, an extension by continuity is possible as $R(x) = 1$ from (iii).

Proof of Proposition 3. (i): Assume for the sake of simplicity that F is continuous and strictly increasing. Let us show that the corresponding $\hat{R}_\infty$ has the same distribution as in the standard uniform case. In view of (3.2), $\hat{R}_\infty$ can be written as

$$\hat{R}_\infty = \sup_{x\in\mathbb{R}} \hat{R}(x) = \sup_{x\in\mathbb{R}} \sqrt{1 + \frac{\sum_{j=1}^m \sum_{k=j+1}^m \left(\hat{F}_j(x) - \hat{F}_k(x)\right)^2}{m\sum_{j=1}^m \hat{F}_j(x)(1-\hat{F}_j(x))}},$$

where for any $j \in \{1, \dots, m\}$,

$$\hat{F}_j(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{\theta^{(i,j)} \leq x\} = \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{F(\theta^{(i,j)}) \leq F(x)\},$$

and where the random variables $F(\theta^{(i,j)})$ are standard uniformly distributed for any $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, m\}$. Finally, observing that

$$\sup_{x\in\mathbb{R}} \hat{R}(x) = \sup_{y\in[0,1]} \hat{R}(F^{-1}(y))$$

concludes the proof.

(ii): Let $x \in \mathbb{R}$ and remark that (3.4) implies that $\hat{F}_j(x) \xrightarrow{P} F(x)$ as $n \to \infty$ for all $j \in \{1, \dots, m\}$. Then it follows from (3.13) in the proof of Proposition 1 that

$$\hat{W}(x) = \frac{1}{m} \sum_{j=1}^{m} \hat{F}_j(x)(1 - \hat{F}_j(x)) = F(x)(1 - F(x)) + o_p(1). \tag{3.16}$$

Besides, from (3.4) and (3.6), for $j = 1, \dots, m$, $\hat{F}_j(x)$ can be expanded as

$$\hat{F}_j(x) = F(x) + \sqrt{\frac{F(x)(1 - F(x))}{\mathrm{ESS}(x)/m}} \xi_{j,n},$$

with $\xi_{j,n} \xrightarrow{d} \mathcal{N}(0, 1)$ as $n \to \infty$. Thus (3.15) in the proof of Proposition 1 entails

$$\hat{B}(x) = \frac{1}{m^2} \sum_{j<k} \left(\hat{F}_j(x) - \hat{F}_k(x) \right)^2 = \frac{F(x)(1 - F(x))}{\mathrm{ESS}(x)} \times \frac{1}{m} \sum_{j<k} (\xi_{j,n} - \xi_{k,n})^2 . \tag{3.17}$$

Introducing the random vector $\boldsymbol{\xi}_n = (\xi_{1,n}, \dots, \xi_{m,n})^\top \xrightarrow{d} \mathcal{N}(0, \boldsymbol{I}_m)$ where $\boldsymbol{I}_m$ is the $m \times m$ identity matrix, one has

$$\frac{1}{m} \sum_{j<k} (\xi_{j,n} - \xi_{k,n})^2 = \frac{m-1}{m} \sum_{j=1}^{m} \xi_{j,n}^2 - \frac{1}{m} \sum_{j \neq k} \xi_{j,n} \xi_{k,n} = \boldsymbol{\xi}_n^\top \boldsymbol{A} \boldsymbol{\xi}_n,$$

with $\boldsymbol{A} = \boldsymbol{I}_m - \boldsymbol{J}_m/m$, and where $\boldsymbol{J}_m$ is the $m \times m$ matrix filled with ones. The symmetric matrix $\boldsymbol{A}$ can be eigen-decomposed as $\boldsymbol{A} = \boldsymbol{Q}^\top \boldsymbol{\Lambda} \boldsymbol{Q}$ with $\boldsymbol{Q}$ an orthogonal matrix and $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \cdots, \lambda_d) = \mathrm{diag}(1, \dots, 1, 0)$. Remark that $\boldsymbol{U}_n \coloneqq \boldsymbol{Q} \boldsymbol{\xi}_n \xrightarrow{d} \mathcal{N}(0, \boldsymbol{I}_m)$ so that

$$\frac{1}{m} \sum_{j<k} (\xi_{j,n} - \xi_{k,n})^2 = (\boldsymbol{Q} \boldsymbol{\xi}_n)^\top \boldsymbol{\Lambda} (\boldsymbol{Q} \boldsymbol{\xi}_n) = \sum_{j=1}^{m} \lambda_j U_{j,n}^2 = \sum_{j=1}^{m-1} U_{j,n}^2 \xrightarrow{d} \chi^2_{m-1}, \tag{3.18}$$

as $n \to \infty$. Collecting (3.16), (3.17) and (3.18) yields

$$\mathrm{ESS}(x)(\hat{R}^2(x) - 1) = \mathrm{ESS}(x) \frac{\hat{B}(x)}{\hat{W}(x)} \xrightarrow{d} \chi^2_{m-1},$$

as $n \to \infty$ and the result is proved.

3.B.2 Proofs in the multivariate case

Lemma 1 (Standardization of margins)**.** *Assume the assumptions of Proposition 1 hold. If the margins of $F_1, \dots, F_m$ coincide, then the multivariate R_∞ is the same as the one calculated on the associated copulas $C_1, \dots, C_m$.*

Proof. Denote by $\phi_1, \dots, \phi_d$ the common margins of the cdf's $F_1, \dots, F_m$, so that $F_j(\boldsymbol{x}) = C_j(\phi_1(x_1), \dots, \phi_d(x_d))$ for any $j \in \{1, \dots, m\}$. Letting $\boldsymbol{y} = (\phi_1(x_1), \dots, \phi_d(x_d))$, we have

$$\sup_{\boldsymbol{x} \in \mathbb{R}^d} R(\boldsymbol{x}) = \sup_{\boldsymbol{y} \in [0,1]^d} R(\phi_1^{-1}(y_1), \dots, \phi_d^{-1}(y_d)).$$

Besides, in view of (3.8), $m(R^2(\phi_1^{-1}(y_1),\dots,\phi_d^{-1}(y_d))-1)$ can be written as

$$\frac{\sum_{j=1}^m\sum_{k=j+1}^m\Big(F_j(\phi_1^{-1}(y_1),\dots,\phi_d^{-1}(y_d))-F_k(\phi_1^{-1}(y_1),\dots,\phi_d^{-1}(y_d))\Big)^2}{\sum_{j=1}^m F_j(\phi_1^{-1}(y_1),\dots,\phi_d^{-1}(y_d))(1-F_j(\phi_1^{-1}(y_1),\dots,\phi_d^{-1}(y_d)))}$$
$$=\frac{\sum_{j=1}^m\sum_{k=j+1}^m\left(C_j(y_1,\dots,y_d)-C_k(y_1,\dots,y_d)\right)^2}{\sum_{j=1}^m C_j(y_1,\dots,y_d)(1-C_j(y_1,\dots,y_d))},$$

which proves the result.

Proof of Lemma 1. Let function $\tilde{R}$ be defined on $[0,1]^2$ by

$$\tilde{R}(c_1,c_2)=\sqrt{1+\frac{1}{2}\frac{(c_1-c_2)^2}{c_1(1-c_1)+c_2(1-c_2)}}. \tag{3.19}$$

For any $c_1\in[0,1]$, $c_2\mapsto\tilde{R}(c_1,c_2)$ is decreasing on $[0,c_1]$, and increasing on $[c_1,1]$. Let $(c_-,c_+)\in[0,1]^2$ and $(c_1,c_2)\in[c_-,c_+]^2$. Without loss of generality, assume that $c_1\le c_2$. Let $c\in[0,c_2]$. Since $c_2\le c_+$ and $\tilde{R}(c,\cdot)$ is increasing on $[c,1]$, we have $\tilde{R}(c,c_2)\le\tilde{R}(c,c_+)$. In particular, for $c=c_-$:

$$\tilde{R}(c_-,c_2)\le\tilde{R}(c_-,c_+). \tag{3.20}$$

Moreover, $\tilde{R}(c_2,\cdot)$ is decreasing on $[0,c_2]$ and $c_-\le c_1\le c_2$, we also have

$$\tilde{R}(c_-,c_2)=\tilde{R}(c_2,c_-)\ge\tilde{R}(c_2,c_1). \tag{3.21}$$

Combining (3.20) and (3.21), we finally obtain $\tilde{R}(c_2,c_1)\le\tilde{R}(c_-,c_+)$, which concludes the proof.

Proof of Proposition 4. By definition of lower and upper Fréchet–Hoeffding bounds (Nelsen, 2006), $W_d(\boldsymbol{u})\le C(\boldsymbol{u})\le M_d(\boldsymbol{u})$ for any copula C and $\boldsymbol{u}\in[0,1]^d$. Thus in view of Lemma 1 it only remains to prove that $R_\infty(W_d,M_d)=\sqrt{(d+1)/2}$. To this end, let p be the index such that $u_p=\min\{u_1,\dots,u_d\}$. Two cases arise:

(i) If $1-d+\sum_{i=1}^d u_i\le 0$, then

$$f(\boldsymbol{u}):=2(R^2(\boldsymbol{u})-1)=\frac{u_p^2}{u_p(1-u_p)}=\frac{1}{1/u_p-1}. \tag{3.22}$$

As a consequence, $R^2(\boldsymbol{u})$ is maximum when u_p is maximum under the constraints

$$\begin{cases}u_p\le u_i & \forall i\ne p,\\ u_p\le d-1-\sum_{i\ne p}u_i.\end{cases}$$

It is easily seen that the maximum occurs in the equality case $u_1=\cdots=u_d=(d-1)/d$ and thus $2(R^2(\boldsymbol{u})-1)=d-1$.

(ii) Conversely, if $1-d+\sum_{i=1}^d u_i\ge 0$, then

$$f(\boldsymbol{u})=\frac{\left(1-d+\sum_{i\ne p}u_i\right)^2}{u_p(1-u_p)+(1-d+\sum_{i=1}^d u_i)(d-\sum_{i=1}^d u_i)}.$$

The problem therefore amounts to maximising $f(\boldsymbol{u})$ under the constraints

$$\begin{cases} \boldsymbol{u} \in [0,1]^d, \\ u_p \leq u_i \quad \forall i \neq p, \\ 1 - d + \sum_{i=1}^d u_i \geq 0. \end{cases} \tag{3.23}$$

The associated Lagrangian can be written with $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \boldsymbol{\lambda}_3 \in \mathbb{R}^d$ and $\lambda_4 \in \mathbb{R}$ as:

$$\mathfrak{L}(\boldsymbol{u}, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \boldsymbol{\lambda}_3, \lambda_4) = f(\boldsymbol{u}) + \sum_{i=1}^d \lambda_{1,i} u_i + \sum_{i=1}^d \lambda_{2,i}(1-u_i) + \sum_{i \neq p} \lambda_{3,i}(u_i - u_p) + \lambda_4 \left(1 - d + \sum_{i=1}^d u_i\right).$$

The first-order conditions are given by (3.23) and $\nabla_{\boldsymbol{u}} \mathfrak{L}(\boldsymbol{u}, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \boldsymbol{\lambda}_3, \lambda_4) = 0$, and the Karush–Kuhn–Tucker conditions are

$$\begin{cases} \lambda_{1,i} \geq 0 \quad \text{with} \quad \lambda_{1,i} u_i = 0, \quad \forall i \in \{1, \dots, d\}, \\ \lambda_{2,i} \geq 0 \quad \text{with} \quad \lambda_{2,i}(1 - u_i) = 0, \quad \forall i \in \{1, \dots, d\}, \\ \lambda_{3,i} \geq 0 \quad \text{with} \quad \lambda_{3,i}(u_i - u_p) = 0, \quad \forall i \in \{1, \dots, d\} \quad \text{s.t.} \quad i \neq p, \\ \lambda_4 \geq 0 \quad \text{with} \quad \lambda_4 \left(1 - d + \sum_{i=1}^d u_i\right) = 0. \end{cases} \tag{3.24}$$

We distinguish different cases:

- If $\lambda_{1,i_0} \neq 0$ for some $i_0 \in \{1, \dots, d\}$, then $u_{i_0} = 0$ and combined with (3.23) we obtain necessarily $i_0 = p$ and $1 - d + \sum_{i=1}^d u_i = 0$, leading to a non-optimal solution for $f(\boldsymbol{u})$.
- If $\lambda_{2,i_0} \neq 0$ for some $i_0 \in \{1, \dots, d\}$, then $u_{i_0} = 1$, and the problem is exactly the same written in dimension $d-1$, and a recurrence proves that the maximum is equal to $\sqrt{(d+1)/2}$ in dimension d. So the maximum is increasing with the dimension and therefore is not reached in this case.

One can thus assume $\boldsymbol{\lambda}_1 = 0$ and $\boldsymbol{\lambda}_2 = 0$. Moreover, for all (i, j) such that $i \neq p$ and $j \neq p$, we have $\frac{\partial f}{\partial u_i} = \frac{\partial f}{\partial u_j}$. Combined with $\nabla_{\boldsymbol{u}} \mathfrak{L} = 0$, we obtain $\lambda_{3,i} = \lambda_{3,j}$, which leads to considering the simplified Lagrangian:

$$\mathfrak{L}(\boldsymbol{u}, \lambda_3, \lambda_4) = f(\boldsymbol{u}) + \lambda_3 \left(\sum_{i=1}^d u_i - d u_p\right) + \lambda_4 \left(1 - d + \sum_{i=1}^d u_i\right).$$

In that form, the function f and the constraints of the Lagrangian can be written only as a function of $(1 - d + \sum_{i=1}^d u_i, u_p, \lambda_3, \lambda_4)$. Let $\tilde{\mathfrak{L}}(x = 1 - d + \sum_{i=1}^d u_i, y = u_p, \lambda_3, \lambda_4) = \mathfrak{L}(\boldsymbol{u}, \lambda_3, \lambda_4)$. Since for all i

$$\frac{\partial \mathfrak{L}}{\partial u_i} = \frac{\partial \tilde{\mathfrak{L}}}{\partial x} \frac{\partial x}{\partial u_i} + \frac{\partial \tilde{\mathfrak{L}}}{\partial y} \frac{\partial y}{\partial u_i} = \frac{\partial \tilde{\mathfrak{L}}}{\partial x} + \frac{\partial \tilde{\mathfrak{L}}}{\partial y} \mathbb{I}\{i = p\},$$

solving $\nabla_{\boldsymbol{u}} \mathfrak{L} = 0$ is equivalent to solving $\nabla_{(x,y)} \tilde{\mathfrak{L}} = 0$. The corresponding problem is therefore

$$\max \frac{(x-y)^2}{x(1-x) + y(1-y)}, \quad \text{under the constraints} \quad \begin{cases} x \geq \max\{0, 1 - d + dy\}, \\ x \leq y \leq 1. \end{cases} \tag{3.25}$$

Combining the constraints with the study of function in the proof of Lemma 1 leads to the solution $(x, y) = (0, \frac{d-1}{d})$. So

$$\sum_{i=1}^{d} u_i = d - 1 \quad \text{and} \quad \min\{u_1, \ldots, u_d\} = \frac{d-1}{d},$$

so necessarily $u_1 = \cdots = u_d = \frac{d-1}{d}$, which concludes the proof.

Proof of Corollary 1 For all $\boldsymbol{u} \in [0,1]^d$, we have

$$\begin{aligned} m\left(R^2(\boldsymbol{u}) - 1\right) &= \sum_{j=1}^{m}\sum_{k=j+1}^{m} \frac{(C_j(\boldsymbol{u}) - C_k(\boldsymbol{u}))^2}{\sum_{\ell=1}^{m} C_\ell(\boldsymbol{u})(1 - C_\ell(\boldsymbol{u}))}, \\ &\leq \sum_{j=1}^{m}\sum_{k=j+1}^{m} \frac{(C_j(\boldsymbol{u}) - C_k(\boldsymbol{u}))^2}{C_j(\boldsymbol{u})(1 - C_j(\boldsymbol{u})) + C_k(\boldsymbol{u})(1 - C_k(\boldsymbol{u}))}, \\ &= \sum_{j=1}^{m}\sum_{k=j+1}^{m} 2\left(R^2(W_d(\boldsymbol{u}), M_d(\boldsymbol{u})) - 1\right), \end{aligned}$$

so that an upper bound on the multivariate R_∞ can be expressed thanks to bivariate as follows:

$$\begin{aligned} m\left(R_\infty^2(C_1, \ldots, C_m) - 1\right) &\leq \sum_{j=1}^{m}\sum_{k=j+1}^{m} 2\left(R_\infty^2(C_j, C_k) - 1\right), \\ &\leq \sum_{j=1}^{m}\sum_{k=j+1}^{m} 2\left(R_\infty^2(W_d, M_d) - 1\right), \\ &= \frac{m(m-1)}{2}(d-1), \end{aligned}$$

using Lemma 1 and Proposition 4. The result is thus proved.

Proof of Corollary 2. One can prove that the maximum of $\boldsymbol{u} \in [0,1]^d \mapsto R(\Pi_d(\boldsymbol{u}), M_d(\boldsymbol{u}))$ is reached at $u_1 = \cdots = u_d \coloneqq u \in [0,1]$, which leads to studying the maximum of

$$f_d(u) \coloneqq 2(R^2(u, \ldots, u) - 1) = \frac{(u^d - u)^2}{u^d(1 - u^d) + u(1 - u)}.$$

The first derivative of f_d is proportional to

$$g_d(u) = -2(d-1)u^{2d-1} + du^{2d-2} - 2(d-1)u^d + 3(d-1)u^{d-1} - 1.$$

Routine calculations show the existence of a unique root in $[0,1]$, but finding the explicit value does not seem possible when $d > 2$ since g_d is a polynomial of order $2d-1$. We thus restrict ourselves to an asymptotic analysis when $d \to \infty$. First, Lemma 1 and Proposition 4 entail that

$$\max_{u \in [0,1]} f_d(u) \leq 2(R_\infty^2(W_d, M_d) - 1) = d - 1.$$

Second, a lower bound can be obtained by letting $u_d = 1 - (\log d)/d$. Indeed, $u_d \to 1$ and $u_d^d = \exp(-(\log d)(1+o(1))) \to 0$ so that the numerator of $f_d(u)$ satisfies $(u_d^d - u_d)^2 \to 1$ as $d \to \infty$. Moreover, the denominator satisfies

$$u_d^d(1-u_d^d)+u_d(1-u_d) = \exp(-\log(d)(1+o(1)))\,(1+o(1))+\frac{\log d}{d}(1+o(1)) = \frac{\log d}{d}(1+o(1)),$$

as $d \to \infty$. As a consequence, $f_d(u_d) = \frac{d}{\log d}(1+o(1))$ and we have proved that

$$\frac{d}{\log d}(1+o(1)) \leq \max_{u\in[0,1]} f_d(u) \leq d-1.$$

The result follows.

Proof of Corollary 3. From Proposition 1 and the definition of a NLOD copula, the proof reduces to calculating $R_\infty(W_d, \Pi_d)$. Two cases are considered.

(i) First, if $1 - d + \sum_{i=1}^d u_i \leq 0$, then $W_d(\boldsymbol{u}) = 0$ and

$$2(R^2(\boldsymbol{u}) - 1) = \frac{1}{\frac{1}{\prod_{i=1}^d u_i} - 1}.$$

The maximisation of $2(R^2(\boldsymbol{u}) - 1)$ is then equivalent to solving:

$$\max \prod_{i=1}^d u_i, \quad \text{under the constraints} \quad \begin{cases} \boldsymbol{u} \in [0,1]^d, \\ 1 - d + \sum_{i=1}^d u_i \leq 0. \end{cases} \tag{3.26}$$

Since the constraints are linear and the objective function is convex, the above optimization problem is convex. The Lagrangian associated with (3.26) can be written with $\lambda_1 \in \mathbb{R}$, $\boldsymbol{\lambda}_2, \boldsymbol{\lambda}_3 \in \mathbb{R}^d$, as:

$$\mathfrak{L}(\boldsymbol{u}, \lambda_1, \boldsymbol{\lambda}_2, \boldsymbol{\lambda}_3) = \prod_{i=1}^d u_i - \lambda_1 \left(1 - d + \sum_{i=1}^d u_i\right) - \sum_{i=1}^d \lambda_{2,i}(u_i - 1) + \sum_{i=1}^d \lambda_{3,i} u_i,$$

The first-order conditions are

$$\begin{cases} \nabla_{\boldsymbol{u}} \mathfrak{L}(\boldsymbol{u}, \lambda_1, \boldsymbol{\lambda}_2, \boldsymbol{\lambda}_3) = 0, \\ 0 \leq u_i \leq 1 \quad \forall i \in \{1, \ldots, d\}, \\ 1 - d + \sum_{i=1}^d u_i \leq 0, \end{cases} \tag{3.27}$$

and the Karush–Kuhn–Tucker conditions are

$$\begin{cases} \lambda_1 \geq 0 \quad \text{with} \quad \lambda_1(1 - d + \sum_{i=1}^d u_i) = 0, \\ \lambda_{2,i} \geq 0 \quad \text{with} \quad \lambda_{2,i}(u_i - 1) = 0, \quad \forall i \in \{1, \ldots, d\}, \\ \lambda_{3,i} \geq 0 \quad \text{with} \quad \lambda_{3,i} u_i = 0, \quad \forall i \in \{1, \ldots, d\}. \end{cases} \tag{3.28}$$

If there exists i_0 such that $u_{i_0} = 0$, then $\prod_{i=1}^d u_i = 0$, which is clearly non-optimal. Thus, (3.28) implies $\lambda_{3,i} = 0$ for all $i \in \{1, \ldots, d\}$. Moreover, for all $i \in \{1, \ldots, d\}$,

$$\frac{\partial \mathcal{L}}{\partial u_i} = 0 \implies \lambda_1 + \lambda_{2,i} = \prod_{j \neq i} u_j,$$

so that, for all $(j,k) \in \{1,\dots,d\}^2$,

$$(\lambda_1 + \lambda_{2,j})u_j = (\lambda_1 + \lambda_{2,k})u_k. \tag{3.29}$$

If there exists i_0 such that $\lambda_{2,i_0} \neq 0$, then $u_{i_0} = 1$ from (3.28) and consequently, for all $k \neq i_0$,

$$u_k = \frac{\lambda_1 + \lambda_{2,i_0}}{\lambda_1 + \lambda_{2,k}}.$$

Taking account of (3.27) yields $u_k \leq 1$ which implies in turn $\lambda_{2,k} \geq \lambda_{2,i_0} \neq 0$ and $u_k = 1$ for all $k \in \{1,\dots,p\}$ from (3.28). The resulting $\boldsymbol{u}$ does not fulfil the third constraint in (3.27). Then, necessarily, $\lambda_{2,i} = 0$ for all $i \in \{1,\dots,d\}$, and combining with (3.29), it follows that the optimum is reached when $u_1 = \cdots = u_d = u$. Replacing in (3.26) yields the optimization problem

$$\max u^d, \quad \text{under the constraints} \quad \begin{cases} 0 \leq u \leq 1, \\ 1 + d(u-1) \leq 0, \end{cases} \tag{3.30}$$

whose solution is $u = (d-1)/d$.

(ii) Second, if $1 - d + \sum_{i=1}^d u_i \geq 0$, the problem to solve is

$$\max f_d(\boldsymbol{u}) \coloneqq \frac{\left(\prod_{i=1}^d u_i - 1 + d - \sum_{i=1}^d u_i\right)^2}{\left(\prod_{i=1}^d u_i\right)\left(1 - \prod_{i=1}^d u_i\right) + \left(1 - d + \sum_{i=1}^d u_i\right)\left(d - \sum_{i=1}^d u_i\right)},$$

$$\text{under the constraints} \quad \begin{cases} \boldsymbol{u} \in [0,1]^d, \\ 1 - d + \sum_{i=1}^d u_i \geq 0. \end{cases}$$

The Lagrangian can be written, with $\lambda_1 \in \mathbb{R}$, $\boldsymbol{\lambda}_2, \boldsymbol{\lambda}_3 \in \mathbb{R}^d$:

$$\mathfrak{L}(\boldsymbol{u}, \lambda_1, \boldsymbol{\lambda}_2, \boldsymbol{\lambda}_3) = f_d(\boldsymbol{u}) + \lambda_1 \left(1 - d + \sum_{i=1}^d u_i\right) + \sum_{i=1}^d \lambda_{2,i}(1 - u_i) + \sum_{i=1}^d \lambda_{3,i} u_i.$$

In the same way as in the proof of Proposition 4, we can focus on the solution such that $\boldsymbol{\lambda}_2 = 0$ and $\boldsymbol{\lambda}_3 = 0$. The first order condition $\nabla_{\boldsymbol{u}} \mathfrak{L} = 0$ leads to the solution

$$u_j = \prod_{i=1}^d u_i \times \frac{2r(\boldsymbol{u}) - \left(\prod_{i=1}^d u_i - 1 + d - \sum_{i=1}^d u_i\right)\left(1 - 2\prod_{i=1}^d u_i\right)}{2r(\boldsymbol{u}) + \left(\prod_{i=1}^d u_i - 1 + d - \sum_{i=1}^d u_i\right)\left(1 - 2(1 - d + \sum_{i=1}^d u_i)\right) - \lambda_1},$$

for $j \in \{1,\dots,d\}$ with $r(\boldsymbol{u}) = \prod_{i=1}^d u_i(1 - \prod_{i=1}^d u_i) + (1 - d + \sum_{i=1}^d u_i)(d - \sum_{i=1}^d u_i)$. Note that from this expression, the maximum verifies $u_1 = \cdots = u_d$, and so the initial d-dimensional optimization problem amounts to the one-dimensional problem:

$$\max f_d(u) \coloneqq \frac{\left(u^d - 1 + d - du\right)^2}{u^d(1-u^d) + (1 - d + du)(d - du)}, \quad \text{under the constraints} \quad \begin{cases} u \in [0,1]^d, \\ u \geq \frac{d-1}{d}. \end{cases} \tag{3.31}$$

Iterated derivative computations allow to show that f_d is a decreasing function on $[\frac{d-1}{d}, 1]$, so the maximum is reached at $u = \frac{d-1}{d}$, which concludes the proof.

Proof of Corollary 4. Let us denote by $R(\mathbb{I}\{\theta_1^{(\cdot)} \leq x_1, \theta_2^{(\cdot)} \leq x_2\})$ and $R(\mathbb{I}\{\theta_1^{(\cdot)} \leq x_1, \theta_2^{(\cdot)} \geq x_2\})$ the two considered versions of $R(\boldsymbol{x})$ in the bivariate case, with standard uniform margins. We clearly have:

$$\begin{aligned} R_\infty^- &= \max_{(x_1,x_2)\in[0,1]^2} R\left(\mathbb{I}\left\{\theta_1^{(\cdot)} \leq x_1, \theta_2^{(\cdot)} \geq x_2\right\}\right) \\ &= \max_{(x_1,x_2)\in[0,1]^2} R\left(\mathbb{I}\left\{\theta_1^{(\cdot)} \leq x_1, 1-\theta_2^{(\cdot)} \leq 1-x_2\right\}\right) \\ &= \max_{(x_1,x_2)\in[0,1]^2} R\left(\mathbb{I}\left\{\theta_1^{(\cdot)} \leq x_1, 1-\theta_2^{(\cdot)} \leq x_2\right\}\right). \end{aligned}$$

Then, remarking that

$$\begin{aligned} (\theta_1^{(\cdot)}, \theta_2^{(\cdot)}) \sim M_2 &\implies (\theta_1^{(\cdot)}, 1-\theta_2^{(\cdot)}) \sim W_2, \\ (\theta_1^{(\cdot)}, \theta_2^{(\cdot)}) \sim W_2 &\implies (\theta_1^{(\cdot)}, 1-\theta_2^{(\cdot)}) \sim M_2, \\ (\theta_1^{(\cdot)}, \theta_2^{(\cdot)}) \sim \Pi_2 &\implies (\theta_1^{(\cdot)}, 1-\theta_2^{(\cdot)}) \sim \Pi_2 \end{aligned}$$

proves the result.

3.C Threshold estimation for $\hat{R}_\infty$

This section details the computation of the empirical quantiles of $\hat{R}_\infty$ that is done to obtain Table 3.2.

3.C.1 Univariate case

Invariance on the underlying distribution under the null hypothesis (all chains have the same distribution). A primary step to compute the quantiles of $\hat{R}_\infty$ when all the chain distributions are identical is to verify that such quantiles are well-defined, in the sense that they do not depend on the choice of chain distribution. This property is expected as using a supremum over the quantiles provides invariance to bijective transformations (see Proposition 3(i)). The first row of Figure 3.11 illustrates the behaviour of $\hat{R}$, rank-$\hat{R}_\infty$, and $\hat{R}_\infty$ on two cases:

- all chains are uniform;
- all chains are Pareto distributed.

The QQ-plots seem to confirm the invariance in distribution of rank-$\hat{R}$ and $\hat{R}_\infty$ for various choices of m (see the yellow and violet dots in the first row of Figure 3.11). This was also expected for rank-$\hat{R}$ because of the use of a rank-normalization step. For the traditional $\hat{R}$, the QQ-plots show a difference of distribution (see the red dots in the first row of Figure 3.11), which makes the quantile estimation ill-defined if the chains distribution is not provided. The rest of Figure 3.11 shows histograms of replications for $\hat{R}$ (second row), rank-$\hat{R}$ (third row) and $\hat{R}_\infty$ (fourth row), when the chains are distributed according to the same uniform and Pareto distributions. Histograms for $\hat{R}$ confirm the difference of behaviour when the chain distribution is uniform or Pareto, while the histograms overlap much more for rank-$\hat{R}$ and $\hat{R}_\infty$.

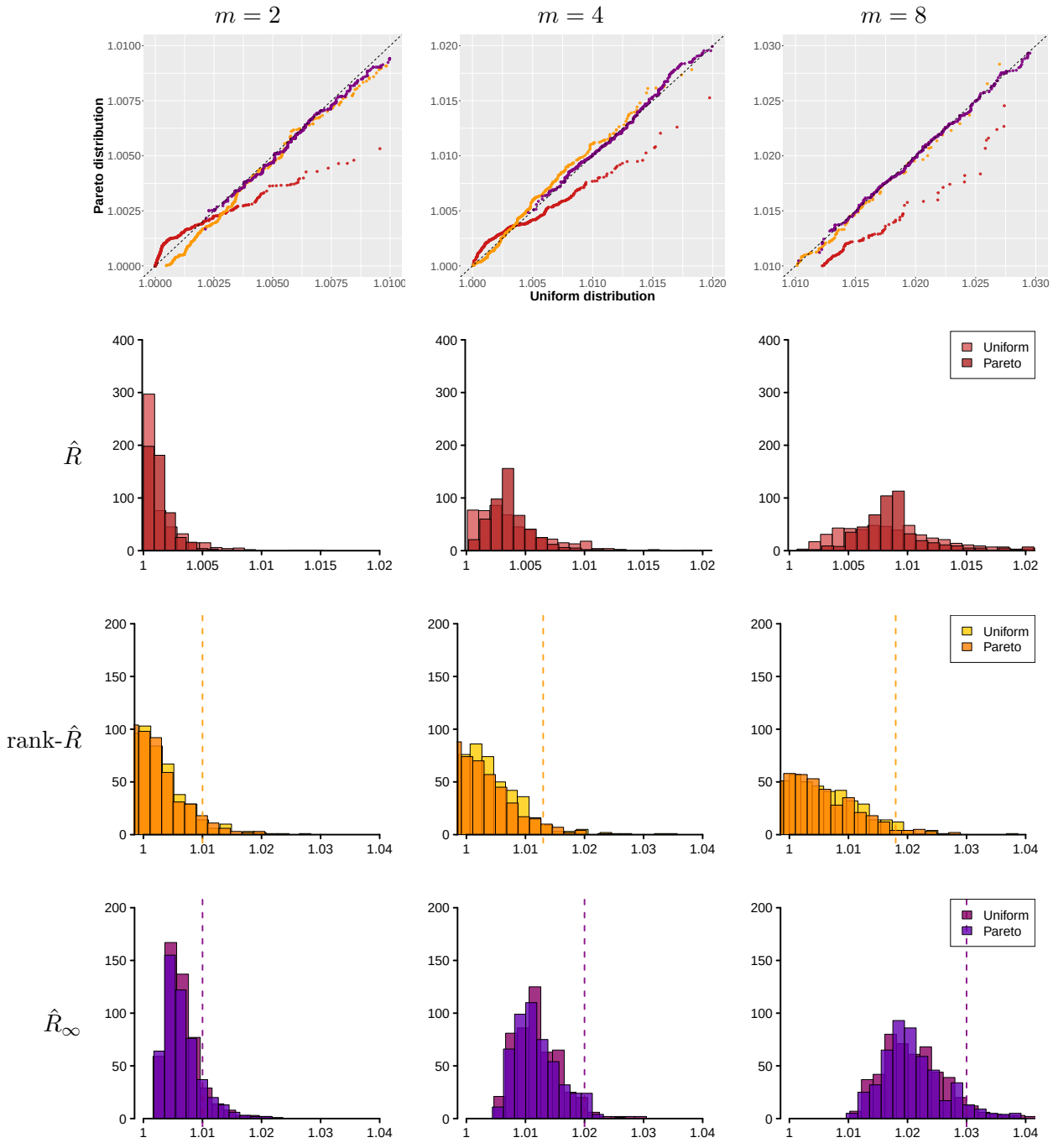


Figure 3.11: Study of $\hat{R}$, rank-$\hat{R}$ and $\hat{R}_\infty$ when all the distributions are the same, for a number of chains $m \in \{2, 4, 8\}$. First row: Q-Q plot that compares 500 replications computed on m uniform chains $\mathcal{U}(-1, 1)$ with m Pareto($\alpha = 0.8, \eta = 1$) ones. Second, third and fourth rows: Histograms of 500 replications with $n \in \{200, 100, 50\}$ for $\hat{R}$ (second row), rank-$\hat{R}$ (third row), and $\hat{R}_\infty$(fourth row) in the case of uniform and Pareto chains. The dashed lines represent the suggested thresholds for rank-$\hat{R}$ and $\hat{R}_\infty$, corresponding to a confidence level of approximately 95%.

Threshold elicitation. As a direct consequence of Proposition 3(i), the empirical quantiles of $\hat{R}_\infty$ only depend on the number of chains m and the length n. The same holds true for rank-$\hat{R}$. Therefore, we are able to estimate empirical quantiles with replications of $\hat{R}_\infty$ using any chain distribution (typically uniform): results are reported in Table 3.2 for a fixed value of $mn = 400$, which is the effective sample size as we are in the i.i.d case. The same estimation can be done for rank-$\hat{R}$, in order to associate a choice of threshold to a confidence level $1 - \alpha$. Consequently, this viewpoint implies a threshold that depends on m for rank-$\hat{R}$ too, as the empirical quantiles change with m (third row of Figure 3.11). Although Vehtari et al. (2021) do not suggest any tuning with respect to the number of chains, using a threshold of 1.01 can lead to large type I error: for example, a threshold of 1.01 leads to approximately 5% when $m = 2$, but to 21% when $m = 8$ (third row of Figure 3.11). Therefore a threshold of 1.01 seems too strong for rank-$\hat{R}$ when m increases: the empirical quantile at $\alpha = 0.05$ suggests to use 1.01 when $m = 2$, 1.013 when $m = 4$ and 1.018 when $m = 8$. The same observation holds for $\hat{R}_\infty$, and looking at Table 3.2 and the fourth row of Figure 3.11 leads to a threshold of 1.01 for $m = 2$, 1.02 for $m = 4$ and 1.03 for $m = 8$ in order to keep a type I error of approximately 5%.

3.C.2 Multivariate case

In the multivariate extension (see Section 3.3), two thresholds have to be elicited: $R^{(M)}_{\infty,\text{lim}}$ for the convergence of margins, and $R^{(C)}_{\infty,\text{lim}}$ for the convergence of the copula. Focusing on $R^{(M)}_{\infty,\text{lim}}$, the quantiles of $\hat{R}_\infty$ for the margins are the same as in the univariate case (given in Table 3.2), but a Bonferroni correction is necessary to take into account the multiplicity of tests. For the copula, $\hat{R}^{(\max)}_\infty$ is a maximum of multiple versions of $\hat{R}_\infty$ computed on different directions of dependence, so its quantiles are different from $\hat{R}_\infty$ ones. However, to reduce the calculation cost, one can compute the quantiles of $\hat{R}_\infty$ under the null hypothesis for the chains, with a Bonferroni correction with 2^{d-1} hypotheses. We estimate the corresponding quantiles using replications to determine the two thresholds $R^{(M)}_{\infty,\text{lim}}$ and $R^{(C)}_{\infty,\text{lim}}$. Here, the computation is done for several values of d, with d relatively small. Results are reported in Table 3.3. Values in bold confirm the rule of thumb given in Section 3.3: $(R^{(M)}_{\infty,\text{lim}}, R^{(C)}_{\infty,\text{lim}}) = (1.03, 1.03)$ for $m = 4$ and $(R^{(M)}_{\infty,\text{lim}}, R^{(C)}_{\infty,\text{lim}}) = (1.04, 1.05)$ for $m = 8$.

3.D Examples of closed-form $R(x)$ and R_∞

We start by a lemma providing useful tools to simplify the calculation of $R(x)$ and R_∞ in the univariate case and when all chains but one have converged. We then review families of distributions for which $R(x)$ and R_∞ can be computed in closed-form.

Lemma 2. *Assume the assumptions of Proposition 1 hold with* $F \coloneqq F_1 = \cdots = F_{m-1} \neq F_m$.

(i) *Then (3.2) can be simplified as*

$$R(x) = \sqrt{1 + \frac{(m-1)(F(x) - F_m(x))^2}{m\left((m-1)F(x)(1-F(x)) + F_m(x)(1-F_m(x))\right)}}. \tag{3.32}$$

		$R^{(M)}_{\infty,\lim}$					$R^{(C)}_{\infty,\lim}$			
		α					α			
d	m	0.005	0.01	0.05	0.1	m	0.005	0.01	0.05	0.1
2	2	1.017	1.017	1.015	1.013	2	1.026	1.026	1.019	1.016
	3	1.024	1.023	1.019	1.018	3	1.029	1.028	1.024	1.021
	4	1.030	1.028	**1.025**	1.022	4	1.033	1.030	**1.026**	1.024
	8	1.041	1.041	**1.037**	1.034	8	1.052	1.050	**1.040**	1.038
3	2	1.022	1.021	1.018	1.016	2	1.022	1.020	1.019	1.018
	3	1.032	1.031	1.023	1.020	3	1.031	1.028	1.025	1.023
	4	1.032	1.031	**1.026**	1.023	4	1.038	1.036	**1.030**	1.027
	8	1.045	1.043	**1.037**	1.036	8	1.052	1.049	**1.047**	1.043
4	2	1.025	1.023	1.016	1.015	2	1.029	1.026	1.022	1.021
	3	1.027	1.026	1.022	1.020	3	1.028	1.028	1.026	1.024
	4	1.027	1.026	**1.025**	1.023	4	1.035	1.034	**1.033**	1.030
	8	1.047	1.044	**1.040**	1.038	8	1.051	1.050	**1.048**	1.047
5	2	1.025	1.020	1.018	1.016	2	1.024	1.024	1.021	1.021
	3	1.027	1.027	1.025	1.022	3	1.028	1.028	1.026	1.024
	4	1.030	1.029	**1.026**	1.024	4	1.043	1.043	**1.040**	1.036
	8	1.053	1.053	**1.038**	1.036	8	1.049	1.049	**1.048**	1.048
6	2	1.022	1.021	1.018	1.015	2	1.020	1.020	1.019	1.018
	3	1.024	1.023	1.021	1.021	3	1.028	1.028	1.025	1.024
	4	1.030	1.030	**1.030**	1.024	4	1.035	1.035	**1.034**	1.033
	8	1.047	1.046	**1.039**	1.037	8	1.059	1.059	**1.058**	1.055

Table 3.3: Left: Empirical quantiles of $\hat{R}_\infty$ for the margins with a Bonferroni correction, under the null hypothesis that all chains have the same distribution, for a fixed value of $mn = 400$. Right: Empirical quantiles of $\hat{R}^{(\max)}_\infty$ for the copula, for a fixed value of $mn = 400$. We have used 500 replications for estimation. Values in bold justify the rule of thumb proposed in Section 3.3.1 for $m \in \{4, 8\}$.

(ii) *If F and F_m are symmetrical distributions wrt 0, then R is an even function.*

(iii) *Let a and $a_m \in \mathbb{R} \cup \{-\infty\}$ be the starting points of F and F_m respectively, and assume $a \leq a_m$. Then, $R(\cdot)$ reaches its supremum on $[a_m, \infty)$:*

$$R_\infty \geq \sqrt{1 + \frac{F(a_m)}{m\,(1 - F(a_m))}}.$$

(iv) *Let b and $b_m \in \mathbb{R} \cup \{+\infty\}$ be the endpoints of F and F_m respectively, with $b \leq b_m$. Then, $R(\cdot)$ reaches its supremum on $(-\infty, b]$ and*

$$R_\infty \geq \sqrt{1 + \frac{(m-1)(1 - F_m(b))}{mF_m(b)}}.$$

Proof. (i) and (ii) are straightforward. For (iii) remark that, when $x \leq a$, $F(x) = F_m(x) = 0$ so that $R(x) = 1$. Besides, for all $x \in [a, a_m]$, one has $F_m(x) = 0$ and thus

$$R^2(x) = 1 + \frac{F(x)}{m(1 - F(x))}.$$

The above defined function is increasing so that the supremum of $R^2(\cdot)$ is reached for $x \geq a_m$ and therefore $R_\infty \geq R(a_m)$. Similarly for (iv), when $x \geq b_m$, $F(x) = F_m(x) = 1$ so that $R(x) = 1$. Besides, for all $x \in [b, b_m]$, one has $F(x) = 0$ and thus

$$R^2(x) = 1 + \frac{(m-1)(1 - F_m(x))}{mF_m(x)}.$$

The above defined function is decreasing so that the supremum of $R(\cdot)$ is reached for $x \leq b$ and therefore $R_\infty \geq R(b)$.

Lemma 3 (Uniform distribution)**.** *Assume that $F_1 = \cdots = F_{m-1}$ are the cdf of the uniform distribution $\mathcal{U}(-\sigma, \sigma)$ while F_m is the cdf of the uniform distribution $\mathcal{U}(-\sigma_m, \sigma_m)$ with $0 < \sigma \leq \sigma_m$. Then,*

$$R^2(x) = \begin{cases} 1 + \frac{\left(\frac{1}{\sigma} - \frac{1}{\sigma_m}\right)^2}{\frac{m^2}{(m-1)x^2} - m\left(\frac{1}{\sigma^2} + \frac{1}{(m-1)\sigma_m^2}\right)} & \text{if } |x| \leq \sigma\,, \\ 1 + \frac{m-1}{m}\left(1 - \frac{2}{1 + \sigma_m/|x|}\right) & \text{if } \sigma \leq |x| \leq \sigma_m\,, \\ 1 & \text{if } |x| \geq \sigma_m. \end{cases}$$

Moreover,

$$R_\infty = R(\pm\sigma) = \sqrt{1 + \frac{m-1}{m}\left(1 - \frac{2}{1 + \frac{\sigma_m}{\sigma}}\right)}.$$

Proof. Recall that

$$F_1(x) = \cdots = F_{m-1}(x) = \frac{x}{2\sigma} + \frac{1}{2}, \quad \forall x \in [-\sigma; \sigma]\,,$$
$$\text{and} \quad F_m(x) = \frac{x}{2\sigma_m} + \frac{1}{2}, \quad \forall x \in [-\sigma_m; \sigma_m]\,.$$

The case $|x| \geq \sigma_m$ is clear, we investigate the two other ones. First, if $\sigma \leq x \leq \sigma_m$, then $F_j(x) = 1$ for $j = 1, \ldots, m-1$ and Lemma 2(i) yields

$$R^2(x) = 1 + \frac{(m-1)\left(\frac{1}{2} - \frac{x}{2\sigma_m}\right)^2}{m\left(\frac{1}{2} + \frac{x}{2\sigma_m}\right)\left(\frac{1}{2} - \frac{x}{2\sigma_m}\right)} = 1 + \frac{m-1}{m}\left(1 - \frac{2}{1 + \frac{\sigma_m}{x}}\right).$$

Using, Lemma 2(ii), allows concluding for $\sigma \leq |x| \leq \sigma_m$. Finally, if $|x| \leq \sigma$, then:

$$\begin{aligned} R^2(x) &= 1 + \frac{x^2\left(\frac{1}{2\sigma} - \frac{1}{2\sigma_m}\right)^2}{m\left(\frac{1}{2} + \frac{x}{2\sigma}\right)\left(\frac{1}{2} - \frac{x}{2\sigma}\right) + \frac{m}{m-1}\left(\frac{1}{2} + \frac{x}{2\sigma_m}\right)\left(\frac{1}{2} - \frac{x}{2\sigma_m}\right)} \\ &= 1 + \frac{\left(\frac{1}{\sigma} - \frac{1}{\sigma_m}\right)^2}{\frac{m^2}{(m-1)x^2} - m\left(\frac{1}{\sigma^2} + \frac{1}{(m-1)\sigma_m^2}\right)}. \end{aligned}$$

Lemma 2(iv) entails that the maximum is reached for $x \in [-\sigma; \sigma]$. The above expression shows that the maximum is located at $x = \pm\sigma$, which gives the result.

Lemma 4 (Pareto distribution)**.** *Assume that $F_1 = \cdots = F_{m-1}$ are the cdf of the Pareto(α, η) distribution with $\alpha > 0$ the shape parameter and $\eta > 0$ the position parameter. Let F_m be the cdf of the Pareto(α, η_m) distribution with $0 < \eta \leq \eta_m$. Then,*

$$R^2(x) = \begin{cases} 1 + \frac{1}{m}\left(\left(\frac{x}{\eta}\right)^\alpha - 1\right) & \text{if } \eta \leq x \leq \eta_m\,, \\ 1 + \frac{1}{m}\frac{(\eta^\alpha - \eta_m^\alpha)^2}{\left(\eta^\alpha + \frac{\eta_m^\alpha}{m-1}\right)x^\alpha - \left(\eta^{2\alpha} + \frac{\eta_m^{2\alpha}}{m-1}\right)} & \text{if } \eta_m \leq x\,, \\ 1 & \text{if } x \leq \eta. \end{cases}$$

$$\text{Moreover,} \quad R_\infty = R(\eta_m) = \sqrt{1 + \frac{1}{m}\left(\left(\frac{\eta_m}{\eta}\right)^\alpha - 1\right)}.$$

Proof. Recall that $F_1(x) = \cdots = F_{m-1}(x) = 1 - (x/\eta)^{-\alpha}\,, \quad \forall x \in [\eta, +\infty)$ and $F_m(x) = 1 - (x/\eta_m)^{-\alpha}\,, \quad \forall x \in [\eta_m, +\infty)$. In the case where $\eta \leq x \leq \eta_m$, $F_m(x) = 0$, and using Lemma 2(iii) entails that $R(\cdot)$ is increasing on $[\eta, \eta_m]$ and

$$R^2(x) = 1 + \frac{1}{m}\frac{F(x)}{1 - F(x)} = 1 + \frac{1}{m}\left(\left(\frac{x}{\eta}\right)^\alpha - 1\right).$$

Moreover, for $\eta_m \leq x$, replacing the Pareto cdf in (3.32) yields

$$\begin{aligned} R^2(x) &= 1 + \frac{1}{m}\frac{x^{-2\alpha}\left(\eta^\alpha - \eta_m^\alpha\right)^2}{x^{-\alpha}\eta^\alpha(1 - x^{-\alpha}\eta^\alpha) + \frac{1}{m-1}x^{-\alpha}\eta_m^\alpha(1 - x^{-\alpha}\eta_m^\alpha)} \\ &= 1 + \frac{1}{m}\frac{(\eta^\alpha - \eta_m^\alpha)^2}{\left(\eta^\alpha + \frac{\eta_m^\alpha}{m-1}\right)x^\alpha - \left(\eta^{2\alpha} + \frac{\eta_m^{2\alpha}}{m-1}\right)}. \end{aligned}$$

Clearly, $R^2(\cdot)$ is decreasing on $[\eta_m, +\infty)$ and is extended by continuity at $x = \eta_m$. In conclusion, $R^2(\cdot)$ is maximum at $x = \eta_m$, and the result is proved.

Lemma 5 (Uniform vs Laplace distribution)**.** *Assume that $m = 2$, F_1 is the cdf of the uniform distribution $\mathcal{U}(-\sigma, \sigma)$ and F_2 is the cdf of the centred Laplace distribution $\mathcal{L}(0, \sigma/2)$ with $\sigma > 0$. Then for any x, $R(x) = R_1(x/\sigma)$ with*

$$R_1^2(x) = \begin{cases} 1 + \frac{\exp(-|x|)}{2(2-\exp(-|x|))} & \text{if } |x| \geq 2\,, \\ 1 + \frac{1}{2}\frac{(|x|/2-1+\exp(-|x|))^2}{1-x^2/4+2\exp(-|x|)(2-\exp(-|x|))} & \text{if } |x| \leq 2. \end{cases}$$

$$\text{Moreover,} \quad R_\infty = R(\pm\sigma) = R_1(\pm 1) = \sqrt{1 + \frac{1}{2(2e^2-1)}}.$$

Proof. In view of Lemma 2(ii,iii), it is sufficient to compare the values of $R_1(x)$ on $(-2, 0]$ with $R_1(-2)$. Then, the derivation of $R_1(x)$ is similar to the ones done in Lemma 3 and Lemma 4. Routine calculations show that R_1 has indeed a local maximum on $(-2, 0)$, but it remains lower than $R_1(-2)$ (see last column of Figure 3.3 for an illustration), which is therefore the value of R_∞.

Chapter 4

Tail behaviour of Bayesian extreme return level estimators

Contents

Résumé

Ce chapitre s'intéresse à l'étude du comportement de différentes grandeurs bayésiennes associées à une vraisemblance GPD pour un nombre fini d'observations. Nous nous concentrons plus précisément sur le comportement en queue de distribution, c'est-à-dire sur des quantiles associés à des probabilités qui tendent vers 1. Le comportement des distributions prédictives *a priori* et *a posteriori*, ainsi que des estimateurs bayésiens des niveaux de retour, sont examinés. L'analyse de la distribution prédictive *a priori* permet une meilleure compréhension des implications d'un choix *a priori* sur les observations extrêmes, tandis que l'étude des grandeurs *a posteriori* nous permet d'explorer les limites de l'extrapolation avec une quantité fixe de données.

Les résultats de toutes les quantités étudiées aboutissent à une conclusion similaire, à savoir que le comportement asymptotique est équivalent à celui de la queue la plus lourde permise *a priori*. En d'autres termes, si l'on place une loi *a priori* uniforme sur $[\xi_1, \xi_2]$ pour le paramètre de forme ξ, alors l'ensemble des quantités bayésiennes étudiées va se comporter asymptotiquement comme une loi ayant pour indice de queue ξ_2.

Une vérification sur simulations permet de confirmer le comportement théorique, mais uniquement pour une taille d'échantillon très faible. Il semblerait donc que la taille d'échantillon n influe sur la vitesse de convergence vers le comportement asymptotique le plus lourd, ce qui ne contredit finalement pas un résultat de consistance pour des quantiles associés à des probabilités de l'ordre de c/n avec $c > 0$. Ces résultats encourageants suggèrent ainsi plusieurs pistes de travaux futurs qui pourraient être explorées pour compléter l'étude.

Après une revue de la quantification de l'incertitude bayésienne pour des observables en Partie 4.1, nous présentons les différents résultats asymptotiques obtenus en Partie 4.2, puis nous illustrons ces résultats par des simulations en Partie 4.3. Des pistes de travaux futurs sont proposées en Partie 4.4, et les preuves des résultats théoriques sont présentées en Partie 4.A.

Abstract

This chapter focuses on studying the behavior of various Bayesian quantities associated with a GPD likelihood for a finite number of observations. Specifically, we focus on tail behaviors, which refers to quantiles associated with probabilities tending to 1. The behavior of prior and posterior predictive distributions, as well as Bayesian estimators of return levels, are examined. Analyzing the prior predictive distribution allows for a better understanding of the implications of a prior choice on extreme observations, while studying posterior quantities enables us to explore the limits of extrapolation with a fixed amount of data.

The results for all the quantities under study lead to a similar conclusion: the asymptotic behavior is equivalent to that of the heaviest tail allowed *a priori*. In other words, if we place a uniform prior distribution on $[\xi_1, \xi_2]$ for the shape parameter ξ, then all Bayesian quantities will asymptotically behave like a distribution with a tail index equal to ξ_2.

Verification through simulations confirms the theoretically obtained behavior, but only for a very small sample size. It appears that the sample size n influences the convergence speed towards the heaviest asymptotic behavior, which ultimately aligns with a consistency result for quantiles associated with probabilities of the order of c/n with $c > 0$. These encouraging results suggest several avenues for future research that could be explored to complement the study.

After reviewing the quantification of Bayesian uncertainty for observable quantities in Section 4.1, we present the different asymptotic results obtained in Section 4.2, followed by illustrating these results through simulations in Section 4.3. Possible directions for future work are proposed in Section 4.4, and the proofs of the theoretical results are presented in Appendix 4.A.

4.1 Introduction

4.1.1 Setting and notation

Let X be a random variable with cumulative distribution function (cdf) F, and M_n the maximum of n i.i.d random variables with cdf F, whose cdf is consequently F^n. Suppose that F belongs to a maximum domain of attraction, which means that there exist two sequences $a_n > 0$, b_n and a cdf H such that $F^n(a_n x + b_n) \to G(x)$ as $n \to \infty$. We denote this property $F \in \mathrm{MDA}(G)$. The extreme value theorem (see for instance Haan and Ferreira, 2006) states that G is necessarily of the same type as the extreme value distribution (EVD) distribution, with cdf:

$$G_\xi(x) = \begin{cases} \exp\left(-\{1+\xi x\}_+^{-1/\xi}\right) & \text{if } \xi \neq 0\,, \\ \exp(-\exp(-x)) & \text{if } \xi = 0, \end{cases} \tag{4.1}$$

where $\{x\}_+ = \max\{0, x\}$ and ξ is called the extreme value index. From this, Pickands theorem (Pickands, 1975) shows that if $F \in \mathrm{MDA}(G_\xi)$, then the distribution of the exceedances $X - u \mid X > u$ is asymptotically, as u converges to the endpoint of F, a generalized Pareto distribution (GPD), with cdf:

$$H(x \mid \sigma, \xi) = \begin{cases} 1 - \{1 + \xi \frac{x}{\sigma}\}_+^{-1/\xi} & \text{if } \xi \neq 0\,, \\ 1 - \exp\left(-\frac{x}{\sigma}\right) & \text{if } \xi = 0. \end{cases} \tag{4.2}$$

This result yields an approximation of the survival function $\bar{F}(x) := 1 - F(x)$ for a given high threshold u and $x \geq u$:

$$\bar{F}(x) \simeq \bar{F}(u)\bar{H}(x - u \mid \sigma, \xi), \tag{4.3}$$

with $\bar{H} := 1 - H$. A traditional framework called peak-over-threshold consists in choosing u as the $(n-k)$th order statistic and consider only the k largest values of the dataset. We are then interested in the T-year return level which is the value exceeded on average once every T years. Denoting by ℓ_α this quantity with $\alpha := 1/(T n_y)$ and n_y the number of observations per year, this is obtained by solving the equation $\bar{F}(\ell_\alpha \mid \sigma, \xi) = \alpha$. Using the approximation in Equation (4.3), we obtain

$$\ell_\alpha \simeq u + H^{-1}\left(1 - \frac{\alpha}{\bar{F}(u)} \mid \sigma, \xi\right) = u + \frac{\sigma}{\xi}\left(\left(\frac{\alpha}{\bar{F}(u)}\right)^{-\xi} - 1\right). \tag{4.4}$$

Estimating this quantity comes with uncertainties around the inference of $(\bar{F}(u), \sigma, \xi)$. In the Bayesian paradigm, the information one has from (σ, ξ) is modeled by a posterior distribution.

4.1.2 Accounting for uncertainty in the estimation

Here, the final aim here is not the estimation of parameters but rather a return level which is an observable quantity. In the general case with parameters $\boldsymbol{\theta} \in \Theta$, density $p(x \mid \boldsymbol{\theta})$ and observations $\boldsymbol{x}^{(n)} = (x_1, \ldots, x_n)$, several approaches exist to deduce an estimation of

an observable from a parameter. A first one by plug-in simply consists in inserting any estimator $\hat{\boldsymbol{\theta}}$ in the model:

$$p_{\text{plug}}(x \mid \boldsymbol{x}^{(n)}) = p(x \mid \hat{\boldsymbol{\theta}}). \tag{4.5}$$

This first point of view only considers a pointwise estimator of θ, and does not take uncertainty around estimation into account. Instead, predictive distributions aim at propagating all the information inferred on parameters in the observable space.

Prior predictive distribution. In the general Bayesian case, suppose that before any observation, the uncertainty around $\boldsymbol{\theta}$ is modeled through a prior $p(\boldsymbol{\theta})$. This uncertainty can be translated at the observation level using the *prior predictive distribution*, also known as the marginal distribution of x:

$$p_{\text{pred}}(x) = \int_{\Theta} p(x \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}. \tag{4.6}$$

See the section corresponding to prior predictive in Mikkola et al. (2023) for a complete review. One advantage of this change from parameters to observations is that it usually corresponds to the space where experts have prior information. It is the case of extreme value models, where there is not an easy way to interpret the parameters: ξ controls the existence of moments and of an upper bound in the distribution, but its interpretation is not direct for someone unfamiliar with extreme value theory. Coles and Tawn (1996) suggests to re-parameterize the EVD likelihood as a function of three quantiles (q_1, q_2, q_3), in order to allow an expert to model prior information directly on quantiles. More precisely, to ensure $q_1 < q_2 < q_3$, a positive prior is chosen on the difference of quantiles, which induces a particular choice of dependence structure that is debatable. Gaioni et al. (2010) provides a simple method to invert the transformation $(\mu, \sigma, \xi) \mapsto (q_1, q_2, q_3)$, and generalizes to the case when more than three quantiles are provided. The author proposes to use a Gaussian prior that is "*the least possible effort in modeling and allows for implementation of user-friendly software*". These priors elicited at the observation level do not help to deduce properties of predictive distribution, typically the theoretical implication of a Gaussian prior for the difference of quantile is not known. In general, translating prior information at the observable level to the parameter level is challenging. Specifically, specifying the predictive distribution $p_{\text{pred}}(x)$ and the conditional distribution $p(x \mid \boldsymbol{\theta})$ requires solving Equation (4.6) to obtain the corresponding distribution $p(\boldsymbol{\theta})$, see Gribok et al. (2004) for an example. Conversely, deriving properties of the predictive distribution $p_{\text{pred}}(x)$ from the conditional distribution $p(x \mid \boldsymbol{\theta})$ and the prior distribution $p(\boldsymbol{\theta})$ often involves calculating the marginal distribution, which is only tractable in specific cases of conjugate distributions.

Posterior predictive distribution. After observing $\boldsymbol{x}^{(n)}$, the distribution of a new observation x can be derived using the posterior predictive distribution:

$$p_{\text{pred}}(x \mid \boldsymbol{x}^{(n)}) = \int_{\Theta} p(x \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{x}^{(n)}) \mathrm{d}\boldsymbol{\theta}. \tag{4.7}$$

This expression integrates out the unobserved variables (the parameters) conditionally on the observed ones (the dataset). Therefore, as it is averaged over the posterior distribution, the uncertainty on the estimation of the parameters is included in the prediction model. Note that a frequentist version of the predictive distribution also exists, see Lawless and

Fredette (2005) for a review and Shen et al. (2018) for a definition unifying frequentist, Bayesian, and fiducial approaches. The use of posterior predictive for an extreme value model can already be found in Davison (1986) and Smith (1999). de Zea Bermudez et al. (2001) uses it for extreme quantile estimation with a Poisson-GPD model. An explicit expression is given in the case when $\xi = 0$. More recently, the properties of posterior predictive for extremes have been studied and compared with other approaches that we will detail in the next section (Fawcett and Green, 2018, Jonathan et al., 2021).

Plug-in versus predictive. Many authors have suggested employing a predictive approach instead of a plug-in one. Indeed, the plug-in method fails to account for the uncertainty associated with parameter estimation and treats the estimates as if they were absolute truth. Note that extreme value problems have intrinsically scarce datasets, so the quantification of uncertainty is all the more crucial for these problems. In the general case, Cox (1975) shows that in the plug-in approach, the coverage probability of prediction intervals should be corrected to consider the uncertainty around the estimation. Davison (1986) also points out the weaknesses of the approach, and suggests to use a Laplace approximation (Tierney and Kadane, 1986) on the posterior predictive instead. Aitchison (1975) shows that the predictive distribution is the optimal choice in terms of KL divergence, as it minimizes the integrated risk (see also Robert, 2007, Chapter 2). An asymptotic comparison of predictive and plug-in estimators is proposed by Smith (1999) for different losses. Here, the conclusion is more nuanced and the authors show that the plug-in method can be asymptotically more accurate on certain losses.

Recently, some empirical comparison have been proposed by Fawcett and Green (2018) and Jonathan et al. (2021) for extreme value models. Fawcett and Green (2018) shows using simulations that the predictive return level estimates are higher than those of the plug-in approach, but argues in favor of this solution as it provides a point summary that considers estimation uncertainty. Jonathan et al. (2021) offers a wide-ranging comparison of several approaches for estimating Bayesian extreme return levels. It includes the plug-in approach, the one using posterior predictive, and a third one which consists in the posterior mean of the return level. The latter is the one recommended by the authors.

In this chapter, we are interested in the properties in the tails of prior and posterior predictive distributions, along with two Bayesian estimators of extreme return level described in Jonathan et al. (2021). The rest of the chapter is organized as follows: in Section 4.2, we present our asymptotic results, and in Section 4.3 we illustrate the behaviour on simulations with several configurations. Further expansions are suggested in Section 4.4 and all the proofs are provided in Appendix 4.A.

4.2 Tail behaviors

In the following section, we are interested in the limits of tail extrapolation for Bayesian quantities with a fixed number of observations. The aim is to get insights on what is happening if one wants to use Bayesian estimators for extreme value inference. We restrict ourselves to the case of $\xi \geq 0$.

4.2.1 Prior predictive

We are first interested in the tail behaviour of the prior predictive associated with a GPD model. Denoting by $\bar{H}_{\text{pred}}$ the associated survival function and $p(\sigma, \xi)$ the prior on the GPD parameters, Equation (4.6) can be written

$$\begin{aligned} \bar{H}_{\text{pred}}(x) &\coloneqq \int_{-\infty}^{+\infty} \int_{0}^{+\infty} \bar{H}(x \mid \sigma, \xi) p(\sigma, \xi) \mathrm{d}\sigma \mathrm{d}\xi, \\ &= \int_{-\infty}^{+\infty} \int_{0}^{+\infty} \left\{ 1 + \xi \frac{x}{\sigma} \right\}_{+}^{-1/\xi} p(\sigma, \xi) \mathrm{d}\sigma \mathrm{d}\xi. \end{aligned} \tag{4.8}$$

Under this formula, it is not clear how the prior uncertainty around σ and ξ is affecting the one at the observation level. In particular in the tail, the choice of $p(\sigma, \xi)$ may induce a maximum domain of attraction for $\bar{H}_{\text{pred}}$.

Equation (4.8) can be seen as a mixture of generalized Pareto survival functions, where the weights accross $\mathrm{d}\sigma$ and $\mathrm{d}\xi$ are governed by the prior $p(\sigma, \xi)$. Similarly to a discrete mixture, survival functions $\bar{H}(x \mid \sigma, \xi)$ with heavier ξ are expected to asymptotically dominate the behaviour in the tail of $\bar{H}_{\text{pred}}$. Typically, if ξ has a prior with a finite upper bound ξ_2, one can expect a prior predictive distribution to be heavy-tailed with index ξ_2. The following proposition details this intuition in the case of a uniform prior for ξ.

Proposition 1. *Suppose $p(\sigma, \xi) = p(\sigma) \otimes \mathcal{U}(\xi_1, \xi_2)$ with $0 \leq \xi_1 < \xi_2$. Depending on the prior $p(\sigma)$, three cases arise:*

(i) *If $\mathbb{E}\left[\sigma^{1/\xi_2 + \epsilon}\right] < \infty$ for some $\epsilon > 0$, then $H_{\text{pred}} \in \mathrm{MDA}(G_{\xi_2})$, and as $x \to \infty$,*

$$\bar{H}_{\text{pred}}(x) \sim \mathbb{E}\left[\sigma^{1/\xi_2}\right] \frac{\xi_2^{2-1/\xi_2}}{\xi_2 - \xi_1} \frac{x^{-1/\xi_2}}{\log x}.$$

(ii) *Otherwise, if $p(\sigma)$ is heavy-tailed with associated cdf $F_\sigma \in \mathrm{MDA}(G_\gamma)$ and $\gamma > \xi_2$, then $H_{\text{pred}} \in \mathrm{MDA}(G_\gamma)$. Moreover, as $x \to \infty$,*

$$\bar{H}_{\text{pred}}(x) \sim c_{\xi_1, \xi_2, \gamma} \bar{F}_\sigma(x), \text{ with } c_{\xi_1, \xi_2, \gamma} = \frac{1}{\xi_2 - \xi_1} \int_{\xi_1}^{\xi_2} \xi^{-1/\gamma} B(1/\xi - 1/\gamma, 1/\gamma) \mathrm{d}\xi,$$

where $B(\cdot, \cdot)$ denotes the Beta function.

(iii) *Finally, if $p(\sigma)$ is heavy-tailed with associated cdf $F_\sigma \in \mathrm{MDA}(G_{\xi_2})$, then $H_{\text{pred}} \in \mathrm{MDA}(G_{\xi_2})$.*

Part of the proof in Appendix 4.A is simplified thanks to Breiman's theorem: see Lemma 1 (Breiman, 1965) and a refinement by Embrechts and Goldie (1980) that handles (iii). A first observation of the result given in Proposition 1 is that the behavior indeed depends on the prior on ξ, but also on the one on σ: if σ has a prior with an extreme index larger than ξ_2, then the prior predictive acts as the one of σ in the tail. Otherwise, assuming that σ is heavy tailed, the extreme index is ξ_2. It should be noted that this distinction does not encompass all possible scenarios, since there are cases where the condition on moments may not be satisfied and having a distribution that does not belong to a maximum domain of attraction. Proposition 1 means that the prior predictive behaves like the most pessimistic case enabled *a priori* (*i.e.* the largest possible ξ). Therefore, parameter uncertainty translates at the observable level the most dispersed case.

4.2.2 Posterior predictive

How is the predictive distribution impacted by the observation of data $\boldsymbol{x}^{(n)}$? The prior distribution $p(\sigma, \xi)$ in Equation (4.8) is replaced by the posterior $p(\sigma, \xi \mid \boldsymbol{x}^{(n)})$ to obtain the survival function of the posterior predictive:

$$\bar{H}_{\text{pred}}(x \mid \boldsymbol{x}^{(n)}) = \iint \bar{H}(x \mid \sigma, \xi) p(\sigma, \xi \mid \boldsymbol{x}^{(n)}) \mathrm{d}\sigma \mathrm{d}\xi.$$

Assuming that the densities $p(x_i \mid \sigma, \xi)$ are exactly GPD for $i = 1, \ldots, n$, the Bayes' rule yields

$$\bar{H}_{\text{pred}}(x \mid \boldsymbol{x}^{(n)}) = \frac{1}{p(\boldsymbol{x}^{(n)})} \iint \left\{1 + \xi \frac{x}{\sigma}\right\}_+^{-1/\xi} p(\sigma, \xi) \sigma^{-n} \prod_{i=1}^{n} \left(1 + \xi \frac{x_i}{\sigma}\right)^{-1/\xi - 1} \mathrm{d}\sigma \mathrm{d}\xi. \quad (4.9)$$

A likelihood term along with the evidence $p(\boldsymbol{x}^{(n)})$ is therefore added in the expression compared to the prior predictive case. The following result can be seen as a generalization to any number of observations n of the Proposition 1 (i) that covers the case $n = 0$.

Proposition 2. *If the extreme value index is a priori uniform so that $\xi \sim \mathcal{U}(\xi_1, \xi_2)$ with $0 \leq \xi_1 < \xi_2$, and the prior on σ is such that $\mathbb{E}\left[\sigma^{(n+1)/\xi_2} \mathbb{I}\{\sigma \leq 1\}\right] < \infty$ and $\mathbb{E}\left[\sigma^{1/\xi_2 - n + \epsilon} \mathbb{I}\{\sigma \geq 1\}\right] < \infty$, then $H_{\text{pred}}(\cdot \mid \boldsymbol{x}^{(n)}) \in \text{MDA}(G_{\xi_2})$, and as $x \to \infty$, we have*

$$\begin{aligned}
\bar{H}_{\text{pred}}(x \mid \boldsymbol{x}^{(n)}) &\sim \frac{x^{-1/\xi_2}}{\log x} \xi_2^{2-1/\xi_2} \int_0^{+\infty} \sigma^{1/\xi_2} p(\sigma, \xi_2 \mid \boldsymbol{x}^{(n)}) \mathrm{d}\sigma \\
&= \frac{x^{-1/\xi_2}}{\log x} \frac{\xi_2^{2-1/\xi_2}}{(\xi_2 - \xi_1) p(\boldsymbol{x}^{(n)})} \int_0^{+\infty} \sigma^{1/\xi_2 - n} \prod_{i=1}^{n} \left(1 + \xi_2 \frac{x_i}{\sigma}\right)^{-1/\xi_2 - 1} p(\sigma) \mathrm{d}\sigma.
\end{aligned}$$

It is important to note that the requirements on the prior moment for σ is not stringent and is less and less as n increases. Typically, if $\xi_2 = 1$ and a distribution with a finite variance is chosen, a value of $n = 1$ is sufficient to satisfy both conditions. This proposition demonstrates that adding observations only affects the constant term in the asymptotic expression, but the extreme index ξ_2 still serve as an upper bound for ξ. Hence, the behavior is still primarily governed by the heaviest tail. This observation reflects the fact that there comes a point where extrapolating in the tail becomes unreasonable, as the associated uncertainty in the estimate becomes too substantial, resulting in predictions that resemble the most extreme cases predicted by the model.

This could have been expected because of the decorrelation of the asymptotic of the quantile x from n. In more conventional extreme value problems, the aim is to estimate the probability associated with a quantile x_n that depends on n, to extend beyond the available data while maintaining a reasonable proximity to it. The consideration of asymptotics with respect to n is not addressed here but can be seen as the next step. Such an analysis could lead to consistency or asymptotic normality results, as discussed in Section 4.4.

4.2.3 Return levels

Recall that the final quantity of interest here is the return level ℓ_α defined in Equation (4.4), which is a random variable in the Bayesian paradigm. From there, three strategies summarized in Jonathan et al. (2021) are possible for obtaining a point estimation:

1. **Plug-in of parameters estimation**: this consists in inserting an estimator of the parameters in Equation (4.4), for example, the posterior mean:
$$\ell_\alpha^{(1)} := u + H^{-1}\left(1 - \frac{\alpha}{\bar{F}(u)} \mid \mathbb{E}\left[\sigma, \xi \mid \boldsymbol{x}^{(n)}\right]\right).$$
2. **Posterior mean of return level**: this consists in considering ℓ_α as a posterior quantity from which a posterior mean can be computed:
$$\ell_\alpha^{(2)} := u + \mathbb{E}\left[H^{-1}\left(1 - \frac{\alpha}{\bar{F}(u)} \mid \sigma, \xi\right) \mid \boldsymbol{x}^{(n)}\right].$$
3. **Posterior predictive quantile**: this consists in inverting the posterior predictive distribution:
$$\ell_\alpha^{(3)} := u + H_{\text{pred}}^{-1}\left(1 - \frac{\alpha}{\bar{F}(u)} \mid \boldsymbol{x}^{(n)}\right), \text{ with } H_{\text{pred}}(\cdot \mid \boldsymbol{x}^{(n)}) \text{ defined in Equation (4.9).}$$

Here, our focus is on the tail behavior of $\ell_\alpha^{(2)}$ and $\ell_\alpha^{(3)}$ when a fixed number of observations is considered. These two methods are selected because, contrary to $\ell_\alpha^{(1)}$, they incorporate parameter uncertainty in the estimation process. However, it is important to note that they differ in terms of when it is included: the second method incorporates it before inverting the cdf, while the third one incorporates it after the inversion. Note that on simulations, Jonathan et al. (2021) observes that $\ell_\alpha^{(2)} \leq \ell_\alpha^{(3)}$ and that the true value ℓ_α verifies $\ell_\alpha \leq \ell_\alpha^{(3)}$. In the subsequent results, we omit the notation of $\bar{F}(u)$ by an abuse of notation, since it is treated as a constant in this context. First, we obtain the following result for the posterior predictive return level:

Proposition 3. *If the extreme value index is a priori uniform so that $\xi \sim \mathcal{U}(\xi_1, \xi_2)$ with $0 \leq \xi_1 < \xi_2$, and the prior on σ is such that $\mathbb{E}\left[\sigma^{1+n/\xi_2}\mathbb{I}\{\sigma \leq 1\}\right] < \infty$ and $\mathbb{E}\left[\sigma^{1-n}\mathbb{I}\{\sigma \geq 1\}\right] < \infty$, then as $\alpha \to 0$, we have*
$$\ell_\alpha^{(2)} \sim \alpha^{-\xi_2}\left(\xi_2 \log(1/\alpha)\right)^{-1} \int_0^{+\infty} \sigma p(\sigma, \xi_2 \mid \boldsymbol{x}^{(n)}) \mathrm{d}\sigma.$$

For $\ell_\alpha^{(3)}$, the asymptotic expression given in Proposition 2 allows us to deduce an asymptotic expression of its inverse:

Proposition 4. *Under the assumptions of Proposition 2 and as $\alpha \to 0$, we have*
$$\ell_\alpha^{(3)} \sim \alpha^{-\xi_2}\left(\xi_2 \log(1/\alpha)\right)^{-\xi_2}\left(\int_0^{+\infty} \sigma^{1/\xi_2} p(\sigma, \xi_2 \mid \boldsymbol{x}^{(n)}) \mathrm{d}\sigma\right)^{\xi_2}.$$

As expected, the predictive return level asymptotically behaves like the heaviest possible case, with an $\alpha^{-\xi_2}$ term, but Proposition 3 shows that it is also the behaviour of the posterior mean of return level $\ell_\alpha^{(2)}$. Only the constant and the power of the log change between the two methods, and we indeed have asymptotically $\ell_\alpha^{(2)} \leq \ell_\alpha^{(3)}$ if $\xi_2 < 1$, but the order changes if $\xi_2 > 1$. When $\xi_2 = 1$, both estimators have the same asymptotic equivalent. The proof methods employed in these approaches results in different assumptions regarding the prior moments of σ. However, it is worth noting that similarly to Proposition 2, both assumptions are rapidly satisfied as the number of data increases.

4.3 Simulations

We validate the results derived in Propositions 3 and 4 through simulations conducted under various configurations. Specifically, we consider different values of $\xi_2 \in \{0.5, 1, 2\}$ as the upper bound for the uniform distribution on ξ, and varying numbers of observations $n \in \{2, 10, 100\}$. In each scenario, we generate samples $\boldsymbol{x}^{(n)}$ from a GPD with parameters $\sigma_0 = 15$, $\xi_0 = 0.1 < \xi_2$, and a threshold $u = 0$. The true return level, as defined in Equation (4.4), simplifies to:

$$\ell_\alpha = \frac{\sigma_0}{\xi_0}\left(\alpha^{-\xi_0} - 1\right).$$

To estimate $\hat{\ell}_\alpha^{(i)}$ for $i \in \{1, 2, 3\}$, we employ the Metropolis–Hastings algorithm implemented in PyMC3 (Salvatier et al., 2016). We retain 2500 iterations across 4 chains, discarding previous 1000 samples as burn-in. Convergence diagnostics such as effective sample size (ESS), autocorrelation, and $\hat{R}_\infty$ (Moins et al., 2023) support the adequacy of this sample size to approximate the posterior distribution. The estimated return levels replicated five times for α ranging from 10^{-2} to 10^{-4} are depicted as dotted lines in Figure 4.1. Plain lines represent the asymptotic formulas derived in Propositions 3 and 4 for $\hat{\ell}_\alpha^{(2)}$ and $\hat{\ell}_\alpha^{(3)}$ respectively. Estimating the two constants involves evaluating an integral that lacks an explicit form. We approximate these constants using an MCMC method on the univariate posterior estimation, assuming a GPD model with known $\xi = \xi_2$. We find that 1000 iterations and 1000 iterations of burn-in are enough here.

As expected, all the asymptotic curves (solid lines) deviate from the actual one (in black), as the slope at the logarithmic scale is ξ_2 instead of ξ_0. Consequently, the discrepancy in the return level asymptotes becomes more pronounced as ξ_2 moves further away from ξ_0. However, we observe that incorporating additional observations in the likelihood causes the estimator to deviate from its asymptotic curve and converge towards the true value. This behavior suggests that the number of observations n influences the convergence rate towards the degenerate asymptotic distribution. At the logarithmic scale, this deviation manifests in both the slope and the intercept, making it challenging to discern that the solid lines actually represent asymptotes of the dotted line, particularly in the case of $n = 100$ and $\xi_2 = 2$. The plug-in method exhibits a similar behavior, but it appears to converge more rapidly towards the true return level compared to the other methods.

4.4 Conclusion and future work

This chapter presents some preliminary findings regarding the tail behaviour of predictive distributions (prior and posterior) and return level estimators when a uniform prior is assumed for the shape parameter ξ. The results demonstrate that incorporating uncertainty into the estimates for a finite number of observations is asymptotically equivalent to the most pessimistic scenario, where the shape parameter is set to the upper bound ξ_2. Therefore, if the cdf of the observation F is such that $F \in \mathrm{MDA}(G_{\xi_0})$ with $\xi_0 < \xi_2$, then the obtained results conflict with the observations.

These findings provide insights into the limits of extrapolation. In the tail asymptotics, the tail index depends solely on the prior through its upper bound, which reflects the degenerate nature of extrapolation that goes beyond the available data. Simulations indicate that substantial tail extrapolation is necessary to observe estimations approaching their

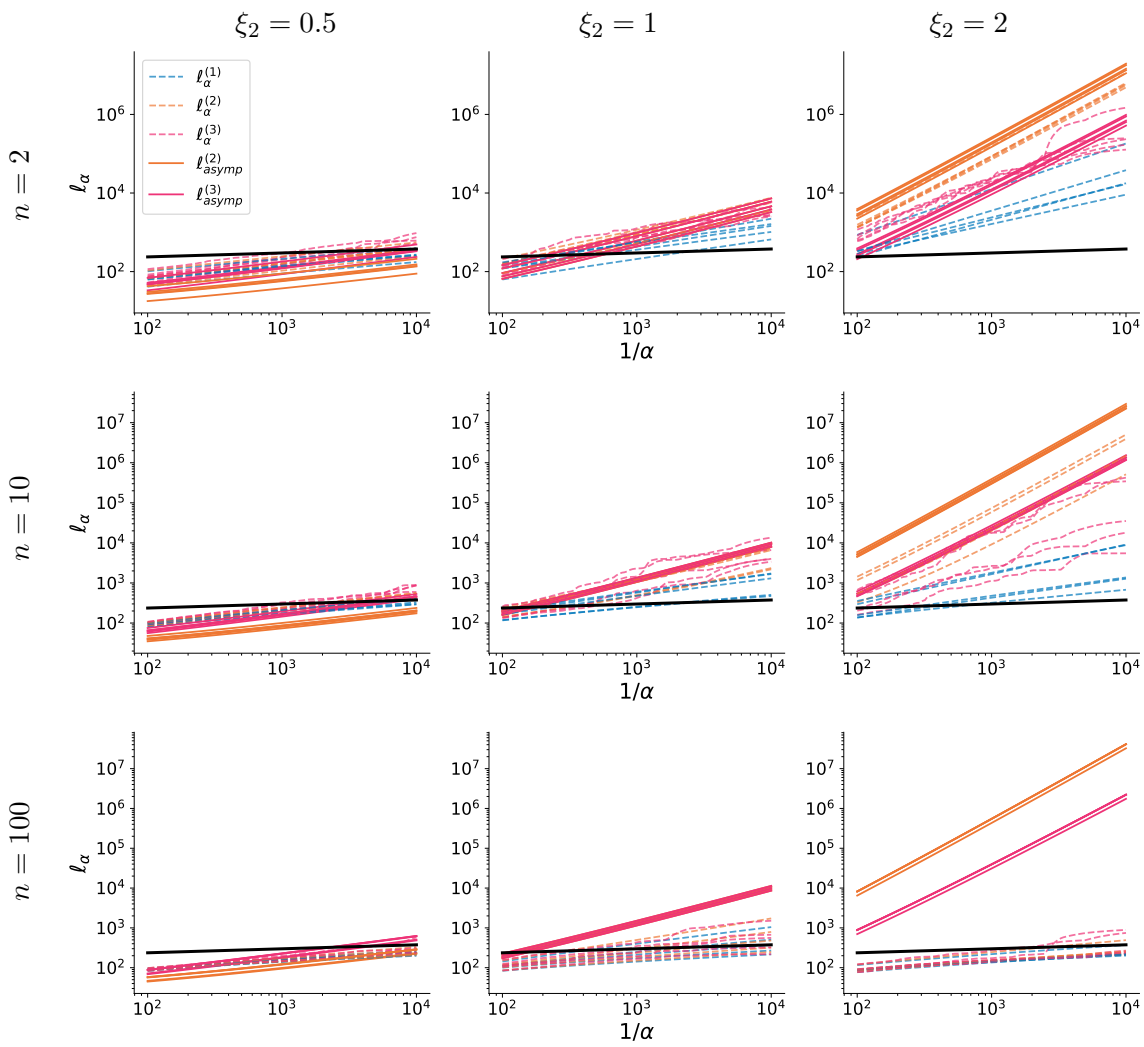


Figure 4.1: On each plot, five replications of return level estimators $\hat{\ell}_\alpha^{(i)}$ for $i \in \{1, 2, 3\}$ (dotted lines), and the asymptotic expressions of $\hat{\ell}_\alpha^{(2)}$ and $\hat{\ell}_\alpha^{(3)}$ (plain lines), for a dataset with n observations following a GPD with $(\sigma_0, \xi_0) = (15, 0.1)$ (the black lines represent the real return level). Each row represents a choice of $n \in \{2, 10, 100\}$, and each column a choice of the support upper bound of the prior on ξ: $\xi_2 \in \{0.5, 1, 2\}$.

asymptotes, typically quantiles associated with probabilities on the order of n^{-2} or even n^{-3}. However, it is worth noting that the influence of the prior becomes more significant in this context, emphasizing the need for careful prior design (Moins et al., 2023). In cases where the prior is uninformative, the choice of an upper bound for a uniform prior on ξ can have a huge impact on posterior estimations.

These results suggest several potential directions for further exploration:

1. **Generalization of the prior on $\xi > 0$.** It is likely that any prior on ξ with finite right endpoint ξ_2 will result in a predictive distribution with an extremal index equal to ξ_2. While the specific constant may vary, it seems that it is not the uniformity of ξ that leads to this outcome, but rather the fact that the support is upper-bounded. In the case of an unbounded support on the prior of ξ, one can expect a super-heavy tailed behavior, where the predictive distribution does not belong to any domain of attraction, with a right tail that decreases extremely slowly to zero, typically exhibiting a logarithmic decrease.

2. **Generalization to the three maximum domains of attraction.** The current study focuses on the case where $\xi \geq 0$, which encompasses two maximum domains of attraction. A natural extension would be to consider the general case for ξ, leading to discussions similar to those in the study by Richards and Tawn (2022), which investigates the tail behavior of aggregated variables.

3. **Derivation of expansions.** Refining the asymptotic results could provide insights into the convergence rate towards its equivalent, particularly the dependence with respect to the sample size n in the posterior results. This information could help us understand the behavior observed in Figure 4.1.

4. **Including an asymptotic with n.** Another important aspect is the asymptotic study of quantiles associated with n-dependent probabilities (e.g. of order c/n with $c > 0$). Increasing the quantile as the sample size grows could yield consistency results, similar to those obtained by Padoan and Rizzelli (2022) for block maxima models. It is worth noting that Padoan and Rizzelli (2022) employ an empirical Bayes method, which yields stronger results by accounting for the misspecification of the maxima distribution approximated by a generalized extreme value distribution.

These future directions hold potential for further exploration and refinement of the presented results.

4.A Proofs

Lemma 1 (Breiman, 1965)**.** *Let $X = YZ$ with Y and Z two independent non-negative random variables such that $F_Y \in MDA(G_\alpha)$ with $\alpha > 0$, and $\mathbb{E}\left[Z^{1/\alpha+\epsilon}\right] < \infty$ for some $\epsilon > 0$. Then, $F_X \in MDA(G_\alpha)$ and*

$$\bar{F}_X(x) \sim \mathbb{E}\left[Z^{1/\alpha}\right] \bar{F}_Y(x),$$

as $x \to \infty$.

Proof of Proposition 1. (i) Letting $p(\sigma, \xi) = p(\sigma) \otimes \mathcal{U}(\xi_1, \xi_2)$, the prior predictive distribution can be written as

$$\bar{H}_{\text{pred}}(x) = \frac{1}{\xi_2 - \xi_1} \int_0^{+\infty} \int_{\xi_1}^{\xi_2} \left(1 + \xi \frac{x}{\sigma}\right)^{-1/\xi} p(\sigma) \mathrm{d}\xi \mathrm{d}\sigma, \; x \geq 0.$$

It appears that $\bar{H}_{\text{pred}}$ can be interpreted as the survival function associated with the product of two random variables: $X_{\sigma=1}$ with survival function

$$\bar{F}_{\sigma=1}(x) = \frac{1}{\xi_2 - \xi_1} \int_{\xi_1}^{\xi_2} (1 + \xi x)^{-1/\xi} \, \mathrm{d}\xi, \; x \geq 0,$$

and σ with density $p(\sigma)$ so that

$$\bar{H}_{\text{pred}}(x) = \int_0^{+\infty} \bar{F}_{\sigma=1}(x/\sigma) p(\sigma) \mathrm{d}\sigma, \; x \geq 0.$$

Without loss of generality, suppose that $\epsilon < \frac{1}{\xi_1} - \frac{1}{\xi_2}$. Let us show that $\bar{F}_{\sigma=1} \in \text{MDA}(G_{\xi_2})$ and more specifically, that

$$\bar{F}_{\sigma=1}(x) \sim \frac{\xi_2^{2-1/\xi_2}}{\xi_2 - \xi_1} \frac{x^{-1/\xi_2}}{\log x}, \tag{4.10}$$

as $x \to \infty$. To this end, consider the expansion

$$\bar{F}_{\sigma=1}(x) = \frac{1}{\xi_2 - \xi_1} \left(I_1(x) + I_2(x)\right), \tag{4.11}$$

with

$$I_1(x) = \int_{\xi_1}^{(1/\xi_2 + \epsilon)^{-1}} (1 + \xi x)^{-1/\xi} \, \mathrm{d}\xi \quad \text{and} \quad I_2(x) = \int_{(1/\xi_2 + \epsilon)^{-1}}^{\xi_2} (1 + \xi x)^{-1/\xi} \, \mathrm{d}\xi.$$

Let us first focus on $I_1(x)$. Since the function $\xi \mapsto (1 + \xi x)^{-1/\xi}$ is increasing for all $x > 0$, one has

$$0 \leq I_1(x) \leq ((1/\xi_2 - \epsilon)^{-1} + \xi_1)(1 + (1/\xi_2 + \epsilon)^{-1} x)^{-1/\xi_2 - \epsilon} = o\left(\frac{x^{-1/\xi_2}}{\log x}\right). \tag{4.12}$$

Second, the change of variable $s = \left(\frac{1}{\xi} - \frac{1}{\xi_2}\right) \log x$ yields

$$\begin{aligned} I_2(x) &= \frac{1}{\log x} \int_0^{\epsilon \log x} \left(1 + \left(\frac{s}{\log x} + \frac{1}{\xi_2}\right)^{-1} x\right)^{-\frac{s}{\log x} - \frac{1}{\xi_2}} \left(\frac{s}{\log x} + \frac{1}{\xi_2}\right)^{-2} \mathrm{d}s, \\ &= \frac{x^{-1/\xi_2}}{\log x} \int_0^{\infty} e^{-s} \left(\frac{1}{x} + \left(\frac{s}{\log x} + \frac{1}{\xi_2}\right)^{-1}\right)^{-\frac{s}{\log x} - \frac{1}{\xi_2}} \left(\frac{s}{\log x} + \frac{1}{\xi_2}\right)^{-2} \mathbb{I}\{s \leq \epsilon \log x\} \mathrm{d}s. \end{aligned}$$

Clearly, for all $s \geq 0$, the integrand is positive and converges to $e^{-s} \xi_2^{2-1/\xi_2}$ as $x \to \infty$. Moreover, it is uniformly bounded on $x \in \mathbb{R}^+$ as follows:

$$\begin{aligned} e^{-s} \left(\frac{1}{x} + \left(\frac{s}{\log x} + \frac{1}{\xi_2}\right)^{-1}\right)^{-\frac{s}{\log x} - \frac{1}{\xi_2}} \left(\frac{s}{\log x} + \frac{1}{\xi_2}\right)^{-2} \mathbb{I}\{s \leq \epsilon \log x\} &\leq e^{-s} \left(1/\xi_2 + \epsilon\right)^{\frac{s}{\log x} + \frac{1}{\xi_2}} \xi_2^2, \\ &\leq e^{-s} \max\left\{1, (1/\xi_2 + \epsilon)^{\epsilon + \frac{1}{\xi_2}}\right\} \xi_2^2. \end{aligned} \tag{4.13}$$

Remarking that the upper bound in (4.13) is integrable, the dominated convergence theorem yields

$$I_2(x) \sim \frac{x^{-1/\xi_2}}{\log x} \int_0^\infty e^{-s} \xi_2^{2-1/\xi_2} \mathrm{d}s = \xi_2^{2-1/\xi_2} \frac{x^{-1/\xi_2}}{\log x}, \tag{4.14}$$

as $x \to \infty$. Collecting (4.11), (4.12) and (4.14) proves (4.10) and Lemma 1 concludes.

(ii) In this case, we use Lemma 1 with the roles of Y and Z reversed. Here, $\alpha = 1/\gamma$, $Y = \sigma$ and $Z = X_{\sigma=1}$, so that it only remains to prove that $\mathbb{E}\left[X^{1/\gamma+\epsilon} \mid \sigma = 1\right] < \infty$ for some $\epsilon > 0$ and compute $\mathbb{E}\left[X^{1/\gamma} \mid \sigma = 1\right]$. For all $\alpha < 1/\xi_2$, we can compute

$$\begin{aligned}
\mathbb{E}\left[X^\alpha \mid \sigma = 1\right] &= \int_0^\infty \alpha x^{\alpha-1} \bar{F}_{X|\sigma=1}(x) \mathrm{d}x \\
&= \frac{1}{\xi_2 - \xi_1} \int_0^\infty \int_{\xi_1}^{\xi_2} \alpha x^{\alpha-1} (1+\xi x)^{-1/\xi} \, \mathrm{d}\xi \mathrm{d}x \\
&= \frac{\alpha}{\xi_2 - \xi_1} \int_0^1 \int_{\xi_1}^{\xi_2} \xi^{-\alpha} (1-y)^{\alpha-1} y^{-\alpha+1/\xi-1} \mathrm{d}\xi \mathrm{d}y \qquad \text{(change of variable } y = (1+\xi x)^{-1}\text{)} \\
&= \frac{\alpha}{\xi_2 - \xi_1} \int_{\xi_1}^{\xi_2} \xi^{-\alpha} B(1/\xi - \alpha, \alpha) \mathrm{d}\xi,
\end{aligned}$$

which is well defined as long as $\alpha < 1/\xi_2$. So if $\epsilon = \frac{1}{2}\left(\frac{1}{\xi_2} - \frac{1}{\gamma}\right) > 0$, we have $\frac{1}{\gamma} + \epsilon = \frac{1}{2}\left(\frac{1}{\xi_2} + \frac{1}{\gamma}\right) < \frac{1}{\xi_2}$, so $\mathbb{E}\left[X^{1/\gamma+\epsilon} \mid \sigma = 1\right] < \infty$ which concludes.

(iii) The proof combines the result proved in (i) that $X_{\sigma=1} \in \mathrm{MDA}(G_{\xi_2})$ with Theorem 3 in Embrechts and Goldie (1980).

Proof of Proposition 2. In the same way as in the proof of Proposition 1, we suppose that $\epsilon < \frac{1}{\xi_1} - \frac{1}{\xi_2}$ and we split the integral in two parts: $\bar{H}_{\text{pred}}(x \mid \boldsymbol{x}^{(n)}) = \frac{1}{(\xi_2-\xi_1)p(\boldsymbol{x}^{(n)})}(I_1(x) + I_2(x))$, with

$$I_1(x) = \int_0^{+\infty} \int_{\xi_1}^{(1/\xi_2+\epsilon)^{-1}} \left(1+\xi\frac{x}{\sigma}\right)^{-1/\xi} \sigma^{-n} \prod_{i=1}^n \left(1+\xi\frac{x_i}{\sigma}\right)^{-1/\xi-1} p(\sigma)\mathrm{d}\xi\mathrm{d}\sigma,$$

$$\text{and} \quad I_2(x) = \int_0^{+\infty} \int_{(1/\xi_2+\epsilon)^{-1}}^{\xi_2} \left(1+\xi\frac{x}{\sigma}\right)^{-1/\xi} \sigma^{-n} \prod_{i=1}^n \left(1+\xi\frac{x_i}{\sigma}\right)^{-1/\xi-1} p(\sigma)\mathrm{d}\xi\mathrm{d}\sigma.$$

For $I_2(x)$, the change of variable $s = \left(\frac{1}{\xi} - \frac{1}{\xi_2}\right)\log x$ leads to

$$I_2(x) = \frac{x^{-1/\xi_2}}{\log x} \int_0^{+\infty} \int_0^{\epsilon \log x} f_x(s,\sigma) p(\sigma) \mathrm{d}s \mathrm{d}\sigma,$$

with

$$\begin{aligned}
f_x(s,\sigma) \coloneqq e^{-s} &\left(\frac{1}{x} + \left(\frac{s}{\log x} + \frac{1}{\xi_2}\right)^{-1} \frac{1}{\sigma}\right)^{-\frac{s}{\log x} - \frac{1}{\xi_2}} \left(\frac{s}{\log x} + \frac{1}{\xi_2}\right)^{-2} \\
&\sigma^{-n} \prod_{i=1}^n \left(1 + \left(\frac{s}{\log x} + \frac{1}{\xi_2}\right)^{-1} \frac{x_i}{\sigma}\right)^{-\frac{s}{\log x} - \frac{1}{\xi_2} - 1}.
\end{aligned}$$

As $x \to \infty$, $f_x(s,\sigma)$ converges to $f(s,\sigma) \coloneqq e^{-s}\xi_2^{2-1/\xi_2}\sigma^{1/\xi_2-n}\prod_{i=1}^n \left(1+\xi_2\frac{x_i}{\sigma}\right)^{-1/\xi_2-1}$. The function $s \mapsto f(s,\sigma)p(\sigma)$ is clearly integrable on $\mathbb{R}^+$, and we have

$$\begin{aligned}
\sigma^{1/\xi_2-n}\prod_{i=1}^n\left(1+\xi_2\frac{x_i}{\sigma}\right)^{-1/\xi_2-1} &\leq \min\left\{\sigma^{1/\xi_2-n}, \sigma^{(n+1)/\xi_2}\prod_{i=1}^n(\xi_2 x_i)^{-1/\xi_2-1}\right\} \\
\text{so}\quad \sigma^{1/\xi_2-n}\prod_{i=1}^n\left(1+\xi_2\frac{x_i}{\sigma}\right)^{-1/\xi_2-1} &\leq \left(\prod_{i=1}^n(\xi_2 x_i)^{-1/\xi_2-1}\right)\sigma^{(n+1)/\xi_2}\mathbb{I}\{\sigma\leq 1\} + \sigma^{1/\xi_2-n}\mathbb{I}\{\sigma\geq 1\},
\end{aligned} \tag{4.15}$$

and so $\sigma \mapsto f(s,\sigma)p(\sigma)$ is integrable on $\mathbb{R}^+$ as we assume that $\mathbb{E}\left[\sigma^{(n+1)/\xi_2}\mathbb{I}\{\sigma\leq 1\}\right] < \infty$ and $\mathbb{E}\left[\sigma^{1/\xi_2-n+\epsilon}\mathbb{I}\{\sigma\geq 1\}\right] < \infty$. To dominate the function $f_x(s,\sigma)$, we start by observing that a similar bound as in Equation (4.13) can be derived to obtain

$$\begin{aligned}
\left(\frac{1}{x}+\left(\frac{s}{\log x}+\frac{1}{\xi_2}\right)^{-1}\frac{1}{\sigma}\right)^{-\frac{s}{\log x}-\frac{1}{\xi_2}} &\leq \max\left\{1,(1/\xi_2+\epsilon)^{\epsilon+\frac{1}{\xi_2}}\right\}\sigma^{\frac{s}{\log x}+\frac{1}{\xi_2}}, \\
&\leq \max\left\{1,(1/\xi_2+\epsilon)^{\epsilon+\frac{1}{\xi_2}}\right\}\left(\sigma^{1/\xi_2}\mathbb{I}\{\sigma\leq 1\}+\sigma^{1/\xi_2+\epsilon}\mathbb{I}\{\sigma\geq 1\}\right).
\end{aligned}$$

Moreover, as $\xi \mapsto (1+\xi x)^{-1/\xi}$ is increasing and $\left(\frac{s}{\log x}+\frac{1}{\xi_2}\right)^{-1} \leq \xi_2 \quad \forall s \in [0, \epsilon\log x]$, we have

$$\begin{aligned}
\sigma^{-n}\prod_{i=1}^n\left(1+\left(\frac{s}{\log x}+\frac{1}{\xi_2}\right)^{-1}\frac{x_i}{\sigma}\right)^{-\frac{s}{\log x}-\frac{1}{\xi_2}-1} &\leq \sigma^{-n}\prod_{i=1}^n\left(1+\xi_2\frac{x_i}{\sigma}\right)^{-1/\xi_2-1}, \\
&\leq \left(\prod_{i=1}^n(\xi_2 x_i)^{-1/\xi_2-1}\right)\sigma^{n/\xi_2}\mathbb{I}\{\sigma\leq 1\}+\sigma^{-n}\mathbb{I}\{\sigma\geq 1\},
\end{aligned}$$

in the same way as in Equation (4.15). Combining both inequalities, we obtain

$$f_x(s,\sigma)p(\sigma) \leq Ce^{-s}p(\sigma)\left(\left(\prod_{i=1}^n(\xi_2 x_i)^{-1/\xi_2-1}\right)\sigma^{(n+1)/\xi_2}\mathbb{I}\{\sigma\leq 1\}+\sigma^{1/\xi_2-n+\epsilon}\mathbb{I}\{\sigma\geq 1\}\right),$$

with $C = \xi_2^2\max\left\{1,(1/\xi_2+\epsilon)^{\epsilon+\frac{1}{\xi_2}}\right\}$. This bound is integrable since $\mathbb{E}\left[\sigma^{(n+1)/\xi_2}\mathbb{I}\{\sigma\leq 1\}\right] < \infty$ and $\mathbb{E}\left[\sigma^{1/\xi_2-n+\epsilon}\mathbb{I}\{\sigma\geq 1\}\right] < \infty$. So by using the dominated convergence theorem, we obtain

$$\begin{aligned}
I_2(x) &\sim \frac{x^{-1/\xi_2}}{\log x}\int_0^{+\infty}\int_0^{+\infty} f(s,\sigma)p(\sigma)\mathrm{d}s\mathrm{d}\sigma \\
&= \frac{x^{-1/\xi_2}}{\log x}\xi_2^{2-1/\xi_2}\int_0^{+\infty}\sigma^{1/\xi_2-n}\prod_{i=1}^n\left(1+\xi_2\frac{x_i}{\sigma}\right)^{-1/\xi_2-1}p(\sigma)\mathrm{d}\sigma.
\end{aligned}$$

To conclude the proof, let us show that $I_1(x) = o\left(\frac{x^{-1/\xi_2}}{\log x}\right)$:

$$\begin{aligned}
I_1(x) &\leq \int_0^{+\infty} \int_{(1/\xi_2+\epsilon)^{-1}}^{\xi_2} \left(1 + (1/\xi_2+\epsilon)^{-1}\frac{x}{\sigma}\right)^{-1/\xi_2-\epsilon} \sigma^{-n} p(\sigma)\mathrm{d}\xi\mathrm{d}\sigma \\
&\leq \int_0^{1} \int_{(1/\xi_2+\epsilon)^{-1}}^{\xi_2} (1/\xi_2+\epsilon)^{1/\xi_2+\epsilon} x^{-1/\xi_2-\epsilon} \sigma^{1/\xi_2-n+\epsilon} p(\sigma)\mathrm{d}\xi\mathrm{d}\sigma \\
&+ \int_1^{+\infty} \int_{(1/\xi_2+\epsilon)^{-1}}^{\xi_2} \left(1 + (1/\xi_2+\epsilon)^{-1}x\right)^{-1/\xi_2-\epsilon} \sigma^{-n} p(\sigma)\mathrm{d}\xi\mathrm{d}\sigma \\
&= C\left(\mathbb{E}\left[\sigma^{(n+1)/\xi_2)}\mathbb{I}\{\sigma \leq 1\}\right] x^{-1/\xi_2-\epsilon} + \mathbb{E}\left[\sigma^{-n}\mathbb{I}\{\sigma \geq 1\}\right]\left(1 + (1/\xi_2+\epsilon)^{-1}x\right)^{-1/\xi_2-\epsilon}\right),
\end{aligned}$$

with $C = (\xi_2 - (1/\xi_2+\epsilon)^{-1})(1/\xi_2+\epsilon)^{1/\xi_2+\epsilon}$. So $I_1(x) = o\left(\frac{x^{-1/\xi_2}}{\log x}\right)$, which concludes.

Proof of Proposition 3 With an abuse of notation, assume that $u = 0$ and $\bar{F}(u) = 1$. The posterior mean of ℓ_α can be written

$$\ell_\alpha^{(2)} := \int_0^{+\infty} \int_{\xi_1}^{\xi_2} \frac{\sigma}{\xi}\left(\alpha^{-\xi} - 1\right)\sigma^{-n} \prod_{i=1}^{n}\left(1 + \xi\frac{x_i}{\sigma}\right)^{-1/\xi-1} p(\sigma)\mathrm{d}\xi\mathrm{d}\sigma$$

Similarly to the proof of Proposition 2, the change of variable $s = -(\xi_2 - \xi)\log\alpha$ leads to

$$\ell_\alpha^{(2)} = \frac{\alpha^{-\xi_2}}{\log 1/\alpha} \int_0^{+\infty} \int_0^{+\infty} f_\alpha(s,\sigma)p(\sigma)\mathrm{d}s\mathrm{d}\sigma,$$

with

$$f_\alpha(s,\sigma) := \frac{\sigma^{-n+1}}{\xi_2 + \frac{s}{\log\alpha}}\left(e^{-s} - \alpha^{\xi_2}\right)\prod_{i=1}^{n}\left(1 + \left(\xi_2 + \frac{s}{\log\alpha}\right)\frac{x_i}{\sigma}\right)^{-\left(\xi_2+\frac{s}{\log\alpha}\right)^{-1}-1} \mathbb{I}\{s \leq -(\xi_2-\xi_1)\log\alpha\}.$$

As $\alpha \to 0$, $f_\alpha(s,\sigma)$ converges to $f(s,\sigma) := e^{-s}\frac{\sigma^{-n+1}}{\xi_2}\prod_{i=1}^{n}(1+\xi_2\frac{x_i}{\sigma})^{-1/\xi_2-1}$. Similarly to Equation (4.15), we have

$$\sigma^{-n+1}\prod_{i=1}^{n}\left(1+\xi_2\frac{x_i}{\sigma}\right)^{-1/\xi_2-1} \leq \left(\prod_{i=1}^{n}(\xi_2 x_i)^{-1/\xi_2-1}\right)\sigma^{1+n/\xi_2}\mathbb{I}\{\sigma\leq 1\} + \sigma^{1-n}\mathbb{I}\{\sigma\geq 1\}, \tag{4.16}$$

and so $(s,\sigma) \mapsto f(s,\sigma)p(\sigma)$ is integrable on $\mathbb{R}^+\times\mathbb{R}^+$ as we assume that $\mathbb{E}\left[\sigma^{1+n/\xi_2}\mathbb{I}\{\sigma\leq 1\}\right] < \infty$ and $\mathbb{E}\left[\sigma^{1-n}\mathbb{I}\{\sigma\geq 1\}\right] < \infty$. Moreover, we have

$$f_\alpha(s,\sigma)p(\sigma) \leq \frac{\sigma^{-n+1}}{\xi_1}e^{-s}\prod_{i=1}^{n}\left(1+\xi_2\frac{x_i}{\sigma}\right)^{-1/\xi_2-1} p(\sigma),$$

which is integrable as we can use the same bound as in Equation (4.16). So by using the dominated convergence theorem, as $\alpha \to 0$, we obtain

$$\begin{aligned}
\ell_\alpha^{(2)} &\sim \frac{\alpha^{-\xi_2}}{\log 1/\alpha}\int_0^{+\infty}\int_0^{+\infty} f(s,\sigma)p(\sigma)\mathrm{d}s\mathrm{d}\sigma, \\
&\sim \frac{\alpha^{-\xi_2}}{\log 1/\alpha}\int_0^{+\infty}\frac{\sigma^{-n+1}}{\xi_2}\prod_{i=1}^{n}\left(1+\xi_2\frac{x_i}{\sigma}\right)^{-1/\xi_2-1} p(\sigma)\mathrm{d}\sigma,
\end{aligned}$$

which concludes.

Proof of Proposition 4 This proof relies on the theory of regularly-varying functions (see Bingham et al., 1989). As $H_{\text{pred}}(\cdot \mid \boldsymbol{x}^{(n)}) \in \text{MDA}(G_{\xi_2})$ using Proposition 2, $1/\bar{H}_{\text{pred}}(x \mid \boldsymbol{x}^{(n)})$ is regularly varying with index $1/\xi_2$. Using Proposition B.1.9 in (Haan and Ferreira, 2006), it is known that if an increasing function is regularly varying with index $1/\nu > 0$ then its inverse is regularly varying with index ν. As $\ell_\alpha^{(3)}$ is defined such that $\bar{H}_{\text{pred}}(\ell_\alpha^{(3)} \mid \boldsymbol{x}^{(n)}) = \alpha$, there exists a slowly-varying function L such that the posterior predictive quantile can be written

$$H_{\text{pred}}^{-1}(1 - 1/\alpha) = \ell_{1/\alpha}^{(3)} = \alpha^{\xi_2} L(\alpha),$$

or equivalently,

$$\ell_\alpha^{(3)} = \alpha^{-\xi_2} L(1/\alpha). \tag{4.17}$$

Moreover, using the expression obtained in Proposition 2, we have

$$\alpha = \bar{H}_{\text{pred}}(\ell_\alpha^{(3)} \mid \boldsymbol{x}^{(n)}) = c_n \frac{{\ell_\alpha^{(3)}}^{-1/\xi_2}}{\log \ell_\alpha^{(3)}} \left(1 + o(1)\right),$$

with

$$c_n := \frac{\xi_2^{2-1/\xi_2}}{(\xi_2 - \xi_1) p(\boldsymbol{x}^{(n)})} \int_0^{+\infty} \sigma^{1/\xi_2 - n} \prod_{i=1}^{n} \left(1 + \xi_2 \frac{x_i}{\sigma}\right)^{-1/\xi_2 - 1} p(\sigma) \mathrm{d}\sigma.$$

Combining with Equation (4.17), we obtain

$$\begin{aligned} c_n \frac{\alpha L(1/\alpha)^{-1/\xi_2}}{\xi_2 \log 1/\alpha + \log L(1/\alpha)} \left(1 + o(1)\right) &= \alpha \\ c_n L(1/\alpha)^{-1/\xi_2} \left(1 + o(1)\right) &= \xi_2 \log 1/\alpha + \log L(1/\alpha). \end{aligned}$$

Since L is slowly varying, we have $\frac{\log L(1/\alpha)}{\log 1/\alpha} \to 0$ as $\alpha \to 0$ (see Bingham et al., 1989, Proposition 1.3.6), and therefore

$$\begin{aligned} & c_n L(1/\alpha)^{-1/\xi_2} \left(1 + o(1)\right) = \left(\xi_2 \log 1/\alpha\right) \left(1 + o(1)\right) \\ \iff & L(1/\alpha)^{-1/\xi_2} = \left(-\tfrac{\xi_2}{c_n} \log \alpha\right) \left(1 + o(1)\right) \\ \iff & L(1/\alpha) = \left(-\tfrac{\xi_2}{c_n} \log \alpha\right)^{-\xi_2} \left(1 + o(1)\right)^{-\xi_2}. \end{aligned}$$

So $\ell_\alpha^{(3)} \sim \left(-\frac{\xi_2}{c_n} \alpha \log \alpha\right)^{-\xi_2}$ as $\alpha \to 0$, which concludes.

Chapter 5

Case studies: Bayesian estimations of extreme river flows and wind speed return levels

Contents

Résumé

Dans ce chapitre, nous illustrons les résultats des chapitres précédents (Chapitres 2, 3 et 4) sur divers ensembles de données environnementales fournis par EDF : trois ensembles de données de débits de cours d'eau et trois ensembles de données de vitesses du vent, à Tours, Reims et Orange.

Les objectifs sont de vérifier le comportement sur données réelles, de fournir une estimation bayésienne des niveaux de retour, de comparer les résultats avec les études internes précédentes d'EDF, ainsi que de répondre à la question principale sur les limites de l'extrapolation dans des exemples pratiques.

Une limite majeure à l'application de nos modèles utilisés jusqu'ici est l'hypothèse d'observations indépendantes et identiquement distribuées que nous avons supposée tout au long du manuscrit. Cette hypothèse n'est pas raisonnable pour les données brutes en raison de la dépendance temporelle journalière, ainsi que de la saisonnalité et des tendances à plus grande échelle de temps. Par conséquent, des étapes de traitement des données sont nécessaires avant l'application des modèles.

Après une brève introduction des motivations et un rappel des notations (Partie 5.1), les différentes étapes de prétraitement des données sont décrites en Partie 5.2, et la configuration des expériences est présentée en Partie 5.3. Les résultats sont ensuite exposés en Partie 5.4, tandis que la conclusion propose différentes pistes d'amélioration (Partie 5.5).

Abstract

In this chapter, we illustrate the results from the previous chapters (Chapters 2, 3 and 4) on various environmental datasets provided by EDF: three river flow datasets and three wind speed datasets, located in Tours, Reims, and Orange.

The objectives are to examine the behavior on different real-world datasets, provide a Bayesian estimate of return levels, compare the results with previous internal studies conducted by EDF, and address the main question regarding the limitations of extrapolation in practical examples.

A major limitation in the application of our models is the assumption of independent and identically distributed data that we have made throughout the manuscript. This assumption is not reasonable for raw data due to daily temporal dependence, as well as seasonality and trends on larger time scales. Therefore, data preprocessing steps are necessary before applying the models.

After a brief introduction of the motivations and a reminder of the notations (Section 5.1), the various data preprocessing steps are described in Section 5.2, and the experimental setup is presented in Section 5.3. The results are then presented in Section 5.4, while the conclusion suggests different avenues for improvement (Section 5.5).

5.1 Introduction

5.1.1 Motivations

After proposing improvements for Bayesian inference of extreme models in the previous chapters, we focus here on the behavior of the Bayesian framework on different environmental datasets provided by EDF. The objectives are multiple:

1. Verify the behavior on real-world datasets, and identify application-driven paths of exploration.
2. Provide a Bayesian estimate of extreme return levels and compare the results with previous EDF studies as well as standard frequentist methods.
3. Investigate experimentally the limits of extrapolation of extreme value modelling.

A first case study has been investigated in Chapter 2 with the estimation of extreme return levels for the Garonne river flow data. We complete this by analysing six other datasets: three of them correspond to river flows in other locations in France: the Loire at Tours, the Meuse at Reims and the Rhône at Orange, and the three others correspond to wind speed in the same places. The datasets come from *Banque Hydro* or has been directly produced internally by EDF. Note that a bivariate study of the observations in Reims is already available in Sibler and Dutfoy (2021). More generally, this analysis belongs to a vast literature of extreme value analysis applied to natural hazard: see Pan et al. (2022) for a recent review on flow analysis, and Walshaw (1994), Holmes and Moriarty (1999), Larsén et al. (2015) for extreme wind estimation.

Each dataset is studied in univariate way following the same steps, described successively in the next sections. After a quick definition of the model in Section 5.1.2, we describe all the preprocessing steps in Section 5.2, and the Bayesian setup in Section 5.3. Results are finally presented in Section 5.4 for the six datasets, and discussed in Section 5.5.

5.1.2 Bayesian extreme value modeling

For each datasets, we use a Poisson process characterisation of extremes to model observations over a given threshold and therefore deduce estimation of extreme return levels. Recall that the likelihood associated with this process for n_u observations $(x_1, \dots, x_{n_u})$ above a threshold u is given by:

$$L(\boldsymbol{x}, n_u \mid \mu, \sigma, \xi) = \exp\left(-m\left(1+\xi\left(\frac{u-\mu}{\sigma}\right)\right)^{-1/\xi}\right)\sigma^{-n_u}\prod_{i=1}^{n_u}\left(1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right)^{-1-1/\xi}, \tag{5.1}$$

with $(\mu, \sigma, \xi) \in \mathbb{R} \times \mathbb{R}^+ \times \mathbb{R}$ and where $m > 0$ is a scaling factor corresponding to the number of years of observation. The objective here is to estimate the return level ℓ_T corresponding to a value exceeded on average once every T years. If G denotes the cdf of the generalized extreme value (GEV) distribution, then:

$$\ell_T = G^{-1}(1-1/T \mid \mu, \sigma, \xi) = \mu - \frac{\sigma}{\xi}\left(1-(-\log(1-1/T))^{-\xi}\right). \tag{5.2}$$

The estimation of ℓ_T through the estimation of the three parameters (μ, σ, ξ) will be done under the Bayesian paradigm. We refer to Chapter 1 for a more detailed introduction on

extreme value theory and Bayesian statistics, and to Chapter 2 for details on the approach that will be used here for a Bayesian estimation of extreme return levels.

5.2 Preprocessing

Several preprocessing steps are necessary to obtain data that can be used for inference. Each of them have already been studied in the literature of extremes applied to environmental data. They lead to hyperparameter choices based here on graphical methods combined with analyses already carried out in internal reports at EDF. We detail here the steps that will be performed for each case.

5.2.1 Jittering

The datasets are obtained with instruments that have a finite precision. Therefore, the obtained measurements are rounded to a given precision, which in our case is to the tenth or even to the nearest unit. This is an issue as we assume that our variable are continuous, and so observing identical data should be a zero probability event, but a lot of rounded measure have the same value here. One way to overcome this issue is to add a continuous noise, which is called a jittering approach (Andreewsky and Bousquet, 2021). Here, we choose a centered uniform noise with a support length equal to the precision of roundness (for example, a uniform $\mathcal{U}(-0.05, 0.05)$ for variables rounded to the nearest tenth). From an extreme value point of view, if the tail index of the observations is higher than the one of a uniform variable (which is $\xi = -0.5$), then its value is unchanged after jittering. However, because of the independence between the noise and the measurements, this step increases slightly the variance of the observations.

5.2.2 Seasonality

Alternating between seasons might change the behaviour of environmental variables, especially when studying extreme events. Therefore, seasonal variability prevents us from considering that the observations are independant and identically distributed (i.i.d).

Boxplots for the different months in the year are represented in Figure 5.1 for the six series. By looking at the flow datasets, we can distinguish two seasons where observations appear to be stationary: one rainy season, mainly during winter, contains the most extreme events, and a drier season in summer. However, this division depends on the location. One can include the spring into the rainy season for the Rhône (Orange), whereas it is more in the dry one for the Meuse (Reims). Concerning the wind speed series, the seasonal variability is less clear. The variance of the observations during summer seems a little lower than over the rest of the year, but in the same time the mean seems to be constant over the months. Moreover, the observation of the maximum in July for the data at Tours indicates that extreme events can be expected during the dry season.

In the extreme value literature, different strategies exist to handle seasonality (more details can be found in Davison and Smith (1990), Fawcett and Walshaw (2016)).

1. The first one consists in keeping only the season concerned by extreme events, or alternatively to fit a model to each season. The latter option assumes that the

	River flow	Wind speed
Tours	December to May	Entire year
Reims	December to February	October to March
Orange	October to May	October to May

Table 5.1: Summary of months kept for inference for each dataset.

observations are governed by the same physical phenomena over the year, but with different intensities depending on the season, which can typically be the case of wind speeds as they are governed by patterns of anticyclone and depressions (Fawcett and Walshaw, 2016).

2. The other option would be to model a time dependence for the parameters of the extreme value distribution: $(\mu(t), \sigma(t), \xi(t))$, that is typically written with a parameter vector and some covariates that depend on t but can also depend on other factors, see Chavez-Demoulin and Davison (2005) for a general additive model example.

These datasets have already been studied by EDF and have led to the choice of keeping only the rainy season on indicated months, even on datasets where no seasonality is really apparent (except for the wind speed in Tours where the entire year is kept). We maintain these choices, for example for the wind speed in Reims, where only the months from October to March are considered, as the justification comes from a physical analysis. Thus, based on the boxplots in Figure 5.1 and on the previous studies, we keep this choice of removing the dry season for all the datasets. Table 5.1 summarizes the season kept for each case.

5.2.3 Temporal dependence

Even with seasonality removed, the i.i.d. assumption for excesses over a given threshold is unrealistic here, as extreme events are due to meteorological conditions that can last several days of observations. This is reflected by the partial autocorrelation graph for the six series in Figure 5.2, which corresponds to the autocorrelation function at each lag $t \in \{0, \dots, 25\}$ after removing the contribution of smaller lags k between 1 and t. Figure 5.2 confirms the correlation between observations separated by only a few days. From there, two strategies exist.

1. **Removing dependence by declustering** (Davison and Smith, 1990). The idea is simply to merge data close in time, as they correspond to the same extreme event. Inference is then performed on clusters of exceedances, with a parameter r_c corresponding to the maximal interval of days between two consecutive excesses inside one cluster. As pointed out in (Coles, 2001, Chapter 5), this method has the limitation to reduce even more the data that will be used for inference, and more generally can be sensitive to the choice of r_c.

2. **Considering dependence by estimating the extremal index**. Under suitable conditions (see Coles, 2001, Chapter 5), one can show that the distribution of the maximum of a stationary sequence of n random variables M_n can be approximated by $F^{n\theta}$ if F is the distribution of the observations, with $\theta \in (0, 1]$ (for independent

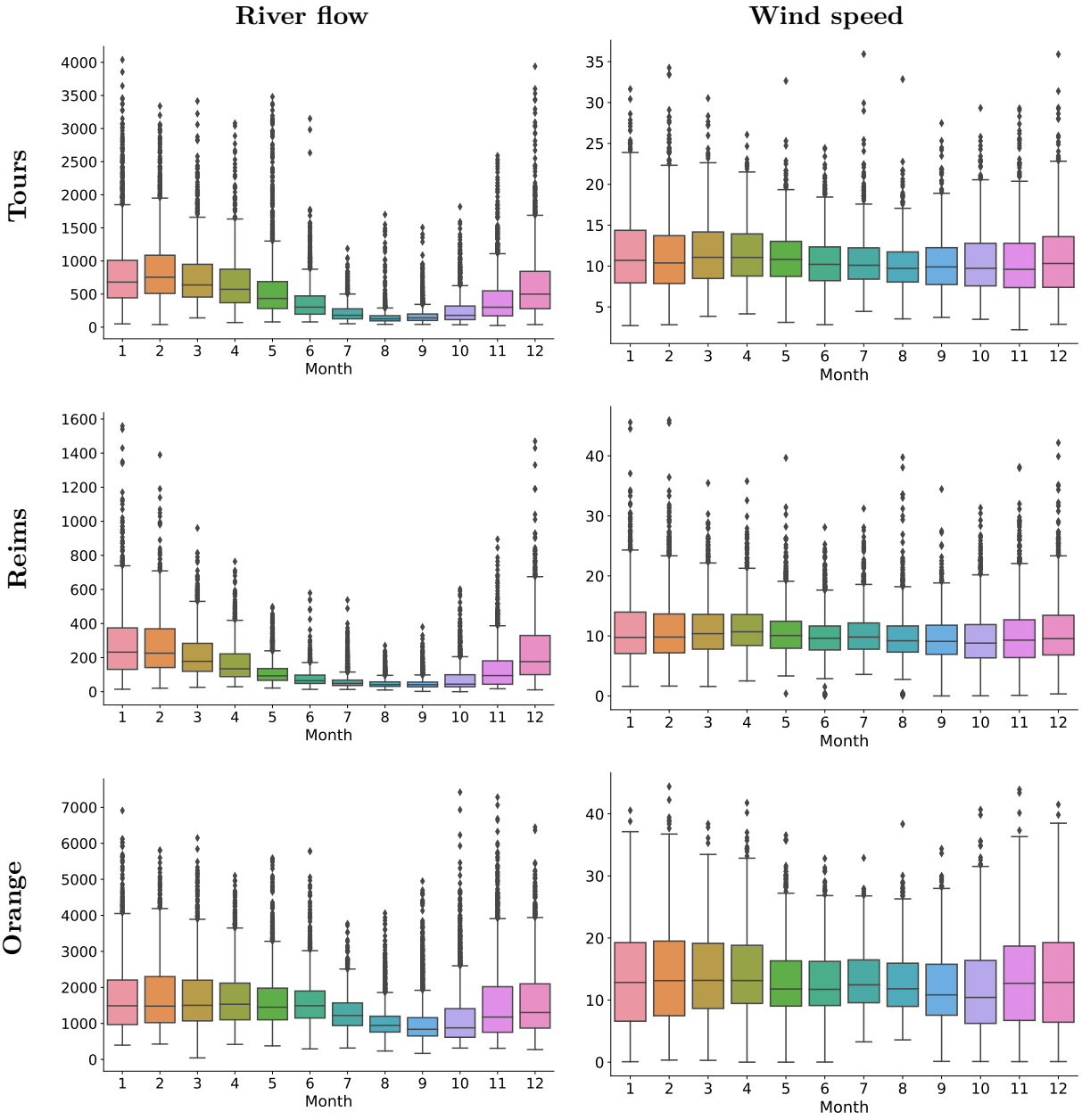


Figure 5.1: Boxplots as functions of the different months of the year (1 = January, ... , 12 = December) for the river flow and the wind speed at Tours, Reims, and Orange.

series, $\theta = 1$). The estimation of θ (known as the extremal index) can be done after fitting an extreme value model: see Ferro and Segers (2003) for an example of an estimator of θ. The dependence can also be modeled by assuming a Markov chain behaviour, see for example Fawcett and Walshaw (2012).

In our case, the declustering method seems satisfactory as the partial autocorrelation graphs in Figure 5.2 provide an easy way to identify a reasonable choice of r_c. Moreover, the excesses are sufficiently sparse so that the sensitivity to r_c for the final number of clusters is reduced. By looking at Figure 5.2, we observe a similar behaviour in the three sites. The cluster size is set to $r_c = 3$ for the three river flow datasets and to $r_c = 2$ for the three wind speed ones.

5.2.4 Threshold elicitation

The choice of threshold is crucial for models using GPD or Poisson process characterisation of extremes. If it is too large, the asymptotic approximation by the limit law will be accurate but the small number of observations will lead to poor estimations of parameters. Conversely, it is too small, the parameter estimation will be improved as the sample size increases, but the asymptotic theoretical behaviour will not be verified. With the Bayesian point of view, the threshold governs the importance given to the prior versus observations.

In the general case, a variety of methods based on graphical methods, non parametric estimation or probabilitic results exist and none of them seems universally accepted, which makes this question still open. A review of threshold selection can be found in Scarrott and MacDonald (2012) and more recently in Pan et al. (2022). In practice, the most common methods are the graphical diagnostics described in (Coles, 2001, Chapter 4): mean residual life plot, Hill plot, QQ plot, or return level plot. In particular, the mean residual life plot has the benefit of being used in all domains of attraction (provided $\xi < 1$). It is based on the property of GPD distributions to have a linear relationship for the mean over the threshold u as a function of u. Therefore, u is chosen as the lowest level that leads to linearity for sample mean excesses. Although this method has already been criticized by Coles (2001) for its difficulty to spot linearity in practice, it is the most popular and still seems to be the most satisfactory according to a recent comparison of Langousis et al. (2016).

Mean residual life plots for the six datasets are shown in Figure 5.3. Some of them allow to draw a clear hypothesis about the choice of threshold, typically the river flow in Tours ($u = 1700$) or the wind speed in Reims ($u = 26$), but others are much less clear. Typically, linearity is far from being verified for the data in Orange. For all datasets, we are helped with previous studies at EDF which suggest a threshold and allow us to have a value when the mean residual life plot is not conclusive, and to confirm the choices when a value is suggested by the plot. A summary of the chosen threshold for each dataset is given in Table 5.2.

A representation of the obtained datasets after these steps is shown in Figure 5.4. The number of observations varies between $n_u = 57$ for the smallest one (river flow in Reims) and $n_u = 272$ for the largest one (river flow in Orange). To assess the robustness of our estimations with respect to the choice of u, we show in Figure 5.6 the evolution of the posterior mean of ξ and the associated credible interval as a function of u.

Note also that looking at the different plots of Figure 5.4, a temporal trend could

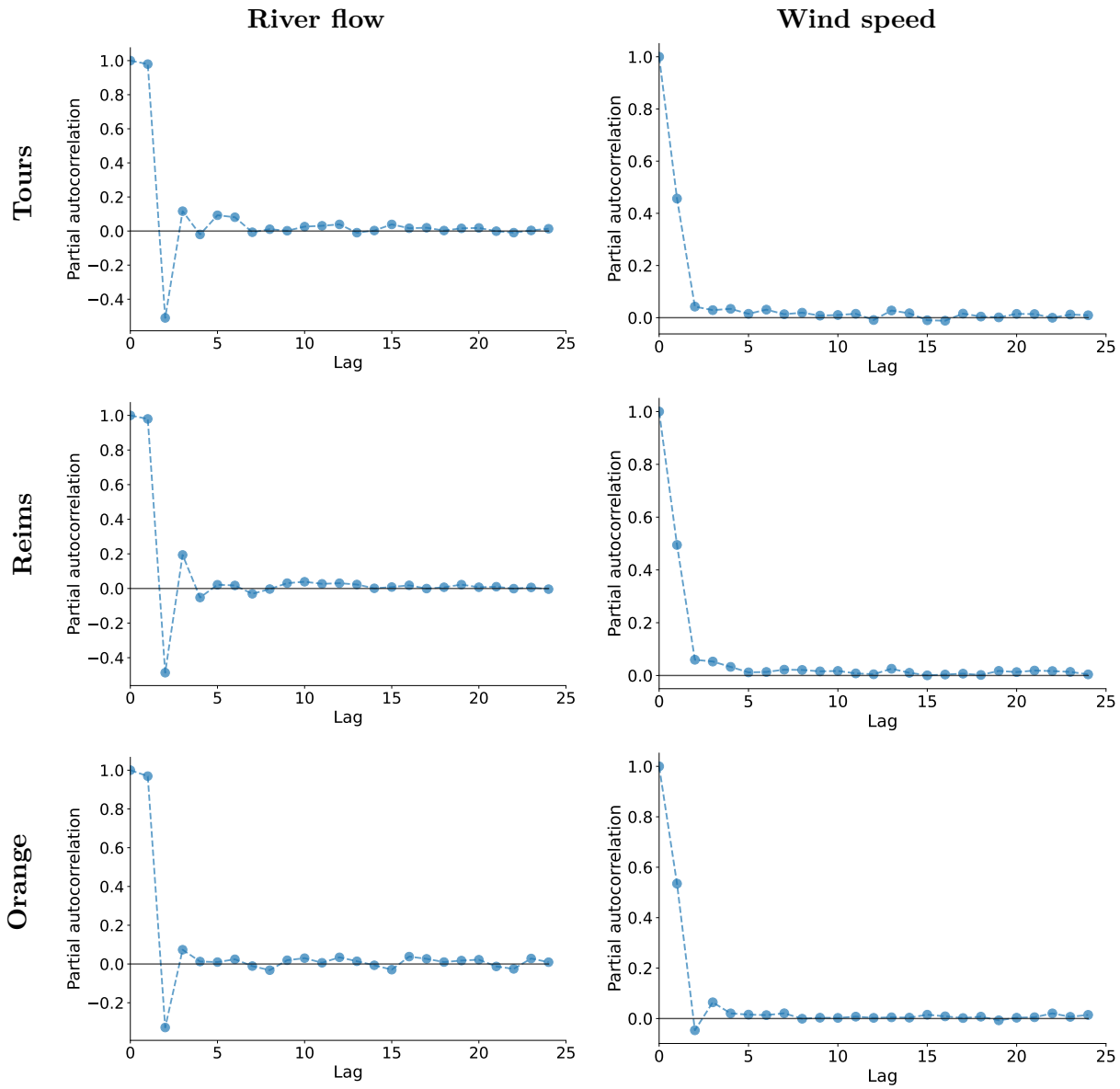


Figure 5.2: Partial autocorrelation graphs as functions of the lag for the river flow and the wind speed at Tours, Reims, and Orange.

	River flow	Wind speed
Tours	1700	22
Reims	620	26
Orange	3000	29

Table 5.2: Summary of chosen threshold for each dataset.

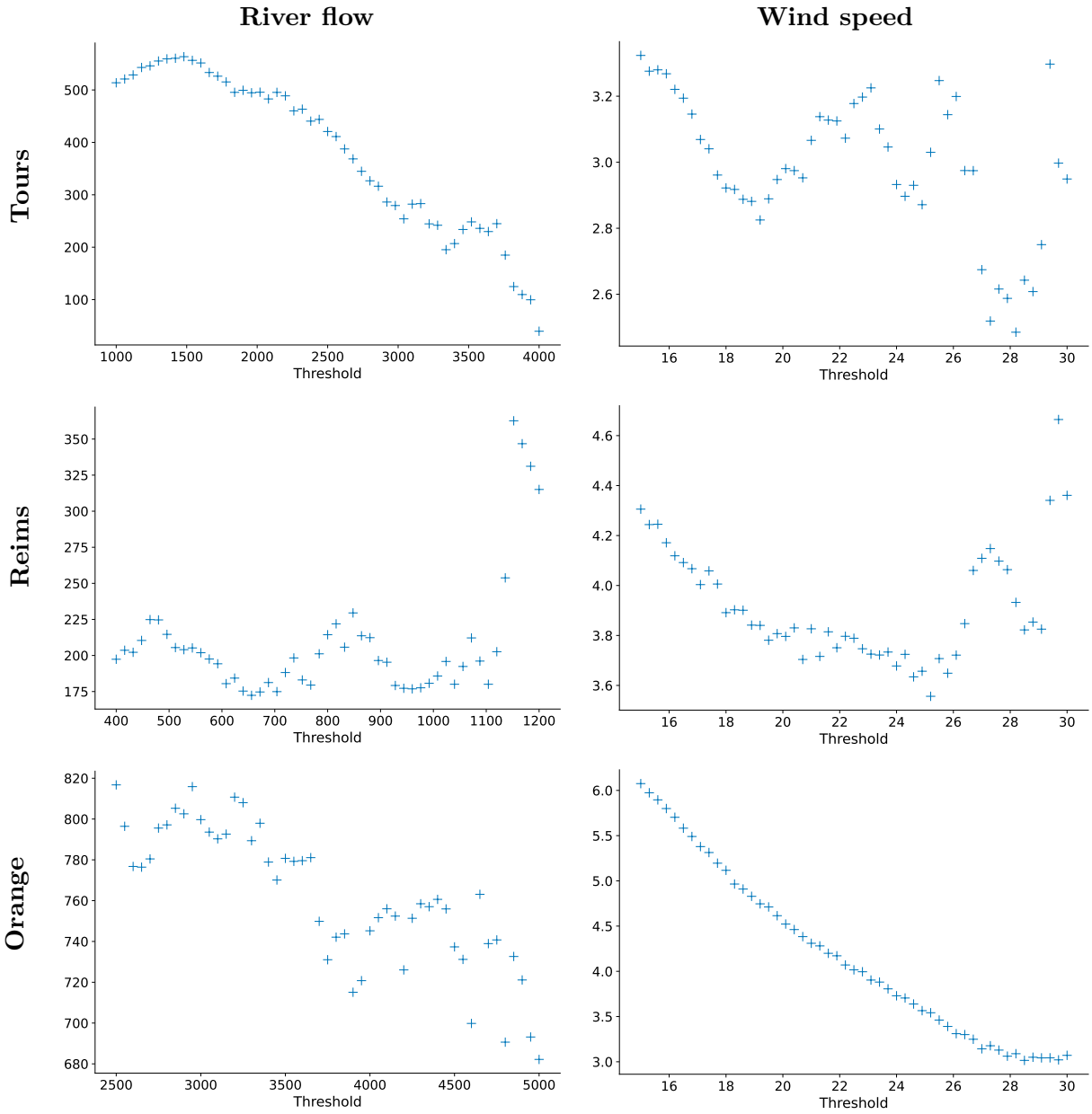


Figure 5.3: Mean residual life plots for the river flow and wind speed at Tours, Reims, and Orange. The threshold choice would be the one for which the curve starts to be linear (good approximation of the excesses by the GPD).

be suspected for some series. A possible way to model such a phenomenon is to include covariates in the extreme value model (See Section 2.2 and Gardes and Girard (2010)).

5.3 Experimental Setup

5.3.1 Prior

In all our cases here, no external information about the parameters is provided. Therefore, the Jeffreys prior on (r, ν, ξ) (see Section 2.3) could be considered, as it is based on an uninformative rule and leads to a proper posterior (see Proposition 2). However, even without any expert information, this kind of dataset has already been studied in the literature:

- It is common to assume that wind speed variables belong to the Gumbel's domain of attraction with $\xi = 0$ (See references in Parey et al., 2021), although there are case studies where inference is improved if one assumes $\xi < 0$ (Walshaw, 1994). As physical models do not suggest any upper bound to wind speed, this observation is justified by a slow convergence of the maxima to a Gumbel distribution.
- Concerning river flows, as in other hydrological applications, the distributions are usually heavy-tailed, with $\xi > 0$ (Langousis et al., 2016).

These two cases, like the majority of natural phenomena, can be modelled by a shape parameter in the range $(-1, 1)$, even without any prior information on ξ. Therefore, a PC prior on ξ defined in Section ?? can be used to include this information. This prior allows the user to navigate between the uninformative case and the deterministic one where $\xi = 0$. In our experiments, the influence of the prior information on the posterior estimation will be checked by comparing three priors:

1. The Jeffreys prior defined in (2.11), supposed to be uninformative.
2. A PC prior with $\lambda = 5$, which corresponds to a prior confidence level of 95% to have $\xi \in [-0.8, 0.4]$. This can be seen as the intermediate case, where the prior information approximately corresponds to our prior knowledge on natural hazards behaviour.
3. A PC prior with $\lambda = 10$, which corresponds to a prior confidence level of 95% to have $\xi \in [-0.3, 0.3]$. This prior corresponds to a case of stronger confidence to fact that $|\xi|$ is near 0. Although in practice it does not correspond to the information we have, this limiting case allows us to see the uncertainty mitigation if one makes such an assumption.

5.3.2 MCMC algorithm

The implementation of the MCMC algorithms uses the Python library PyMC3 (Salvatier et al., 2016). Four chains of 5000 Metropolis–Hastings iterations are performed on the (r, ν, ξ) parameterization, defined in Section 2.2. A burn-in period of 1000 iterations is done beforehand. Different convergence diagnostics such as the effective sample size

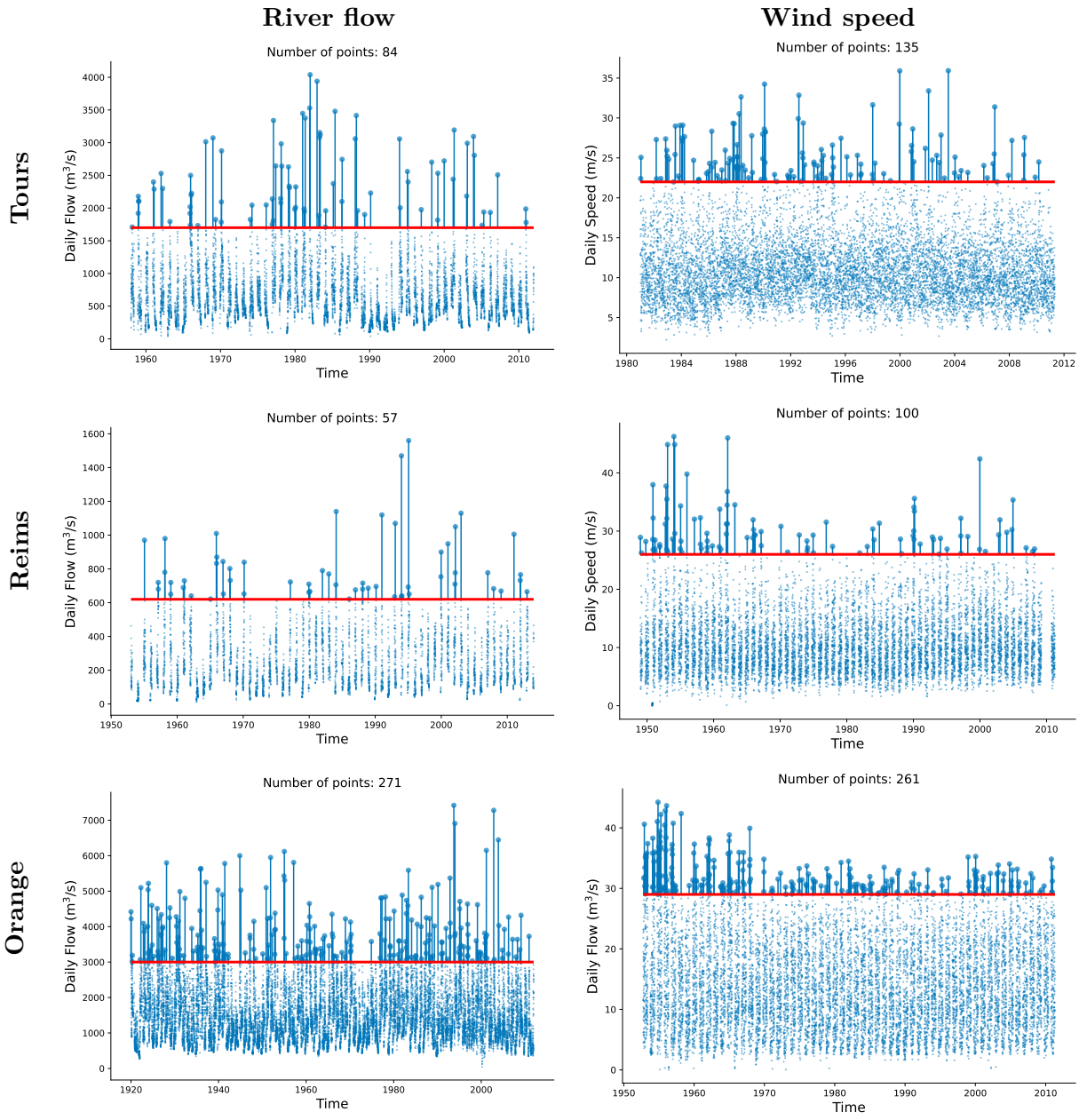


Figure 5.4: Data above the threshold (in red) obtained after preprocessing for the river flow and wind speed at Tours, Reims, and Orange.

(ESS), autocorrelation graphs, and the local version of the Gelman–Rubin diagnostic $\hat{R}_\infty$ proposed in Chapter 3, indicate a convergence to the posterior distribution on each of the datasets.

5.3.3 Return level estimation

From a Bayesian perspective, the parameters $\boldsymbol{\theta} := (\mu, \sigma, \xi)$ resulting from the transformation of (r, ν, ξ) follow a posterior distribution. Various methods can be employed to combine these parameters and estimate the return level as defined in Equation (5.2), as detailed in Chapter 4. In this study, we will utilize the second method outlined in Section 4.2, where the return level is treated as a random variable (a function of $\boldsymbol{\theta}$). The posterior mean of this variable is then determined as follows:

$$\hat{\ell}_T^{(2)} = \mathbb{E}_{p(\cdot|\mathbf{x})}(G^{-1}(1 - 1/T \mid \boldsymbol{\theta})).$$

Subsequently, the quantification of uncertainty is achieved by calculating the posterior quantiles of the return level:

$$\mathrm{CI}^{(2)} = [q_{2.5}(G^{-1}(1 - 1/T \mid \boldsymbol{\theta})), q_{97.5}(G^{-1}(1 - 1/T \mid \boldsymbol{\theta}))].$$

A comparative analysis, both theoretical and experimental, of this method along with two others is presented in Chapter 4. However, no definitive conclusion regarding the preferred choice can be drawn from this investigation. It is worth noting that, according to the empirical comparison by Jonathan et al. (2021), the posterior mean of the return level is often the preferred option.

5.4 Results

5.4.1 Parameter estimation

The results of the Bayesian estimation of (μ, σ, ξ) are reported in Table 5.3 and Figure 5.5 for the river flow and wind speed datasets. For each case study, the results with the different priors are given with a frequentist estimation using the maximum likelihood estimator (MLE) as implemented in the extRemes package (Gilleland and Katz, 2016). Note that despite its asymptote in -0.5, Jeffreys prior seems to yield the closest estimation to the MLE.

As expected, the credible intervals are tighter when information is added in the prior, i.e. when a PC prior is used and when λ increases. The posterior mean of ξ is approaching zero, and the posterior mean of σ compensates this constraint in ξ due to the prior (if ξ increases with λ, σ decreases, and conversely). Except for the river flow in Tours, all the estimations indicate a value of ξ near 0. For the river flow in Reims and Orange, the assumption that $\xi > 0$ is plausible but one cannot reject the hypothesis that $\xi \leq 0$. For the wind speed series, the assumption that $\xi = 0$ seems to be confirmed on the three sites. More generally, this confirms the information included in the PC prior that ξ will be close to zero. Therefore, its influence is relatively weak as it does not contradict the data, and also because of the sufficient number of observations: the estimations in Reims are those that vary the most as the sample size is the smallest ($n_u = 57$ for the river flow and $n_u = 107$ for the wind speed).

River flow

		Jeffreys			**PC**($\lambda = 5$)			**PC** ($\lambda = 10$)			**MLE**	
		Mean	SD	95%-CI	Mean	SD	95%-CI	Mean	SD	95%-CI	MLE	SD
Tours	μ	2113	108	[1904, 2324]	2072	96	[1881, 2253]	2054	94	[1871, 2239]	2275	106
	σ	845	98	[652, 1032]	784	87	[620, 959]	760	83	[603, 921]	754	72
	ξ	−0.34	0.11	[−0.50, −0.14]	−0.23	0.12	$[-0.43, 10^{-3}]$	−0.160	0.12	[−0.38, 0.05]	−0.33	0.10
Reims	μ	612	23.2	[566, 659]	610	22.7	[565, 655]	610	24.1	[562, 657]	619	21.3
	σ	169	36.1	[103, 242]	172	32.0	[111, 235]	178	30.7	[117, 236]	159	32.9
	ξ	0.12	0.16	[−0.17, 0.43]	0.09	0.12	[−0.12, 0.34]	0.05	0.09	[−0.12, 0.25]	0.11	0.15
Orange	μ	3860	79.1	[3709, 4016]	3856	77.2	[3709, 4013]	3856	76.1	[3713, 4010]	3989	0.31
	σ	811	53.3	[710.1, 912.4]	808	51.8	[705, 906]	805	51.1	[704, 903]	818	0.32
	ξ	0.03	0.07	[−0.09, 0.18]	0.04	0.06	[−0.10, 0.16]	0.02	0.06	[−0.08, 0.14]	-1.10^{-9}	0.06

Wind speed

		Jeffreys			**PC**($\lambda = 5$)			**PC** ($\lambda = 10$)			**MLE**	
		Mean	SD	95%-CI	Mean	SD	95%-CI	Mean	SD	95%-CI	MLE	SD
Tours	μ	26.9	0.49	[26.0, 27.9]	26.9	0.50	[26.0, 27.9]	26.8	0.49	[25.9, 27.8]	26.6	0.48
	σ	3.15	0.29	[2.66, 3.73]	3.20	0.29	[2.66, 3.79]	3.20	0.30	[2.65, 3.80]	3.10	0.26
	ξ	−0.08	0.09	[−0.25, 0.10]	−0.05	0.08	[−0.21, 0.11]	−0.03	0.07	[−0.18, 0.11]	−0.10	0.09
Reims	μ	28.0	0.47	[27.1, 29.0]	28.0	0.46	[27.1, 28.9]	28.0	0.46	[27.1, 29.0]	28.7	0.52
	σ	4.02	0.48	[3.15, 5.00]	4.04	0.46	[3.20, 4.98]	4.07	0.44	[3.24, 4.96]	4.02	0.44
	ξ	0.09	0.11	[−0.11, 0.32]	0.07	0.09	[−0.09, 0.27]	0.05	0.08	[−0.10, 0.20]	0.08	0.11
Orange	μ	33.5	0.35	[32.8, 34.2]	33.5	0.35	[32.9, 34.2]	33.5	0.35	[32.9, 34.2]	33.2	0.34
	σ	3.13	0.24	[2.69, 3.60]	3.12	0.23	[2.69, 3.57]	3.10	0.21	[2.67, 3.51]	3.08	0.22
	ξ	0.04	0.08	[−0.10, 0.19]	0.03	0.06	[−0.09, 0.17]	0.02	0.06	[−0.08, 0.15]	0.03	0.08

Table 5.3: Posterior mean and 95% credible interval (CI) for River flow and wind speed studies with Jeffreys prior and PC prior with $\lambda \in \{5, 10\}$. An estimation by maximum likelihood (MLE) and the associated standard deviation estimate (SD) are also given for each case.

The case of the river flow data in Tours is more questionable: it is clear that we are in the Weibull domain of attraction ($\xi < 0$), and that the assumption of having $\xi = 0$ can be rejected at 95% even when some prior information is added to penalize values far from 0 (for $\lambda = 5$, 0 is at the frontier of the 95% credible interval but almost all the values inside are negative). However, ξ seems also too negative for Jeffreys prior, defined for $\xi > -0.5$: this support constraint is also applied to the posterior and the 95% credible interval hits this value of -0.5. This border effect can be seen on the top right plot in Figure 5.5: it is clear that this behaviour is not due to the data but on a poor choice of prior and its asymptote that favors -0.5. This confirms that our choice of a PC prior with $\lambda = 5$ is the most reasonable for all our case studies.

5.4.2 Influence of threshold selection

To assess the robustness of our estimation with respect to the choice of threshold u, Figure 5.6 shows the evolution of the posterior mean of ξ and the associated credibility interval at 95% for each dataset and three priors: Jeffreys prior, PC prior with $\lambda = 1$ which is almost flat for $\xi < 1$, and PC prior with $\lambda = 10$ which concentrates around zero. A first remark that can be made is that estimators using PC priors are more stable than those using Jeffreys' one, even PC prior with $\lambda = 1$ which can be seen as uninformative.

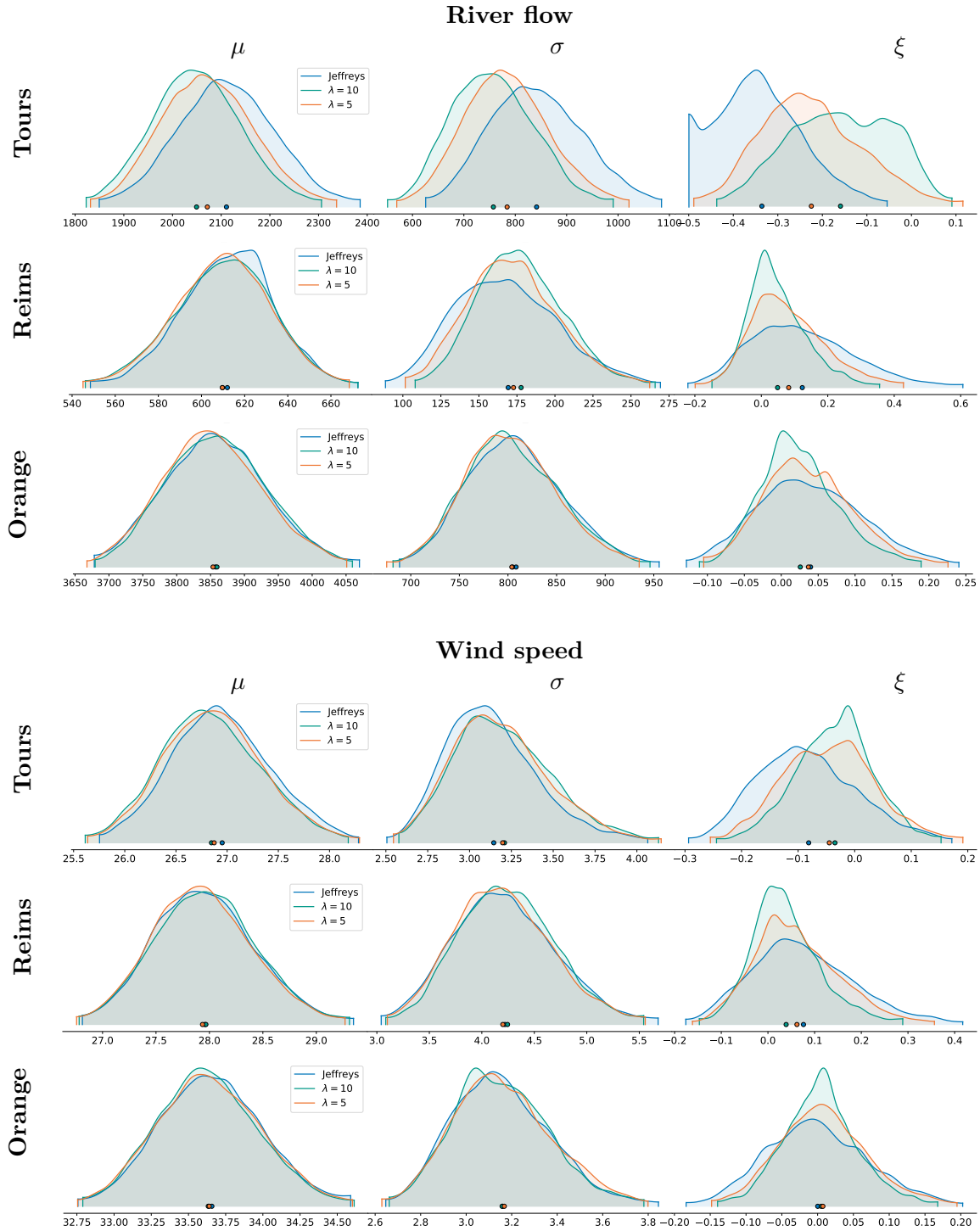


Figure 5.5: Posterior distibutions of (μ, σ, ξ) for Jeffreys prior and PC prior with $\lambda \in \{5, 10\}$, applied on the three River flow and the three wind speed datasets.

This is especially apparent for the wind speed a Orange and Reims, and the river at Reims where the estimation of ξ decreases with the threshold. This regularization effect comes with a bias that can be added for values near 0, which is the case for all datasets with the PC prior with $\lambda = 10$. However, Jeffreys prior puts a significant part of its mass near -0.5, which can also induce negative bias in particular for the estimation of a negative ξ like the River flow at Tours. Priors have therefore an influence on posterior inference even in the case of Jeffreys, supposed to be uninformative.

5.4.3 Return level estimation

The extrapolation of the return level curve is shown in Figure 5.7 for the three priors. The associated centennial, millennial, and deca-millennial return levels are reported in Table 5.4 for the river flow and the wind speed cases.

There are cases where the results differ significantly with the prior, but the plots in Figure 5.7 show that the observed annual maxima are always included in all credible intervals, which prevents us from excluding a potential wrong choice of prior. For the river flow at Tours, the disagreement between the priors was expected in view of the differences of posterior for ξ (see Figure 5.5). However, we also observe strong differences between return level curves for the datasets at Reims. This corresponds to studies with the lowest number of observations, and even if the posterior estimations for the parameters seemed to agree, this difference here can be explained by the difference in variability for ξ (see Figure 5.5). Typically, the credible intervals for ξ are wider with Jeffreys prior, and adding high values of ξ increases the estimation of the posterior mean of the return level. This intuition corroborates the results found in Chapter 4, where we show that the highest values of ξ determine the tail behaviour of the posterior return level. For the river flow and wind speed in Orange, the prior influence is less important due to a higher number of observations.

Note that despite some differences with the three priors for some datasets, the estimation of the return levels using MLE are always in the credible intervals at 95%.

5.4.4 Limits of extrapolation

An important question is that of the limit of extrapolation of return levels for these studies: looking at the uncertainties around the estimation, is it reasonable to provide a deca-millennial return level? To this end, several sources of uncertainty must be considered:

- **Uncertainty related to data collection/preprocessing.** Here we focus on a statistical model using data that are supposed to be i.i.d. This is clearly not the case of the environmental variables under study, where the randomness comes both from the lack of knowledge of the underlying physical model and from the uncertainty associated with the collection of measurements. The i.i.d. assumption requires therefore the use of several steps that are described in Section 5.2. However, all of these steps are questionable in their ability to really solve the problem, in particular for these quantities that are likely to be impacted by climate change.
- **Uncertainty related to model misspecification.** In addition to the approximations associated with the data assumptions, there are also those associated with the asymptotic model: the Poisson process characterization of extremes comes from

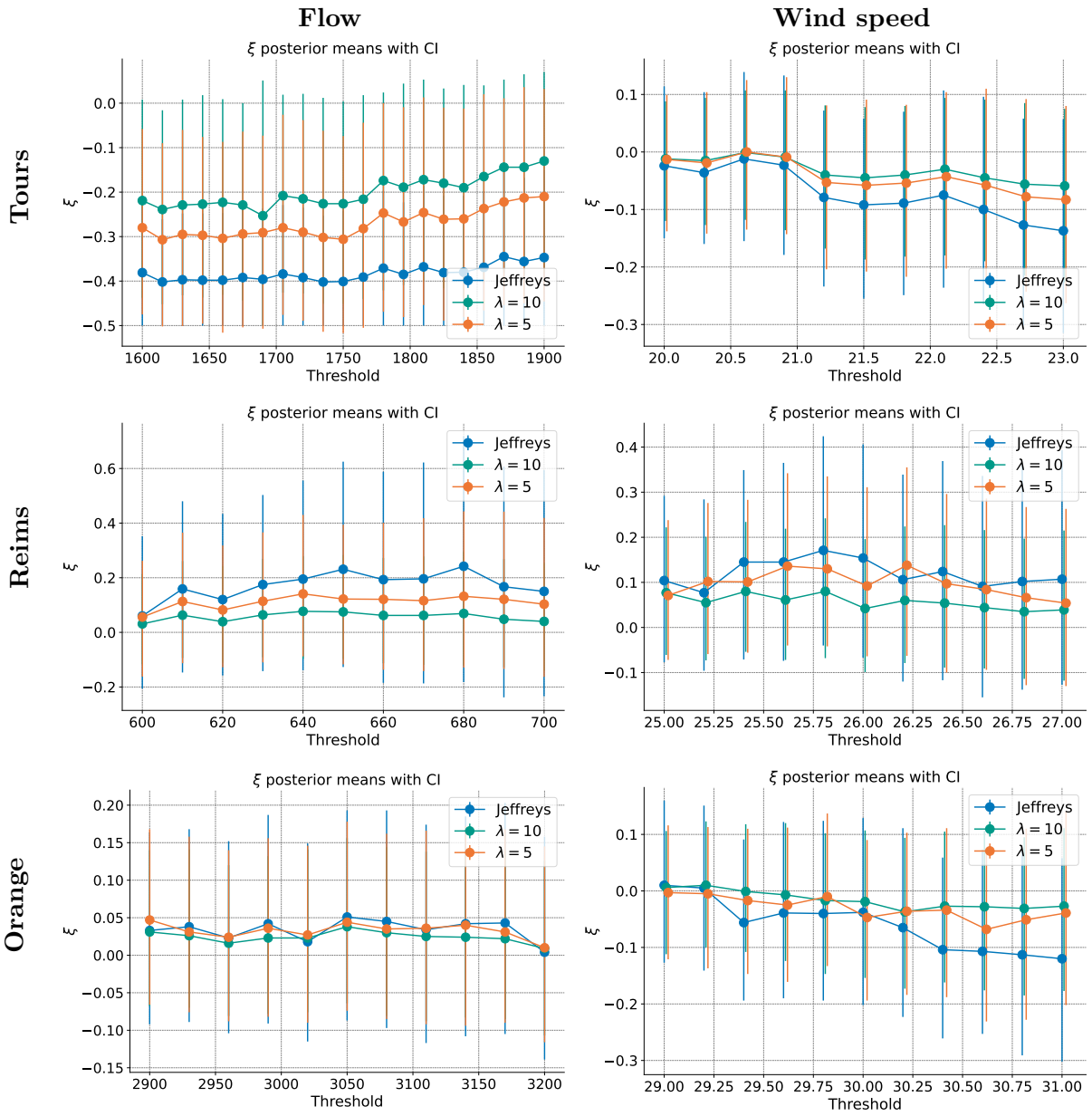


Figure 5.6: Evolution of the posterior mean of ξ and the associated credible interval at 95% as a function of the threshold for the River flow and wind speed at Tours, Reims, and Orange. For each plot, the three curves correspond to different choices of prior: Jeffreys prior (in blue), PC prior with $\lambda = 1$ (in green) and $\lambda = 10$ (in red).

River flow return levels

	Return Period	Jeffreys		PC ($\lambda = 5$)		PC ($\lambda = 10$)		MLE
Tours	Centennial	4127	[3745, 4678]	4395	[3729, 5377]	4575	[3780, 5608]	4031
	Millennial	4479	[3951, 5397]	5049	[3979, 7005]	5432	[3974, 7415]	4285
	Deca-millennial	4670	[4023, 5913]	5529	[4021, 8503]	6116	[4024, 9197]	4400
Reims	Centennial	1696	[1224, 2473]	1594	[1213, 2107]	1529	[1236, 1915]	1573
	Millennial	2886	[1394, 5768]	2416	[1409, 4100]	2131	[1500, 3150]	2271
	Deca-millennial	5638	[1510, 14019]	3780	[1554, 8048]	2864	[1556, 4965]	3171
Orange	Centennial	8001	[6543, 9681]	7943	[6707, 9507]	7841	[6737, 9177]	7659
	Millennial	10653	[7513, 14955]	10506	[7627, 14251]	10211	[7891, 13397]	9546
	Deca-millennial	13726	[8030, 22164]	13418	[8341, 21022]	12782	[8498, 18630]	11429

Wind speed return levels

	Return Period	Jeffreys		PC ($\lambda = 5$)		PC ($\lambda = 10$)		MLE
Tours	Centennial	39.31	[34.46, 46.39]	40.48	[34.68, 47.52]	40.66	[35.22, 46.65]	38.00
	Millennial	44.76	[36.05, 59.25]	47.19	[36.29, 61.75]	47.48	[37.17, 59.28]	41.97
	Deca-millennial	49.86	[36.67, 74.03]	53.81	[37.06, 79.19]	54.09	[37.95, 73.71]	45.10
Reims	Centennial	53.94	[42.44, 71.27]	49.97	[42.63, 59.52]	51.53	[42.21, 63.67]	51.03
	Millennial	79.77	[47.44, 139.11]	70.94	[47.12, 108.24]	65.57	[48.76, 92.87]	65.72
	Deca-millennial	124.98	[48.55, 279.70]	99.35	[49.91, 188.62]	85.50	[51.99, 146.06]	83.33
Orange	Centennial	46.95	[42.75, 52.44]	47.60	[43.28, 52.69]	47.73	[43.38, 52.39]	48.52
	Millennial	52.90	[44.87, 63.90]	54.27	[46.03, 64.58]	54.51	[46.38, 64.07]	57.19
	Deca-millennial	58.35	[46.29, 76.89]	60.55	[47.69, 77.78]	60.89	[48.13, 76.73]	66.57

Table 5.4: Estimation of the centennial, millennial, and deca-millennial return levels for the three river flow and the three wind speed studies. The posterior mean and 95% credible interval (CI) of the return level are computed with Jeffreys prior and PC prior with $\lambda \in \{5, 10\}$. An estimation using maximum likelihood (MLE) is also given for each case.

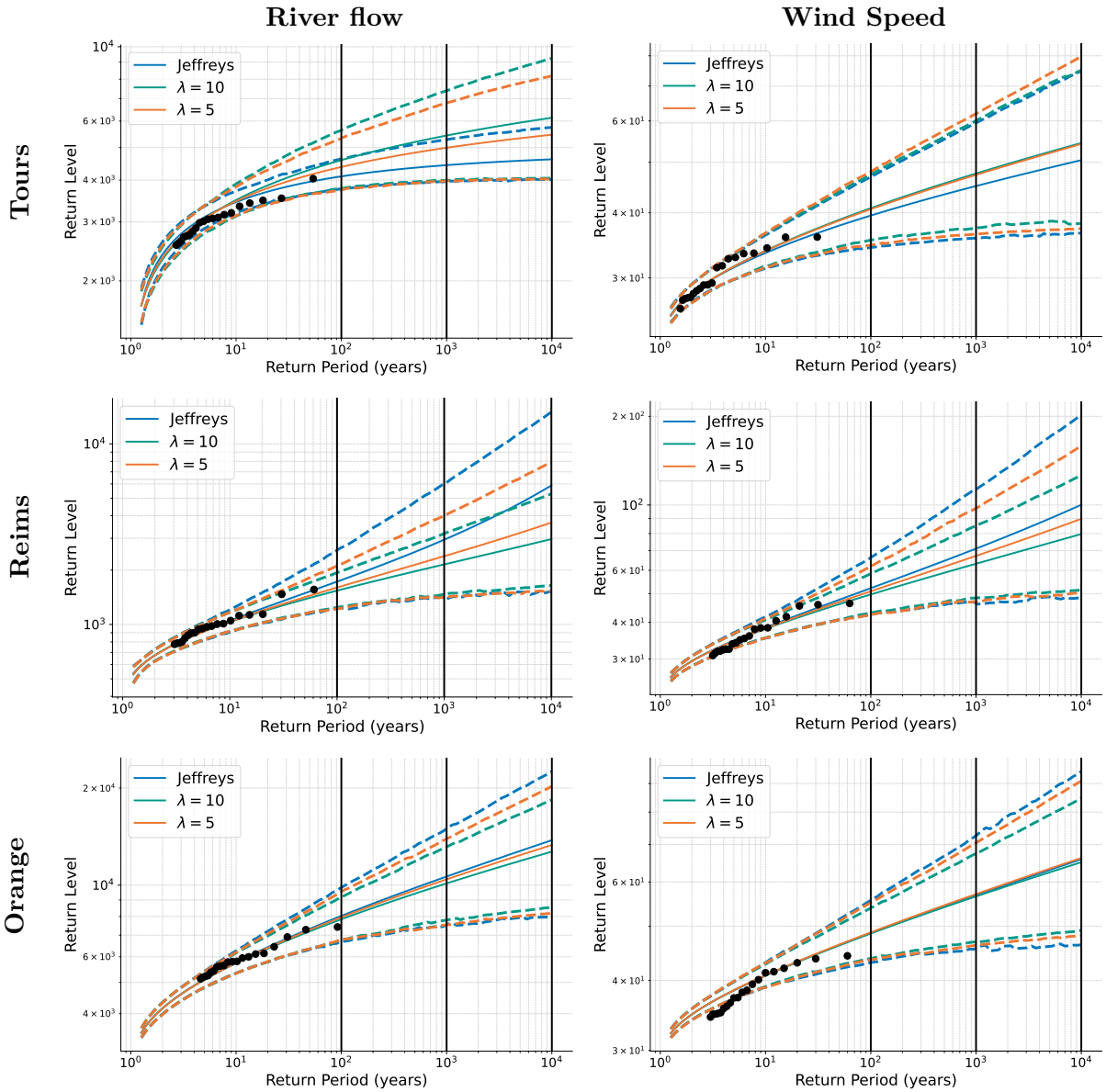


Figure 5.7: Return levels for annual maxima for the river flow and wind speed at Tours, Reims, and Orange. The three curves correspond to different choices of prior: Jeffreys (in blue), PC prior with $\lambda = 10$ (in green) and $\lambda = 5$ (in orange).

a result of convergence of point process, and therefore is only true in the limit. In addition to the error associated with the estimation of the parameters of the process, there is also an error associated with approximating the true process with its limit, with a finite number of data and a fixed threshold. See the thesis of Albert (2018) for a study of this extrapolation error for the same purpose. Thus, any credible interval for return levels obtained here assumes that the model is well specified, and do not consider this approximation error term.

- **Uncertainty related to MCMC approximation.** To have an access to the posterior distribution of parameters or return level, we use an MCMC method that generates samples that converge to the target distribution. Therefore, for a finite number of iterations of the algorithm, all the quantities of interested derived from the posterior are obtained by a Monte Carlo approximation, which adds another form of uncertainty. See Krüger et al. (2021) for results of consistency in the estimation of posterior predictive distribution using MCMC methods. Note that in view of all the MCMC convergence diagnostic we use (autocorrelation, ESS, $\hat{R}$), it seems that convergence to the posterior has been achieved, and it is reasonable to assume that this error is negligible compared with the others cited.

- **Uncertainty in the model itself.** Finally, one can look at the uncertainty modelled by our Bayesian method, in particular at the credible intervals of the return levels. In other terms, if we assume that our data are i.i.d., exactly distributed according to a model of extremes, that the MCMC algorithm indeed returns exact posterior samples, and that the prior effectively models our prior knowledge about the data, etc., what information do we have about a return level associated with a given period? What uncertainty encompass the estimation? By looking at the results in Table 5.4, and in particular the credible intervals for deca-millenial return levels, the size of the intervals are in the order of the estimate. If the interval is centered (which is not the case here, the average value is closer to the right bound than to the left one), this means an error of around 50% of the estimated value at 95%. For example, if we consider the deca-millenial estimation of the river flow at Orange ($13418 m^3/s$), the obtained credible interval at 95% ([8341, 21022]) represents a relative error of 38% for the lower bound, and of 57% for the upper one. In such a case that is supposed to favor the estimation error as it corresponds to the largest dataset, and without even considering all the other sources of uncertainty, these values seem too high to give a return level that can be exploited.

Therefore, considering all these remarks, one can discuss the relevance of estimating deca-millenial return levels. The millenial ones, although reduced, seem to suffer from the same issues too.

5.5 Conclusion and future work

Using the results obtained in the previous chapters, we applied a Bayesian framework for the estimation of a Poisson process model for extremes on datasets from different locations in France: wind speed and river flow in Tours, Reims, and Orange. Based on graphical methods, prior knowledge from the litterature on environmental studies as well as internal EDF reports, we discussed all the steps and choices of our model.

In particular, prior elicitation plays a crucial role as it adds information even with an uninformative prior. Our results show that it can strongly influence the posterior inference and the estimation of return levels. The PC prior is interesting as it can reduce uncertainty and add coherent information on the parameter ξ. Looking at the estimation, Jeffreys prior seems to behave the most like a non-informative prior, in the sense that its uncertainty associated with the estimate is the greatest in most cases, and because its value is the closest to the frequentist MLE. However, its asymptote in $-1/2$ may add bias to the estimation when $\xi < 0$ and may constrain the support of the posterior distribution. Moreover, as the variability in the posterior of ξ is wider with Jeffreys prior, the associated posterior return levels may be superior to those using PC prior, as a higher mass on bigger ξ increases the posterior mean of return values (see Chapter 4). As a first future exploration track, it could be interesting to quantify the prior influence by using recent methods like the estimation of a prior effective sample size for extreme value models (see for example Jones et al., 2022).

Some future reaserch directions have already been mentionned in the previous chapters, and we complete here by those concerning the preprocessing steps before applying the model:

- **Dealing with seasonality**: keeping only the rainy season is a simple method but has various flaws: the reduction of the effective number of observations, the empirical choice of months to keep, and also the non-consideration of longer time trend in the series. Adding a time covariate may improve the model in that sense, and could be tested on datasets in the future.
- **Dealing with data dependence**: the way we decorrelate the data could be improved in various ways. One interesting method would be to generalize the Poisson process model with a Hawkes process (Hawkes, 1971). This allows the model to have an intensity with a self-exiciting part, which means that observing an occurrence may influence the observation of another one. Several works already apply this for modelling the excesses over a threshold (Chavez-Demoulin and McGill, 2012, Dissanayake et al., 2021), and it would be intersting to study how to estimate this process in a Bayesian way, and how to connect with the asymptotic results of Ferro and Segers (2003).
- **Threshold elicitation**: although crucial, the question of choosing the number of excesses is still open, and the graphical methods commonly used may be unsatisfactory. In the Bayesian paradigm, this issue can be associated to the one of quantifying the prior impact, since it governs the number of data for inference.

After all these steps, our method provides an estimate of the return level associated with return periods up to deca-millennial, but one can also bring a partial answer to the original problem: what is the reasonable limit of extrapolation? This question is, of course partly subjective, and requires clarification of what “reasonable“ means. However, we have shown that the uncertainties associated with a deca-millennial estimate, modeled here by credible intervals, are very large without even including all sources of errors. Consequently, an estimation of a level as far in the tail as a deca-millenial one seems unreasonable, at least from experimental insights.

Conclusion & perspectives

Conclusion

This thesis has made several contributions that enhance Bayesian methods for estimating extreme events, with a particular focus on characterizing extremes using the Poisson process. Specifically, the research has concentrated on addressing computational challenges and various steps involved in the so-called Bayesian workflow (Gelman et al., 2020).

In Chapter 2, two key steps of the workflow are addressed. Firstly, prior elicitation is tackled by suggesting a Jeffreys prior for the uninformative case and a PC prior for the informative case regarding the shape parameter ξ. The posterior propriety is demonstrated in both cases, and the impact of incorporating additional information on posterior uncertainty is investigated. The second step concerns the computational issues associated with MCMC algorithms in extreme models, and are resolved by employing an orthogonal reparameterization. The proposed change of variable facilitates the convergence of MCMC algorithms for observations in the three domains of attraction.

A third step is then addressed in Chapter 3, which is the improvement of MCMC convergence diagnostic, and in particular the potential scale reduction factor $\hat{R}$. A local version of the diagnostic is demonstrated, both theoretically and experimentally, to be effective in diagnosing a variety of convergence issues, and a multivariate extension is suggested.

Lastly, a final step concerning the evaluation of the model is done in Chapter 4, where the aim is to investigate different posterior quantities of interest. Preliminary results highlight the degenerate behavior in the tail for all of them for a finite number of observations.

To provide empirical validation and practical application, all these suggestions are applied to various environmental datasets of interest for EDF in Chapter 5. This empirical validation complements the overall study and opens up new avenues for future exploration.

Perspectives

The various research axes developed during the thesis and described in the previous chapters all offer multiple perspectives. Although each chapter mentions possible future work, the different topics are summarized here, for which it would be interesting to carry out extensions.

Priors elicitation for extreme value models (Chapter 2). In general, when it comes to Bayesian statistics, the choice of prior is still an open question and remains closely related to the underlying problem. Only a few uninformative priors, such as Jeffreys, are

constructed based on general rules for any likelihood. However, the specific case of extreme value models does not seem to be resolved.

Firstly, other priors could be calculated. In the case of uninformative priors for extreme models, it would be interesting to explicitly specify reference priors (Bernardo, 1979) which, to our knowledge, is not explored in the literature, and also extend the calculations of Ho (2010) for matching priors (Datta and Sweeting, 2005).

On the predictive scale, a more detailed study on the implications of a prior distribution on quantiles, and reciprocally, studying the distribution of quantiles after a prior assumption on (μ, σ, ξ), could extend the work of Coles and Tawn (1996) and Gaioni et al. (2010). These works could also be complemented by a study of a multivariate distribution on three quantiles (q_1, q_2, q_3) instead of the three univariate priors, which would explain the underlying dependence structure.

Finally, to discriminate between all these choices of priors, both in the informative and uninformative cases, it would be interesting to use recent advances on quantifying prior impact (Nott et al., 2020, Jones et al., 2022) for extreme value cases.

Improving MCMC convergence diagnostic (Chapter 3). Concerning convergence diagnostics, multiple tracks could be explored for expanding our proposition of a local $\hat{R}$.

First, a significant contribution would be to obtain a stronger convergence result on the empirical process $\hat{R}(\cdot)$ that is studied in Section 3.2. In our work, only a result of weak convergence is proved, *i.e.* the convergence in distribution of $\hat{R}(x)$ for all x. This approach prevents us from obtaining any results on $\hat{R}_\infty$, which is the supremum over all quantiles x. Therefore, a stronger convergence result on the empirical process could allow an explicit formulation of the threshold as a function of a confidence level α for $\hat{R}_\infty$, in the same way it is already obtained for $\hat{R}(x)$.

This work could also be extended by linking the proposition made by Margossian et al. (2022) in parallel to ours. In particular, the authors suppose to have groups of chains called *superchains* where all the chains inside are initizialized in the same way. This allows the authors to refine their estimation of W and B, and this idea could be exploited for our local version, but also in order to link with other convergence diagnostics like ESS (Vats and Knudson, 2021).

Finally, one can see our approach to dealing with the $R(x)$ estimation 'too frequentist', in the sense that we provide a scalar estimator $\hat{R}(x)$ and study the behavior asymptotically with the number of observations. Instead, a Bayesian point of view for dealing with the diagnostic is an idea that could be explored.

Bridging the gap between Bayesian and frequentist extremes (Chapter 4). There exist various approaches aimed at reconciling the Bayesian and frequentist perspectives, and it would be insightful to compare these methods asymptotically, following the path set by Smith (1999).

Firstly, as mentioned in the previous paragraph on prior elicitation, deriving matching priors is a way to obtain frequentist properties through prior modeling. Also, the Bernstein–von Mises theorem (Van der Vaart, 2000) shows that under certain conditions, frequentist and Bayesian approaches asymptotically lead to the same result. However, these results do not apply to extreme value models as extreme value distributions like

GEV or GPD are not regular, since the support depends on the parameter. Moreover, the models are misspecified in the sense that the block maxima are only asymptotically GEV, and the excesses are asymptotically GPD. The recent work by Padoan and Rizzelli (2022) brings very promising results in the GEV case, as it manages to achieve asymptotic results for misspecified models using an empirical Bayesian model. This work could be extended and connected with the results in Chapter 4.

In particular, Chapter 4 delves into the behavior of Bayesian quantities in the tail, such as posterior predictives or return levels. Examining the tail behavior with probabilities dependent on the sample size n could potentially yield asymptotic properties of Bayesian estimators for extreme return levels, which could then be compared with frequentist estimators, such as maximum likelihood or probability weighted moments.

A more adapted extreme model for environmental application (Chapter 5). Extreme value models are extensively employed in the analysis of environmental data, and so there is room for improvement in considering the inherent characteristics of these data to enhance the accuracy of inference. Several existing studies have addressed important aspects such as data dependence, seasonality, and non-stationarity, as referenced in Chapter 5.

While this thesis primarily focused on improving models based on i.i.d. observations, further exploration of more general cases, especially from a Bayesian perspective, could be pursued. Currently, the preprocessing steps preceding model application assume marginal i.i.d. observations within the Bayesian framework. Chapter 5 highlights the potential value of incorporating assumptions regarding the temporal nature of events, such as generalizing to a Hawkes process (Hawkes, 1971), incorporating covariates, or performing Bayesian estimation of the extremal index (Ferro and Segers, 2003).

Also, it is important to address scenarios in which historical data are only available above a threshold that decreases over time. Such cases occur with seismic data, and a relevant improvement would involve extending the Poisson model to incorporate a piecewise constant threshold. This advancement would have direct practical relevance and enhance the applicability of the model in these contexts.

Other extensions. In addition to the extensions proposed in the previous chapters, there are several unexplored ideas that could further enhance this work.

Firstly, in this thesis, when it comes to eliciting a threshold as in Chapter 2 or Chapter 5, only a simple graphical method recommended by Coles (2001) has been employed, even when it does not yield clear conclusions. For these cases, previous investigations conducted at EDF, along with an assessment of estimation stability with respect to the threshold (e.g. in Figure 5.6), have been carried out. However, in the general case, this approach is not optimal, and developing a comprehensive method for prior selection in the univariate context remains an open challenge. Alternative methods including those suggested in Scarrott and MacDonald (2012), Pan et al. (2022), could be explored, and in particular on a Bayesian framework.

Secondly, in the environmental context, considering the datasets in a multivariate manner could be beneficial. This can be achieved either by incorporating spatial aspects, i.e., simultaneously analyzing data from multiple stations for a specific variable of interest (e.g., wind speed or river flow), or by considering multiple related variables at the same location.

The field of multivariate extreme value theory has witnessed significant advancements over the years, see Chapter 8 in Beirlant et al. (2006) for an introduction and Davison et al. (2012), Dutfoy et al. (2014) for some reviews.

Lastly, considering semi-parametric models for extreme value inference, particularly within a Bayesian framework, could offer alternative estimators for the tail index with appealing asymptotic properties. A first conjugate case with Pareto data has been introduced in Chapter 11 of Beirlant et al. (2006), and further extensions have been proposed recently (Beirlant et al., 2018, Li et al., 2019). These methods provide other opportunities to enhance the estimation of extreme values in the Bayesian paradigm.

Bibliography

Aitchison, J. (1975). Goodness of prediction fit. Biometrika 62(3), 547–554. 106

Albert, C. (2018). Estimation des limites d'extrapolation par les lois de valeurs extrêmes. Application à des données environnementales. PhD thesis (in French), Université Grenoble Alpes. 28, 29, 45, 138

Albert, C., A. Dutfoy, L. Gardes, and S. Girard (2020). An extreme quantile estimator for the log-generalized Weibull-tail model. Econometrics and Statistics 13, 137–174. 46

Ameraoui, A., K. Boukhetala, and J.-F. Dupuy (2016). Bayesian estimation of the tail index of a heavy tailed distribution under random censoring. Computational Statistics & Data Analysis 104, 148–168. 24

Andreewsky, M. and N. Bousquet (2021). Collecting and Analyzing Data. In Extreme Value Theory with Applications to Natural Hazards: From Statistical Theory to Industrial Practice, pp. 59–81. Springer, Cham. 123

Balkema, A. A. and L. De Haan (1974). Residual life time at great age. The Annals of Probability 2(5), 792–804. 4

Baudin, M., A. Dutfoy, B. Iooss, and A.-L. Popelin (2017). OpenTURNS: An industrial software for uncertainty quantification in simulation. In D. H. R. Ghanem and H. Owhadi (Eds.), Handbook of uncertainty quantification. Springer. 80

Behrens, C. N., H. F. Lopes, and D. Gamerman (2004). Bayesian analysis of extreme events with threshold estimation. Statistical modelling 4(3), 227–244. 28

Beirlant, J., Y. Goegebeur, J. Segers, and J. L. Teugels (2006). Statistics of Extremes: Theory and Applications. John Wiley & Sons. 4, 24, 144

Beirlant, J., G. Maribe, and A. Verster (2018). Penalized bias reduction in extreme value estimation for censored Pareto-type data, and long-tailed insurance applications. Insurance: Mathematics and Economics 78, 114–122. 24, 144

Belzile, L. R., C. Dutang, P. J. Northrop, and T. Opitz (2022). A modeler's guide to extreme value software. arXiv:2205.07714. 25, 36, 37

Beranger, B., S. A. Padoan, and S. A. Sisson (2021). Estimation and uncertainty quantification for extreme quantile regions. Extremes 24, 349–375. 25

Berger, J. O. (1988). Statistical decision theory and Bayesian analysis. Springer Science & Business Media. 14, 17

Berger, J. O., J. M. Bernardo, and D. Sun (2009). The formal definition of reference priors. The Annals of Statistics 37(2), 905–938. 18

Berger, J. O. and R. L. Wolpert (1988). The likelihood principle. Institute of Mathematical Statistics. 18

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. Journal of the Royal Statistical Society: Series B 41(2), 113–128. 18, 142

Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. arXiv:1701.02434. 21

Betancourt, M. (2019). Incomplete Reparameterizations and Equivalent Metrics. arXiv:1910.09407. 19, 36, 53

Betancourt, M. and M. Girolami (2015). Hamiltonian Monte Carlo for hierarchical models. In Current trends in Bayesian methodology with applications, pp. 79–101. CRC Press. 36

Bingham, E., J. P. Chen, M. Jankowiak, F. Obermeyer, N. Pradhan, T. Karaletsos, R. Singh, P. Szerlip, P. Horsfall, and N. D. Goodman (2019). Pyro: Deep universal probabilistic programming. The Journal of Machine Learning Research 20(1), 973–978. 64

Bingham, N. H., C. M. Goldie, and J. L. Teugels (1989). Regular variation. Cambridge: Cambridge University Press. 117

Blei, D. M., A. Kucukelbir, and J. D. McAuliffe (2017). Variational inference: A review for statisticians. Journal of the American statistical Association 112(518), 859–877. 22

Bottolo, L., G. Consonni, P. Dellaportas, and A. Lijoi (2003). Bayesian analysis of extreme values by mixture modeling. Extremes 6, 25–47. 25, 26

Bousquet, N. (2021). Bayesian Extreme Value Theory. In Extreme Value Theory with Applications to Natural Hazards: From Statistical Theory to Industrial Practice, pp. 271–325. Springer, Cham. 25, 36

Bousquet, N. and P. Bernardara (2021). Extreme Value Theory with Applications to Natural Hazards. Springer. 13

Bousquet, N. and M. Keller (2017). Bayesian prior elicitation and selection for extreme values. arXiv:1712.00685. 25, 26

Breiman, L. (1965). On some limit theorems similar to the arc-sin law. Theory of Probability & Its Applications 10(2), 323–331. 107, 112

Brooks, S., A. Gelman, G. Jones, and X.-L. Meng (2011). Handbook of Markov chain Monte Carlo. CRC press. 19

Brooks, S. P. and A. Gelman (1998). General methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics 7(4), 434–455. 66, 78

Brooks, S. P. and G. O. Roberts (1999). On quantile estimation and Markov chain Monte Carlo convergence. Biometrika 86(3), 710–717. 69

Browne, W. J., F. Steele, M. Golalizadeh, and M. J. Green (2009). The use of simple reparameterizations to improve the efficiency of Markov chain Monte Carlo estimation for multilevel models with applications to discrete time survival models. Journal of the Royal Statistical Society: Series A 172(3), 579–598. 37

Butler, A., J. E. Heffernan, J. A. Tawn, R. A. Flather, and K. J. Horsburgh (2007). Extreme value analysis of decadal variations in storm surge elevations. Journal of Marine Systems 67(1-2), 189–200. 13

Cabras, S. (2013). Default priors based on pseudo-likelihoods for the Poisson-GPD model. In Advances in Theoretical and Applied Statistics, pp. 3–12. Springer. 26

Carpenter, B., A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell (2017). Stan: A probabilistic programming language. Journal of Statistical Software 76(1), 1–32. 19, 64, 77

Casson, E. and S. Coles (2000). Simulation and extremal analysis of hurricane events. Journal of the Royal Statistical Society: Series C 49(3), 227–245. 13

Castellanos, M. E. and S. Cabras (2007). A default Bayesian procedure for the generalized Pareto distribution. Journal of Statistical Planning and Inference 137(2), 473–483. 26, 27, 40, 41

Castillo, E. (1988). Extreme Value Theory in Engineering. San Diego: Academic Press. 13

Castro-Camilo, D., R. Huser, and H. Rue (2021). Practical strategies for GEV-based regression models for extremes. arXiv:2106.13110. 13, 26, 27

Chavez-Demoulin, V. and A. C. Davison (2005). Generalized additive modelling of sample extremes. Journal of the Royal Statistical Society: Series C 54(1), 207–222. 38, 39, 56, 124

Chavez-Demoulin, V. and J. McGill (2012). High-frequency financial data modeling using Hawkes processes. Journal of Banking & Finance 36(12), 3415–3426. 139

Chopin, N. and O. Papaspiliopoulos (2020). An introduction to sequential Monte Carlo, Volume 4. Springer. 22

Clarke, B. S. and A. R. Barron (1994). Jeffreys' prior is asymptotically least favorable under entropy risk. Journal of Statistical Planning and Inference 41(1), 37–60. 18

Coles, S. (2001). An Introduction to Statistical Modeling of Extreme Values. Springer Series in Statistics. London: Springer-Verlag. 4, 9, 12, 24, 34, 35, 45, 124, 126, 143

Coles, S. (2003). The use and misuse of extreme value models in practice. In Extreme values in finance, telecommunications, and the environment, pp. 98–119. Chapman and Hall/CRC. 24

Coles, S. G. and E. A. Powell (1996). Bayesian methods in extreme value modelling: a review and new developments. International Statistical Review 64(1), 119–136. 24, 36

Coles, S. G. and J. A. Tawn (1996). A bayesian analysis of extreme rainfall data. Journal of the Royal Statistical Society: Series C 45(4), 463–478. 13, 17, 24, 25, 26, 105, 142

Cox, D. R. (1975). Prediction Intervals and Empirical Bayes Confidence Intervals. Journal of Applied Probability 12(S1), 47–55. 106

Cox, D. R. and N. Reid (1987). Parameter orthogonality and approximate conditional inference. Journal of the Royal Statistical Society: Series B 49(1), 1–18. 37

Datta, G. S. and T. J. Sweeting (2005). Probability matching priors. Handbook of Statistics 25, 91–114. 18, 142

Davison, A. (1986). Approximate predictive likelihood. Biometrika 73(2), 323–332. 27, 106

Davison, A. C., S. A. Padoan, and M. Ribatet (2012). Statistical Modeling of Spatial Extremes. Statistical Science 27(2), 161–186. 144

Davison, A. C. and R. L. Smith (1990). Models for exceedances over high thresholds. Journal of the Royal Statistical Society: Series B 52(3), 393–425. 8, 123, 124

De Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. In Annales de l'institut Henri Poincaré, Volume 7, pp. 1–68. 16

de Valpine, P., D. Turek, C. J. Paciorek, C. Anderson-Bergman, D. T. Lang, and R. Bodik (2017). Programming with models: writing statistical algorithms for general model structures with nimble. Journal of Computational and Graphical Statistics 26(2), 403–413. 19, 64

de Zea Bermudez, P., A. Turkman, and K. Turkman (2001). A predictive approach to tail probability estimation. Extremes 4(4), 295–314. 27, 106

de Zea Bermudez, P. and M. A. Turkman (2003). Bayesian approach to parameter estimation of the generalized Pareto distribution. Test 12, 259–277. 26, 27

Dey, D. K. and J. Yan (2016). Extreme value modeling and risk analysis: methods and applications. CRC Press. 24

Diebolt, J., M. A. El-Aroui, M. Garrido, and S. Girard (2005). Quasi-conjugate Bayes estimates for GPD parameters and application to heavy tails modelling. Extremes 8, 57–78. 25, 26, 37

Dissanayake, P., T. Flock, J. Meier, and P. Sibbertsen (2021). Modelling short-and long-term dependencies of clustered high-threshold exceedances in significant wave heights. Mathematics 9(21), 2817–2850. 139

do Nascimento, F. F., D. Gamerman, and H. F. Lopes (2012). A semiparametric Bayesian approach to extreme value estimation. Statistics and Computing 22, 661–675. 28

Dombry, C., S. Engelke, and M. Oesting (2017). Bayesian inference for multivariate extreme value distributions. Electronic Journal of Statistics 11(2), 4813–4844. 25

Doss, C. R., J. M. Flegal, G. L. Jones, and R. C. Neath (2014). Markov chain Monte Carlo estimation of quantiles. Electronic Journal of Statistics 8(2), 2448–2478. 69

Dutfoy, A., S. Parey, and N. Roche (2014). Multivariate extreme value theory - a tutorial with applications to hydrology and meteorology. Dependence Modeling 2(1), 30–48. 144

Embrechts, P. and C. M. Goldie (1980). On closure and factorization properties of subexponential and related distributions. Journal of the Australian Mathematical Society 29(2), 243–256. 107, 114

Embrechts, P., C. Klüppelberg, and T. Mikosch (2013). Modelling extremal events, Volume 33. Springer Science & Business Media. 5, 12, 72

Engeland, K., H. Hisdal, and A. Frigessi (2004). Practical extreme value modelling of hydrological floods and droughts: a case study. Extremes 7, 5–30. 13

Engelund, S. and R. Rackwitz (1992). On predictive distribution functions for the three asymptotic extreme value distributions. Structural Safety 11(3-4), 255–258. 27

Erhardt, R. and S. A. Sisson (2016). Modelling extremes using approximate Bayesian computation. In Extreme Value Modelling and Risk Analysis, pp. 281–306. Chapman and Hall/CRC Press Boca Raton, FL. 24

Fawcett, L. and A. C. Green (2018). Bayesian posterior predictive return levels for environmental extremes. Stochastic Environmental Research and Risk Assessment 32(8), 2233–2252. 27, 49, 106

Fawcett, L. and D. Walshaw (2012). Estimating return levels from serially dependent extremes. Environmetrics 23(3), 272–283. 126

Fawcett, L. and D. Walshaw (2016). Sea-surge and wind speed extremes: optimal estimation strategies for planners and engineers. Stochastic Environmental Research and Risk Assessment 30(2), 463–480. 27, 123, 124

Ferro, C. A. and J. Segers (2003). Inference for clusters of extreme values. Journal of the Royal Statistical Society: Series B 65(2), 545–556. 126, 139, 143

Finkenstadt, B. and H. Rootzén (2003). Extreme values in finance, telecommunications, and the environment. CRC Press. 12

Fisher, R. A. and L. H. C. Tippett (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. Mathematical proceedings of the Cambridge philosophical society 24(2), 180–190. 4

Fosdick, L. D. (1959). Calculation of order parameters in a binary alloy by the Monte Carlo method. Physical Review 116(3), 565. 70

Fréchet, M. (1927). Sur la loi de probabilité de l'écart maximum. Annales de la Société Polonaise de Mathématique 6, 93–116. 5

Frigessi, A., O. Haug, and H. Rue (2002). A dynamic mixture model for unsupervised tail estimation without threshold selection. Extremes 5, 219–235. 27

Gaioni, E., D. Dey, and F. Ruggeri (2010). Bayesian modeling of flash floods using generalized extreme value distribution with prior elicitation. Chilean Journal of Statistics 1(1), 75–90. 25, 26, 105, 142

Gardes, L. and S. Girard (2010). Conditional extremes from heavy-tailed distributions: An application to the estimation of extreme rainfall return levels. Extremes 13(2), 177–204. 129

Gelfand, A. E., S. K. Sahu, and B. P. Carlin (1995). Efficient parametrisations for normal linear mixed models. Biometrika 82(3), 479–488. 37

Gelfand, A. E., S. K. Sahu, and B. P. Carlin (1996). Efficient parametrizations for generalized linear mixed models. Bayesian Statistics 5, 48–74. 37

Gelfand, A. E. and A. F. Smith (1990). Sampling-based approaches to calculating marginal densities. Journal of the American Statistical Association 85(410), 398–409. 70

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013). Bayesian Data Analysis (3rd ed.). New York: CRC Press. 14, 16, 23, 24, 43, 44, 45, 47, 54, 55, 56, 57, 65, 69, 82

Gelman, A. and D. B. Rubin (1992). Inference from iterative simulation using multiple sequences. Statistical Science 7(4), 457–472. 21, 24, 62, 63, 65, 67, 70

Gelman, A., D. Simpson, and M. Betancourt (2017). The prior can often only be understood in the context of the likelihood. Entropy 19(10), 555–568. 18

Gelman, A., A. Vehtari, D. Simpson, C. C. Margossian, B. Carpenter, Y. Yao, L. Kennedy, J. Gabry, P.-C. Bürkner, and M. Modrák (2020). Bayesian workflow. arXiv:2011.01808. 18, 34, 36, 141

Geman, S. and D. Geman (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence 6(6), 721–741. 19

Geyer, C. J. (1992). Practical Markov chain Monte Carlo. Statistical Science 7(4), 473–483. 21

Gilks, W. R., S. Richardson, and D. Spiegelhalter (1995). Markov Chain Monte Carlo in Practice. Boca Raton: CRC Press. 19, 36, 37, 58

Gilleland, E. and R. W. Katz (2016). extRemes 2.0: an extreme value analysis package in R. Journal of Statistical Software 72, 1–39. 60, 131

Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. Annals of Mathematics 44(3), 423–453. 4

Gong, L. and J. M. Flegal (2016). A practical sequential stopping rule for high-dimensional Markov chain Monte Carlo. Journal of Computational and Graphical Statistics 25(3), 684–700. 69

Gribok, A. V., A. M. Urmanov, J. Wesley Hines, and R. E. Uhrig (2004). Backward specification of prior in Bayesian inference as an inverse problem. Inverse Problems in Science and Engineering 12(3), 263–278. 17, 105

Gumbel, E. J. (1958). Statistics of Extremes. Columbia University Press. 5, 6

Haan, L. and A. Ferreira (2006). Extreme value theory: an introduction. New York: Springer. 4, 34, 104, 117

Hartmann, M., G. Agiashvili, P. Bürkner, and A. Klami (2020). Flexible prior elicitation via the prior predictive distribution. Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI) 124(36), 1129–1138. 17

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57(1), 97–109. 14, 20

Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. Biometrika 58(1), 83–90. 139, 143

Ho, K.-W. (2010). A matching prior for extreme quantile estimation of the generalized Pareto distribution. Journal of Statistical Planning and Inference 140(6), 1513–1518. 26, 27, 142

Hoffman, M. D. and A. Gelman (2014). No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. Journal of Machine Learning Research 15(1), 1593–1623. 21, 36, 43, 53, 66

Holmes, J. and W. Moriarty (1999). Application of the generalized Pareto distribution to extreme value analysis in wind engineering. Journal of Wind Engineering and Industrial Aerodynamics 83(1), 1–10. 122

Hosking, J. R. and J. R. Wallis (1987). Parameter and quantile estimation for the generalized Pareto distribution. Technometrics 29(3), 339–349. 9

Hosking, J. R. M., J. R. Wallis, and E. F. Wood (1985). Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. Technometrics 27(3), 251–261. 6

Hundecha, Y., M. Pahlow, and A. Schumann (2009). Modeling of daily precipitation at multiple locations using a mixture of distributions to characterize the extremes. Water Resources Research 45(12), 1–15. 28

Huzurbazar, V. S. (1950). Probability distributions and orthogonal parameters. Mathematical Proceedings of the Cambridge Philosophical Society 46(2), 281–284. 37

Jeffreys, H. (1939). The Theory of Probability (1st ed.). Oxford: Oxford Univ. Press. 14, 17, 32, 33, 37, 48

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. Proceedings of the Royal Society of London: Series A 186(1007), 453–461. 40

Jóhannesson, Á. V., S. Siegert, R. Huser, H. Bakka, and B. Hrafnkelsson (2022). Approximate Bayesian inference for analysis of spatiotemporal flood frequency data. The Annals of Applied Statistics 16(2), 905–935. 25, 37

Jonathan, P., D. Randell, J. Wadsworth, and J. Tawn (2021). Uncertainties in return values from extreme value analysis of peaks over threshold using the generalised Pareto distribution. Ocean Engineering 220, 107725. 27, 46, 106, 108, 109, 131

Jones, D. E., R. N. Trangucci, and Y. Chen (2022). Quantifying observed prior impact. Bayesian Analysis 17(3), 737–764. 16, 139, 142

Kass, R. E. and L. Wasserman (1996). The selection of prior distributions by formal rules. Journal of the American statistical Association 91(435), 1343–1370. 18

Kotz, S. and S. Nadarajah (2000). Extreme Value Distributions: Theory and Applications. London: Imperial College Press. 40

Krüger, F., S. Lerch, T. Thorarinsdottir, and T. Gneiting (2021). Predictive inference based on Markov chain Monte Carlo output. International Statistical Review 89(2), 274–301. 138

Lambert, B. and A. Vehtari (2022). R^*: A robust MCMC convergence diagnostic with uncertainty using decision tree classifiers. Bayesian Analysis 17(2), 353–379. 66, 79

Langousis, A., A. Mamalakis, M. Puliga, and R. Deidda (2016). Threshold detection for the generalized Pareto distribution: Review of representative methods and application to the NOAA NCDC daily rainfall database. Water Resources Research 52(4), 2659–2681. 126, 129

Larsén, X. G., J. Mann, O. Rathmann, and H. E. Jørgensen (2015). Uncertainties of the 50-year wind from short time series using generalized extreme value distribution and generalized Pareto distribution. Wind Energy 18(1), 59–74. 122

Lawless, J. F. and M. Fredette (2005). Frequentist prediction intervals and predictive distributions. Biometrika 92(3), 529–542. 105

Leadbetter, M., G. Lindgren, and H. Rootzén (1983). Extremes and Related Properties of Random Sequences and Processes. New York: Springer. 10, 35

Lee, J., Y. Fan, and S. A. Sisson (2015). Bayesian threshold selection for extremal models using measures of surprise. Computational Statistics & Data Analysis 85, 84–99. 28

Lemoine, N. P. (2019). Moving beyond noninformative priors: why and how to choose weakly informative priors in Bayesian analyses. Oikos 128(7), 912–928. 18

Li, C., L. Lin, and D. B. Dunson (2019). On posterior consistency of tail index for Bayesian kernel mixture models. Bernoulli 25(3), 1999–2028. 24, 144

Lopes, R. H., I. Reid, and P. R. Hobson (2007). The two-dimensional Kolmogorov-Smirnov test. In XI International Workshop on Advanced Computing and Analysis Techniques in Physics Research, Amsterdam, the Netherlands. 77

MacDonald, A., C. J. Scarrott, D. Lee, B. Darlow, M. Reale, and G. Russell (2011). A flexible extreme value mixture model. Computational Statistics & Data Analysis 55(6), 2137–2157. 28

Margossian, C. C., M. D. Hoffman, P. Sountsov, L. Riou-Durand, A. Vehtari, and A. Gelman (2022). Nested $\hat{R}$: Assessing the convergence of Markov chain Monte Carlo when running many short chains. arXiv:2110.13017. 71, 142

Martin, G. M., D. T. Frazier, and C. P. Robert (2020). Computing Bayes: Bayesian computation from 1763 to the 21st century. arXiv:2004.06425. 14

Martín, J., M. I. Parra, M. M. Pizarro, and E. L. Sanjuán (2022). Baseline methods for the parameter estimation of the generalized Pareto distribution. Entropy 24(2), 178. 28

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equation of state calculations by fast computing machines. The Journal of Chemical Physics 21(6), 1087–1092. 14, 20

Mikkola, P., O. A. Martin, S. Chandramouli, M. Hartmann, O. A. Pla, O. Thomas, H. Pesonen, J. Corander, A. Vehtari, S. Kaski, et al. (2023). Prior Knowledge Elicitation: The Past, Present, and Future. Bayesian Analysis, 1–33. 16, 105

Moins, T., J. Arbel, A. Dutfoy, and S. Girard (2021a). Contributed discussion: "Rank-Normalization, Folding, and Localization: An Improved $\widehat{R}$ for Assessing Convergence of MCMC". Bayesian Analysis 16(2), 711–712. 29, 66

Moins, T., J. Arbel, A. Dutfoy, and S. Girard (2021b). On reparameterisations of the Poisson process model for extremes in a Bayesian framework. In JDS 2021-52èmes Journées de Statistique de la Société Française de Statistique (SFdS), pp. 1–6. 29

Moins, T., J. Arbel, A. Dutfoy, and S. Girard (2022a). localrhat: a local $\widehat{R}$ to improve MCMC convergence diagnostic (R package). https://github.com/TheoMoins/localrhat. 67, 71, 80

Moins, T., J. Arbel, A. Dutfoy, and S. Girard (2022b). On the use of a local $\hat{R}$ to improve MCMC convergence diagnostic. In JDS 2022-53èmes Journées de Statistique de la Société Française de Statistique (SFdS), pp. 1–6. 29

Moins, T., J. Arbel, A. Dutfoy, and S. Girard (2023). On the use of a local $\hat{R}$ to improve MCMC convergence diagnostic. Bayesian Analysis to appear, 1–25. 29, 43, 44, 47, 54, 55, 56, 57, 62, 63, 110

Moins, T., J. Arbel, S. Girard, and A. Dutfoy (2023). Reparameterization of extreme value framework for improved Bayesian workflow. Computational Statistics & Data Analysis 187, 107807. 26, 29, 32, 33, 81, 112

Neal, R. M. (1996). MCMC using Hamiltonian dynamics. In Handbook of Markov Chain Monte Carlo, pp. 113–162. Chapman and Hall/CRC. 21, 36, 43, 53, 66, 80

Nelsen, R. B. (2006). An introduction to copulas. New York: Springer. 75, 76, 89

Northrop, P. J. (2022a). revdbayes: Ratio-of-Uniforms Sampling for Bayesian Extreme Value Analysis. https://paulnorthrop.github.io/revdbayes/, https://github.com/paulnorthrop/revdbayes. 37, 45, 58

Northrop, P. J. (2022b). rust: Ratio-of-Uniforms Simulation with Transformation. https://paulnorthrop.github.io/rust/, https://github.com/paulnorthrop/rust. 58

Northrop, P. J. and N. Attalides (2016). Posterior propriety in Bayesian extreme value analyses using reference priors. Statistica Sinica 26(2), 721–743. 25, 26, 27, 41, 42, 53

Northrop, P. J., N. Attalides, and P. Jonathan (2017). Cross-validatory extreme value threshold selection and uncertainty with application to ocean storm severity. Journal of the Royal Statistical Society: Series C 66(1), 93–120. 28

Nott, D. J., X. Wang, M. Evans, and B.-G. Englert (2020). Checking for prior-data conflict using prior-to-posterior divergences. Statistical Science 35(2), 234–253. 142

Opitz, T., R. Huser, H. Bakka, and H. Rue (2018). INLA goes extreme: Bayesian tail regression for the estimation of high spatio-temporal quantiles. Extremes 21(3), 441–462. 26, 27, 37, 41

O'Hagan, A. (2019). Expert knowledge elicitation: subjective but scientific. The American Statistician 73(sup1), 69–81. 16

Padoan, S. A. and S. Rizzelli (2022). Empirical Bayes inference for the block maxima method. arXiv:2204.04981. 26, 27, 112, 143

Pan, X., A. Rahman, K. Haddad, and T. B. Ouarda (2022). Peaks-over-threshold model in flood frequency analysis: A scoping review. Stochastic Environmental Research and Risk Assessment 36(9), 2419–2435. 13, 122, 126, 143

Papaspiliopoulos, O., G. O. Roberts, and M. Sköld (2003). Non-centered parameterisations for hierarchical models and data augmentation. Bayesian Statistics 7, 307–326. 37, 82

Parent, E. and J. Bernier (2003). Bayesian POT modeling for historical data. Journal of Hydrology 274(1-4), 95–108. 25, 26

Parey, S., T.-T.-H. Hoang, and N. Bousquet (2021). Stochastic and Physics-Based Simulation of Extreme Situations. In Extreme Value Theory with Applications to Natural Hazards: From Statistical Theory to Industrial Practice, pp. 229–270. Springer, Cham. 129

Perepolkin, D., B. Goodrich, and U. Sahlin (2021). Hybrid elicitation and indirect Bayesian inference with quantile-parametrized likelihood. osf.io/paby6. 17

Pickands, J. (1975). Statistical inference using extreme order statistics. Annals of Statistics 3(1), 119–131. 4, 7, 34, 104

Prescott, P. and A. Walden (1983). Maximum likeiihood estimation of the parameters of the three-parameter generalized extreme-value distribution from censored samples. Journal of Statistical Computation and Simulation 16(3-4), 241–250. 6

Raftery, A. E. and S. Lewis (1992). How many iterations in the Gibbs sampler? Bayesian Statistics 4, 763–773. 67, 68, 69

Resnick, S. I. (2008). Extreme Values, Regular Variation, and Point Processes. Springer Science & Business Media. 4

Richards, J. and J. A. Tawn (2022). On the tail behaviour of aggregated random variables. Journal of Multivariate Analysis 192, 105065. 112

Robert, C. P. (1995). Convergence control methods for Markov chain Monte Carlo algorithms. Statistical Science 10(3), 231–253. 64

Robert, C. P. (2007). The Bayesian Choice: from Decision-Theoretic Foundations to Computational Implementation. New York: Springer. 14, 16, 17, 18, 40, 106

Robert, C. P. and G. Casella (2004). Monte Carlo statistical methods. Springer Verlag. 19, 64, 69

Roberts, G. O. and N. G. Polson (1994). On the geometric convergence of the Gibbs sampler. Journal of the Royal Statistical Society: Series B 56(2), 377–384. 20, 36

Roberts, G. O. and J. S. Rosenthal (2001). Optimal scaling for various Metropolis-Hastings algorithms. Statistical Science 16(4), 351–367. 21

Roberts, G. O. and S. K. Sahu (1997). Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. Journal of the Royal Statistical Society: Series B 59, 291–317. 36

Roberts, G. O. and R. L. Tweedie (1996). Exponential convergence of Langevin distributions and their discrete approximations. Bernoulli 2(4), 341–363. 21

Ross, S. M. (1996). Stochastic Processes. New York: John Wiley & Sons. 10

Rousseau, J. and B. Szabo (2017). Asymptotic behaviour of the empirical Bayes posteriors associated to maximum marginal likelihood estimator. The Annals of Statistics 45(2), 833–865. 17

Roy, V. (2020). Convergence diagnostics for Markov chain Monte Carlo. Annual Review of Statistics and Its Application 7, 387–412. 64

Rue, H., S. Martino, and N. Chopin (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society: Series B 71(2), 319–392. 22

Sabourin, A., P. Naveau, and A.-L. Fougères (2013). Bayesian model averaging for multivariate extremes. Extremes 16(3), 325–350. 25

Salvatier, J., T. V. Wiecki, and C. Fonnesbeck (2016). Probabilistic programming in Python using PyMC3. PeerJ Computer Science 2, e55. 19, 42, 64, 110, 129

Savage, L. J. (1954). The foundations of statistics. Courier Corporation. 14

Scarrott, C. and A. MacDonald (2012). A review of extreme value threshold estimation and uncertainty quantification. REVSTAT-Statistical Journal 10(1), 33–60. 126, 143

Sharkey, P. and J. A. Tawn (2017). A Poisson process reparameterisation for Bayesian inference for extremes. Extremes 20(2), 239–263. 25, 28, 32, 33, 38, 39, 44, 45, 49, 50, 53, 54, 55, 56, 60

Sharma, P., M. Khare, and S. Chakrabarti (1999). Application of extreme value theory for predicting violations of air quality standards for an urban road intersection. Transportation Research Part D: Transport and Environment 4(3), 201–216. 13

Shen, J., R. Y. Liu, and M.-g. Xie (2018). Prediction with confidence—a general framework for predictive inference. Journal of Statistical Planning and Inference 195, 126–140. 106

Sibler, A. and A. Dutfoy (2021). Conjunction of a Flood and a Storm. In Extreme Value Theory with Applications to Natural Hazards: From Statistical Theory to Industrial Practice, pp. 393–408. Springer, Cham. 122

Simpson, D., H. Rue, A. Riebler, T. G. Martins, and S. H. Sørbye (2017). Penalising model component complexity: A principled, practical approach to constructing priors. Statistical Science 32(1), 1–28. 26, 27, 32, 33, 38, 41

Sisson, S. A., Y. Fan, and M. Beaumont (2018). Handbook of approximate Bayesian computation. CRC Press. 22

Smith, R. L. (1989). Extreme value analysis of environmental time series: An application to trend detection in ground-level ozone. Statistical Science 4(4), 367–377. 12

Smith, R. L. (1999). Bayesian and frequentist approaches to parametric predictive inference. In Bayesian Statistics 6, Eds. J.M. Bernardo, J.O. Berger, A.P. Dawid & A.F. Smith, pp. 589–612. Oxford Press UK. 27, 106, 142

Smith, R. L. (2003). Statistics of extremes, with applications in environment, insurance, and finance. In Extreme values in finance, telecommunications, and the environment, pp. 20–97. Chapman and Hall/CRC. 13, 24, 26, 27

Solari, S. and M. Losada (2012). A unified statistical model for hydrological variables including the selection of threshold for the peak over threshold method. Water Resources Research 48(10), 1–15. 28

Stan Development Team (2021). RStan: the R interface to Stan. R package version 2.21.3. 80

Stephenson, A. (2016). Bayesian Inference for Extreme Value Modelling. In Extreme Value Modeling and Risk Analysis: Methods and Applications, pp. 257–280. Chapman & Hall/CRC: Boca Raton, Florida. 24, 36

Stephenson, A. and J. Tawn (2004). Bayesian inference for extremes: Accounting for the three extremal types. Extremes 7(4), 291–307. 25, 26, 41

Tancredi, A., C. Anderson, and A. O'Hagan (2006). Accounting for threshold uncertainty in extreme value estimation. Extremes 9(2), 87–106. 28

Tibshirani, R. and L. Wasserman (1994). Some aspects of the reparametrization of statistical models. Canadian Journal of Statistics 22, 163–173. 37

Tierney, L. and J. B. Kadane (1986). Accurate approximations for posterior moments and marginal densities. Journal of the American Statistical Association 81(393), 82–86. 14, 22, 106

Van der Vaart, A. W. (2000). Asymptotic Statistics. Cambridge: Cambridge University Press. 37, 142

Vats, D. and J. M. Flegal (2021). Invited discussion: "Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC". Bayesian Analysis 16(2), 695–701. 64

Vats, D., J. M. Flegal, and G. L. Jones (2019). Multivariate output analysis for Markov chain Monte Carlo. Biometrika 106(2), 321–337. 66, 69, 79, 83

Vats, D. and C. Knudson (2021). Revisiting the Gelman–Rubin Diagnostic. Statistical Science 36(4), 518 – 529. 65, 67, 78, 142

Vehtari, A., A. Gelman, D. Simpson, B. Carpenter, and P.-C. Bürkner (2021). Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC (with discussion). Bayesian Analysis 16(2), 667–718. 29, 65, 66, 68, 69, 70, 71, 72, 82, 96

Wadsworth, J. and J. Tawn (2012). Likelihood-based procedures for threshold diagnostics and uncertainty in extreme value modelling. Journal of the Royal Statistical Society: Series B 74(3), 543–567. 28

Wadsworth, J. L., J. A. Tawn, and P. Jonathan (2010). Accounting for choice of measurement scale in extreme value modeling. Annals of Applied Statistics 4(3), 1558–1578. 12, 13, 28, 35, 38, 44, 45, 50, 53, 54, 55, 56, 58, 59

Walshaw, D. (1994). Getting the most from your extreme wind data: a step by step guide. Journal Of Research of the National Institute Of Standards And Technology 99(4), 399–399. 122, 129

Walshaw, D. (2000). Modelling extreme wind speeds in regions prone to hurricanes. Journal of the Royal Statistical Society: Series C 49(1), 51–62. 25, 26

Weibull, W. (1951). A statistical distribution function of wide applicability. Journal of Applied Mechanics 103(730), 293–297. 5

Woutersen, T. (2011). Consistent estimation and orthogonality. Advances in Econometrics 27(1), 155–178. 37

Zhou, C. (2010). The extent of the maximum likelihood estimator for the extreme value index. Journal of Multivariate Analysis 101(4), 971–983. 8

Zorzetto, E., A. Canale, and M. Marani (2020). Bayesian non-asymptotic extreme value models for environmental data. arXiv:2005.12101. 25, 26