Submitted to Bernoulli

Asymptotic theory for Bayesian inference and prediction: from the ordinary to a conditional Peaks-Over-Threshold method

CLÉMENT DOMBRY[1,a], SIMONE A. PADOAN[2,b] and STEFANO RIZZELLI[3,c]

[1]Université Marie et Louis Pasteur, CNRS, LmB (UMR6623), F-25000 Besançon, France, [a]clement.dombry@univ-fcomte.fr
[2]Department of Decision Sciences, Bocconi University, [b]simone.padoan@unibocconi.it
[3]Department of Statistical Sciences, University of Padova, [c]stefano.rizzelli@unipd.it

The Peaks Over Threshold (POT) method is the most popular statistical method for the analysis of univariate extremes. Even though there is a rich applied literature on Bayesian inference for the POT, the asymptotic theory for such proposals is missing. Even more importantly, the ambitious and challenging problem of predicting future extreme events according to a proper predictive statistical approach has received no attention to date. In this paper we fill this gap by developing the asymptotic theory of posterior distributions (consistency, contraction rates, asymptotic normality, and asymptotic coverage of credible intervals) and prediction in the Bayesian POT framework. We extend this asymptotic theory to account for cases where the focus is on the tail properties of the conditional distribution of a response variable given a vector of random covariates. To enable accurate prediction of extreme events more severe than those previously observed, we derive the posterior predictive distribution as an estimator of the conditional distribution of an out-of-sample random variable, given that it exceeds a sufficiently high threshold. We establish Wasserstein consistency of the posterior predictive distribution under both the unconditional and covariate-conditional approaches and derive its contraction rates. Simulations show the good performance of the proposed Bayesian inferential methods. The analysis of the change in the frequency of financial crises over time shows the utility of our methodology.

Keywords: Bernstein-von Mises; contraction rates; extreme quantiles; extreme value index; non-parametric estimation; posterior consistency; probabilistic forecast; Wasserstein distance

1. Introduction

1.1. Statistical model and its challenges

The mission of Extreme Value Theory (EVT) is modeling and predicting future events that are much more exceptional than those experienced in the past. Accomplishing this goal is undoubtedly important in many applied fields—including finance [19], actuarial science [3], and environmental sciences [8]—where EVT provides tools for supporting risk assessment. Here we focus on the Peaks Over Threshold (POT) method which is arguably the most popular approach in the univariate case. In the first part, we work with a random variable Y with a generic distribution F. Under weak conditions the distribution of $Y-t \mid Y>t$, for a high threshold t that goes to the upper end-point of F, is approximately a Generalized Pareto (GP) distribution $H_{\gamma_0}(\cdot/\sigma_0)$, which depends on a shape parameter $\gamma_0 \in \mathbb{R}$, called the extreme value index (EVI), and a scale parameter $\sigma_0 > 0$ [2].

The key benefits of this result are twofold: given a sample of n independent and identically distributed (i.i.d.) observations from an unknown distribution F, the distribution H_γ provides an approximation to the distribution of rescaled excesses above a high threshold. By leveraging the quantile expression of

H_γ, we derive an approximate formula for the extreme quantiles of F, a crucial tool in applications for assessing future risks.

The POT method is highly practical, yet its inferential theory remains complex (see [11] for examples). Regardless of the inferential approach used, the theoretical analysis must account for the fact that the GP distribution is an inherently misspecified model for excesses, as the threshold must be fixed in practice. Additionally, the scale parameter is threshold-dependent, varying with the chosen threshold rather than being a fixed parameter as in classical statistical literature (e.g., [36]). Moreover, the GP family is an irregular model, as the sign of the shape parameter influences its support, making likelihood-based inference—including Bayesian methods—notoriously challenging. In this paper, we develop inferential methods and the corresponding asymptotic theory while carefully addressing these complexities.

One of the most widely used statistical frameworks in applications is conditional inference, where the objective is to assess specific characteristics of a response variable Y (such as its mean or quantiles) based on available information about covariates. In the second part of this work, we consider a response variable Y following a generic distribution F and a covariate vector $\boldsymbol{X}$ with law $P(\mathrm{d}\boldsymbol{x})$. Our focus is on modeling and analyzing the conditional distribution $F_{\boldsymbol{x}}$ of Y given that $\boldsymbol{X} = \boldsymbol{x}$. Several methodologies have been proposed for modeling and statistically analyzing conditional extremes, including various conditional extreme value models and non- or semi-parametric regression approaches (see, e.g., [23, 33]). In this work, we establish a link between the extremes of a response variable and its associated covariates by adopting a conditional distribution framework that complies with what we call the *proportional tail model*. Similar to the triangular array approach in [16], this model allows the tail of $F_{\boldsymbol{x}}$ to vary according to a scale function, known as the *scedasis* function, while maintaining a constant extreme value index (EVI) (see also [17]). This approach enables the semi-parametric estimation of conditional tail probabilities and a non-parametric assessment of covariate effects using peaks above a high threshold and their associated *concomitant* covariates.

1.2. Objectives and contributions

Over the last decades, numerous inference methods for both unconditional and conditional extreme events have been developed (see, e.g., [3, 10, 11, 16, 23, 37, 38]). While several studies have explored Bayesian inference for the POT method (see Ch. 11 in [3] and [9, 12, 20, 30, 35]), establishing a rigorous inferential theory remains a significant challenge. To the best of our knowledge, no asymptotic results have been derived for these Bayesian approaches. More importantly, the ambitious and challenging problem of predicting future extreme events within a proper statistical predictive framework has received little attention, with the notable exception of [25]. A practical and accessible approach to forecasting such events can be achieved through the Bayesian paradigm, which naturally yields the posterior predictive distribution—an estimator of the conditional distribution of an out-of-sample random variable, given that it exceeds a sufficiently high threshold, representative of future extreme events. The main contributions of this article can be summarized as follows (with a detailed discussion to follow): (i) establishing a rigorous theoretical foundation for Bayesian inference in both unconditional and conditional settings within the POT framework; (ii) providing mathematical guarantees on the accuracy of forecasts based on the posterior predictive distribution.

In the first part of this paper, we develop the asymptotic theory for Bayesian inference within the classical POT framework. Specifically, we provide general and simple conditions on the prior distribution of the GP model's parameters under which we derive key results for the corresponding posterior distribution: consistency with a $\sqrt{k}$-contraction rate (where $k \equiv k(n)$ is the number of peaks above the selected high threshold in the sample), the celebrated Bernstein-von Mises (BvM) theorem, and the asymptotic coverage probability of credible intervals (see Theorem 2.8). These results notably differ from those obtained in the block maximum context by [31], as our approach allows for a broader selection of prior distributions and more general conditions. In particular, we accommodate families of informative proper priors for the shape parameter γ, and both non-informative improper priors and informative proper data-dependent priors for the scale parameter σ, as the prior specification for the latter is more nuanced. In contrast, [31] only considers the case of data-dependent priors. Moreover, while the theory in [31] relies on certain conditions on the density of F, which may be seen as restrictive, our results are derived under weaker, more standard conditions (see Ch. 2 in [11]).

To develop a more comprehensive theory, we deeply investigate the frequentist properties of the empirical log-likelihood process, specifically in the context of the misspecified GP model. In this analysis, we derive its uniform convergence, the convergence of its derivatives (discussed in the Supplementary Material [13]), as well as local and global bounds and a local asymptotically Gaussian expansion (see Propositions 2.1 and 2.6, and Theorem 2.4). Additionally, we obtain the contraction rates for the corresponding Maximum Likelihood Estimator (MLE). Our results make several important contributions to statistical inference. First, they address a longstanding question (see Corollary 2.5): Is the MLE, computed over the entire parameter space, consistent, and is there a unique global maximizer of the likelihood? Second, our findings are essential for developing the asymptotic theory of Bayesian methods (see Theorem 2.8). While a version of the BvM theorem exists for misspecified models [26], this result does not directly apply here. The GP distribution is an irregular statistical model that violates the smoothness conditions outlined in [26], and the degree of misspecification in our setting is not fixed but instead varies with n. This makes the testability assumptions in [26] difficult to verify.

Beyond estimating the tail of F, practitioners are particularly interested in quantifying the quantiles of F for exceptionally small exceedance probabilities, as these correspond to events more extreme than any observed so far. To address this, we extend the asymptotic theory of Bayesian methods (see Corollary 2.10) by deriving consistency, contraction rates, asymptotic normality, and the asymptotic coverage probability of credible intervals for the posterior distribution of the so-called extreme quantiles (Chs. 3–4 in [11]). This is achieved through the development of a general Bayesian delta method, a widely applicable result that extends beyond EVT. The posterior distribution of extreme quantiles is a valuable tool for assessing the intensities of future extreme events, as it quantifies the uncertainty of their magnitude. However, to fully account for the uncertainty in predicting such events, we propose a genuine statistical predictive approach. We conclude the first part by introducing the posterior predictive distribution as an estimator for the conditional distribution of an out-of-sample random variable, given that it exceeds a sufficiently high threshold, representative of a future extreme event. We derive conditions under which this predictive distribution is Wasserstein consistent and quantify its contraction rate (see Theorem 2.12).

In the second part, we address the problem of Bayesian inference for the tail properties of the conditional distribution $F_{\boldsymbol{x}}$ of Y given $\boldsymbol{X} = \boldsymbol{x}$. Building on the *tail proportionality* condition, we show that this inference can be achieved in two steps: first, Bayesian inference of the GP distribution parameters using peaks above a high threshold, and second, estimation of the scedasis function using the concomitant covariates. The first step is already described in the first part of the paper. For the second step, we specify a Dirichlet Process (DP) prior (e.g., Ch. 4.1 in [22]) for the conditional law of the concomitant covariate $\boldsymbol{X}$ given $Y > t$, which induces a prior on the scedasis function at $\boldsymbol{x}$. This function depends on the unknown marginal probability measure $\mathcal{P}(\mathrm{d}\boldsymbol{x})$, which we treat as a nuisance parameter for simplicity, and we estimate it using two methods: a kernel-based method and a K-nearest neighbours approach. For the corresponding posterior distribution, we derive the same type of asymptotic theory established in the first part of the paper (see Theorem 3.1 and Corollary 3.2). With the kernel method, we also provide contraction rates for the posterior distribution of the scedasis function, which hold uniformly in $\boldsymbol{x}$ (see Theorem 3.3). We then turn to the functional estimation of the marginal law of the concomitant covariates. Under the same DP prior, we show that the posterior distribution satisfies the BvM theorem over an infinite-dimensional Skorohod space (see Theorem 3.4). This result forms the basis for deriving a Kolmogorov-Smirnov-type statistical test to assess whether the concomitant covariates significantly affect the extremes of the response variable.

Extreme conditional quantiles are essential tools for assessing the risks associated with extreme events in a phenomenon, particularly when other concomitant dynamics reach certain levels. Under the proportional tail model, the posterior distribution of these quantiles is readily available, as it is induced by the posterior distributions of the GP distribution's parameters and the scedasis function. To further refine our analysis, we also develop the asymptotic behavior of the posterior distribution. The final, but perhaps most significant, statistical problem addressed in this article is forecasting within an extreme regression framework. In this context, we define the posterior predictive distribution as an estimator of the conditional distribution of an out-of-sample random variable, given that it exceeds an extreme conditional quantile and the concomitant covariates reach certain levels. For this predictive distribution, we establish minimal conditions to prove its Wasserstein consistency and quantify its contraction rate.

1.3. Workflow

Section 2 presents key concepts and notation for the POT method, outlines the theoretical properties of the log-likelihood empirical process, develops the asymptotic theory for Bayesian inference within the POT framework, and establishes the Wasserstein contraction rates for the corresponding posterior predictive distribution. Section 3 introduces the proportional tail model to link the extremes of a response variable to covariates, provides the posterior asymptotic theory for tail-related quantities in the conditional distribution, and derives the Wasserstein contraction rates for the associated posterior predictive distribution. Section 4 offers a comprehensive simulation study demonstrating the finite-sample performance of the proposed methodology. All proofs are provided in the Supplementary Material [13], which also includes additional results, details on posterior computations, simulations, and an application to real financial data, analyzing the change in crisis frequency over time.

2. The Peaks-Over-Threshold approach

The POT method is one of the most widely used approaches for modeling univariate extremes and has been applied in a broad range of contexts, including sea-level analysis, financial indices, and lifespan modeling, among others [11]. In this section, we develop a set of results—covering uniform convergence, local and global bounds for the empirical log-likelihood process, as well as contraction rates and uniqueness of the maximum likelihood estimator—that extend the existing likelihood theory for the misspecified GP model and provide a foundation for the Bayesian asymptotic theory developed in this work. For a list of symbols used throughout the article, see Table 1 in the Supplementary Material [13].

2.1. Background

Let Y be a random variable with distribution function F_0 belonging to the domain of attraction of a Generalized Extreme Value (GEV) distribution G_{γ_0}, denoted by $F_0 \in \mathcal{D}(G_{\gamma_0})$, where $\gamma_0 \in \mathbb{R}$ is the EVI that describes the tail heaviness of F_0 [11, Theorem 1.1.3]. This means that for any integer $m \geq 1$, there are norming constants $a_m > 0$ and $b_m \in \mathbb{R}$ such that $F_0^m(a_m y + b_m) \to G_{\gamma_0}(y) = \exp(-(1+\gamma_0 y)^{-1/\gamma_0})$, as $m \to \infty$, for all $y \in \mathbb{R}$ that are continuity points of G_{γ_0}. Examples of distributions in the domain of attraction include: $F_0(y) = 2\pi^{-1}\arctan(y)$, i.e. Half-Cauchy distribution, then $\gamma_0 = 1$, corresponding to the Fréchet domain of attraction; (ii) $F_0(y) = \mathrm{Ga}(y;2,2)$, where $\mathrm{Ga}(y;\alpha,\beta)$ is a Gamma distribution function with shape $\alpha > 0$ and rate $\beta > 0$ parameters, then $\gamma_0 = 0$, corresponding to the Gumbel domain of attraction; and (iii) $F_0(y) = \mathrm{Be}(y;1,3)$, where $\mathrm{Be}(y;\alpha,\beta)$ is a Beta distribution function with shape parameters $\alpha,\beta > 0$, then $\gamma_0 = -1/3$, corresponding to the Weibull domain of attraction. The domain of attraction condition is quite general, as it is satisfied by most continuous distributions encountered in practice—such as Pareto, Student-t, Gaussian, exponential, and bounded distributions (see, e.g., [11, Chs. 1.1–1.2] and [32, Chs. 1.2–1.3])—except for distributions with highly irregular or oscillating tails (see the counterexamples in [32, Ch. 1.5] and [11, Appendix B]). The domain of attraction condition (or first-order condition) leads to an important result: if, for $t < y^*$ there exists a positive scaling function $s_0(t) > 0$ such that

$$\lim_{t\uparrow y^*} \mathbb{P}(Y \le s_0(t)z \mid Y > t) = H_{\gamma_0}(z), \tag{1}$$

where $y^* = \sup\{y : F_0(y) < 1\}$ is the right end-point of F_0, then the limiting distribution H_{γ_0} must be a GP distribution with unit scale ([2], [11, Theorem 1.1.6]). One possible choice for the norming constants is $b_m = U_0(m)$, where $U_0(t) = F_0^{\leftarrow}(1-1/t)$ with $F_0^{\leftarrow}(y) = \inf\{x : F_0(x) \ge y\}$ is the so-called *tail quantile* (or the $(1-1/t)$-quantile of F_0), $a_m = a_0(m)$ for a suitable positive function $a_0(\cdot)$, and for the scaling function one can set $s_0(t) = a_0(1/(1-F_0(t)))$, see [11, Ch. 1.2]. In the sequel we consider this choice. In the case that F_0 is differentiable, a possible selection for a_0 is $a_0(t) = tU_0'(t)$. With this choice we obtain with the above illustrative examples: (i) $a_0(t) = \pi/(2t)\csc^2(\pi/(2t))$ and $s_0(t) = t^2\arctan(1/t)$, where csc is the cosecant function and arctan is the arctangent function; (ii) $a_0(t) = (1+2U_0(t))/(4U_0(t))$ and $s_0(t) = (1+2t)/(4t)$; (iii) $a_0(t) = 1/(3t^{1/3})$ and $s_0(t) = (1-t)/3$. The GP is a family of two-parameter distributions defined as $H_{\boldsymbol{\vartheta}}(z) = H_\gamma(z/\sigma)$ for all $z \in \mathcal{S}_{\boldsymbol{\vartheta}}$, where $\boldsymbol{\vartheta} = (\gamma,\sigma)^\top \in \mathbb{R}\times(0,\infty)$,

$$H_\gamma(z) = \begin{cases} 1-(1+\gamma z)_+^{-1/\gamma}, & \text{if } \gamma \neq 0, \\ 1-\exp(-z), & \text{if } \gamma = 0, \end{cases}$$

with $(z)_+ = \max(0,z)$, and $\mathcal{S}_{\boldsymbol{\vartheta}}$ is $[0,\infty)$ if $\gamma \ge 0$ while it is $[0,-\sigma/\gamma]$ if $\gamma < 0$. The GP density is $h_{\boldsymbol{\vartheta}}(z) = h_\gamma(z/\sigma)/\sigma$, where

$$h_\gamma(z) = \begin{cases} (1+\gamma z)_+^{-(1/\gamma+1)}, & \text{if } \gamma \neq 0, \\ \exp(-z), & \text{if } \gamma = 0. \end{cases}$$

The log-likelihood of the density $h_{\boldsymbol{\vartheta}}$, corresponding to a single observation, is defined for $z \ge 0$ as

$$\ell_{\boldsymbol{\vartheta}}(z) = \begin{cases} -\log\sigma - \left(1+\dfrac{1}{\gamma}\right)\log\left(1+\dfrac{\gamma}{\sigma}z\right), & \text{if } 1+\dfrac{\gamma}{\sigma}z > 0, \\ -\infty, & \text{otherwise.} \end{cases}$$

Observe that the log-likelihood is unbounded when $\gamma < -1$, more precisely for any $y \ge 0$

$$\lim_{\sigma\downarrow-\gamma z} \ell_{\boldsymbol{\vartheta}}(z) = \infty.$$

We recall that given the function

$$I_{\boldsymbol{\vartheta}'} = -\int_0^1 \left.\frac{\partial^2 \ell_{\boldsymbol{\vartheta}}}{\partial\boldsymbol{\vartheta}\partial\boldsymbol{\vartheta}^\top}\right|_{\boldsymbol{\vartheta}=\boldsymbol{\vartheta}'} \left(\sigma'\frac{v^{-\gamma'}-1}{\gamma'}\right) \mathrm{d}v, \quad \forall\boldsymbol{\vartheta}' \in \mathbb{R}\times(0,\infty), \tag{2}$$

the Fisher information matrix corresponding to the GP log-likelihood with true parameter $\boldsymbol{\vartheta}_0$ is defined as $I_{\boldsymbol{\vartheta}_0}$ (see also Formula (2.11) in the Supplementary Material [13]), which is positive definite as soon as $\gamma_0 > -1/2$. For this reason in the sequel we restrict the parameter space to $\Theta = (-1/2, \infty) \times (0, \infty)$. As the GP model is irregular, $I_{\boldsymbol{\vartheta}_0}$ cannot be derived employing the second Bartlett's identity [e.g., 36, Ch. 7.2].

2.2. Empirical log-likelihood process asymptotics

In this section, we examine key frequentist properties of the empirical log-likelihood process associated with the GP density function. We derive its uniform convergence, along with that of its derivatives (detailed in the Supplementary Material [13]), its local asymptotic expansion, and both local and global bounds. These results provide a comprehensive understanding of the GP likelihood theory, serving as a foundation for the asymptotic theory of Bayesian methods and offering valuable insights from a frequentist perspective. As a by-product, we establish the uniqueness and contraction rates of the MLE of the GP likelihood function for large samples, a new and noteworthy contribution.

For $n \geq 1$, let $(Y_1, \dots, Y_n)$ be i.i.d. copies of Y whose distribution satisfies $F_0 \in \mathcal{D}(G_{\gamma_0})$. The first-order condition (1) implies that $\mathbb{P}(Y \leq y \mid Y > t) \approx H_{\boldsymbol{\vartheta}_0}(y - t)$ for high enough threshold t, with $\boldsymbol{\vartheta}_0 = (\gamma_0, \sigma_0)^\top$ where the scale parameter σ_0 is representative of the true scale function $s_0(t)$ and $y \equiv y_t = t + s_0(t)z$, for all $z \geq 0$. Let $k = k(n)$ be a so-called *intermediate sequence*, with $k = o(n)$ and $k \to \infty$ as $n \to \infty$. A practical way of defining a high threshold is $t = U_0(n/k)$ and estimating $U_0(n/k)$ by the order statistic $Y_{n-k,n}$, where $Y_{1,n} \leq \cdots \leq Y_{n,n}$ are the n order statistics, and F_0 by the empirical distribution F_n, we obtain $s_0(U_0(n/k)) \approx a_0(1/(1 - F_n(Y_{n-k,n}))) = a_0(n/k)$. Accordingly, σ_0 is representative of $a_0(n/k)$. A possible criterion for selecting k is discussed in Section 6.1 of the Supplementary Material [13]. We focus on the normalized excess over-high threshold variables, also known as "pseudo-observations",

$$Z_i = \frac{Y_{n-i+1,n} - Y_{n-k,n}}{a_0(n/k)}, \quad 1 \leq i \leq k.$$

For an arbitrary Borel set B the empirical probability measures relative to observed unrescaled peaks and pseudo-observations are denoted by $\mathbb{P}_n(B) = k^{-1} \sum_{i=1}^k \mathbb{1}(Y_{n-k+i,n} - Y_{n-k,n} \in B)$ and $\mathbb{P}_n^{\text{pse}}(B) = k^{-1} \sum_{i=1}^k \mathbb{1}(Z_i \in B)$, respectively. Given a probability measure P on a measurable space $(\mathcal{X}, \mathcal{B})$ and a measurable function $f : \mathcal{X} \mapsto \mathbb{R}$ we use Pf to denote $\int f \mathrm{d}P$. Accordingly, $\mathbb{P}_n f = k^{-1} \sum_{i=1}^k f(Y_{n-i+1,n} - Y_{n-k,n})$ and $\mathbb{P}_n^{\text{pse}} f = k^{-1} \sum_{i=1}^k f(Z_i)$.

Based on this, we define the empirical log-likelihood process relative to the observed unnormalised excesses over a high threshold as $\mathscr{L}_n(\boldsymbol{\vartheta}) = \mathbb{P}_n \ell_{\boldsymbol{\vartheta}}$, $\boldsymbol{\vartheta} \in \Theta$. Since σ is not a fixed parameter and depends on the sample size n, to stabilize it as n increases, we introduce the reparametrization $\boldsymbol{\theta} := r(\boldsymbol{\vartheta}) = (\gamma, \sigma/a_0(n/k))^\top$ for all $\boldsymbol{\vartheta} \in \Theta$, yielding $\boldsymbol{\theta}_0 = r(\boldsymbol{\vartheta}_0) = (\gamma_0, 1)^\top$. Our theory is developed using the empirical log-likelihood process $L_n(\boldsymbol{\theta}) = \mathbb{P}_n^{\text{pse}} \ell_{\boldsymbol{\theta}}$, which is defined through the GP log-likelihood $\ell_{\boldsymbol{\theta}}$. Note that $L_n(\boldsymbol{\theta}) = \mathscr{L}_n(\boldsymbol{\vartheta}) + \log a_0(n/k)$. For convenience, we refer to $\mathscr{L}_n$ and L_n as the "realistic" and "theoretical" empirical log-likelihood processes, respectively. The former is the version used for practical inference, while the latter is employed to study the asymptotic properties of the process. Because these log-likelihoods are derived from an asymptotically justified GP model, which is for this result a misspecified model for threshold exceedances, this feature must be taken into account when establishing the corresponding asymptotic properties. Finally, we define the (theoretical) score and information processes as $S_n(\boldsymbol{\theta}) = (\partial/\partial\boldsymbol{\theta}) L_n(\boldsymbol{\theta})$ and $J_n(\boldsymbol{\theta}) = (\partial^2/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top) L_n(\boldsymbol{\theta})$. In the following, given two vectors $\boldsymbol{x}, \boldsymbol{y}$ of equal size, $\boldsymbol{x}^\top \boldsymbol{y}$ denotes the usual vector product, while componentwise multiplication and division are denoted as $\boldsymbol{x}\boldsymbol{y} = (x_1 y_1, \dots, x_q y_q)^\top$ and $\boldsymbol{x}/\boldsymbol{y} = (x_1/y_1, \dots, x_q/y_q)^\top$, respectively.

We now present several key results that are crucial for the asymptotic theory of the MLE and serve as the foundation for deriving the main findings in the subsequent section on the Bayesian approach. Let $B(\boldsymbol{\theta}_0,\varepsilon)=\{\boldsymbol{\theta}\in\Theta:\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|<\varepsilon\}$ denote the open ball centered at $\boldsymbol{\theta}_0$ with radius ε, and let $B(\boldsymbol{\theta}_0,\varepsilon)^{\complement}$ be its complement in Θ. We define

$$\widehat{\boldsymbol{\theta}}_n \in \underset{\boldsymbol{\theta}\in B(\boldsymbol{\theta}_0,\varepsilon)}{\operatorname{argmax}}\ L_n(\boldsymbol{\theta})$$

as a local maximizer of the empirical log-likelihood process. Note that the local maximizer of the realistic empirical log-likelihood process $\mathscr{L}_n(\boldsymbol{\vartheta})$ satisfies $\widehat{\boldsymbol{\vartheta}}_n=r^{-1}(\widehat{\boldsymbol{\theta}}_n)$. Note that a different neighborhood of $\boldsymbol{\theta}_0$ with compact closure in Θ might be considered for the definition of local MLE $\widehat{\boldsymbol{\theta}}_n$, though $B(\boldsymbol{\theta}_0,\varepsilon)$ is a natural choice. Note also that the first-order condition is equivalent to [11, Ch. 1]

$$\lim_{t\to\infty}\frac{U_0(ty)-U_0(t)}{a_0(t)}=\frac{y^{\gamma_0}-1}{\gamma_0},\quad \forall\, y>0. \tag{3}$$

To quantify how the right-hand side accurately approximates the left-hand side in (3) for $t=n/k$ that goes to infinity, and to control the asymptotic behavior of generic estimation procedures based on this approximation, we consider the following so-called second-order condition. This condition is a standard assumption in EVT for quantifying the asymptotic bias in the estimation of the extreme value index and extreme quantile. For technical details and examples, see Sections 1.2 and 6.1 in the Supplementary Material [13], and for a comprehensive discussion see Ch. 2 and Appendix B in [11].

Condition 1. *There is a rate function A, i.e. a positive or negative function satisfying $A(t)\to 0$ as $t\to\infty$, such that:*

(a) A is regularly varying with index $\rho\le 0$, i.e. $\lim_{t\to\infty}A(xt)/A(t)=x^{\rho}$ for all $x>0$, and

$$\lim_{t\to\infty}\frac{\frac{U_0(tx)-U_0(t)}{a_0(t)}-\frac{x^{\gamma_0}-1}{\gamma_0}}{A(t)}=\int_1^x v^{\gamma_0-1}\int_1^v u^{\rho-1}\mathrm{d}u\mathrm{d}v. \tag{4}$$

(b) $\sqrt{k}A(n/k)\to\lambda\in\mathbb{R}$ as $n\to\infty$.

With the examples (i)–(iii) presented at the beginning of Section 2.1, we have $\rho=-2$ and $A(t)=-t^{-2}\pi/3$, $\rho=0$ and $A(t)=-(-\log t)^{-2}$, and $\rho=-\infty$ and $A(t)=0$, for all $t>0$, respectively. These examples cover the cases $\rho<0$ and $\rho=0$ (mild and strong bias), furthermore, the case $\rho=-\infty$ and $A(t)=0$, corresponds to a setting with no bias. Further examples are provided in Table 2 of the Supplementary Material [13]. Note that when $\rho<0$, the no-bias condition is attained provided that $\sqrt{k}(n/k)^{\rho+\delta}\to 0$ as $n\to\infty$, for any small $\delta>0$.

Proposition 2.1. *Under Condition 1 there exist $\varepsilon_0>0$ and constants $c_1,c_2,c_3>0$ such that the following three properties hold with probability tending to 1 as $n\to\infty$:*

- *When $\gamma_0>0$, L_n is strictly concave on $B(\boldsymbol{\theta}_0,\varepsilon_0)$ with a unique maximizer $\widehat{\boldsymbol{\theta}}_n$ and*

$$-c_1\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_n\|^2\le L_n(\boldsymbol{\theta})-L_n(\widehat{\boldsymbol{\theta}}_n)\le -c_2\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_n\|^2,\quad \boldsymbol{\theta}\in B(\boldsymbol{\theta}_0,\varepsilon_0). \tag{5}$$

- *More generally, if $\gamma_0>-1/2$, $S_n(\boldsymbol{\theta}_0)=O_{\mathbb{P}}(1/\sqrt{k})$ and*

$$L_n(\boldsymbol{\theta})-L_n(\boldsymbol{\theta}_0)\le(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^\top S_n(\boldsymbol{\theta}_0)-c_3\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2,\quad \boldsymbol{\theta}\in B(\boldsymbol{\theta}_0,\varepsilon_0), \tag{6}$$

$$\sup_{\boldsymbol{\theta}\in B(\boldsymbol{\theta}_0,\epsilon_n)} \left\| J_n(\boldsymbol{\theta}) + \boldsymbol{I}_{\boldsymbol{\theta}_0} \right\| = o_{\mathbb{P}}(1), \text{ for any } \epsilon_n = R/\sqrt{k} \text{ with } R = o(\sqrt{k}/\log k \vee k^{1/2+\gamma_0^-}), \tag{7}$$

where $\boldsymbol{I}_{\boldsymbol{\theta}_0}$ is the Fisher information of the GP log-likelihood (defined as in (2) but with $\boldsymbol{\theta}$ in the place of $\boldsymbol{\vartheta}$) at $\boldsymbol{\theta}_0$.

Next result uses Proposition 2.1 to establish the contraction rates of the local MLE $\widehat{\boldsymbol{\vartheta}}_n$.

Corollary 2.2. *For any $R \to \infty$ satisfying $R = o(\sqrt{k})$ we have $\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| = O_{\mathbb{P}}(R/\sqrt{k})$. Accordingly, the normalized local MLE sequence $(\widehat{\boldsymbol{\vartheta}}_n)_{n\geq 1}$ is $\sqrt{k}$-consistent, i.e.*

$$\widehat{\gamma}_n = \gamma_0 + O_{\mathbb{P}}(1/\sqrt{k}) \quad \text{and} \quad \widehat{\sigma}_n = a_0(n/k)\big(1 + O_{\mathbb{P}}(1/\sqrt{k})\big), \tag{8}$$

and unique with probability tending to one, as $n \to \infty$.

Remark 2.3. Surprisingly, this result has not been established before. Previous works have mainly focused on the existence of a consistent local maximizer (potentially among several others) and the asymptotic behavior of the solutions to the likelihood equations, assuming they lie within a narrow neighborhood of the true parameter values. This neighborhood was defined by the conditions $|\gamma/(\sigma/a_0(n/k)) - \gamma_0| = O_{\mathbb{P}}(1/\sqrt{k})$ and $\sigma/a_0(n/k) = e^{O_{\mathbb{P}}(1)}$ [e.g., 15, Proposition 3.1]. Recently, [18] showed that the log-likelihood is strictly concave within a slightly larger, but still shrinking, neighborhood of the form $\{\boldsymbol{\vartheta} : |\gamma - \gamma_0| + |\sigma/a_0(n/k) - 1| < R/\sqrt{k}\}$, with probability tending to one as $n \to \infty$, ensuring that there is a unique maximizer within this region. However, this result does not rule out the possibility of other local maximizers outside this neighborhood. Corollary 2.2 addresses this gap, providing the contraction rates for a generic local maximizer. Nevertheless, it does not guarantee that the MLE, obtained by maximizing the likelihood over the entire parametric space, is consistent and represents the global unique likelihood maximizer.

In Theorem 2.4, we derive global upper bounds for the theoretical empirical log-likelihood process, which serve as a crucial tool in answering the open question: Is the MLE computed over the entire parameter space consistent, and is it the unique global likelihood maximizer? The affirmative answer is given in Corollary 2.5.

Theorem 2.4. *Under Condition 1 there exist $\varepsilon_0 > 0$ and constants $c_1, c_2 > 0$ and, for all large enough $\overline{\tau} > 0$, there exist constants $c_3 > 0$, $c_4, c_5 \in \mathbb{R}$ such that, with probability tending to 1 as $n \to \infty$,*

$$L_n(\boldsymbol{\theta}) - L_n(\boldsymbol{\theta}_0) \leq (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top S_n(\boldsymbol{\theta}_0) - c_1\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2 \quad \text{if } \boldsymbol{\theta} \in B(\boldsymbol{\theta}_0, \varepsilon_0), \tag{9}$$

$$L_n(\boldsymbol{\theta}) - L_n(\boldsymbol{\theta}_0) \leq -c_2 \quad \text{if } \boldsymbol{\theta} \in B(\boldsymbol{\theta}_0, \varepsilon_0)^{\complement}, \tag{10}$$

$$L_n(\boldsymbol{\theta}) - L_n(\boldsymbol{\theta}_0) \leq -\log\sigma - \frac{c(\tau)}{\sigma} - d(\tau), \quad \text{for all } \boldsymbol{\theta} \in \Theta, \tag{11}$$

with $\tau = \gamma/\sigma$, $c(\tau) = c_3\mathbb{1}_{\tau\leq\overline{\tau}} + (2\tau)^{-1}\log\tau\mathbb{1}_{\tau>\overline{\tau}} > 0$, $d(\tau) = c_4\mathbb{1}_{\tau\leq\overline{\tau}} + (\log\tau + c_5)\mathbb{1}_{\tau>\overline{\tau}}$.

Corollary 2.5. *Any element of the set $\operatorname{argmax}_{\boldsymbol{\theta}\in\Theta} L_n(\boldsymbol{\theta})$ is a $\sqrt{k}$-consistent MLE and, with probability tending to one, there is a unique maximizer coinciding with the local MLE, i.e. $\widehat{\boldsymbol{\theta}}_n$ is the unique global maximizer of the likelihood.*

The global bounds presented in Theorem 2.4 are essential for deriving contraction rates and the BvM theorem for the posterior distribution of the GP distribution's parameters (see the next section). In particular, the BvM result is derived from the asymptotic properties of the theoretical empirical log-likelihood process and its associated quantities (e.g., Ch. 7 in [36]). We further extend the analysis of the theoretical empirical log-likelihood process by providing its local asymptotic expansion, from which we can deduce the asymptotic normality of the MLE. In the following, we denote a multivariate normal distribution by $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{\Sigma}$ is the covariance matrix, which reduces to $\mathcal{N}(\mu, \sigma^2)$ in the univariate case. The corresponding probability measure is denoted as $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Proposition 2.6. *Assume that Condition 1 is satisfied. Then, for all fixed $c > 0$ and $\epsilon_n = c/\sqrt{k}$ we have*

$$\sup_{\boldsymbol{\theta} \in B(\boldsymbol{\theta}_0, \epsilon_n)} \left| L_n(\boldsymbol{\theta}) - L_n(\boldsymbol{\theta}_0) - (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top S_n(\boldsymbol{\theta}_0) + \frac{1}{2k}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \boldsymbol{I}_{\boldsymbol{\theta}_0}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \right| = o_{\mathbb{P}}\left(\frac{1}{k}\right).$$

In particular, $\sqrt{k} S_n(\boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(\lambda \boldsymbol{\mu}, \boldsymbol{V})$, where $\boldsymbol{\mu} = (\mu_1, \mu_2)^\top$ is defined through

$$\mu_1 = \begin{cases} \dfrac{1}{\gamma_0 \rho}\left[-\dfrac{\gamma_0}{(1+\gamma_0)(1+2\gamma_0)} + \dfrac{2 + 5\gamma_0 + 3\gamma_0^2 - 2\rho - 4\gamma_0\rho - \gamma_0^2\rho}{(1+\gamma_0)(1+2\gamma_0)(1+\gamma_0-\rho)(1-\rho)} \right], & \rho < 0 \neq \gamma_0, \\ \dfrac{1}{\rho}\left[-1 + \dfrac{1}{(1-\rho)^2} \right], & \rho < 0 = \gamma_0, \\ \dfrac{1}{\gamma_0}\left[-\dfrac{1}{(1+\gamma_0)(1+2\gamma_0)} + \dfrac{1}{1+\gamma_0} \right], & \rho = 0 \neq \gamma_0, \\ 2 & \rho = 0 = \gamma_0, \end{cases}$$

$$\mu_2 = \begin{cases} \dfrac{1}{\rho}\left[-\dfrac{1}{1+2\gamma_0} + \dfrac{1+\gamma_0}{(1+2\gamma_0)(1+\gamma_0-\rho)} \right], & \rho < 0 \neq \gamma_0, \\ \dfrac{1}{\rho}\left[-1 + \dfrac{1}{1-\rho} \right], & \rho < 0 = \gamma_0, \\ \dfrac{1}{\gamma_0}\left[-\dfrac{1}{1+2\gamma_0} + \dfrac{1}{1+\gamma_0} \right], & \rho = 0 \neq \gamma_0, \\ 1, & \rho = 0 = \gamma_0, \end{cases}$$

and

$$\boldsymbol{V} = \begin{pmatrix} \dfrac{5\gamma_0^2 + 6\gamma_0 + 2}{(1+2\gamma_0)^2(1+\gamma_0)^2}, & \dfrac{1+\gamma_0}{(1+2\gamma_0)^2} \\ \dfrac{1+\gamma_0}{(1+2\gamma_0)^2}, & \dfrac{(1+\gamma_0)^2}{(1+2\gamma_0)^2} \end{pmatrix}. \tag{12}$$

Accordingly, with $\boldsymbol{b} = \boldsymbol{I}_{\boldsymbol{\theta}_0}^{-1}\boldsymbol{\mu}$ and $\boldsymbol{\Sigma} = \boldsymbol{I}_{\boldsymbol{\theta}_0}^{-1}\boldsymbol{V}\boldsymbol{I}_{\boldsymbol{\theta}_0}^{-1}$, the MLE in Corollary 2.5 satisfies

$$\sqrt{k}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(\lambda \boldsymbol{b}, \boldsymbol{\Sigma}).$$

Remark 2.7. Recently, [18] derived similar asymptotic results for the GP log-likelihood in a more complex nonstationary space-time framework, assuming no bias. In contrast, we work in a simpler setup but derive the theory under the more general assumption that bias may exist. While the asymptotic bias and variance of the MLE in our study match those reported in [11, Theorem 3.4.2], our result is the first to establish asymptotic normality for the global likelihood maximizer.

2.3. Asymptotic theory of the posterior distribution

In this section we study the asymptotic properties of a Bayesian procedure for inference with the POT approach. We assume to work with a prior distribution on $\boldsymbol{\vartheta} \in \Theta$ with density of the following form

$$\pi(\boldsymbol{\vartheta}) = \pi_{\mathrm{sh}}(\gamma)\pi_{\mathrm{sc}}^{(n)}(\sigma), \quad \boldsymbol{\vartheta} \in \Theta, \tag{13}$$

where π_{sh} is a prior density on γ and for each $n = 1, 2, \ldots, \pi_{\mathrm{sc}}^{(n)}$ is a prior density on σ, whose expression may or may not depend on n. Although the prior in (13) assumes independence between γ and σ, it enables the specification of fairly flexible forms of their joint density. To establish our main results on the posterior distribution of these parameters, we need to work with a (genuine or empirical) prior density that satisfies the following weak conditions.

Condition 2. *The densities π_{sh} and $\pi_{sc}^{(n)}$ are such that:*

(a) For each $n = 1, 2, \ldots, \pi_{sc}^{(n)} : \mathbb{R}_+ \to \mathbb{R}_+$ and

(a.1) there is a constant $\delta > 0$ such that $\pi_{sc}^{(n)}(a_0(n/k))a_0(n/k) > \delta$ and for any constant $\eta > 0$ there is $\epsilon > 0$ such that

$$\sup_{1-\epsilon<\sigma<1+\epsilon} \left| \frac{\pi_{sc}^{(n)}(a_0(n/k)\sigma)}{\pi_{sc}^{(n)}(a_0(n/k))} - 1 \right| < \eta;$$

(a.2) there is $C > 0$ such that $\sup_{\sigma>0} \sigma a_0(n/k)\pi_{sc}^{(n)}(a_0(n/k)\sigma) \leq C$;

Inequalities (a.1)-(a.2) hold with probability tending to 1, for fixed $\delta, \eta, \epsilon, C$, if $\pi_{sc}^{(n)}$ is data-dependent.

(b) π_{sh} is a positive and continuous function at γ_0 such that: $\int_{-1/2}^{0} \pi_{sh}(\gamma)\,\mathrm{d}\gamma < \infty$, $\sup_{\gamma>0} \pi_{sh}(\gamma) < \infty$.

Below we provide concrete examples of informative and improper non-informative prior distributions that satisfy such conditions.

Example (Informative data-dependent prior). Let $\pi_{\mathrm{sc}}^{(n)}(\cdot) = \pi(\cdot/\widehat{\sigma}_n)/\widehat{\sigma}_n$, where π is a proper prior density on $(0, \infty)$ and $\widehat{\sigma}_n$ is an estimator of $a_0(n/k)$. Then,

$$\pi_{\mathrm{sc}}^{(n)}(\sigma a_0(n/k))a_0(n/k) = \pi(\sigma a_0(n/k)/\widehat{\sigma}_n)a_0(n/k)/\widehat{\sigma}_n.$$

Now, if $a_0(n/k)/\widehat{\sigma}_n \xrightarrow{\mathbb{P}} 1$ as $n \to \infty$ (examples are the MLE, generalized probability weighted moment estimator, etc.) and π is continuous and $\sigma \mapsto \pi(\sigma t)$ is uniformly integrable in σ for t in a neighborhood of 1 (examples are Gamma, Inverse-Gamma, Weibull, Pareto, etc.), then Condition 2(a) is satisfied. The informative joint prior density is completed setting $\pi_{\mathrm{sh}}(\gamma) = \pi(\gamma)$, where π is a probability density on $(-1/2, \infty)$ assumed continuous and bounded away from infinity.

Example (Improper non-informative prior). Consider a uniform distribution on $\log\sigma$ so that $\pi_{\mathrm{sc}}^{(n)}(\sigma) = \pi(\sigma) \propto 1/\sigma$, for any given $n \geq 1$. Accordingly, well known non-informative prior densities are: the uniform prior $\pi(\boldsymbol{\vartheta}) \propto \sigma^{-1}$, the maximal data information $\pi(\boldsymbol{\vartheta}) \propto \sigma^{-1}\exp(-(\gamma+1))$ and the Jeffreys prior $\pi(\boldsymbol{\vartheta}) \propto \sigma^{-1}((1+\gamma)(1+2\gamma)^{1/2})^{-1}$, with $\sigma > 0$ and $\gamma > -1/2$ (e.g., [30]). In these cases

$$\pi(a_0(n/k))a_0(n/k) = 1, \quad \frac{\pi(a_0(n/k)\sigma)}{\pi(a_0(n/k))} = \frac{1}{\sigma},$$

and so Condition 2(a) is trivially satisfied.

Given a prior density π (proper informative or improper non-informative), accordingly, the posterior distribution on the parameters $\boldsymbol{\vartheta}$ of the GP distribution is defined as

$$\Pi_n(B) = \frac{\int_B \exp(k\mathscr{L}_n(\boldsymbol{\vartheta}))\pi(\boldsymbol{\vartheta})\mathrm{d}\boldsymbol{\vartheta}}{\int_\Theta \exp(k\mathscr{L}_n(\boldsymbol{\vartheta}))\pi(\boldsymbol{\vartheta})\mathrm{d}\boldsymbol{\vartheta}}, \tag{14}$$

for all measurable sets $B \subset \Theta$. Consistently with the frequentist context, the asymptotic theory of the posterior distribution is derived working with the theoretical empirical log-likelihood process L_n. The corresponding posterior distribution is denoted by $\Upsilon_n = \Pi_n \circ r^{-1}$ (see Section 2.2).

Theorem 2.8. *Assume $F_0 \in \mathcal{D}(G_{\gamma_0})$ and that Condition 1 is satisfied and that the prior density $\pi(\boldsymbol{\vartheta})$, $\boldsymbol{\vartheta} \in \Theta$, satisfies Condition 2. Then:*

- *(Contraction rate) Υ_n is consistent with $\sqrt{k}$-contraction rate, namely there is a $R > 0$ such that for all sequences $\varepsilon_n \to 0$, satisfying $\sqrt{k}\varepsilon_n \to \infty$ as $n \to \infty$,*

$$\Upsilon_n\left(B(\boldsymbol{\theta}_0, \varepsilon_n)^{\complement}\right) = O_{\mathbb{P}}\left(\exp\left(-Rk\varepsilon_n^2\right)\right).$$

- *(Bernstein-von Mises) As $n \to \infty$,*

$$\sup_{B\subset\mathbb{R}^2} |\Upsilon_n(\{\boldsymbol{\theta} : \sqrt{k}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \in B\}) - \mathcal{N}(B; \boldsymbol{I}_{\boldsymbol{\theta}_0}^{-1}\sqrt{k}S_n(\boldsymbol{\theta}_0), \boldsymbol{I}_{\boldsymbol{\theta}_0}^{-1})| = o_{\mathbb{P}}(1),$$

  where B represents any Borel set in $\mathbb{R}^2$.
- *(Coverage probability) If $\lambda = 0$, for any $\alpha \in (0, 1)$, as $n \to \infty$,*

$$\mathbb{P}\left(\left\{\gamma_0 \in (\Upsilon_{n;sh}^{\leftarrow}(\alpha/2), \Upsilon_{n;sh}^{\leftarrow}(1 - \alpha/2))\right\}\right) = 1 - \alpha + o(1),$$

  where $\Upsilon_{n;sh}^{\leftarrow}(1 - a)$ is the $(1 - a)$-quantile of the posterior distribution of γ.

Theorem 2.8 implies that the posterior distribution Π_n, together with its marginal posterior distributions, asymptotically concentrates around the true parameter values and is asymptotically Gaussian. Moreover, the coverage probabilities of credible intervals for γ_0 achieve asymptotically the nominal level. This is possible only when $\lambda = 0$, that is, when there is no asymptotic bias. In this case, the limiting distribution of $\boldsymbol{I}_{\boldsymbol{\theta}_0}^{-1}\sqrt{k}S_n(\boldsymbol{\theta}_0)$ is Gaussian with zero mean vector (see Proposition 2.6). Beyond estimating γ_0 to assess the heaviness of the data distribution's tail, one of the primary objective of EVT is to infer extreme events beyond those observed in the past. This can be achieved by examining the $(1 - p)$-quantile of F_0 for an exceedance probability $p = p_n$ such that $np \to c \geq 0$ as $n \to \infty$ (or, more generally, $p = o(k/n)$). The key case occurs when $c \leq 1$, implying that in a sample of size n, at most one observation is expected to exceed such an extreme quantile. Since F_0 is unknown in applications, we can use the approximation $1 - p = F_0(y_p) \approx 1 - (k/n)(1 - H_{\boldsymbol{\vartheta}_0}(y_p - U_0(n/k)))$ for large n (derived from the first-order condition (1)), and $y_p = F_0^{\leftarrow}(1 - p)$, to obtain as $n \to \infty$ the quantile approximation

$$F_0^{\leftarrow}(1 - p) \approx U_0\left(\frac{n}{k}\right) + H_{\boldsymbol{\vartheta}_0}^{\leftarrow}\left(1 - \frac{np}{k}\right) = U_0\left(\frac{n}{k}\right) + \sigma_0\frac{\left(\frac{k}{np}\right)^{\gamma_0} - 1}{\gamma_0}, \tag{15}$$

[e.g., 11, Ch. 3]. For each given n, k and p, the right hand-side of (15) is a continuous map $T_n : \Theta \to \mathbb{R}$: $\boldsymbol{\vartheta} \mapsto U_0(n/k) + H_{\boldsymbol{\vartheta}}^{-1}(1 - np/k)$ at $\boldsymbol{\vartheta}_0$ and therefore Π_n induces a posterior distribution $\Pi_n \circ T_n^{-1}$ on the approximate extreme quantile. In studying its asymptotic properties, we explicitly account for the fact that $U_0(n/k)$ is not estimated within the Bayesian framework but is instead estimated by the frequentist estimator $Y_{n-k,n}$. We refer to the map, the posterior distribution and the extreme quantile as $\widetilde{T}_n(\boldsymbol{\vartheta}) := Y_{n-k,n} + H_{\boldsymbol{\vartheta}}^{-1}(1 - np/k)$, $\widetilde{\Pi}_n := \Pi_n \circ \widetilde{T}_n^{-1}$ and $Q(p) = Y_{n-k,n} + H_{\boldsymbol{\vartheta}}^{\leftarrow}(1 - np/k)$, respectively.

We introduce a Bayesian delta method, which is useful for deriving the asymptotic theory of the posterior distribution of extreme quantiles. To the best of our knowledge, such a Bayesian delta method has not previously appeared in the literature and is of independent interest, as it can be applied beyond the EVT context in other settings.

Theorem 2.9. *Consider a statistical model with parameter $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^m$. Let $\pi(\boldsymbol{\theta})$ be a prior density on $\boldsymbol{\theta}$, which can possibly be data-dependent as described e.g. in (13), and $\Pi_n(\boldsymbol{\theta} \in \cdot)$ be the posterior distribution of $\boldsymbol{\theta}$. Assume the following conditions:*

(a) Let $\boldsymbol{v}_n \in (0, \infty)^m, \boldsymbol{\theta}_n \in \mathbb{R}^m$ be sequences such that $\boldsymbol{v}_n \to \boldsymbol{\infty}$ and $\boldsymbol{\theta}_n \xrightarrow{\mathbb{P}} \boldsymbol{\theta}_0$ and

$$\sup_{B \subset \mathbb{R}^m} |\Pi_n(\{\boldsymbol{\theta} : \boldsymbol{v}_n(\boldsymbol{\theta} - \boldsymbol{\theta}_n) \in B\}) - \mathcal{N}(B; \mathbf{0}, \boldsymbol{D})| = o_{\mathbb{P}}(1),$$

where B represents any measurable subset of $\mathbb{R}^m$ and $\boldsymbol{D}$ is positive definite.

(b) Let $l \leq m$ and $T_n : \Theta \to \mathbb{R}^l$ be a sequence of continuously differentiable maps such that, for a sequence $\boldsymbol{w}_n \in \mathbb{R}^l$ and a $(l \times m)$ matrix $\boldsymbol{J}$ of full rank, we have $\nabla \overline{T}_n(\boldsymbol{\Delta}) \to \boldsymbol{J}$ uniformly on compact sets as $n \to \infty$, where

$$\overline{T}_n(\boldsymbol{\Delta}) = \boldsymbol{w}_n(T_n(\boldsymbol{\theta}_n + \boldsymbol{v}_n^{-1}\boldsymbol{\Delta}) - T_n(\boldsymbol{\theta}_n)), \quad \boldsymbol{\Delta} \in \mathbb{R}^m.$$

Then,

$$\sup_{B \subset \mathbb{R}^l} |\Pi_n(\{\boldsymbol{\theta} : \boldsymbol{w}_n(T_n(\boldsymbol{\theta}) - T_n(\boldsymbol{\theta}_n)) \in B\}) - \mathcal{N}(B; 0, \boldsymbol{J}\boldsymbol{D}\boldsymbol{J}^\top)| = o_{\mathbb{P}}(1),$$

where B represents any measurable subset of $\mathbb{R}^l$.

Finally, the following corollary presents the asymptotic properties of the posterior distribution for extreme quantiles, derived directly from Theorems 2.8 and 2.9. In view of the next result, we introduce the following function (e.g., Ch. 4.3 in [11]) for any $t > 1$ and $\gamma > -1/2$,

$$q_\gamma(t) := \frac{\partial H_\gamma^{\leftarrow}}{\partial \gamma}(1 - 1/t) = \int_1^t v^{\gamma - 1} \log v \mathrm{d}v.$$

Corollary 2.10. *Assume that the conditions of Theorem 2.8 are satisfied and, in the special subcase where $\rho = 0$, further assume that $\gamma_0 < 0$. Then, for $p = o(k/n)$ such that $\log(k/np) = o(\sqrt{k})$:*

- *(Contraction rate) There is a $R > 0$ such that for all sequences $\varepsilon_n \to 0$, satisfying $\sqrt{k}\varepsilon_n \to \infty$ and $\varepsilon_n \log(k/np) \to 0$ as $n \to \infty$,*

$$\widetilde{\Pi}_n\left(\left\{Q(p) \in \mathbb{R} : \left|\frac{Q(p) - F_0^{\leftarrow}(1 - p)}{q_{\gamma_0}(k/(np))a_0(n/k)}\right| > \varepsilon_n\right\}\right) = O_{\mathbb{P}}\left(\exp\left(-Rk\varepsilon_n^2\right)\right).$$

- *(Bernstein–von Mises) For $v_n = \sqrt{k}/(q_{\gamma_0}(k/(np))a_0(n/k))$, as $n \to \infty$*

$$\sup_{B\subset\mathbb{R}} \left| \widetilde{\Pi}_n\left(\left\{Q(p)\in\mathbb{R} : v_n\left(Q(p) - F_0^{\leftarrow}(1-p)\right)\in B\right\}\right) - \mathcal{N}(B; \widetilde{\Delta}_n, \widetilde{V}) \right| = o_{\mathbb{P}}(1),$$

where B is any measurable subset of $\mathbb{R}$, see Section 3.7 and Equation (3.32) of the Supplementary Material [13] for the explicit expression of $\widetilde{\Delta}_n$ and $\widetilde{V}$.
- *(Coverage probability) If $\lambda = 0$, for any fixed $\alpha \in (0,1)$, as $n \to \infty$*

$$\mathbb{P}\left(\left\{F_0^{\leftarrow}(1-p) \in (\widetilde{\Pi}_n^{\leftarrow}(\alpha/2), \widetilde{\Pi}_n^{\leftarrow}(1-\alpha/2))\right\}\right) = 1 - \alpha + o(1)$$

where $\widetilde{\Pi}_n^{\leftarrow}(1-a)$ is the $(1-a)$-quantile of $\widetilde{\Pi}_n$, for $a \in (0,1)$.

Remark 2.11. The restriction $\gamma_0 < 0$, imposed in the special case where Condition 1 holds with $\rho = 0$, helps to control bias when applying formula (15). This assumption is also common in the frequentist framework [e.g., 11, Theorem 4.3.1].

2.4. Asymptotic theory of predictive distribution

The posterior distribution of extreme quantiles is undoubtedly a valuable tool for addressing the challenging task of assessing yet-to-occur extreme events, as it inherently provides a measure of uncertainty. However, at its core, it remains a method for inferring a distributional parameter. A more comprehensive approach to quantifying uncertainty in forecasting future extreme events involves employing a genuine statistical predictive framework.

A practical way to achieve statistical prediction is through the posterior predictive distribution. Given a past sample $\boldsymbol{Y}_n = (Y_1, \dots, Y_n)$, we consider an independent out-of-sample random variable $Y^\star$ as a representative of a future event. In the present setup, $\boldsymbol{Y}^\star$ and $\boldsymbol{Y}_n$ have the same univariate marginal distribution. We then focus on the conditional distribution $F_{0,n}^\star(y) = \mathbb{P}(Y^\star \le y \mid Y^\star > U_0(n/k), \boldsymbol{Y}_n)$, which we refer to as the predictive distribution of an extreme event. While conditioning on $\boldsymbol{Y}_n$ is technically unnecessary—since $Y^\star$ and $\boldsymbol{Y}_n$ are independent—we retain it to emphasize the crucial role of past data in defining its estimator. Following the Bayesian inferential framework outlined in Section 2.3, a natural estimator of the predictive distribution of an extreme event is given by the posterior predictive distribution

$$\widehat{F}_n^\star(y) = \int_\Theta H_{\boldsymbol{\vartheta}}(y - Y_{n-k,n})\Pi_n(\mathrm{d}\boldsymbol{\theta}). \tag{16}$$

The posterior predictive distribution can serve as a powerful tool for forecasting future extreme peaks. For instance, one can obtain a point forecast by computing the quantile $\widehat{F}_n^{\star\leftarrow}(1-p^\star)$ for a small $p^\star \in (0,1)$, or derive a more comprehensive prediction by identifying an entire region of plausible future values based on the highest posterior predictive density (e.g., [34]). Given the significant societal impact of extreme events, it is crucial to assess the accuracy of the proposed forecasting method. The next result establishes that forecasts derived from our posterior predictive distribution are asymptotically reliable, as the latter approaches the true predictive distribution as the sample size grows. To quantify the closeness between two distributions F and G, we use the Wasserstein distance of order v for $v \ge 1$, i.e. $W_v(F,G) = (\int_0^1 |F^{\leftarrow}(p) - G^{\leftarrow}(p)|^v \mathrm{d}p)^{1/v}$. Moreover, we recall that by the scaling property of the Wasserstein distance, for $a_0(n/k) > 0$, we have

$$W_v(\widehat{F}_n^\star, F_{0,n}^\star) = a_0(n/k) W_v(\widehat{F}_n^\star(a_0(n/k)\cdot), F_{0,n}^\star(a_0(n/k)\cdot)).$$

Theorem 2.12. *Assume that the conditions of Theorem 2.8 are satisfied. Assume also that Condition 2(b) changes as: π_{sh} is positive and continuous at γ_0 and there is $B \subseteq (-1/2, 1/\nu)$ such that $\pi_{sh}(\gamma) = 0$, for all $\gamma \notin B$, and*

$$\int_{B\cap(-1/2,0)} \pi_{sh}(\gamma)\mathrm{d}\gamma < \infty, \quad \int_B (1-\gamma\nu)^{-1/\nu}\pi_{sh}(\gamma)\mathrm{d}\gamma < \infty.$$

Then, for all sequences $\varepsilon_n \to 0$, satisfying $k\varepsilon_n^2/\log(k) \to \infty$ as $n \to \infty$, we have

$$\frac{W_\nu(\widehat{F}_n^\star, F_{0,n}^\star)}{a_0(n/k)} = O_{\mathbb{P}}(\varepsilon_n).$$

Remark 2.13. The alternative condition included in Theorem 2.12 is a requirement to work with the Wasserstein metric of order ν. Specifically, ensuring the integrability of the ν-moment of the GP distribution with respect to the prior density π_{sh} is necessary for the theory to hold. This condition is relatively weak and is naturally satisfied by several common prior distributions. Examples include: A Uniform prior on $(-1/2, 1/\nu - \varepsilon)$ for an arbitrary small $\varepsilon \geq 0$; Beta prior on the transformed parameter $(1-\gamma\nu)/(1+\nu/2)$, where the Beta shape parameters are $(\alpha + 1/\nu, \beta)$, for $\alpha, \beta > 0$.

3. Extreme regression

In this section, we introduce a Bayesian framework for inference and prediction of extremes of a response variable that is linked to some covariates that we model through the *proportional tail model* for conditional extremes. The tail proportionality condition has already been adopted in extreme value analysis in various forms to address a range of statistical problems. It was first introduced by [16] by means of a triangular arrays framework for modeling heteroscedastic heavy-tailed variables, motivated by the analysis of non-stationary extremes over time, and was later applied to trend detection in heteroscedastic settings by [28]. Although initially formulated for independent observations, extensions to serial dependence were subsequently developed by [5]. More recently, this framework has been further extended to conditional extremes and extreme quantile regression (see, e.g., [1, 4, 7]), as well as to multivariate settings (e.g., [14]).

A key assumption underlying our proportional tail model is that the upper tail of the conditional distribution of the response, given the covariates, changes according to a scaling factor, while the EVI remains unchanged. This framework offers a tractable approximation to the tail behavior of a wide class of conditional models (see Section 3.1); moreover, its simplicity facilitates inference (Section 3.2) and supports its use in applied domains such as climate science. In this context, it is reasonable to assume that the tail behavior of a response variable—such as temperature or wind speed—may change over time or with respect to covariates as a consequence of climate change. At the same time, it is also plausible that the tail heaviness, as measured by γ_0, remains unchanged due to physical constraints: temperatures and wind speeds cannot increase indefinitely. Models with a finite upper end-point or light-tailed ($\gamma_0 \leq 0$) are appropriate, and a transition to models with an infinite upper end-point ($\gamma_0 \geq 0$) would be unrealistic. Nevertheless, the validity of this modeling assumption in a given application can be readily evaluated by means of appropriate statistical tests, including those proposed by [16] and the test introduced at the end of Section 3.2. Precisely, our method leverages peaks above a high threshold, along with the corresponding concomitant covariates, to estimate the marginal tail probability parametrically and the effect of covariates non-parametrically. This foundation enables semi-parametric inference on extreme conditional quantiles and facilitates forecasting of conditional extremes. To the best of our knowledge, this is the first approach that jointly models and estimates both the marginal extremal properties of the response and the dependence structure induced by covariates.

3.1. Proportional tail model

Let $(Y, \boldsymbol{X})$ be a random vector on $\mathbb{R} \times [0,1]^d$. We denote the marginal distribution of Y by F_0, the marginal law of $\boldsymbol{X}$ by $\mathcal{P}_0$, the conditional distribution of Y given that $\boldsymbol{X} = \boldsymbol{x}$ by $F_{\boldsymbol{x}}^{(0)}$ and the corresponding $(1-p)$-quantile by $F_{\boldsymbol{x}}^{(0)\leftarrow}(1-p)$. We assume that F_0 is absolutely continuous and that satisfies $F_0 \in \mathcal{D}(G_{\gamma_0})$ and the conditional distribution $F_{\boldsymbol{x}}^{(0)}$ satisfies the *tail proportionality* condition, i.e. there is a positive bounded function c_0 on $[0,1]^d$, named *scedasis function* [16], such that

$$\lim_{y \to y_0^*} \frac{1 - F_{\boldsymbol{x}}^{(0)}(y)}{1 - F_0(y)} = c_0(\boldsymbol{x}), \quad \boldsymbol{x} \in [0,1]^d, \tag{17}$$

where, we recall, y_0^* is the right end-point of F_0. Note that Condition (17) together with the first-order condition in (3) implies

$$\lim_{t \to \infty} \frac{U_{\boldsymbol{x}}^{(0)}(ty) - U_{\boldsymbol{x}}^{(0)}(t)}{a_0(t)} = (c_0(\boldsymbol{x}))^{\gamma_0} \frac{y^{\gamma_0} - 1}{\gamma_0}, \quad \forall\, y > 0,$$

where $U_{\boldsymbol{x}}^{(0)}(t) = F_{\boldsymbol{x}}^{(0)\leftarrow}(1 - 1/t)$. This means that the conditional distributions $F_{\boldsymbol{x}}^{(0)}$ changes according to the scaling function $(c_0(\boldsymbol{x}))^{\gamma_0}$, while the heaviness of its tail remains unchanged, since its tail index is equal to γ_0 no matter what is $\boldsymbol{x}$. Let $\varsigma : [0,1]^d \mapsto (0,\infty)$ and $\mu : [0,1]^d \mapsto \mathbb{R}$ be uniformly bounded away from 0 and ∞, and from $-\infty$ and ∞, respectively. As illustrative examples: (i) if $F_{\boldsymbol{x}}^{(0)}(y) = t_{\nu_0}(y/\varsigma(\boldsymbol{x}))$, where t_{ν_0} is a Student-t distribution with ν_0 degrees of freedom and scale parameter $\varsigma(\boldsymbol{x})$, then $F_0(y) = \int t_{\nu_0}(y/\varsigma(\boldsymbol{x}))\mathcal{P}_0(\mathrm{d}\boldsymbol{x})$ satisfies $F_0 \in \mathcal{D}(G_{1/\nu_0})$ by [32, Proposition 0.5] and (17) with $c_0(\boldsymbol{x}) = (\varsigma(\boldsymbol{x}))^{\nu_0} / \int (\varsigma(\boldsymbol{x}))^{\nu_0}\mathcal{P}_0(\mathrm{d}\boldsymbol{x})$; (ii) if $F_{\boldsymbol{x}}^{(0)}(y) = 1/(1 + \exp(-(y - \mu(\boldsymbol{x}))))$, i.e. a logistic distribution with location parameter $\mu(\boldsymbol{x})$, then by basic calculations we have $F_0 \in \mathcal{D}(G_0)$ and (17) is satisfied with $c_0(\boldsymbol{x}) = \exp(\mu(\boldsymbol{x}))/\int \exp(\mu(\boldsymbol{x}))\mathcal{P}_0(\mathrm{d}\boldsymbol{x})$; (iii) if $F_{\boldsymbol{x}}^{(0)}(y) = \exp(-(-y/\varsigma(\boldsymbol{x}))^{\alpha_0})$, i.e. reverse-Weibull distribution with scale parameter $\varsigma(\boldsymbol{x})$, then $F_0 \in \mathcal{D}(G_{\gamma_0})$, with $\gamma_0 = -1/\alpha_0$, and satisfies (17) with $c_0(\boldsymbol{x}) = (\varsigma(\boldsymbol{x}))^{-\alpha_0} / \int (\varsigma(\boldsymbol{x}))^{-\alpha_0}\mathcal{P}_0(\mathrm{d}\boldsymbol{x})$. The proportional tail assumption is not restrictive: under this assumption, the tail behavior of $F_{\boldsymbol{x}}^{(0)}$ is asymptotically influenced by the covariates only through the scaling function c_0, which can be derived for location–scale families as those above but also for more general models.

Under this framework and assuming convergence in (17) to be uniform, we obtain the following asymptotic approximations. First, according to the univariate case, for all $z > 0$ and with $y = U_0(n/k) + z$ we have $\mathbb{P}(Y \le y \mid Y > U_0(n/k)) \approx H_{\boldsymbol{\vartheta}_0}(y - U_0(n/k))$ as $n \to \infty$. By leveraging this and (17) we obtain that for all measurable $B \subset [0,1]^d$

$$\begin{aligned}
\mathbb{P}(Y > y, \boldsymbol{X} \in B) &= \int_B (1 - F_{\boldsymbol{x}}^{(0)}(y))\mathcal{P}_0(\mathrm{d}\boldsymbol{x}) \\
&\approx \int_B c_0(\boldsymbol{x})(1 - F_0(y))\mathcal{P}_0(\mathrm{d}\boldsymbol{x}) \\
&\approx \frac{k}{n}\mathcal{P}_0^*(B)(1 - H_{\boldsymbol{\vartheta}_0}(y - U_0(n/k))),
\end{aligned} \tag{18}$$

where $\mathcal{P}_0^*(\mathrm{d}\boldsymbol{x}) := c_0(\boldsymbol{x})\mathcal{P}_0(\mathrm{d}\boldsymbol{x})$ and the approximations in the last two lines hold for $n \to \infty$. The above result entails that the conditional distribution of $\boldsymbol{X}$ given that $Y > U_0(n/k)$ is asymptotically approximated by the probability measure $\mathcal{P}_0^*$, as $n \to \infty$ (see Lemma 3.5 in [13]). Second, as a direct consequence, the conditional tail probability $\mathbb{P}(Y > y \mid \boldsymbol{X} \in B)$ can be approximated in turn by the right-hand side of (18) divided by $\mathcal{P}_0(B)$, as $n \to \infty$. As a result one obtains for the conditional distribution $\mathbb{P}(Y \leq y \mid \boldsymbol{X} \in B) = 1 - \mathbb{P}(Y > y \mid \boldsymbol{X} \in B)$ a useful approximation that we refer to as the conditional proportional tail model. Third, the result

$$\begin{aligned} &\mathbb{P}(\, Y \leq y, \boldsymbol{X} \in B \mid Y > U_0(n/k)) \\ &\approx \mathcal{P}_0^*(B) - \mathcal{P}_0^*(B)(1 - H_{\boldsymbol{\vartheta}_0}(y - U_0(n/k))) = \mathcal{P}_0^*(B) H_{\boldsymbol{\vartheta}_0}(y - U_0(n/k)), \end{aligned} \tag{19}$$

suggests that the joint conditional distribution of $(Y, \boldsymbol{X})$ given that $Y > U_0(n/k)$ factorises asymptotically as a product of its marginal distributions, which is useful in the next section to derive a Bayesian procedure for the inference on the parameters $(\boldsymbol{\vartheta}_0, \mathcal{P}_0^*)$, which in turn allows inference about $\mathbb{P}(Y \leq y \mid \boldsymbol{X} \in B)$, through the conditional proportional tail model. In the next section we study functional estimation of $\mathcal{P}_0^*(B)$, namely for an infinite collection of sets B, as it allows to perform hypothesis testing to verify the effect of the covariates $\boldsymbol{X}$ on Y, given that $Y > U_0(n/k)$, and both pointwise and functional estimation of its density $\mathrm{d}\mathcal{P}_0^*(\boldsymbol{x})/\mathrm{d}\mathcal{P}_0(\boldsymbol{x})$. Note that $\mathcal{P}_0^*(B)/\mathcal{P}_0(B) \approx c_0(\boldsymbol{x})$ when $B \downarrow \{\boldsymbol{x}\}$, and the conditional proportional tail model approximation of conditional distribution becomes

$$\mathbb{P}(Y \leq y | \boldsymbol{X} = \boldsymbol{x}) \approx 1 - c_0(\boldsymbol{x}) \frac{k}{n} \left(1 + \gamma_0 \frac{y - U_0(n/k)}{a_0(n/k)}\right)_+^{-1/\gamma_0}, \tag{20}$$

as $n \to \infty$. In this context, the aim is to estimate the scedasis function c_0 and more importantly for applications, the extreme quantiles of the conditional distribution. Assuming that $p = o(k/n)$ as $n \to \infty$, then exploiting the right-hand side of formula (20) one obtains, for any $\boldsymbol{x} \in [0,1]^d$, the following approximation for the $(1-p)$-quantile $F_{\boldsymbol{x}}^{(0)\leftarrow}(1-p)$ of the conditional distribution,

$$F_{\boldsymbol{x}}^{(0)\leftarrow}(1-p) \approx U_0(n/k) + H_{\boldsymbol{\vartheta}_0}^{\leftarrow}\left(1 - \frac{np}{k} \frac{1}{c_0(\boldsymbol{x})}\right), \tag{21}$$

as $n \to \infty$. Therefore, inference on $F_{\boldsymbol{x}}^{(0)\leftarrow}(1-p)$ can be achieved leveraging that on $(c_0, \boldsymbol{\vartheta}_0)$.

3.2. Asymptotic theory of the posterior distribution

Let $(Y_i, \boldsymbol{X}_i)_{1 \leq i \leq n}$ be a sample of i.i.d. copies of $(Y, \boldsymbol{X})$. The Bayesian inference for the proportional tail model and related quantities is grounded on the joint statistical model $\{\mathcal{H}_{\boldsymbol{\vartheta}}^k \times \mathcal{P}^{*k}, \boldsymbol{\vartheta} \in \Theta, \mathcal{P}^* \in \mathscr{P}\}$, which is motivated by the approximation (19). In particular, $\mathcal{H}_{\boldsymbol{\vartheta}}^k$ and $\mathcal{P}^{*k}$ are the probability measures of k independent GP variables and concomitant covariates. Moreover, $\mathscr{P}$ is the family of Borel probability measures on $[0,1]^d$. The model is fitted to the subsample $(Y_{n-i+1,n} - Y_{n-k,n}, \boldsymbol{X}_{n-i+1,n})_{1 \leq i \leq k}$ of peaks $(Y_{n-i+1,n} - Y_{n-k,n})_{1 \leq i \leq k}$ above a high threshold $Y_{n-k,n}$ and concomitant covariates $(\boldsymbol{X}_{n-i+1,n})_{1 \leq i \leq k}$, with $\boldsymbol{X}_{1,n}, \ldots, \boldsymbol{X}_{n,n}$ that are the covariates associated to the order statistics $Y_{1,n} < \cdots < Y_{n,n}$. Note that the continuity of the distribution F_0 ensures that there are almost surely no ties. In this way there are several sources of misspecification: the exceedances are dependent and only approximately distributed according to the Pareto distribution $H_{\boldsymbol{\vartheta}_0}$, the corresponding concomitant covariates are only approximately distributed according to the law $\mathcal{P}_0^*$, exceedances and concomitant covariates are dependent and only approximately independent from each other. Despite that, we can show that the posterior distribution of the proportional tail model parameters and the conditional extreme quantiles enjoy good asymptotic properties.

We specify the prior distribution for the proportional tail model parameters as $\Lambda(\mathrm{d}\boldsymbol{\vartheta}, \mathrm{d}\mathcal{P}^*) = \Pi(\mathrm{d}\boldsymbol{\vartheta})\Phi(\mathrm{d}\mathcal{P}^*)$, where the prior distribution $\Pi(\mathrm{d}\boldsymbol{\vartheta})$ on the GP parameters is defined as in formula (13) and the prior distribution on the law $\mathcal{P}^*$ is defined as $\Phi(\mathrm{d}\mathcal{P}^*) = \mathrm{DP}(\mathrm{d}\mathcal{P}^*; \tau)$, namely a Dirichlet process, where τ is a finite positive measure on Borel sets $B \subset [0,1]^d$ [see e.g. 22, Ch 4.1] which we hereafter assume absolutely continuous. According to the approximate joint model in (19) we have that the posterior distribution for the parameters $(\boldsymbol{\vartheta}, \mathcal{P}^*)$ is for all measurable sets $(B \times C) \subset \Theta \times \mathscr{P}$ given by

$$\Lambda_n(B \times C) = \Pi_n(B)\Phi_n(C).$$

Since the approximate statistical model arising from (19) postulates independence among the exceedances and the concomitant covariates and given the independence between the prior distributions, then the posterior distribution Λ_n splits into the product between the posterior distribution Π_n of $\boldsymbol{\vartheta}$, given in (14), and the posterior distribution of $\mathcal{P}^*$, which due to conjugacy property of the DP prior (see Ch 4.6 in [22]) becomes $\Phi_n(C) = \mathrm{DP}(C; \tau + k\mathbb{P}_n^*)$, that is a Dirichlet process with parameter $\tau + k\mathbb{P}_n^*$, where $\mathbb{P}_n^*(\cdot) = k^{-1}\sum_{i=1}^k \mathbb{1}(\boldsymbol{X}_{n-i+1,n} \in \cdot)$ is the empirical measure associated to the covariates concomitant to peaks. For this reason we can initially handle the two posterior distributions separately. The inference about $\boldsymbol{\vartheta}$ via Π_n is fully described in Section 2. Then, we are left here to describe the inference on $\mathcal{P}^*$ via Φ_n, discuss its asymptotic properties and, most importantly, determine the resulting theory for inference on the extreme quantiles of the conditional distribution. This is done by explicitly accounting for the fact that Π_n and Φ_n are dependent random measures which, however, become increasingly close to Gaussian measures with asymptotically independent random means and deterministic variance as $n \to \infty$ (see Corollary 3.18, Remark 3.19 in the Supplementary Material [13]).

The asymptotic theory from the available Bayesian non-parametric literature (see Ch. 6–12 in [22]) cannot be directly applied to the posterior distribution Φ_n, although it is a standard DP, as $\mathcal{P}^{*k}$ is a misspecified model for the concomitant covariates associated to the peaks. We establish here the asymptotic theory of Φ_n under misspecification, provided that some weak conditions are satisfied. Estimation results rely on the control of the convergence speed of the first-order condition in (17) through the following second-order condition.

Condition 3. *There is a nonincreasing $A_1(t)$, such that $A_1(t) \downarrow 0$ as $t \to \infty$ and*

$$\sup_{\boldsymbol{x}\in[0,1]^d} \left| \frac{1 - F_{\boldsymbol{x}}^{(0)}(y)}{1 - F_0(y)} - c_0(\boldsymbol{x}) \right| = O\left(A_1\left(\frac{1}{1 - F_0(y)} \right) \right), \quad y \to y_0^*.$$

In the proportional tail model examples presented at the beginning of Section 3.1, Conditions 1(a) and 3 are satisfied with rate functions A and A_1 such that $A(t) \asymp A_1(t) \asymp t^\rho$ as $t \to \infty$, where $\rho = -2/\nu_0$, $\rho = -1$ and $\rho = -1$ for the models at points (i)–(iii), respectively. Condition 3 is required to establish the high-accuracy asymptotic properties of the posterior distribution of the scedasis function stated in Theorem 3.1. When this condition is not satisfied, the proposed methodology may still be used as a practical inferential tool. However, its asymptotic properties are no longer guaranteed by our theoretical results.

A crucial step is the estimation of the scedasis function $c_0(\boldsymbol{x})$. For this purpose, we use the fact that whenever c_0 is positive and continuous we have

$$c_0(\boldsymbol{x}) = \lim_{n\to\infty} \frac{\mathcal{P}_0^*(B_n)}{\mathcal{P}_0(B_n)},$$

where B_n is a sequence of sets containing $\boldsymbol{x}$ and with a decreasing volume. Accordingly, if $\mathcal{P}_0(B_n)$ was known, the Dirichlet Process prior on $\mathcal{P}^*$ would induce a prior on $\mathcal{P}^*(B_n)/\mathcal{P}_0(B_n)$ and its posterior could be used to infer the scedasis function at $\boldsymbol{x}$. However, $\mathcal{P}_0(B_n)$ is unknown in practice. In the sequel we regard it as a nuisance parameter and assess it by the estimator $\widehat{p}_n \equiv \widehat{p}_n(\boldsymbol{x}) = n^{-1}\sum_{i=1}^n \mathbb{1}(\boldsymbol{X}_i \in B_n)$. We obtain then a data-dependent prior on $c(\boldsymbol{x}) := \mathcal{P}^*(B_n)/\widehat{p}_n$, for a given $\boldsymbol{x}$, and we establish the asymptotic properties of its posterior distribution Ψ_n to guarantee a high accuracy of Bayesian inference on $c_0(\boldsymbol{x})$.

We first focus on the situation where $B_n = B(\boldsymbol{x}, r_n) = \{\boldsymbol{y} \in [0,1]^d : \|\boldsymbol{y} - \boldsymbol{x}\| \le r_n\}$ is the ball of center $\boldsymbol{x}$ and radius r_n and we consider two possible ways of selecting the radius r_n: the deterministic one where ball volume is the same for all $\boldsymbol{x} \in [0,1]^d$, we call the resulting estimation procedure *kernel-based method with bandwidth* $bw = r_n$ of order $(K/n)^{1/d}$, for some positive sequence $K \equiv K(n) \ll n$; the data-dependent one where the ball volume changes depending on $\boldsymbol{x}$ in order to estimate $c(\boldsymbol{x})$ with a prespecified number $K \equiv K(n)$ of surrounding points, we call the resulting estimation procedure *K-nearest neighbours* (KNN) based method. Their definition is made precise in the next result.

Theorem 3.1. *Assume $\mathcal{P}_0$ is absolutely continuous with a density p_0 which is positive and locally Lipschitz continuous on $(0,1)^d$. Let c_0 be a locally Lipschitz continuous function. Let $B_n = B(\boldsymbol{x}, r_n)$, where $\boldsymbol{x} \in (0,1)^d$ and*

$$r_n = R\left(\frac{K}{n}\right)^{1/d}\left(1 + o\left(\sqrt{\frac{n}{Kk}}\right)\right) \text{ or } r_n = \min\left\{h > 0 \colon \sum_{i=1}^n \mathbb{1}(X_i \in B(\boldsymbol{x}, h)) \ge K\right\},$$

with $R > 0$. Assume $k \to \infty$, $K \to \infty$, $k = o(n)$, $K = o(n)$, $n = o(kK)$ and $(K/n)^{1/d}(kK/n)^{1/2} = o(1)$. Assume also that Condition 3 is satisfied and $kA_1(n/2k) \to 0$ as $n \to \infty$.

- *(Contraction rate) There is a $R' > 0$ such that for all sequences ε_n, satisfying $\varepsilon_n \to 0$ and $(kK/n)^{1/2}\varepsilon_n \to \infty$ as $n \to \infty$, we have that*

$$\Psi_n\left(\{c(\boldsymbol{x}) : |c(\boldsymbol{x}) - c_0(\boldsymbol{x})| > \varepsilon_n\}\right) = O_{\mathbb{P}}\left(e^{-R'(kK/n)\varepsilon_n^2}\right).$$

- *(Bernstein-von Mises) We have*

$$\sup_{B \subset \mathbb{R}} \left|\Psi_n\left(\{c(\boldsymbol{x}) : v_n\left(c(\boldsymbol{x}) - c_0(x)\right) \in B\}\right) - \mathcal{N}(B; \Delta_n, c_0(\boldsymbol{x}))\right| = o_{\mathbb{P}}(1),$$

  where B is any Borel subset of $\mathbb{R}$, $v_n = \sqrt{R'' kK/n}$, $\Delta_n = v_n(\widehat{p}_n^*/\widehat{p}_n - c_0(\boldsymbol{x})) \xrightarrow{d} \mathcal{N}(0, c_0(\boldsymbol{x}))$, $\widehat{p}_n^* = k^{-1}\sum_{i=1}^k \mathbb{1}(\boldsymbol{X}_{n-k+i,n} \in B_n)$ and $R'' = R^d \pi^{d/2} p_0(\boldsymbol{x})/\Gamma(1 + d/2)$ and $R'' = 1$ with the above left-hand and right-hand side definition of r_n, respectively.
- *(Coverage probability) For any $\alpha \in (0,1)$ we have*

$$\mathbb{P}\left(\left\{c_0(\boldsymbol{x}) \in (\Psi_n^{\leftarrow}(\alpha/2), \Psi_n^{\leftarrow}(1-\alpha/2))\right\}\right) = 1 - \alpha + o(1),$$

  where, for $a \in (0,1)$, $\Psi_n^{\leftarrow}(1-a)$ is the $(1-a)$-quantile of Ψ_n.

Similarly to the unconditional case, assessing the risk associated to extreme events of a certain phenomenon that takes place when other concomitant dynamics reach certain levels is of primary interest in applications. This task can be achieved computing extreme conditional quantiles, namely the quantiles of $F_{\boldsymbol{x}}^{(0)}$ corresponding to a small exceedance probability $p = o(k/n)$, which can be approximated for large n by the right-hand side of formula (21). That expression, viewed as a function of $\boldsymbol{\vartheta}_0$ and $c_0(\boldsymbol{x})$, is a continuous map $T_n : \Theta \times \mathbb{R} \to \mathbb{R}$: $(\boldsymbol{\vartheta}, c(\boldsymbol{x})) \mapsto U_0(n/k) + H_{\boldsymbol{\vartheta}}^{\leftarrow}(1 - np/(kc(\boldsymbol{x})))$, for any given n, k, p and $\boldsymbol{x}$, and therefore Λ_n induces a posterior distribution $\Lambda_n \circ T_n^{-1}$ on the approximate extreme conditional quantile. Since our procedure is based on the frequentist estimation of $U_0(n/k)$ via $Y_{n-k,n}$, next we are going to denote the map, the posterior distribution and the extreme conditional quantile as $\widetilde{T}_n(\boldsymbol{\vartheta}, c(\boldsymbol{x})) := Y_{n-k,n} + H_{\boldsymbol{\vartheta}}^{\leftarrow}(1 - np/(kc(\boldsymbol{x})))$, $\widetilde{\Lambda}_n := \Lambda_n \circ \widetilde{T}_n^{-1}$ and

$$Q_{\boldsymbol{x}}(p) = Y_{n-k,n} + H_{\boldsymbol{\vartheta}}^{\leftarrow}(1 - np/(kc(\boldsymbol{x}))). \tag{22}$$

Corollary 3.2. *Assume that conditions of Theorems 2.8 and 3.1 are satisfied and, in the special subcase where $\rho = 0$, further assume that $\gamma_0 < 0$. Let $p = o(k/n)$ be such that $\log(k/np) = o(\sqrt{k})$ and*

$$q_{\gamma_0}(k/np)(k/np)^{-\gamma_0}\sqrt{K/n} \to \omega \in [0, \infty], \quad n \to \infty.$$

If $\omega = \infty$, define v_n as in Corollary 2.10, otherwise set

$$v_n = (c_0(\boldsymbol{x}))^{1-\gamma_0} \frac{\sqrt{R'' kK/n}}{(k/np)^{\gamma_0} a_0(n/k)}.$$

Then, for all $\boldsymbol{x} \in (0,1)^d$:

- *(Contraction rate) There is a $R''' > 0$ such that, for all sequences $\varepsilon_n \to 0$ satisfying $\sqrt{kK/n}\varepsilon_n \to \infty$ and $\sqrt{K/n}\varepsilon_n \log(k/np) \to 0$ as $n \to \infty$, we have*

$$\widetilde{\Lambda}_n\left(\left\{Q_{\boldsymbol{x}}(p) \in \mathbb{R} : \left|\frac{Q_{\boldsymbol{x}}(p) - F_{\boldsymbol{x}}^{(0)\leftarrow}(1-p)}{a_0(n/k) q_{\gamma_0}(c_0(\boldsymbol{x})k/(np))}\right| > \widetilde{\varepsilon}_n\right\}\right) = O_{\mathbb{P}}\left(\exp\left(-R'''(kK/n)\varepsilon_n^2\right)\right)$$

  where $\widetilde{\varepsilon}_n = \varepsilon_n (np/k)^{-\gamma_0} / q_{\gamma_0}(c_0(\boldsymbol{x})k/(np))$ if $\omega < \infty$, while $\widetilde{\varepsilon}_n = \varepsilon_n \sqrt{K/n}$ if $= \infty$.
- *(Bernstein-von Mises) We have*

$$\sup_{B \subset \mathbb{R}} \left|\widetilde{\Lambda}_n\left(\left\{Q_{\boldsymbol{x}}(p) \in \mathbb{R} : v_n\left(Q_{\boldsymbol{x}}(p) - F_{\boldsymbol{x}}^{(0)\leftarrow}(1-p)\right) \in B\right\}\right) - \mathcal{N}(B; \widetilde{\Xi}_n, \widetilde{\Omega})\right| = o_{\mathbb{P}}(1),$$

  where B is any Borel subset of $\mathbb{R}$, see Section 3.8 and (3.39) of [13] for $\widetilde{\Xi}_n$ and $\widetilde{\Omega}$.
- *(Coverage probability) For any $\alpha \in (0,1)$, if $\lambda = 0$ we have*

$$\mathbb{P}\left(\left\{F_{\boldsymbol{x}}^{(0)\leftarrow}(1-p) \in (\widetilde{\Lambda}_n^{\leftarrow}(\alpha/2), \widetilde{\Lambda}_n^{\leftarrow}(1-\alpha/2))\right\}\right) = 1 - \alpha + o(1),$$

  where, for $a \in (0,1)$, $\widetilde{\Lambda}_n^{\leftarrow}(1-a)$ is the $(1-\alpha)$-quantile of $\widetilde{\Lambda}_n$.

We conclude this section by discussing the more general problem of functional estimation. Let $\mathcal{G}_n \subset (0,1)^d$ be a grid of $N \equiv N(n) \to \infty$ points such that $(B(\boldsymbol{x}, r_n), \boldsymbol{x} \in \mathcal{G}_n)$ covers $[0,1]^d$, where r_n is deterministic with $R > 0$. Define the piecewise constant function $c(\boldsymbol{x}') = \mathcal{P}^*(B(\boldsymbol{x}, r_n))/\widehat{p}_n(\boldsymbol{x})$ for $\boldsymbol{x} = \operatorname{argmin}_{\boldsymbol{x}'' \in \mathcal{G}_n} \|\boldsymbol{x}' - \boldsymbol{x}''\|_\infty$, with the convention $c_0(\boldsymbol{x}') = 0$ if $\widehat{p}_n(\boldsymbol{x}) = 0$. For simplicity, we continue to denote its posterior distribution by Ψ_n.

Theorem 3.3. *Work under Theorem 3.1 conditions, with deterministic radius r_n and Lipschitz continuous functions p_0, c_0 on $[0,1]^d$. Set $\delta_0 = \inf_{\boldsymbol{x}' \in [0,1]^d} p_0(\boldsymbol{x}')$. Then, there is a $R'''' > 0$ such that, for* *all positive* $\delta_n = \delta_0 + o(1)$, $\varepsilon_n = o(1)$ *satisfying* $r_n = o(\delta_n)$, $\varepsilon_n \sqrt{\delta_n k K/(n \log N)} \to \infty$ *as* $n \to \infty$*, the posterior satisfies*

$$\Psi_n\left(\{c : \|c - c_0\|_{\delta_n} > \varepsilon_n\}\right) = O_{\mathbb{P}}\left(\exp\left(-R'''' \delta_n (kK/n)\varepsilon_n^2\right)\right)$$

where $\|c - c_0\|_{\delta_n} = \sup_{\boldsymbol{x}' : p_0(\boldsymbol{x}') > \delta_n} |c(\boldsymbol{x}') - c_0(\boldsymbol{x}')|$.

Theorem 3.3 provides uniform contraction rates of the posterior distribution of $c(\boldsymbol{x})$ over a collection of covariate values $\boldsymbol{x}$, for which the density $p_0(\boldsymbol{x})$ is positive, and is therefore a stronger result than Theorem 3.1. Whenever p_0 is bounded away from 0, our findings cover the L^∞-functional contraction rates, ensuring high accuracy of Ψ_n-based inference on the scedasis function. Despite model misspecification for given n and k, an appropriate choice of K allows one to attain uniform convergence rates of order $(k/\log k)^{1/(d+2)}$, which are known to be optimal for the estimation of Lipschitz functions in the basic setting where the model is correctly specified and the k observations are i.i.d. [see, e.g., 6, 27]. This result provides practical guidance for selecting K, as illustrated in Section 6.2 of [13].

We finally discuss inference on the distribution function. For any $\boldsymbol{x} \in [0,1]^d$, let $P(\boldsymbol{x})$ and $P^*(\boldsymbol{x})$ be the cumulative distribution functions of $\mathcal{P}$ and $\mathcal{P}^*$, respectively, and $P_0(\boldsymbol{x})$ and $P_0^*(\boldsymbol{x})$ their true counterparts. We accordingly set $\mathbb{P}_n^\circ(\boldsymbol{x}) = n^{-1}\sum_{i=1}^n \mathbb{1}(\boldsymbol{X}_i \le \boldsymbol{x})$ and $\mathbb{P}_n^*(\boldsymbol{x}) = k^{-1}\sum_{i=1}^k \mathbb{1}(\boldsymbol{X}_{n-k+i,n} \le \boldsymbol{x})$. Finally, let $\mathcal{B}(\boldsymbol{x})$ be a P_0^*-Brownian bridge, i.e. a zero-mean Gaussian process with covariance

$$\mathbb{E}[\mathcal{B}(\boldsymbol{x}_1)\mathcal{B}(\boldsymbol{x}_2)] = P_0^*(\min(\boldsymbol{x}_1, \boldsymbol{x}_2)) - P_0^*(\boldsymbol{x}_1)P_0^*(\boldsymbol{x}_2), \quad \boldsymbol{x}_1, \boldsymbol{x}_2 \in [0,1]^d$$

and $\widetilde{\Phi}$ be its probability law. The following result establishes that the posterior distribution $\widetilde{\Phi}_n$ on P^* converges to $\widetilde{\Phi}$ in the functional sense and is therefore also asymptotically Gaussian. Note that $\widetilde{\Phi}_n$ and $\widetilde{\Phi}$ are Borel probability measures on the complete, separable Skorohod space $D([0,1]^d, d_0)$, defined in [29]. We denote with $\nu(\cdot;\cdot)$ the Lévy-Prohorov metric [22, p. 488], metrizing weak convergence over separable spaces.

Theorem 3.4. *(Bernstein-von Mises) Work under Condition 3. Let* $\mathcal{P}_0$*,* τ *be absolutely continuous, with* τ *having positive density over* $[0,1]^d$*. Then, if* $kA_1(n/2k) = o(1)$ *and* $k = o(n)$ *as* $n \to \infty$*, it holds that*

$$\nu\left(\widetilde{\Phi}_n\left(\{P^* : \sqrt{k}(P^* - \mathbb{P}_n^*) \in \cdot\,\}\right); \widetilde{\Phi}\right) = o_{\mathbb{P}}(1).$$

As a consequence, as $n \to \infty$

$$\nu\left(\widetilde{\Phi}_n\left(\{P^* : \|\sqrt{k}(P^* - \mathbb{P}_n^*)\|_\infty \in \cdot\,\}\right); \widetilde{\Phi}\left(\{\widetilde{P} : \|\widetilde{P}\|_\infty \in \cdot\,\}\right)\right) = o_{\mathbb{P}}(1).$$

One practical application of Theorem 3.4 is to derive a test statistic for assessing whether the concomitant covariates have a significant effect on the extremes of the response variable. When the covariates have no effect on the extremes of the response variable we have that $c_0(\boldsymbol{x}) = 1$ for all $\boldsymbol{x} \in [0,1]^d$. On this basis we consider then the system of hypotheses

$$\mathscr{H}_0 : P_0^* = P_0 \quad \text{versus} \quad \mathscr{H}_1 : P_0^* \neq P_0,$$

where again P_0 and P_0^* are the cumulative distribution functions of $\mathcal{P}_0$ and $\mathcal{P}_0^*$, respectively. For testing the validity of $\mathscr{H}_0$ we consider a Kolmogorov-Smirnov-type test as in [16] and we rely on the Bernstein-von Mises result in Theorem 3.4 to draw many samples from the posterior distribution and estimate the corresponding critical value. Specifically, in order to perform our hypothesis testing we consider the following scheme:

1. Compute the test statistic $\mathbb{S} = \sqrt{k}\|\mathbb{P}_n^* - \mathbb{P}_n^\circ\|_\infty$;
2. Draw independent samples $(\mathcal{P}_m^*)_{1\leq m\leq M}$ from the posterior distribution $\mathrm{DP}(\tau + k\mathbb{P}_n^*)$, for a large value M and then compute the corresponding distributions $(P_m^*)_{1\leq m\leq M}$ and statistics

$$\mathbb{S}_m = \sqrt{k}\|P_m^* - \mathbb{P}_n^*\|_\infty, \quad m = 1, \dots M;$$

3. For any $\alpha \in (0, 1)$, compute the $(1-\alpha)$-quantile of $(\mathbb{S}_m)_{1\leq m\leq M}$ denoted by $\widehat{Q}_{\mathbb{S}}(1-\alpha)$. Finally, reject $\mathscr{H}_0$ if $\mathbb{S} > \widehat{Q}_{\mathbb{S}}(1-\alpha)$.

Thanks to Theorem 3.4 we have that asymptotically the significance level of such a hypothesis test is α, as $n \to \infty$ and for $M \to \infty$. We remark that the empirical quantile $\widehat{Q}_{\mathbb{S}}(1-\alpha)$ is an estimate of the $(1-\alpha)$-quantile of the sup-norm $\|\mathcal{B}\|_\infty$ of the P_0^*-Brownian bridge $\mathcal{B}$.

3.3. Asymptotic theory of predictive distribution

One of the most prominent statistical problems is the prediction of certain events in a regression framework. Accurate predictions of more severe extreme events than those occurred in the past is far from being trivial. Whenever this is possible the resulting benefit is huge, because of the strong impact on real life that such events have. Here we want to go beyond the inference obtained by the posterior distribution $\widetilde{\Lambda}_n$ on an extreme conditional quantile $Q_{\boldsymbol{x}}(p)$, for a very small p, see Section 3.2.

The aim of this section is to propose a probabilistic forecasting method in an extreme regression type of framework. Given a past sample $(\boldsymbol{Y}_n, \boldsymbol{X}^{(n)})$, where $\boldsymbol{X}^{(n)} = (\boldsymbol{X}_1, \dots, \boldsymbol{X}_n)$, let $(Y^\star, \boldsymbol{X}^\star)$ be an independent out-of-sample response variable and covariate vector, representative of future events. We consider the conditional distribution $F_{0,n}^\star(y \mid \boldsymbol{x}) = \mathbb{P}(Y^\star \leq y \mid Y^\star > U_{\boldsymbol{x}}^{(0)}(n/k), \boldsymbol{X}^\star = \boldsymbol{x}, \boldsymbol{X}^{(n)}, \boldsymbol{Y}_n)$, for all $y > U_{\boldsymbol{x}}^{(0)}(n/k)$ and $\boldsymbol{x} \in (0, 1)^d$. Similar to Section 2.4, we leverage the results from Section 3.2 in a regression setting to perform forecasting using the posterior predictive distribution. A possible estimator of $F_{0,n}^\star(\cdot \mid \boldsymbol{x})$ is given by

$$\widehat{F}_n^\star(y \mid \boldsymbol{x}) = \int_{\Theta\times\mathscr{P}} H_\gamma\left(\frac{y - Y_{n-k,n}}{\sigma\,(c(\boldsymbol{x}))^\gamma} - \frac{1-(c(\boldsymbol{x}))^{-\gamma}}{\gamma}\right)\Phi_n(\mathrm{d}c(\boldsymbol{x}))\Pi_n(\mathrm{d}\boldsymbol{\vartheta}), \tag{23}$$

see also Remark 4.3 in the Supplementary Material [13] for further technical details justifying its construction. We recall again that for any $v \geq 1$ the Wasserstein distance of order v satisfies the scaling property

$$W_v(\widehat{F}_n^\star, F_{0,n}^\star) = a_0(n/k) W_v(\widehat{F}_n^\star(a_0(n/k)\cdot \mid \boldsymbol{x}), F_{0,n}^\star(a_0(n/k)\cdot \mid \boldsymbol{x})),$$

where on the left-hand side we omit conditioning on $\boldsymbol{x}$ for brevity. Next result establishes the Wasserstein consistency of the predictive distribution. This is important as it guarantees that predictions based on the posterior predictive distribution $\widehat{F}_n^\star(\cdot \mid \boldsymbol{x})$ are increasingly accurate for increasing sample size, regardless is the value of $\boldsymbol{x}$ on $(0, 1)^d$.

Theorem 3.5. *Assume that the conditions of Theorem 2.12 and 3.1 are satisfied. Then, for all sequences $\varepsilon_n \to 0$, satisfying $(kK/n)\varepsilon_n^2/\log(kK/n) \to \infty$ as $n \to \infty$, we have*

$$\frac{W_v(\widehat{F}_n^\star, F_{0,n}^\star)}{a_0(n/k)} = O_{\mathbb{P}}(\varepsilon_n).$$

4. Simulation experiments

We assess the finite-sample performance of Bayesian inference based on the posterior distributions introduced in Sections 2 and 3. First, we provide a brief overview of the computational methods used to sample from these posteriors. Given that unconditional analysis is typically of narrower scope than conditional analysis in applications, we then summarize the key findings for the posterior Π_n of the parameter $\boldsymbol{\vartheta}$ and $\widetilde{\Pi}_n$ of the extreme quantile $Q(p)$ for brevity, with a full description available in Section 6.1 of the Supplementary Material [13]. Finally, we present a detailed analysis of the posterior distributions Ψ_n for the scedasis function $c(\boldsymbol{x})$ and $\widetilde{\Lambda}_n$ for the extreme conditional quantile $Q_{\boldsymbol{x}}(p)$.

4.1. Posterior distribution computation

The analytical expression of the posterior Π_n is unknown in closed-form. Sampling from it is however viable using MCMC computational methods. The adaptive random-walk Metropolis-Hastings algorithm [24] and its Gaussian random-walk version with Robbins–Monro process optimal scaling (see [21]), is readily implementable and computationally efficient. It has already been successfully exploited by [31], with the block maxima approach, where extensive simulation experiments demonstrate that an accurate inference is achievable through the posterior distribution, is consistent with the corresponding theoretical results. To save space we provide the full description of such sampling procedure in Section 5 of the Supplementary Material [13]. The sampling $Q \sim \widetilde{\Pi}_n$ is achieved as a by-product of first sampling $\boldsymbol{\vartheta} \sim \Pi_n$ and exploiting the transformation $Q(p) = Y_{n-k,n} + H_{\boldsymbol{\vartheta}}^{\leftarrow}(1 - np/k)$, for a small $p \in (0,1)$. Let $\boldsymbol{\vartheta}_1^*, \dots, \boldsymbol{\vartheta}_N^*$ be a sample from Π_n, then a Monte Carlo approximation of the posterior predictive distribution in (16) is

$$\widehat{F}_n^{\star}(y) \approx \frac{1}{N} \sum_{i=1}^{N} H_{\boldsymbol{\vartheta}_i^*}(y - Y_{n-k,n}).$$

Since $\boldsymbol{\vartheta}$ and $\mathcal{P}^*$ are independent with distribution Π_n and Φ_n, see Sections 3.1 and 3.2 for details, and given that Φ_n is a Dirichlet process, the computation of the density and other related quantities of the Dirichlet-multinomial distribution or sampling from it, is readily done using the `R` package `extraDistr` [39].

Conditionally to the data sample, $\boldsymbol{\vartheta}$ and $c(\boldsymbol{x})$ are independent, then sampling $Q_{\boldsymbol{x}} \sim \widetilde{\Phi}_n$ is achieved sampling first $\boldsymbol{\vartheta} \sim \Pi_n$ and independently $c(\boldsymbol{x}) \sim \Psi_n$, for any given $\boldsymbol{x} \in [0,1]^d$, and then transforming them by the formula (22). Finally, let $\boldsymbol{\vartheta}_1^*, \dots, \boldsymbol{\vartheta}_N^*$ be a sample from Π_n and $c_1^*(\boldsymbol{x}), \dots, c_N^*(\boldsymbol{x})$ be a sample from Ψ_n, then an approximation of the posterior predictive distribution in (23) is obtained as

$$\widehat{F}_n^{\star}(y|\boldsymbol{x}) \approx \frac{1}{N} \sum_{i=1}^{N} H_{\gamma_i^*}\left(\frac{y - Y_{n-k,n}}{\sigma_i^* \, (c_i^*(\boldsymbol{x}))^{\gamma_i^*}} - \frac{1 - (c_i^*(\boldsymbol{x}))^{-\gamma_i^*}}{\gamma_i^*} \right), \quad \boldsymbol{x} \in [0,1]^d.$$

4.2. Unconditional POT setting

We investigate the behavior of the posterior distributions Π_n and $\widetilde{\Pi}_n$ and the performance of the resulting inference. The investigation relies on a simulation experiment involving nine distributions, three for each domain of attraction: Fréchet, Pareto and Half-Cauchy in the Fréchet one, Exponential, Gumbel and Gamma in the Gumbel one and Beta, (reverse) Weibull and Power-law in the Weibull one. To save space, we refer to Section 6.1 of the Supplementary Material [13] for a complete description of: the simulation setup, the computational aspects and the collected results. We summarize our findings below. Firstly, we study the concentration properties of the posterior distribution Π_n, theoretically implied by the consistency result in Theorem 2.8. With all the considered distributions, the empirical posterior distribution is already well concentrated around the true parameter value with only $k = 20$ exceedances from a sample of size $n = 155$. In the Fréchet domain of attraction the posteriors are more spread than those obtained with the other two domains. However, increasing the sample size to $n = 303, 699, 2146$ and the number of exceedances to $k = 30, 50, 100$ the posterior distribution shrinks considerably and in the last case concentrates very much in proximity to the true parameter values. These results support the asymptotic concentration properties in Theorem 2.8.

Secondly, we compute the Monte Carlo coverage probability of symmetric- and asymmetric-95% credible intervals for γ_0 and $a_0(n/k)$ and the extreme quantile $F_0^{\leftarrow}(0.999)$ and of symmetric- and asymmetric-95% credible regions for $\boldsymbol{\vartheta}_0$. Overall, with all the models in the three domains of attraction the performance is very good, with coverage probabilities that are close to the 95% nominal level already with the smallest intermediate sequence $k = 20$ and sample size $n = 155$. With increasing n and k the coverage probabilities get even closer. In Fréchet and Gumbel domains of attraction, the coverage probabilities relative to γ_0 and $a_0(n/k)$ are almost the same. In the Weibull domain of attraction, the symmetric intervals for γ_0 are slightly larger than expected. Differently, symmetric intervals for $a_0(n/k)$ have coverage probability slightly smaller than the nominal level. In the Fréchet and Gumbel domains of attraction, the symmetric intervals for $F_0^{\leftarrow}(0.999)$ are slightly larger than expected, while in the Weibull domain a coverage probability above nominal level is expected. Overall, the asymmetric ones perform much better. In the three domains of attraction the worst coverage probabilities are obtained with the credible region for $\boldsymbol{\vartheta}_0$, since a higher-dimensional parameter is more difficult to estimate. However, also in this case the coverage probabilities approach the nominal level as k and n increase. Concluding, the valuable theoretical properties are actually verifiable in practice already with moderate sample sizes.

4.3. Extreme regression setting

We study the behavior of the posterior distributions Φ_n and $\widetilde{\Lambda}_n$ and the performance of the resulting inference. We consider two experiments where the data are generated according to the following mechanism. First, we sample $x_1, \dots, x_n$ observations from $X_1, \dots, X_n$ i.i.d. random covariates. To account for cases in which the covariate is scattered over whole $[0, 1]$, concentrated on 1/2, concentrated to the left near zero and concentrated to the right near one, we consider the following options for the distribution of X: $\mathcal{U}(0, 1)$, i.e. uniform on $[0, 1]$, Beta$(2, 2)$, Beta$(2, 5)$ and Beta$(5, 2)$, where Beta(a, b) is a Beta distribution with shape parameters a and b. Second, similarly to [16], the ith observation y_i is generated from $Y|(X_i = x_i)$ whose distribution is the rescaled Fréchet distribution $F_{x_i}(y) = \exp(-\varsigma(x_i)/y)$, $y > 0$. We consider three possible models for the scedasis function:

(i) [scedasis straight line] $\varsigma(x) = (1 + \beta x)\mathbb{1}(0 \le x \le 1)$;
(ii) [scedasis broken line] $\varsigma(x) = (1 + 2\beta x)\mathbb{1}(0 \le x \le 0.5) + (1 + 2\beta(1 - x))\mathbb{1}(0.5 < x \le 1)$;
(iii) [scedasis bump function] $\varsigma(x) = \mathbb{1}((0 \le x \le 0.4) \cup (0.6 \le x \le 1)) + (1 + 10\beta(x - 0.4))\mathbb{1}(0.4 < x \le 0.5) + (1 + 10\beta(0.6 - x))\mathbb{1}(0.5 < x < 0.6)$.

Note that these data generating processes satisfy the proportional tail assumption with scedasis function c_0 related to ς by $c_0(x) = \varsigma(x)/\int_0^1 \varsigma(z)\mathcal{P}_0(\mathrm{d}z)$, where $\mathcal{P}_0$ is the measure pertaining to the distribution of the covariate. In the first experiment, β is taken to be a sequence of 100 equally spaced values in $[-1, 1]$. For each value of it we simulate $n = 5000$ observations, according to the sampling scheme above described, and we perform the hypothesis testing introduced below Theorem 3.4, where we

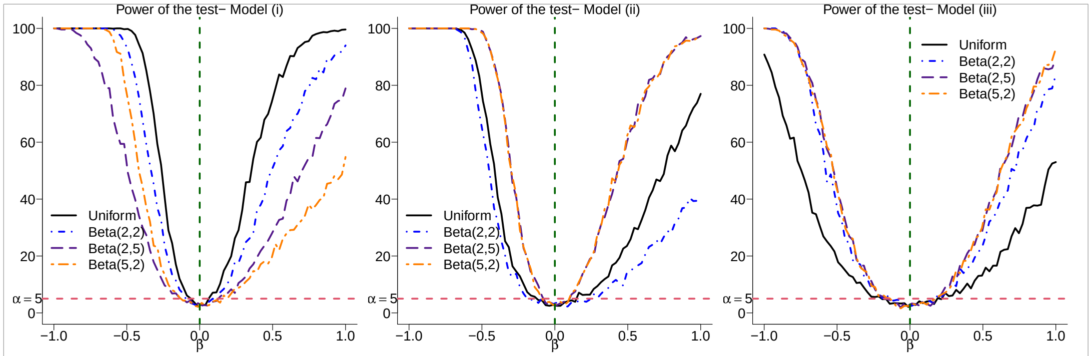

Figure 1. Estimated power functions. Lines report the empirical proportion of simulated samples under $H_1 : P_0^* \neq P_0$ that rejected $H_0 : P_0^* = P_0$ as a function β. Dotted red horizontal line is the 5% significance level of the test.

use the setting: $k = 400$, DP prior with parameter $\tau = 5 \cdot \mathcal{U}(\cdot; 0, 1)$, where $\mathcal{U}(\cdot; a, b)$ is the uniform measure on a, b, $M = 1000$ independent samples from the DP prior and significance level $\alpha = 0.05$. We repeat the sampling and testing steps $N = 1000$ times and we compute the rejection rates. The estimated significance level, as the proportion of simulated samples under $\mathscr{H}_0 : P_0^*(x) = P_0$ that rejects $\mathscr{H}_0$ is: 3.8% if $X \sim \mathcal{U}(0, 1)$, 4.6% if $X \sim$Beta$(2, 2)$, 3.7% if $X \sim$Beta$(2, 5)$ and 4.1% if $X \sim$Beta$(5, 2)$ with a scedasis straight line; 3.9% if $X \sim \mathcal{U}(0, 1)$, 3.9% if $X \sim$Beta$(2, 2)$, 4.2% if $X \sim$Beta$(2, 5)$ and 3.7% if $X \sim$Beta$(5, 2)$ with a scedasis broken line; 4.6% if $X \sim \mathcal{U}(0, 1)$, 4.5% if $X \sim$Beta$(2, 2)$, 4.0% if $X \sim$Beta$(2, 5)$) and 4.2% if $X \sim$Beta$(5, 2)$ with a scedasis bump function. Figure 1 displays the estimated powers of the test, as the proportion of samples simulated under $\mathscr{H}_1 : P_0^* \neq P_0$ that rejects $\mathscr{H}_0$, obtained with different covariate distributions by the black solid line, the blue dotdashed line, the violet dashed line, and the yellow twodashed, respectively, and with the different scedasis models (i)–(iii) from left to the right panel. Results highlight accurate estimation of α and a good power of the test. The best results are obtained with a scedasis straight line, in which case α is better estimated when the covariate distribution is right-skewed. The largest power of the test is obtained when the covariate is uniformly scattered. As expected, the test is less performing with a scedasis broken line and a scedasis bump function, since they are much more complicated functions to estimate. In these cases, α is better estimated if the covariate's distribution is concentrated around $1/2$. While the largest power is obtained if the covariate's distribution is concentrated near the endpoints 0 and 1 with model (ii) and uniformly scattered instead with model (iii).

In the second experiment, we simulate $n = 5000$ observations from a rescaled Fréchet distribution, with $\varsigma(x)$ specified as in models (i)–(iii), and we take $\beta = 1, 2, 10$ and we apply the sampling procedures of Section 4.1 to simulate $N = 20,000$ realizations of $\boldsymbol{\vartheta}$ and $c(\boldsymbol{x})$ from Π_n and Ψ_n for 100 equally spaced values of $x \in [0, 1]$. In the first case we use the informative prior on γ and the data-dependent prior on σ described in Section 5 of [13] and in the second case we use both the kernel-based method with bandwidth $bw = 0.1$ and the KNN-based method with $K = 750$ neighbours (see Section 3.2 for details). Again, the DP prior is set with parameter $\tau = 5 \cdot \mathcal{U}(\cdot; 0, 1)$. Combining those samples by the transformation (22) we obtain a sample from $\widetilde{\Lambda}_n$. We repeat these steps $M = 1000$ times and compute a Monte Carlo approximation of the root mean integrated relative squared error (RMIRSE), i.e.

$$\mathrm{RMIRSE} = \left(\mathbb{E} \left(\int_0^1 \left(\frac{\widehat{f}_n(x))}{f_0(x)} - 1 \right)^2 \mathrm{d}x \right) \right)^{1/2},$$

Table 1. Monte Carlo approximation of the RMIRSE for the posterior mean estimators $\overline{c}_n$ and $\overline{Q}_{x,n} \equiv \overline{Q}_{x,n}(0.001)$. Posterior distributions are computed using the kernel and KNN methods, different scedasis models and different covariate distribution. The sixth and eighth columns report the relative gain of KNN compared to kernel one.

Model	X's distribution	n	k	RMIRSE - $\overline{c}_n$ kernel	RMIRSE - $\overline{c}_n$ KNN	RMIRSE - $\overline{Q}_{x,n}$ kernel	RMIRSE - $\overline{Q}_{x,n}$ KNN
(i)	$\mathcal{U}(0,1)$	5000	400	0.780	−52.8%	3.841	−3.7%
	Beta(2,2)	–	–	1.043	−21.4%	3.853	−2.8%
	Beta(2,5)	–	–	1.334	−0.7%	4.190	4.5%
	Beta(5,2)	–	–	2.864	22.4%	5.632	8.5%
(ii)	$\mathcal{U}(0,1)$	5000	400	1.465	15.4%	2.064	−2.0%
	Beta(2,2)	–	–	2.012	17.3%	2.054	−11.6%
	Beta(2,5)	–	–	3.419	8.6%	7.074	41.1%
	Beta(2,5)	–	–	3.493	9.6%	7.820	39.8%
(iii)	$\mathcal{U}(0,1)$	5000	400	1.316	0.9%	2.143	5.1%
	Beta(2,2)	–	–	1.580	20.1%	2.278	11.8%
	Beta(2,5)	–	–	5.714	41.9%	5.105	11.8%
	Beta(5,2)	–	–	4.777	31.7%	5.303	12.8%

where the true function $f_0(x)$ is either $c_0(x)$ or $F_x^{(0)\leftarrow}(0.999)$ and the corresponding estimator $\widehat{f}_n(x)$ is either the scedasis posterior mean $\overline{c}_n(x)$ or the extreme conditional quantile posterior mean $\overline{Q}_{x,n}(p)$. Table 1 reports the results split according to the scedasis model (vertical sections), the different covariate's distributions (along the rows), the function to be estimated ($c_0(x)$ in the fifth and sixth column and $F_x^{(0)\leftarrow}(0.999)$ in the seventh and eight column) and the estimation method for prior and posterior construction (kernel and KNN). The RMIRSE obtained with the kernel-based method shows greater precision when estimating a scedasis straight line in comparison to the other two cases. The difference among results is also fairly small when estimating $F_x^{(0)\leftarrow}(0.999)$ but the RMIRSE does not highlight a clear superiority obtained with a specific scedasis form. Regardless of the scedasis form, the posterior mean is more accurate when the covariate is uniformly distributed or symmetrically concentrated around $1/2$ than in the other cases. Overall, results suggest that the posterior mean is an accurate estimator for $c_0(x)$ and $F_x^{(0)\leftarrow}(0.999)$. The sixth and eighth columns of Table 1 report the relative gain (in percentage) of using the KNN method in place of the kernel one, i.e. (RMIRSE(kernel)-RMIRSE(KNN))/RMIRSE(kernel) $\cdot 100\%$. When estimating $c_0(x)$ ($F_x^{(0)\leftarrow}(0.999)$), apart from three (four) cases the remaining ones highlight better performance of the KNN method, with a gain that ranges between 0.9% (4.5%) to 41.9% (41.1%) and therefore on balance it is preferable.

In the sequel we focus on the results obtained with the KNN method. The top panels of Figure 2 display, for three specific values $x = 0.1, 0.5, 0.9$, the Monte Carlo distribution of $\overline{c}_n(x) - c_0(x)$, obtained with the $M = 1000$ data samples. Results obtained with the model (i)-(iii) are displayed from left to right panel. Each panel reports the results obtained with the different covariate distributions. Since the boxplots are almost all centred around zero with a small dispersion we can conclude that $\overline{c}_n(x)$ is an accurate estimator of $c_0(x)$. Note that, when the covariate distribution is left-skewed (right-skewed), e.g. the Beta(5,2) (Beta(2,5)), $\overline{c}_n(x)$ overestimates a bit $c_0(0.1)$ ($c_0(0.9)$) as expected, since only few data are available around $x = 0.1$ ($x = 0.9$). With model (iii), $\overline{c}_n$ underestimates a bit $c_0(0.5)$ when the covariate's distribution are not concentrated around $x = 1/2$. Similar results are obtained when estimating $F_x^{(0)\leftarrow}(0.999)$, see the bottom panels of Figure 2. The distribution of $\overline{Q}_{x,n}(0.001) -$

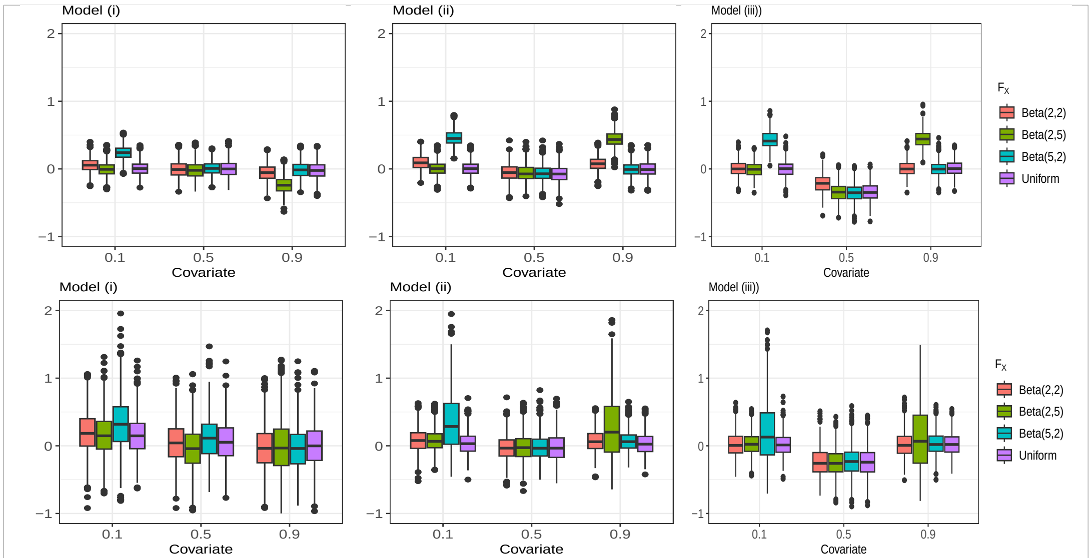

Figure 2. Boxplot of $\overline{c}_n(x) - c_0(x)$ (top panels) and $\log(\overline{Q}_{x,n}(0.001)) - \log(F_x^{(0)\leftarrow}(0.999))$ (bottom panels), obtained with the scedasis model (i)-(iii) and different covariate's distributions and setting the covariate value $x = 0.1, 0.5, 0.9$.

$F_x^{(0)\leftarrow}(0.999)$ is more spread as expected since the estimation of the conditional extreme quantile is harder.

Finally, we compute the Monte Carlo coverage probability of the credible intervals for $c_0(x)$ and $F_x^{(0)\leftarrow}(0.999)$. To save space, results are reported in Table 4 of Section 6.2 of the Supplementary Material [13]. Results are reported separately according to the different scedasis form (vertical sections) and covariate's distributions (along the rows). The column Type indicates with the letter “A" the coverage probability of an asymmetric-95% credible interval (obtained with the quantiles of the posterior distribution) and with the letter “S" the symmetric version (obtained as $[\widehat{f}(x)_n \pm z_{\alpha/2}\widehat{s}_n(x)]$, where $\widehat{s}_n(x)$ is the posterior standard deviation and $z_{\alpha/2}$ is the standard normal $(1-\alpha/2)$-quantile). With model (i) the coverage probabilities are very close to the nominal level apart for the case of $c_0(0.9)$ ($c_0(0.1)$) when the covariate is Beta(2,5) (Beta(5,2)) distributed, but this is expected as there are few observations close to one (zero). Therefore, a larger sample size is needed to achieve the nominal level all over $[0.1]$. Similar conclusions hold with model (ii) and (iii).

Acknowledgments

The authors thank the Editor, Associate Editor, and three anonymous reviewers for their valuable comments on an earlier version of this paper.

Funding

S. A. Padoan is supported by the Bocconi Institute for Data Science and Analytics and financially supported by MUR–PRIN Bando 2022–prot. 20229PFAX5, European Union - Next Generation EU, Mission 4 Component 2 CUP J53D23004260001.

Supplementary Material

Supplement to "Asymptotic theory for Bayesian inference and prediction: from the ordinary to a conditional Peaks-Over-Threshold method" [13]
The Supplementary Material document contains a detailed discussion of important frequentist properties of the empirical log-likelihood process relative to the GP density function, all the proofs, further numerical details and results concerning the simulation study and further details on the real application.

References

[1] AHMAD, A. A., DEME, E. H., DIOP, A., GIRARD, S. and USSEGLIO-CARLÈVE, A. (2020). Estimation of extreme quantiles from heavy-tailed distributions in a location-dispersion regression model. *Electron. J. Stat.* **14** 4421–4456.

[2] BALKEMA, A. A. and DE HAAN, L. (1974). Residual life time at great age. *Ann. Probab.* **2** 792–804.

[3] BEIRLANT, J., GOEGEBEUR, Y., SEGERS, J. and TEUGELS, J. L. (2004). *Statistics of Extremes: Theory and Applications*. John Wiley & Sons.

[4] BOBBIA, B., DOMBRY, C. and VARRON, D. (2021). Extreme quantile regression in a proportional tail framework. *Transactions of A. Razmadze Mathematical Institute* **175** 13–32.

[5] BÜCHER, A. and JENNESSEN, T. (2024). Statistics for heteroscedastic time series extremes. *Bernoulli* **30** 46–71.

[6] CASTILLO, I. (2014). On Bayesian supremum norm contraction rates. *Ann. Statist.* **42** 2058—2091.

[7] CHEN, Y., MA, M. and SUN, H. (2023). Statistical inference for extreme extremile in heavy-tailed heteroscedastic regression model. *Insurance Math. Econom.* **111** 142–162.

[8] COLES, S. (2001). *An introduction to statistical modeling of extreme values*. Springer.

[9] COLES, S. and PERICCHI, L. (2003). Anticipating catastrophes through extreme value modelling. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **52** 405–416.

[10] DAOUIA, A., GIJBELS, I. and STUPFLER, G. (2022). Extremile regression. *J. Amer. Statist. Assoc.* **117** 1579–1586.

[11] DE HAAN, L. and FERREIRA, A. (2006). *Extreme Value Theory: An Introduction*. Springer.

[12] DO NASCIMENTO, F. F., GAMERMAN, D. and DAVIS, R. (2016). A Bayesian semi-parametric approach to extreme regime identification. *Braz. J. Probab. Stat.* **30** 540–561.

[13] DOMBRY, C., PADOAN, S. A. and RIZZELLI, S. (2026). Supplement to "Asymptotic theory for Bayesian inference and prediction: from the ordinary to a conditional Peaks-Over-Threshold method". https://doi.org/DOI:10.1214/[providedbytypesetter]

[14] DREES, H. (2023). Statistical inference on a changing extreme value dependence structure. *Ann. Statist.* **51** 1824–1849.

[15] DREES, H., FERREIRA, A. and DE HAAN, L. (2004). On maximum likelihood estimation of the extreme value index. *Ann. Appl. Probab.* **14** 1179–1201.

[16] EINMAHL, J. H., HAAN, L. and ZHOU, C. (2016). Statistics of heteroscedastic extremes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 31–51.

[17] EINMAHL, J. H. and HE, Y. (2022). Extreme value estimation for heterogeneous data. *J. Bus. Econ. Stat.* **41** 255–269.

[18] EINMAHL, J. H. J., FERREIRA, A., DE HAAN, L., NEVES, C. and ZHOU, C. (2022). Spatial dependence and space–time trend in extreme events. *Ann. Statist.* **50** 30–52.

[19] EMBRECHTS, P., KLÜPPELBERG, C. and MIKOSCH, T. (1997). *Modelling extremal events: for insurance and finance*. Springer Science & Business Media.

[20] FÚQUENE PATIÑO, J. A. (2015). A semi-parametric Bayesian extreme value model using a Dirichlet process mixture of gamma densities. *J. Appl. Stat.* **42** 267–280.

[21] GARTHWAITE, P. H., FAN, Y. and SISSON, S. A. (2016). Adaptive optimal scaling of Metropolis–Hastings algorithms using the Robbins–Monro process. *Comm. Statist. Theory Methods* **45** 5098–5111.

[22] GHOSAL, S. and VAN DER VAART, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press.
[23] GOEGEBEUR, Y., GUILLOU, A. and SCHORGEN, A. (2014). Nonparametric regression estimation of conditional tails: the random covariate case. *Statistics* **48** 732–755.
[24] HAARIO, H., SAKSMAN, E. and TAMMINEN, J. (2001). An adaptive Metropolis algorithm. *Bernoulli* **7** 223–242.
[25] HALL, P., PENG, L. and TAJVIDI, N. (2002). Effect of extrapolation on coverage accuracy of prediction intervals computed from Pareto-type data. *Ann. Statist.* **30** 875–895.
[26] KLEIJN, B. and VAN DER VAART, A. (2012). The Bernstein-Von-Mises theorem under misspecification. *Electron. J. Stat.* **6** 354–381.
[27] LIU, L., LI, D. and WONG, W. H. (2023). Convergence Rates of a Class of Multivariate Density Estimation Methods Based on Adaptive Partitioning. *J. Mach. Learn. Res.* **24** 1–64.
[28] MEFLEH, A., BIARD, R., DOMBRY, C. and KHRAIBANI, Z. (2020). Trend detection for heteroscedastic extremes. *Extremes* **23** 85–115.
[29] NEUHAUS, G. (1971). On Weak Convergence of Stochastic Processes with Multidimensional Time Parameter. *Ann. Math. Statist.* **42** 1285–1295.
[30] NORTHROP, P. J. and ATTALIDES, N. (2016). Posterior propriety in Bayesian extreme value analyses using reference priors. *Statist. Sinica* 721–743.
[31] PADOAN, S. A. and RIZZELLI, S. (2024). Empirical Bayes inference for the block maxima method. *Bernoulli* **30** 2154–2184.
[32] RESNICK, S. I. (2007). *Extreme Values, Regular Variation, and Point Processes* **4**. Springer Science & Business Media.
[33] RESNICK, S. I. and ZEBER, D. (2014). Transition kernels and the conditional extreme value model. *Extremes* **17** 263–287.
[34] ROBERT, C. P. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer.
[35] TANCREDI, A., ANDERSON, C. and O'HAGAN, A. (2006). Accounting for threshold uncertainty in extreme value estimation. *Extremes* **9** 87–106.
[36] VAN DER VAART, A. (2000). *Asymptotic Statistics*. Cambridge University Press.
[37] WANG, H. and DEYUAN, L. (2015). Estimation of Extreme Conditional Quantiles. In *Extreme Value Modeling and Risk Analysis: Methods and Applications* (D. Dey and J. Yan, eds.) Book chapter 15. Chapman and Hall/CRC.
[38] WANG, H. J., LI, D. and HE, X. (2012). Estimation of high conditional quantiles for heavy-tailed distributions. *J. Amer. Statist. Assoc.* **107** 1453–1464.
[39] WOLODZKO, T. (2020). `extraDistr`: Additional Univariate and Multivariate Distributions `R` package version 1.9.1.