Abstract

The thesis concerns Bayesian statistics, the sub-domain of statistics which can be further splitted into Bayesian parametric and Bayesian non-parametric. The former compromises classic methodology for prior and posterior distributions in models with a finite number of parameters. The latter contains the methods characterised by 1) big parameter spaces and 2) construction of probability measures over big spaces. By saying "Bayesian" we acknowledge two mainstreams of thought in statistics : being Bayesian means placing a prior over the parameter space while frequentists deal with fixed unknown in place of parameters.

The second part discusses Extreme Value theory and proposes new method of estimation of the tail of distribution by approximating it with Fisher distribution, a generalisation of the commonly used Generalised Pareto Distribution. As a new result we propose a Bayesian inference of extreme quantiles based on aforementioned approximation. We compare and discuss the Bayesian approach with non-Bayesian methods, namely method of moments, probability weighted moments, maximum likelihood estimation.

1 Background

In this chapter we introduce some basic probability definitions. For most of the definitions we use [Billingsley, 2012].

2 Basic probability facts

Definition 2.1. *If $\mathcal{F}$ is a σ-field in Ω and p is a probability measure of $\mathcal{F}$ then the triple $(\Omega, \mathcal{F}, p)$ is called a probability space.*

We will briefly review the properties of the conditional probabilities on which Bayesian statistics is heavily basing. From now on we will assume that sets appearing in formulas lie in the basic σ-field of $\mathcal{F}$.

Definition 2.2. *For sets A, B such that $p(B) > 0$ the conditional probability of B given A is defined as*

$$p(A|B) = \frac{p(A, B)}{p(B)}$$

In Bayesian statistics one sets a *prior* on parameters $\boldsymbol{\theta}$ in considered parameters space $\mathcal{Q} \ni \boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$ and updates those priors with incoming data $p(\boldsymbol{\theta}|X_1, \ldots, X_n) \sim \pi(\boldsymbol{\theta})p(X|\boldsymbol{\theta})$, where $p(X|\boldsymbol{\theta})$ is the *likelihood function* of $\boldsymbol{\theta}$. If $p(\boldsymbol{\theta}|X_1, \ldots, X_n)$ lies in the same probability density family as $\pi(\boldsymbol{\theta})$ then it is called *conjugate posterior distribution* to the $\pi(\boldsymbol{\theta})$, or equivalently $\pi(\boldsymbol{\theta})$ in called *conjugate prior* for the posterior distribution $p(\boldsymbol{\theta}|X_1, \ldots, X_n)$.

We give a brief overview of probability distributions we worked with.

Example 2.1. *The Fisher probability density function is given by*

$$f(x; \alpha_1, \alpha_2, \beta) = \frac{1}{\beta} \frac{1}{B(\alpha_1, \alpha_2)} \frac{(x/\beta)^{\alpha_1 - 1}}{(1 + x/\beta)^{\alpha_1 + \alpha_2}}, \;\; x > 0, \;\; \alpha_1, \alpha_2, \beta > 0., \tag{2-1}$$

where $B(\alpha_1, \alpha_2) = \Gamma(\alpha_1 + \alpha_2)/\Gamma(\alpha_1)\Gamma(\alpha_2)$ is a beta function and Γ is gamma function.

While describing discrete probability distributions one often speaks about *probability mass function*, giving the probability that the discrete random variable is exactly equal to some given value.

Example 2.2. *The probability mass function of multinomial distribution is*

$$P(X_1 = x_1, \ldots, X_k = x_k) = \frac{n!}{x_1! \cdots x_k!} \prod_{i=1}^{k} x_i^{p_i} = \frac{\Gamma(\sum_{i=1}^{k} x_i + 1)}{\prod_{i=1}^{k} \Gamma(x_i + 1)} \prod_{i=1}^{k} x_i^{p_i}.$$

Example 2.3. *The Dirichlet probability density function is*

$$f(x; \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{K} x_i^{\alpha_i - 1}, \tag{2-2}$$

where $\boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_K)$, $\alpha_i > 0$ and $x_i > 0$, $\sum_{i=1}^{K} x_i = 1$. Such sets of $(x_i)_{i \in \{1, \ldots, K\}}$ are called simplex and are often denoted by Δ_K.

Remark 2.1. *Dirichlet distribution given in 2.3 is a conjugate prior for 2.2.*

Another useful tools is cumulative distribution function which is closely related to probability density function :

$$F_X(x) = p(X \leq x) = \int_{-\infty}^{x} f(u)du \Rightarrow f(x) = \frac{dF_X(x)}{dx}. \tag{2-3}$$

Hence one may quickly access one more property of cdf

$$p(a \leq x \leq b) = F(b) - F(a) = \int_a^b f(u)du, \;\; a, b \in \mathbb{R}. \tag{2-4}$$

A closely connected and widely used in extreme value theory is *survival function* defined as $S_X(x) := 1 - F_X(x)$.

Example 2.4. *The cdf and sf of Generalised Pareto Distribution (GPD) with pdf*

$$f(x; \alpha, \beta) = \frac{\alpha}{\beta}\left(1 + \frac{x}{\beta}\right)^{-\alpha-1}, \;\; x > 0, \;\; \alpha, \beta > 0 \tag{2-5}$$

are

$$F_X(x) = 1 - \left(1 + \frac{x}{\beta}\right)^{-\alpha}, \quad S_X(x) = \left(1 + \frac{x}{\beta}\right)^{-\alpha}. \tag{2-6}$$

Example 2.5. *The cdf of Fréchet distribution is*

$$F(x) = \exp(-x^{-1/\beta}), \;\; \beta > 0. \tag{2-7}$$

Simulations in the second part of this thesis were done on the data simulated from this distribution with $\beta = 2$.

It is often the case that one has at his disposal data $(x_1, \dots, x_n)$ with unknown cdf F. A common practice in that case is to work with model-free estimate of F which can be obtained directly from sampled data in increasing order $x_{(1)}, \dots, x_{(n)}$.

Definition 2.3. *For an ordered sample of iid observations* $x_{(1)}, \dots, x_{(n)}$ *we define the empirical distribution function as*

$$\tilde{F}(x) = \frac{i}{n+1} \quad \textit{for} \quad x_{(i)} \leq x \leq x_{(i+1)}. \tag{2-8}$$

We denote the joint cdf $F_{XY}(x, y) = p(X \leq x, Y \leq y)$. The conditional cdf reads as $F_{XY}(x|y) = p(X \leq x \mid Y \leq y) = F_{XY}(x, y)/F_Y(y)$. The Bayesian rule also applies $F(x|y) = F(y|x)F(x)/F(y)$.

Example 2.6. *Suppose we have data sorted in increasing order* $x_{1:n} := (x_1, \dots, x_n)$ *and suppose that we are interested in* $p(Y > x + u \mid Y > u)$*, where* u *is* k*-positional statistic of ordered data* $u = \max\{x_{1:k}\}$*. Then*

$$p(Y > x + u \mid Y > u) = \frac{p(Y > x + u, \; Y > u)}{p(Y > u)} = \frac{1 - F_Y(x + u)}{1 - F(u)} = \frac{S_Y(x + u)}{S(u)}. \tag{2-9}$$

It is convenient to summarise a considered probability distribution by two first statistics (also referred to as first and second moment) that characterise its main features. The most common are the expectation and variance.

Definition 2.4. *Suppose X has a distribution μ. For any $g : \mathbb{R} \to \mathbb{R}$ we have*

$$\mathbb{E}(g(X)) = \int_{\mathbb{R}} g(x)\mu(dx).$$

For $g(x) = x$ we obtain definition of expected value. The variance of the random variable X is defined as

$$Var(X) = \mathbb{E}\Big(X - \mathbb{E}X\Big)^2 = \mathbb{E}[X^2] - [\mathbb{E}(X)]^2.$$

Remark 2.2. *Moments of order $\tau \in \mathbb{R}$ may be estimated from the data sample $(x_1, \ldots, x_n)$ using*

$$\mathbb{E}X^\tau = \frac{1}{n}\sum_{i=1}^{n} x_i^\tau \tag{2-10}$$

Example 2.7. *Let X be sampled from the Dirichlet distribution* (2-2)*. Hence*

$$\mathbb{E}(X_i) = \alpha_i / \sum_j \alpha_j, \quad Var(X_i) = \frac{\alpha_i(\sum_j \alpha_j - \alpha_i)}{(\sum_j \alpha_j)^2(\sum_j \alpha_j + 1)}.$$

Example 2.8. *Gamma distribution $\Gamma(\alpha, \beta)$ has pdf given by*

$$f(x; \alpha, \beta) = \frac{\beta^\gamma}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x},\ \ \alpha > 0, \beta > 0,\ \ x > 0. \tag{2-11}$$

Therefore for $X \sim \Gamma(\alpha, \beta)$ we have two first moments $\mathbb{E}X = \alpha/\beta,\ \ VarX = \alpha/\beta^2$.

Example 2.9. *The Beta distribution $Be(\alpha, \beta)$ has pdf*

$$f(x; \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)},\ \ x \in (0, 1). \tag{2-12}$$

For $X \sim Be(\alpha, \beta)$ the mean and variance are respectively

$$\mathbb{E}X = \frac{\alpha}{\alpha + \beta}, \quad VarX = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}. \tag{2-13}$$

Remark 2.3. *The expected value does not have to exist.*

As a standard example of such a pathology serves Cauchy distribution with pdf :

$$f(x; \gamma) = \left(\pi\gamma\left[1 + \frac{(x - x_0)^2}{\gamma^2}\right]\right)^{-1},\ \ \gamma > 0,\ \ x \in \mathbb{R}.$$

In theory of extreme value the particular interest of the researcher lies in the quantiles as the provide good overall overview about the distribution one is investigating.

Definition 2.5. *We say that x is the quantile of cdf F at the level p if*

$$q(p) = \inf\{x : \; F(x) \geq p\}. \tag{2-14}$$

Example 2.10. *For Fréchet distribution with cdf as in* (2-7) *we obtain*

$$q(p) = \inf\{x : \; F(x) \geq p\} = \inf\{x : \; \exp(-x^{-1/\beta}) \geq p\} = -\log(p)^{-\beta}.$$

There is plenty of methods which aim at providing a goodness-of-fit evaluation. In our work we will use a *quantile plots* :

Definition 2.6. *For an ordered sample of iid observations $x_{(1)}, \ldots, x_{(n)}$ from a probability distribution with estimated cdf $\hat{F}$ a quantile plot consists of the points*

$$\left\{ \left(\hat{F}^{-1}\left(\frac{i}{n+1} \right), x_i \right), \quad \forall i \right\}. \tag{2-15}$$

3 Choosing a prior

The role of the prior in the Bayesian analysis is to encode possessed information into the problem under consideration. Choosing an adequate prior is not a trivial task. One is often set before a dilemma : set a *weakly informative prior* when one is strongly biased to some belief about the data or on the contrary, set a *strongly informative prior* for the fear of not influencing the posterior probability too heavily.

In general, weakly informative priors help control inference computationally and statistically. Computationally, a prior increases the curvature around the volume where the solution is expected to lie, which guides gradient-based Hamiltonian Monte Carlo sampling by not allowing them to stray too far from the location of a surface.

3.1 Considered priors

There is infinite number of possible priors. In our simulations in the second part we used so called *Jeffrey priors* [Jeffreys, 1946].

1. for GPD as in (2-5) Jeffrey priors are :

$$\pi^J(\alpha) \propto \frac{1}{(\alpha+1)\sqrt{\alpha(\alpha+2)}}, \; \alpha > 1$$
$$\pi^J(\beta) \propto \frac{1}{\beta}, \; \beta > 0.$$

2. for Fisher distribution as in (2-1) Jeffrey priors are

$$\pi^J(\alpha_1, \alpha_2) \propto \Big(\zeta(\alpha_1)\zeta(\alpha_2) - \zeta(\alpha_1 + \alpha_2) \Big)^{1/2} \Big(\zeta(\alpha_1) + \zeta(\alpha_2) \Big)^{1/2},$$
$$\pi^J(\beta) \propto \frac{1}{\beta}, \; \beta > 0,$$

where $\psi(x)$ is digamma function defined as $\psi(x) := \frac{d}{dx}\log\Gamma(x) = \Gamma'(x)/\Gamma(x)$.

4 Extreme value theory - introduction

5 Motivations

Assessing the probability of extreme events is an important issue in many domains, ranging from the risk management to forecasting natural catastrophes. Extreme value theory supplies us with the solid fundamentals needed for the statistical modelling of such events. The focus of this part is put on the use of extreme value theory to compute extreme quantiles from a distribution by approximating its cumulative distribution function with the GPD and Fisher family and to compare the fit of those two distributions. The method is implemented on the Fréchet distribution. All the code used for simulations is available at GitHub repository [Lewandowski, 2018]. We worked with Stan software [Stan Development Team, 2015].

6 Model formulation

Extreme value theory is a branch of statistics dealing with extreme deviations from the median of a probability distribution. We let $X_{1:n}$ be a sequence of iid random variables with cumulative distribution function F, we denote the maximum value $M_n = \max(X_{1:n})$ and the upper point of F by $w(F) = \sup\{x : F(x) < 1\}$. In theory the exact distribution of the maximum can be derived and is given by

$$\mathbb{P}(M_n \leq z) = \mathbb{P}(X_{1:n} \leq z) = \prod_{i=1}^{n} \mathbb{P}(X_i \leq z) = F^n(z),$$

which converges almost surely to $w(F)$ becoming a point mass there. The problem with this approach is that usually F is unknown - one may suggest estimating $\tilde{F}$ from the data as in (2-8), but then small discrepancies in the estimation of F strongly influence F^n.

Therefore standard approach is to seek non-degenerate H such that for constants $a_n > 0, b_n$ the cumulative distribution function of the normalised M_n converges to H :

$$\mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq z\right) = F^n(a_n z + b_n) \xrightarrow{n\to\infty} H(z). \tag{6-1}$$

This is an extreme value analogue of the Central Limit Theorem. We say that H is an *extreme value cumulative distribution function* and call H an *attractor* of F, denoted as $F \in D(H)$. Several interesting properties of H in correspondence with the survival function of X_i may be proven.

Proposition 6.1. *Let $X_{1:n}$ be iid random variables with survival function S and cumulative distribution function F. Then $F \in D(H)$ if and only if for $n \to \infty$ and all $x \in \mathbb{R}$ we have*

$$nS(a_n x + b_n) \to -\log H(x)$$

Démonstration. (⇒) If $F \in D(H)$ then for all $x \in \mathbb{R}$ there exist sequences $a_n > 0,\ b_n$ satisfying $F^n(a_n x + b_n) \xrightarrow{n\to\infty} H(x)$. We know the asymptotic behaviour of the survival function : $S(x) \xrightarrow{x\to\infty} 0$. Thus taking logarithm of the former expression and as $\log(1+y) \approx y$ for $y \approx 0$ one obtains

$$nS(a_n x + b_n) \xrightarrow{n\to\infty} -\log H(x).$$

(⇐) If there exist sequences $a_n > 0,\ b_n$ such that $nS(a_n x + b_n) \xrightarrow{n\to\infty} -\log H(x)$ and since $H(x) \xrightarrow{x\to\infty} 1$ after repeating in reverse order previous steps we conclude the proof. □

Following definition allows us to characterise cumulative distribuant function with respect to their similarity. We remind that a *degenerated* cumulative distribution function is of the form $F(x) = \mathbb{1}_{x>k_0}$ for some $k_0 \in \mathbb{R}$.

Definition 6.1. *Let F_1 and F_2 be two non-degenerated cumulative distribution functions. F_1 and F_2 are of the same type iff there exists $a > 0$ and $b \in \mathbb{R}$ satisfying $F_1(x) = F_2(ax + b)$ for all $x \in \mathbb{R}$.*

In estimation of quantiles one has to be sure about their uniqueness and hence uniqueness of the corresponding cumulative distribution function function. This is provided by

Lemma 6.1. *Let F_1 be non-degenerate cumulative distribuant function. Then for all $x \in \mathbb{R}$*

$$F_1(x) = F_1(ax + b) \Leftrightarrow a = 1,\ b = 0.$$

The extreme value theorem characterises three classes of the limit of cumulative distribution function F as in (6-1). They can be written compactly as a single distribution with cumulative distribution function [G. Coles, 2001]

$$H(z) = \exp\left(-\left(1 + \gamma\frac{z-\mu}{\sigma}\right)^{-1/\gamma}\right), \tag{6-2}$$

defined for $\mu, \gamma \in \mathbb{R},\ 1 + \gamma(z-\mu)/\sigma > 0,\ \sigma > 0$. This distribution is called *generalised extreme value (GEV)* distribution.

Example 6.1. *In the example 2.5 we provide a specific form of Fréchet cumulative distribution function, which in general reads as*

$$F(x) = \exp\left(-\left(1 + \frac{x-m}{s}\right)^{-\alpha}\right),\ x > m,$$

where m is a location parameter and $s > 0$ is a scale parameter. After reparametrization of (6-2) $\sigma = s/\alpha,\ \gamma = 1/\alpha,\ \mu = m$ we obtain exactly the general Fréchet distribution and hence it belongs to GEV family of distributions.

6.1 GEV and GPD

The expression (6-2) provides the asymptotic behaviour of a cumulative distribution function of a data sample. However, in extreme value theory one is usually interested in statistical behaviour of small number of excesses. In this section we derive a cumulative distribution function corresponding to such a behaviour, which is usually referred to as GPD, the Generalized Pareto Distribution family. Joining (6-1) with (6-2) we have

$$F^n(z) \approx \exp\left(- \left(1 + \gamma \frac{z-\mu}{\sigma} \right)^{-1/\gamma} \right).$$

Then from a Taylor expansion of $\log F(z) \approx -(1 - F(z))$ for large values of z one obtains

$$1 - F(u) \approx \frac{1}{n} \left(1 + \gamma \left(\frac{u-\mu}{\sigma} \right)^{-1/\gamma} \right). \tag{6-3}$$

In extreme value modelling one is usually interested in modelling the probability of values $\mathbb{P}(X > u + y \mid X > u)$, where u is some fixed threshold. This can be compactly written using survival functions as in example 2.6 :

$$\mathbb{P}(X > u + y | X > u) = \frac{1 - F(u+y)}{1 - F(u)} = \frac{S(u+y)}{S(u)}, \ y > 0,$$

and using (6-3) can be approximated with

$$\begin{aligned} \mathbb{P}(X > u + y \mid X > u) &\approx \frac{(1/n)[1 + \gamma(u + y - \mu)/\sigma]^{-1/\gamma}}{(1/n)[1 + \gamma(u - \mu)/\sigma]^{-1/\gamma}} \\ &= \left(\frac{1 + \gamma(u-\mu)/\sigma + \gamma y/\sigma}{1 + \gamma(u-\mu)/\sigma} \right)^{-1/\gamma} \\ &= \left(1 + \frac{\gamma y}{\tilde{\sigma}} \right)^{-1/\gamma}, \end{aligned} \tag{6-4}$$

where $\tilde{\sigma} = \sigma + \gamma(u - \mu)$. The survival function of the form (6-4) is usually referred to as a GPD family.

Example 6.2. *For Fréchet distribution with cumulative distribution function as in* (2-7) *after Taylor approximation we have*

$$\mathbb{P}(X > u + x | X > u) \approx \left(1 + \frac{x}{u} \right)^{-1/\alpha}. \tag{6-5}$$

This corresponds to the GPD presented in (6-4) *with $\beta = \gamma$ and $u = \tilde{\sigma}/\beta$. As a remark about this distribution we note that it has heavy tails, i.e. its survival distribution function decrease as a power function.*

Above suggests following framework for modelling of extreme values : consider threshold excesses $y_i := x_i - u \mid x_i > u$. Presented theory implies that y_i's may be seen as realisations of iid random variables sampled from (6-4). We will pursue this approach, also referred to as *Peaks Over Threshold (POT)* approach.

It should be noted that when choosing a threshold k to consider highest $x_{n-k,n}$ values, one is faced with bias-variance trade-off - too low a threshold may violate the asymptotic behaviour of the model which leads to the bias, while with too high a threshold one risks having too few excesses to estimate the model, which leads to high variance.

6.2 GPD vs. Fisher distribution

We worked with a specific parametrization of GPD (2-5), with which we approximated the distribution of the excesses of Fréchet distribution. As our main contribution we consider using Fisher distribution as an approximation of the tail of distribution which outperforms the one of GPD. In this section we justify this choice - GPD is the special case of the Fisher distribution :

$$\text{Fisher}(1, \alpha_2, \beta) = \text{GPD}(\alpha_2, \beta). \tag{6-6}$$

Therefore one is justified to hope that whatever values are to be obtained via GPD, they may be improved via Fisher distribution, even if for the sole reason of greater generality of the latter. On the figure 1 we present plots where we compare $\text{GPD}(\alpha, \beta)$ with $\text{Fisher}(x, \alpha, \beta)$, for 15 evenly distributed different values of $x \in (0, 2)$ and $\alpha \in \{1, 5, 10\},\ \beta \in \{1/2, 1, 2\}$.

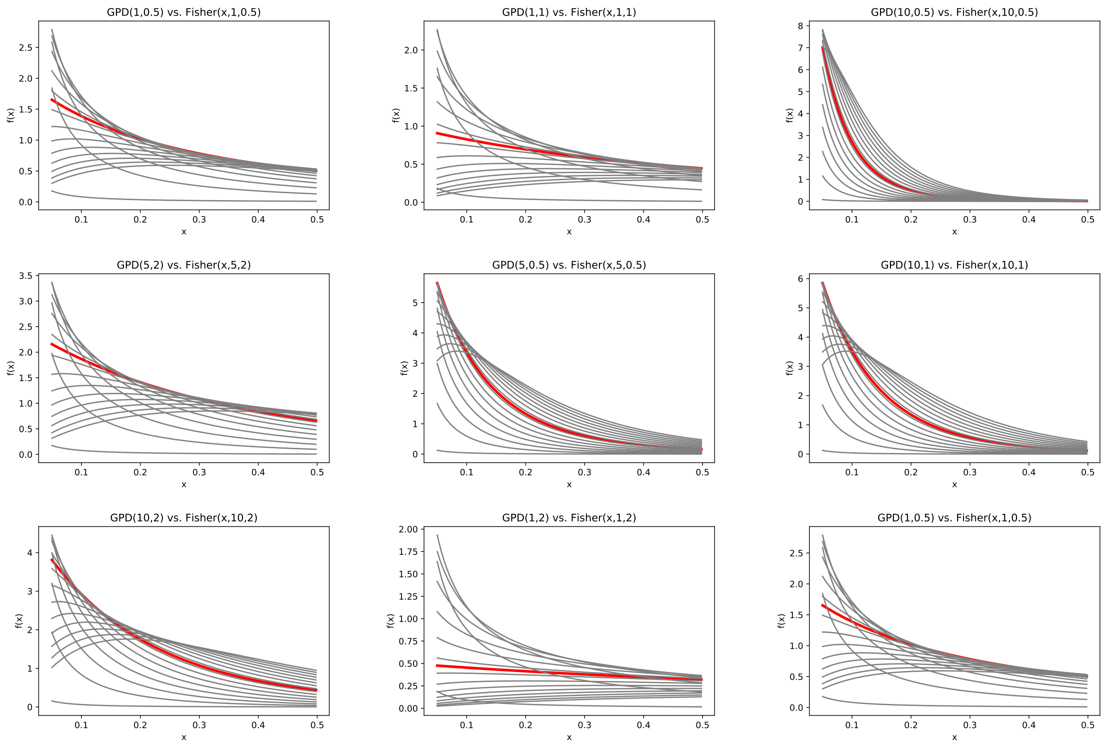

FIGURE 1 – Comparison of GPD(α_2, β) with Fisher$(\alpha_1, \alpha_2, \beta)$, $x \in (0, 2)$. The red line on each plot corresponds to Fisher$(\alpha_1 = 1, \alpha_2, \beta) = $ GPD(α_2, β), the grey lines correspond to Fisher$(\alpha_1 \neq 1, \alpha_2, \beta)$

7 Body

8 Estimation of GEV quantiles

Due to the justification given in section 6.1 we considered POT approach, where only observations x_i exceeding a sufficiently high threshold u_n are taken into account. According to the theorem of [Balkema and de Haan, 1974] as well as and motivations above the probability distribution of the k excesses $y_j = x_{n-j+1,n} - u_n,\ j = 1,\dots,k$ and $u_n = x_{n-k,n}$ may be approximated by GPD(α, β) with cumulative distribution function as presented in (2-6).

The question of inferring the values of quantiles of the underlying distribution comes down to the one of how exactly one is able to estimate the distribution of observed excesses, assuming them to be distributed according to GPD. Parameters of GPD may be estimated in many different ways. Common approaches may be splitted into two mainstreams : Bayesian and non-Bayesian methods. We give description of each streamline, later on we pay more attention to those used in our work. We begin with non-Bayesian methods. In the separate figures we compare (a) non-Bayesian methods, (b) Bayesian methods, (c) the best from (a) and (b). All the plots are in the section 9.

8.1 Non-Bayesian methods

8.1.1 Maximum likelihood estimation

We will present a MLE (see e.g. [Smith, 1987]) for GPD with survival function as in (6-4). From the survival function we read cumulative distribution function to be

$$F_Y(y) = 1 - \left(1 + \frac{\gamma y}{\sigma}\right)^{-1/\gamma},$$

and hence pdf

$$f_Y(y) = \frac{1}{\sigma}\left(1 + \frac{\gamma y}{\sigma}\right)^{-1/\gamma - 1}.$$

Negative log-likelihood of joint pdf for k excesses assuming their independence is given by

$$-\ell(\sigma,\gamma) = -\log\prod_{i=1}^{k} f_Y(y) = k\log\sigma + \left(1 + \frac{1}{\gamma}\right)\sum_{i=1}^{k}\log\left(1 + \gamma\frac{Y_i}{\sigma}\right), \tag{8-1}$$

with the support $\{(\sigma,\gamma) : 1+\gamma Y_i/\sigma > 0,\ \forall i\}$. There is no analytic solution to optimisation in (8-1) ; the common tactic is to reparametrize with $(\tau,\gamma) := (\gamma/\sigma,\gamma)$. Hence one can compute derivative of newly obtained $\tilde{\ell}(\tau,\gamma)$ with respect to γ, which is maximised at $\hat{\gamma} = 1/k\sum_i \log(1+\tau Y_i)$ and then perform numerical optimisation of $\tilde{\ell}(\tau,\gamma=\hat{\gamma})$.

8.1.2 Probability weighted moments

Introduced in [Hosking et al., 1985], Probability Weighted Moments (PWM) are sometimes used when maximum likelihood estimates are unavailable or difficult to compute. They may also be used as starting values for maximum likelihood estimates.

To infer parameters we worked with (6-4) with $\tilde{\sigma} = 1$, in which we identified (2-6) by a change of variables $\sigma = \beta/\alpha,\ \gamma = 1/\alpha$. Probability weighted moments of order $s \in \mathbb{N}$ are given by :

$$\nu_s = \mathbb{E}(Y^s S_{\gamma,\sigma}(Y)), \tag{8-2}$$

where Y are excesses and $S_{\gamma,\sigma}(Y)$ is a sufrvival function of GPD. This algebra may be simplified thanks to

Lemma 8.1. *The PWM of order $s \geq 0$ exists provided that $\gamma < 1$ and is given by*

$$\nu_s = \frac{\sigma}{(s+1)(s+1-\gamma)}. \tag{8-3}$$

Démonstration. Follows from integration by parts integral in (8-2). □

Therefore one may estimate moments having observed excesses Y with (8-2), then using (8-3) extract the values of γ, σ and using GPD approximation of the tail obtain quantiles. In this work we used PWM method to calculate quantiles of GPD which we assumed as an approximation of excesses. The results are presented in section 9.

8.1.3 Method of moments

One more non-Bayesian method which we tried was method of moments (MOM). The advantage of this method with comparison to MLE is its computational simplicity, however the obtained estimators are often biased. We started with computing general expression for k-th moment of Fisher distribution :

$$\mathbb{E}X^k = \int_{\mathbb{R}_+} \frac{1}{\beta B(\alpha_1, \alpha_2)} \frac{(x/\beta)^{\alpha_1 - 1}}{(1 + x/\beta)^{\alpha_1+\alpha_2}} = \beta^k \frac{B(\alpha_1 + k, \alpha_2 - k)}{B(\alpha_1, \alpha_2)},\ k \in (-\alpha_1, \alpha_2).$$

We considered ratio of the form

$$\frac{\mathbb{E}X^\tau}{\mathbb{E}X^{\tau-1}} = \beta \frac{\alpha_1 + \tau - 1}{\alpha_2 - \tau},\ \tau \in (0,1),$$

which was obtained using the relations between beta and gamma functions. Moments were estimated using formula (2-10). We considered $\tau = \{\frac{1}{2}, \frac{3}{4}, 1\}$:

$$\begin{cases} \mathbb{E}X = \beta \dfrac{\alpha_1}{\alpha_2 - 1} \\ \dfrac{\mathbb{E}X^{1/2}}{\mathbb{E}X^{-1/2}} = \beta \dfrac{\alpha_1 - 1/2}{\alpha_2 - 1/2} \\ \dfrac{\mathbb{E}X^{3/4}}{\mathbb{E}X^{-1/4}} = \beta \dfrac{\alpha_1 - 1/4}{\alpha_2 - 3/4}. \end{cases} \tag{8-4}$$

The equation was solved by changing variables $(\alpha_1', \alpha_2', \beta) := (\alpha_1 - 1/2, \alpha_2 - 1/2, \beta)$.

Similarly, to obtain MOM estimates for GPD one needs to substitute $\alpha_1 = 1$ in (8-4) and hence obtain :

$$\begin{cases} \mathbb{E}X = \dfrac{\beta}{\alpha_2 - 1} \\ \dfrac{\mathbb{E}X^{1/2}}{\mathbb{E}X^{-1/2}} = \dfrac{\beta}{2} \dfrac{1}{\alpha_2 - 1/2}. \end{cases} \tag{8-5}$$

We present results in section 9.

8.1.4 Further non-Bayesian methods

We described three non-Bayesian methods we used, others include :

1. estimators based on the elemental percentile method (EPM) [Castillo and Hadi, 1997], which involve intensive computations,
2. another viable alternative is an estimation method based on minimising either a certain distance measure between the empirical and the fitted distribution or minimising the Andersen-Darling statistics. This method is known as the quantile distance estimation (QDE) [Soni et al., 2012].

8.2 Bayesian methods

In this part we pay attention to Bayesian procedures to make statistical inference on the parameters of both Generalised Pareto and Fisher distributions.

From the conditional survival function formulae as in (2-9) we have $S_u(y) = S(y+u)/S(u) \approx S_{GPD}(y)$, where last approximation is justified by the theory developed before. One may change the variables $y := x - u$ and as $S(u) = k/N$ from the fact that u is chosen to be k-positional statistics we have :

$$S_u(x) \approx S(u)S_{GPD}(x-u) = \frac{k}{N}\left(1 + \frac{x-u}{\beta}\right)^{-1/\alpha}.$$

We can easily read the cumulative distribution function given survival function :

$$F_u(x) = 1 - \frac{k}{N}\left(1 + \frac{x-u}{\beta}\right)^{-1/\alpha}$$

and from the quantile definition

$$q(p) = \{\inf x : F_u(x) \geq p\} = u + \beta\left[\left(\frac{N(1-p)}{k}\right)^{-1/\alpha} - 1\right]. \tag{8-6}$$

In this case we were able to recover analytic formula for quantiles, which simplifies the computations. For Fisher distribution after similar steps we end up with non-analytical formulae :

$$q(p) = u + S^{-1}_{Fisher}\left(\frac{N(1-p)}{k}\right). \tag{8-7}$$

We would like to indicate that formulas (8-6) and (8-7) may be used in plenty of ways. Our approach was the following :

1. **plug-in GPD / Fisher** : for each dataset of GEV we fitted GPD/Fisher and plugged in the formulas : (a) averaged, (b) median values of parameters (α, β) in the case of GPD and $(\alpha_1, \alpha_2, \beta)$ in the case of Fisher,
2. **Bayes GPD / Fisher** : for each parameter value reached by MCMC chain we calculated the values of parameters and then took (a) an average, (b) a median over the sum. The advantage of this method over the plug-in method lies in ability to access uncertainty from posterior and that is why we refer to it as a Bayesian method.

The motivation of using median instead of averaging in the third method was to minimise the influence of possible outliers.

9 Discussion

There are two standard ways of assessing goodness of fit of probability distribution : probability and quantile plots, the latter described in definition 2.6 (for more details see [G. Coles, 2001]). In our work we decided to work with quantile plots.

We present results obtained in the extreme value part. For the simulation of Bayesian methods which were performed in Stan software [Stan Development Team, 2015] we used priors as described in section 3.1. We point attention of the reader that for the present moment according to our knowledge there is no digamma function implemented in Stan - hence for the parameters in the Fisher distribution we used $\pi(\alpha_1) \propto \Gamma(2, 2)$, $\pi(\alpha_2) \propto 1/\alpha_2^2$, the Pareto distribution with shape parameter $\alpha = 1$ and $\pi(\beta) \propto 1/\beta$ as in GPD.

The idea was to put the prior as low dependent on the possessed information as possible. This approach may lead to justified doubts and we believe that using an expert knowledge may improve results.

Firstly we present quantiles obtained via non-Bayesian method (figure 2). As one can see, the best values of quantiles are obtained with Fisher MOM method, especially for the more extreme quantiles.

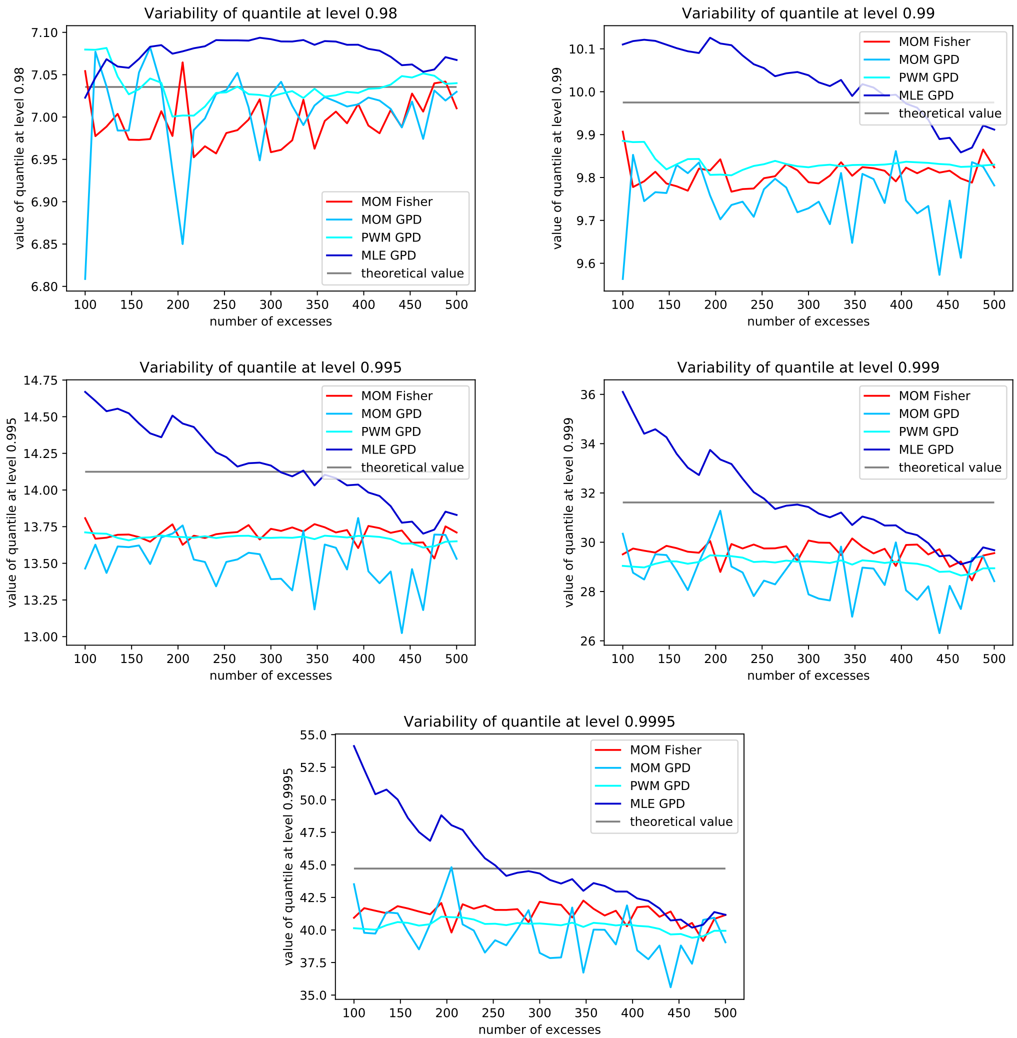

FIGURE 2 – Comparison of variability of non-Bayesian quantiles with $k = 35$ different threshold levels averaged over $N = 30$ datasets with $n = 1000$ samples from Fréchet(1/2).

In the figure 3 we compare fits obtained with Bayesian methods.

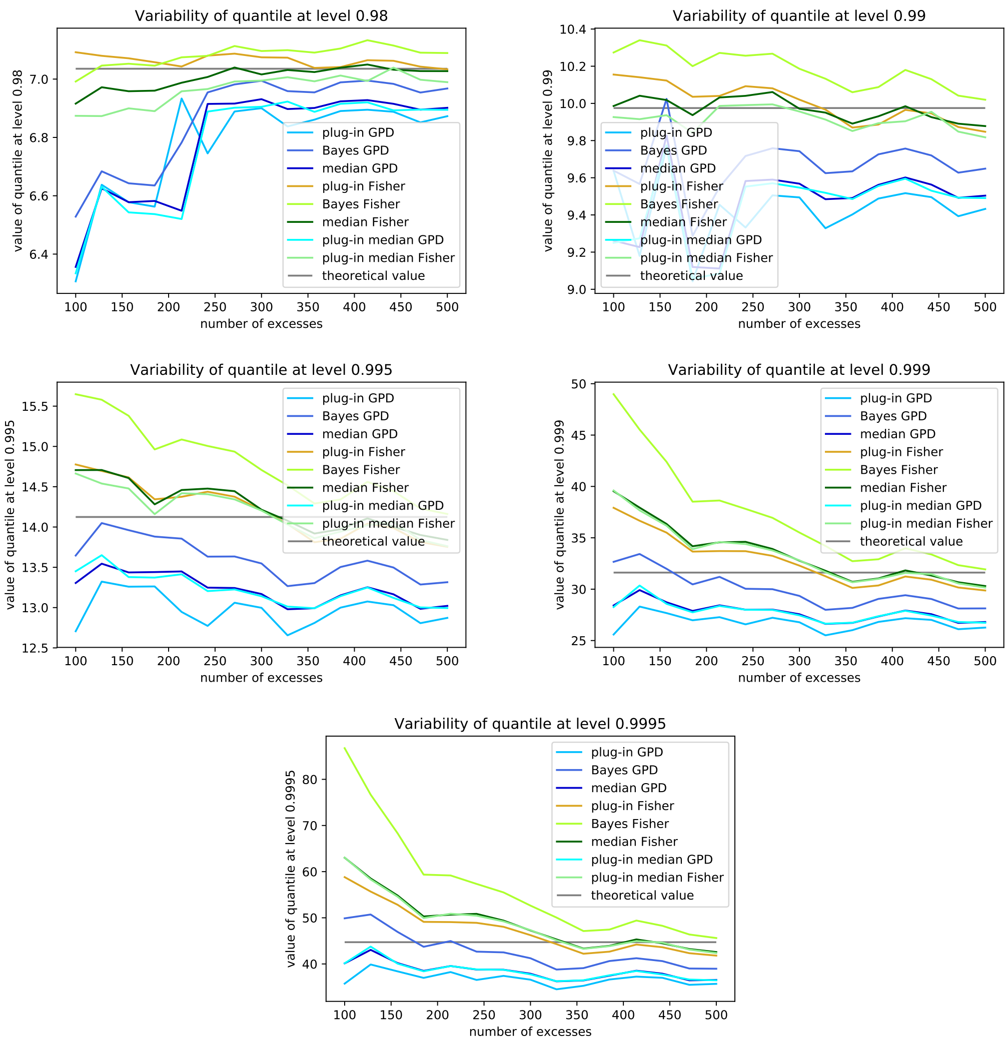

FIGURE 3 – Comparison of variability of Bayesian quantiles with $k = 15$ different threshold levels averaged over $N = 20$ datasets with $n = 1000$ samples from Fréchet(1/2). The MCMC specifications are 500 warm-up runs out of total 2000 runs with 1 chain.

As can be seen on the plot 3 at lowest considered quantiles all the proposed methods perform similarly. This changes when we increase quantile level : the quantiles obtained with Fisher are outperforming those provided by GPD with increasing number of excesses. There is no clear advantage of any given Fisher method over remaining Fishers, and hence as the leading method to compare with non-Bayesian method we chose plug-in median Fisher.

Finally, in the figure 4 we present comparison of best non-Bayesian and Bayesian method. As can be seen, for lowest quantiles the method of moments for Fisher distribution performs better, for the quantile at the level 0.995 there is no clear advantage of any of methods over the other, for the highest quantiles with number of excesses $k > 350$ it is the Bayesian method which is better.

Due to above one may believe that given better priors indicated by an expert knowledge it will be Bayesian method which will outperform the non-Bayesian one.

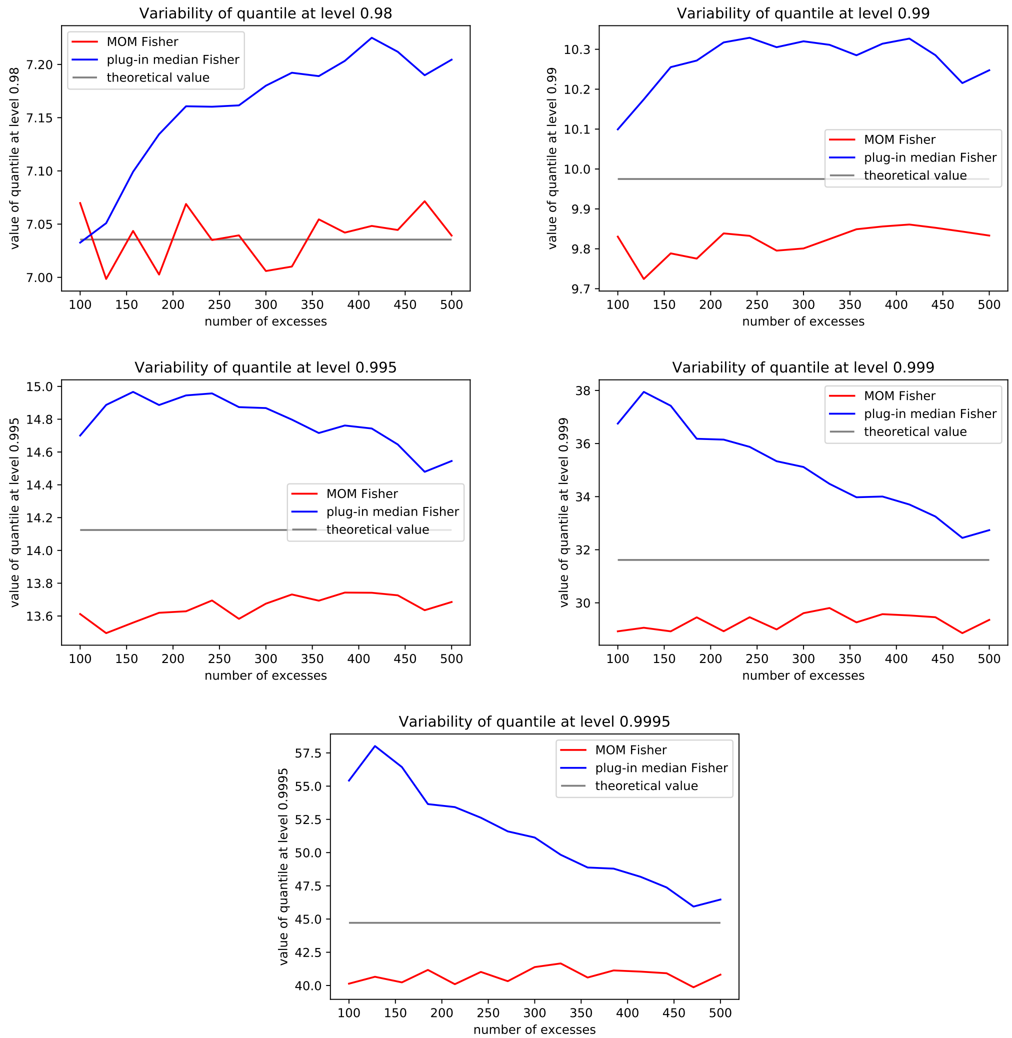

FIGURE 4 – Comparison of best non-Bayesian method MOM Fisher with best Bayesian method plug-in median Fisher. The simulation setup is $k = 15$ different threshold levels averaged over $N = 20$ datasets with $n = 1000$ samples from Fréchet(1/2). The MCMC specifications are 200 warm-up runs out of total 1000 runs with 1 chain.

9.1 Future research

As the quality of obtained quantiles depends on chosen priors, the results may be improved by choosing more adequate priors. Further, we believe it is worthwhile to try to infer the quantiles in the described way of other probability distribution which lies in the domain of attraction of a GEV family, e.g. Burr, log-Gamma distribution.

Références

[Balkema and de Haan, 1974] Balkema, A. A. and de Haan, L. (1974). Residual life time at great age. *Ann. Probab.*, 2(5) :792–804.

[Billingsley, 2012] Billingsley, P. (2012). *Probability and Measure*. Wiley Series in Probability and Statistics. Wiley.

[Castillo and Hadi, 1997] Castillo, E. and Hadi, A. S. (1997). Fitting the generalized pareto distribution to data. *Journal of the American Statistical Association*, 92(440) :1609–1620.

[G. Coles, 2001] G. Coles, S. (2001). *An introduction to statistical modeling of extreme values. - springer*, volume 208. Springer.

[Hosking et al., 1985] Hosking, J. R. M., Wallis, J. R., and Wood, E. F. (1985). Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics*, 27(3) :251–261.

[Jeffreys, 1946] Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London A : Mathematical, Physical and Engineering Sciences*, 186(1007) :453–461.

[Lewandowski, 2018] Lewandowski, M. (2018). Improvements of gpd fitted to gev. `https://github.com/michalmariuszlewandowski/extreme_values`.

[Smith, 1987] Smith, R. L. (1987). Estimating tails of probability distributions. *Ann. Statist.*, 15(3) :1174–1207.

[Soni et al., 2012] Soni, P., Dewan, I., and Jain, K. (2012). Nonparametric estimation of quantile density function. *Computational Statistics & Data Analysis*, 56(12) :3876 – 3886.

[Stan Development Team, 2015] Stan Development Team (2015). Stan : A C++ Library for Probability and Sampling, Version 2.7.0.