Master Project in Mathematical Sciences with Orientation Statistics and Financial Mathematics

Measures of Surprise and Threshold Selection in Extreme Value Statistics

Professor : Anthony C. Davison
Supervisor : Scott Sisson

Irene Vicari

January 15, 2010

Contents

Introduction

Measures of surprise have been recently studied in statistics. This new concept can be used as the first exploratory tool to verify if a model under the null hypothesis fits appropriately. As no alternative models are necessary, the use of the measures of surprise is considered very simple. At the same time, this new alternative to test the goodness of a model cannot replace the full Bayesian analysis.

The aim of this project is threshold selection for threshold models. The estimate of the threshold could be investigated by using the measures of surprise. The reason is that no alternative models are specified for non-extreme data, for which the distribution in extreme value theory is unknown, and thus the surprise would be a credible tool.

In order to quantify the measures of surprise, predictive marginal likelihoods are computed. The purpose of these quantities is to observe if data are surprising under a given model. For this reason we calculate the p-values with respect to the predictive marginal likelihoods. Section 1 describes all these measures and gives their respectively p-values.

In Section 2 the basic concepts of the extreme value theory are reviewed. Firstly, the generalized extreme value distribution is definite to allow the introduction of the generalized Pareto distribution and its properties. Finally, different methods to estimate the threshold u of a dataset having a generalized Pareto distribution are studied.

In order to analyse the measures of surprise more profoundly, a simulation study is carried out in Section 3 and two particular predictive marginal likelihoods are considered: the posterior and the prior predictive marginal likelihoods. Given different datasets, the aim is to estimate the threshold u using these two measures of surprise.

Three different samples (uniform and generalized Pareto data, gamma and generalized Pareto data, generalized Pareto data only) are generated and the behaviour of the posterior predictive measures of surprise is analysed. Because of numerical reasons, we use the mean of the generalized

Pareto density of each observation rather than the product of the densities. This approximation allows to obtain some interesting results which give the estimation of the true threshold u.

The prior predictive measures of surprise are estimated using the Laplace approximation method. Once more, the likelihood is replaced by the mean of the densities for numerical reasons. For this approach the study is carried out only for the sample generated by uniform and generalized Pareto data.

Finally, we discuss the results and the problems that we have had concerning some computations of the measures of surprise. Furthermore, some suggestions are also presented in order to improve and ride over these computational difficulties.

Chapter 1

Measures of Surprise

"Once a (null) model (or hypothesis) H_0 is formulated and x_{obs} is observed, are data *surprising*?" (Bayarri and Berger, 1997). In Statistics this is a very old question and to answer it we introduce the notion of **measure of surprise**.

Definition 1 (Bayarri and Berger (1997))**.** The **measure of surprise** indicates the need of modification of the model. It gives the incompatibility degree of data with an hypothesized model H_0 without any reference to alternative models.

This means that there is no way to compare the model under the null hypothesis without any other models.

The use of the measure of surprise is considered extremely interesting because it is very simple. No alternative models with their priors over their parameters exist. A "surprise" analysis cannot replace a full Bayesian one but it plays an important role as exploratory tool. This means that if the data x_{obs} can be explained by H_0, we might not need to carry out also the full analysis which corresponds to compare the null model with different alternative models with their associated priors over their parameters. On the other hand, if x_{obs} is "surprising", then we have to indicate an alternative model to H_0 and we have to carry out a Bayesian analysis without rejecting directly the model under the null hypothesis.

Under it we usually have $X \sim f(x \mid \theta)$ and $\theta \sim \pi(\theta)$ but since there is no explicit H_1, no prior is assigned. Once we introduce an alternative model, we have $X \sim f_1(x \mid \eta)$ and $\eta \sim \pi_1(\eta)$.

We often use a statistic $T(X)$ to investigate the compatibility of the model under H_0 with the observed data. If we know the parameter θ, the

distribution under H_0 will be $f(t \mid \theta) = f(t)$. The best-known way to measure the compatibility of the model is p-values or tail area probabilities defined by Bayarri and Morales (2003) as follows,

$$p = \Pr^{f(\cdot)}\{T(X) \geq T(x_{\text{obs}})\}.$$

On the other hand, it is extremely rare to know the parameter θ. Therefore, in the next sections we will focus different kind of probability distributions used to compute the p-values.

1.1 Weaver's and Good's Surprise Indices

The surprise index is based on the probability $f(x_{\text{obs}})$ of observing data that eventually occurred. Weaver underlined that a small probability is not necessary surprising unless it is small if compared with the probability $f(x)$ of the other possible results (Weaver (1948) and Weaver (1963)).

The basic idea consists in the comparison between $f(x_{\text{obs}})$ and the average (expected) probability. Let X be a random variable or vector having a discrete distribution. Let $x_1, x_2, \ldots$ have probabilities $f_1, f_2, \ldots$ respectively. Then the **surprise index** associated with the observed value x_{obs} is

$$\lambda_1 = \frac{E\{f(X)\}}{\Pr\{X = x_{\text{obs}}\}} = \frac{\sum_i f_i^2}{f_{\text{obs}}},$$

where $\sum_i f_i^2$ corresponds to the Gini's homogeneity index (Good, 1988).

The surprise index generalized to continuous random variables is

$$\lambda_1 = \frac{E\{f(X)\}}{\Pr\{X = x_{\text{obs}}\}} = \frac{\int f(x)^2 dx}{f(x_{\text{obs}})}. \tag{1.1}$$

We notice that the **Weaver's index** (1.1) is multiplicative. This means that if X and Y are independent random variables, then

$$\lambda_1(x_{\text{obs}}, y_{\text{obs}}) = \lambda_1(x_{\text{obs}})\lambda_1(y_{\text{obs}}). \tag{1.2}$$

When we use Weaver's surprise index, two possible difficulties could arise. The first one concerns its invariance: it is invariant only under linear transformations. The second one refers to the standard chosen to compare the observed $f(x_{\text{obs}})$ with its expected value $E\{f(X)\}$ which might be considered somewhat arbitrary. A single-parameter generalization of (1.1) is suggested by Good (1953) and Good (1956) that would also possess the property (1.2).

These measures of surprise compare $f(x_{\text{obs}})$ with some sort of geometric expectation. Two different cases are considered. First, for $c > 0$ the index is

$$\lambda_c = \frac{[E\{f(X)^c\}]^{1/c}}{f(x_{\text{obs}})}. \tag{1.3}$$

Second, the limiting case, as $c \to 0$ gives

$$\lambda_0 = \frac{\exp[E\{\log(f(X))\}]}{f(x_{\text{obs}})}. \tag{1.4}$$

Notice that for $c = 1$, equation (1.3) corresponds to Weaver's index.

Another generalization has been proposed by Good (1988),

$$\lambda_0 = \frac{\phi^{-1}[E\{\phi(f(X))\}]}{f(x_{\text{obs}})}.$$

In this case ϕ is a monotonic increasing function that is multiplicative only in the case in which ϕ is a power or logarithm (so that it reduces to either (1.3) or (1.4)).

If an additive index is required, it could be possible to use the **logarithmic surprise index**. This was proposed by Good (1956) using the logarithm of (1.3),

$$\Lambda_c = \log(\lambda_c), \quad c \geq 0.$$

This index has many connections with information theory. In particular, $\Lambda_c + \log\{f(x_{\text{obs}})\}$ is also called **Renyi's generalized entropy** (Renyi, 1961). Then, we have

$$\begin{aligned} \Lambda_1 &= \log[E\{f(X)\}] - \log\{f(x_{\text{obs}})\} \\ \Lambda_0 &= E[\log\{f(X)\}] - \log\{f(x_{\text{obs}})\}. \end{aligned}$$

Measures λ_1, λ_0 and Λ_1, Λ_0 are considered to be the most "natural" by Good basing on properties of the expected indices of surprise before the experiment is performed.

We need to detail the distribution of the observations under H_0 if we want to compute these indices. Unfortunately it is not always possible. Then we introduce tail areas or Bayesian p-values which allow to compute the measure of surprise, as we work on suitable predictive distributions.

1.2 Prior Predictive p-values

Under H_0 data are distributed as $X \sim f(x \mid \theta)$ and the prior distribution is $\theta \sim \pi(\theta)$. Then the **prior predictive distribution** for Bayesians is

$$m(x) = \int f(x \mid \theta)\pi(\theta)d\theta, \tag{1.5}$$

which is the natural tool to quantify surprise. Equation (1.5) corresponds to the probability of observing data x. This means that a small value of $m(x)$ would indicate data that are unlikely to be observed. If the observation x_{obs} produces a "small" $m(x_{\text{obs}})$, then there is evidence of "surprise".

In order to understand how small is $m(x_{\text{obs}})$, we have to compare it with some standard. For example, we compare $m(x_{\text{obs}})$ with some "possible" $m(x)$ (see Section 1.5). Box (1980) proposed to compute the associated tail area of $m(x_{\text{obs}})$ in the prior predictive $m(x)$ to measure the smallness of $m(x_{\text{obs}})$. He defined

$$\alpha = \Pr^{m(\cdot)}\{m(X) < m(x_{\text{obs}})\}$$

as an overall predictive check of a given model, where the probability is computed with respect to the prior predictive distribution (1.5). So we can use $1 - \alpha$ or $1/\alpha$ as measures of surprise. In the same way, we can compute the surprise for some functions $D(x_{\text{obs}})$ by

$$\Pr^{m(\cdot)}\{m(D(X)) < m(D(x_{\text{obs}}))\}. \tag{1.6}$$

As these measures of surprise are very close to classical p-values, they violate the conditionality principle and the likelihood principle, too. These probabilities are also based on values of X that provide a much stronger evidence against the null model than the observed one, so we obtain an exaggerated measure of surprise. Another negative feature is that the prior predictive p-values are not invariant under one-to-one transformation (see the example of Evans (1997)).

To remove some of these difficulties, it is necessary to use directly a statistic $T = T(X)$ to compute the p-value (Bayarri and Berger, 2000). The most natural and simple T statistic for the prior predictive is $T(X) = 1/m(X)$. Thus, the **prior predictive p-value** can be written as

$$p_{\text{prior}} = \Pr^{m(\cdot)}\{T(X) \geq T(x_{\text{obs}})\}, \tag{1.7}$$

which is more used than (1.6) and which is invariant under one-to-one transformation.

We notice that $m(X)$ measures the likelihood of x relative to both the model and the prior. Therefore we could get an excellent model where the prior is of poor quality because we often use non-informative prior for the parameters. Unfortunately this prior is often improper, so that the computation of (1.7) will be impossible because also the prior predictive $m(x)$ is improper.

1.3 Posterior Predictive p-values

The **posterior predictive p-values** allow us to compute the p-values for a predictive distribution, and the measure of surprise is defined as

$$m(x \mid x_{\text{obs}}) = \int f(x \mid \theta)\pi(\theta \mid x_{\text{obs}})d\theta, \tag{1.8}$$

where θ has the proper posterior distribution $\pi(\theta \mid x_{\text{obs}})$.

Guttam (1967) was the first to propose this measure of surprise based on posterior predictive distribution to check a model. The idea is based on a comparison between the observed empirical frequencies in a partition of the sample space with the "theoretical" frequencies computed from the posterior predictive distribution of a future observation. A χ^2 procedure is used for the comparison and the "surprise" is based on p-values.

Rubin (1984) generalized the use of the **posterior predictive p-values** which is based on the use of tail area probabilities corresponding to the observed value of some test statistics $T = T(X)$ as

$$p_{\text{post}} = \Pr\{T(X) \geq T(x_{\text{obs}}) \mid x_{\text{obs}}\}, \tag{1.9}$$

where the probability is computed with respect to the posterior predictive distribution $m(x \mid x_{\text{obs}})$ defined in (1.8).

More studies have been carried out by Meng (1994) and Gelman *et al.* (1996). They replace the statistic $T(X)$ by a function $T(X, \theta)$. Furthermore, $f(x \mid \theta)$ used in equation (1.8) becomes $f(x \mid \theta, A)$, where A is an "auxiliary" statistic. So, the posterior predictive p-value has the following form,

$$p_{\text{post}} = \Pr\{T(X, \theta) \geq T(x_{\text{obs}}, \theta) \mid x_{\text{obs}}, A(x_{\text{obs}})\},$$

where the probability is computed with respect to the joint distribution

$$\Pr\{\theta, x \mid x_{\text{obs}}, A(x_{\text{obs}})\} = f(x \mid \theta, A(x) = A(x_{\text{obs}}))\pi(\theta \mid x_{\text{obs}}).$$

We use posterior predictive distributions to compute tail areas and we obtain posterior predictive p-values when no alternative models exist. This

method has some problems which are similar to those of Aitkin (1991), who needs posterior predictive distributions to calculate Bayes factors in the presence of alternative models.

Unlike the prior predictive p-values, improper and non-informative priors can be used to compute posterior predictive p-values since $\pi(\theta \mid x_{\text{obs}})$ will be proper. Furthermore, $m(x \mid x_{\text{obs}})$ will be much more influenced by the model than by the prior. Finally, the posterior predictive p-values are easy to compute using the outputs from Bayesian analyses.

Unfortunately, there are two weaknesses when posterior predictive p-values are computed.

The first weakness concerns the observed data x_{obs}, which are used twice to compute the full posterior predictive distribution $m(x \mid x_{\text{obs}})$: first to modify the improper prior $\pi(\theta)$ into a proper distribution $\pi(\theta \mid x_{\text{obs}})$ and second to measure the surprise in the posterior predictive distribution $m(x \mid x_{\text{obs}})$.

The second weakness is that posterior p-values are very similar to classical p-values, so they have the same inadequacies of the latter. In order to understand this, we look at (1.9). We can rewrite the posterior predictive p-value in the following way,

$$p_{\text{post}} = \int \Pr^{m(\cdot|x_{\text{obs}})}\{T(X) \geq T(x_{\text{obs}})\}\pi(\theta \mid x_{\text{obs}})d\theta, \tag{1.10}$$

so p_{post} corresponds to the expected value of the classical tail probability

$$p_c(\theta) = \Pr\{T(X) \geq T(x_{\text{obs}}) \mid \theta\},$$

with respect to the posterior distribution. For a large sample, we have $p_{\text{post}} \approx p_c(\hat{\theta})$, where $\hat{\theta}$ is the MLE of θ, and then the behaviour of both measures (posterior predictive p-values and classical p-values) will be similar.

1.4 Other Proposals to Measure Surprise

Another interesting measure of surprise is due to Evans (1997), who proposed a measure invariant under one-to-one transformation.

Suppose that $\varphi = \varphi(\theta)$ is a parametric function of interest. Then the **observed relative surprise** for testing the null hypothesis $H_0 : \varphi = \varphi_0$ against each alternative $H_1 : \varphi = \varphi_1$ has been defined by Evans (1997) as

$$\Pr\left\{\frac{\pi(\varphi_1 \mid x_{\text{obs}})}{\pi(\varphi_1)} > \frac{\pi(\varphi_0 \mid x_{\text{obs}})}{\pi(\varphi_0)} \mid x_{\text{obs}}\right\}, \tag{1.11}$$

where the probability is computed with respect to the posterior distribution $\pi(\varphi_1 \mid x_{\text{obs}})$. This measure is invariant under one-to-one transformation given the presence of the Jacobians in both numerator and denominator. The use of (1.11) has been suggested also for estimation (minimizing the observed relative surprise) and for confidence regions ($\alpha-$relative surprise regions) (Evans, 1997).

However, there are two difficulties when we use (1.11) as a measure of surprise. The first, once more, concerns the use of the data twice: once to obtain the ratio of the posterior to the prior and the second to compute the probability that this ratio is larger than its hypothesized value. This problem can be related to Aitkin's posterior Bayes factors (Aitkin, 1991). The probability given in (1.11), also called **Evans' relative surprise**, can be rewritten as follows,

$$\Pr^{\varphi_1|x_{\text{obs}}}\left\{\frac{f(x_{\text{obs}} \mid \varphi_1)}{f(x_{\text{obs}} \mid \varphi_0)} > 1\right\} = \Pr^{B|x_{\text{obs}}}\{B > 1\}.$$

The expected value of this distribution

$$\frac{\int f(x_{\text{obs}} \mid \varphi_1)\pi(\varphi_1 \mid x_{\text{obs}})d\varphi_1}{f(x_{\text{obs}} \mid \varphi_0)}$$

corresponds to the **Aitkin's posterior Bayes factor** for H_1.

The second difficulty is that we have to assess carefully the alternatives to φ_0 and for each alternative we have to specify a prior distribution.

Evans (1997) proposed to use the surprise to check a model by defining the observed relative surprise in the following way,

$$\Pr\left\{\frac{m(T(X) \mid x_{\text{obs}})}{m(T(X))} > \frac{m(T(x_{\text{obs}}) \mid x_{\text{obs}})}{m(T(x_{\text{obs}}))}\right\}, \tag{1.12}$$

where $m(T(X) \mid x_{\text{obs}})$ is the posterior predictive density of $T(X)$, $m(T(X))$ is the prior predictive density of $T(X)$ and $T(X)$ is a function with a Lebesgue measure on the appropriate space.

Once more there is no invariance. This probability can be also used for prediction. However, if the ratio $m(T(x_{\text{obs}}) \mid x_{\text{obs}})/m(T(x_{\text{obs}}))$ used in (1.12) is very large, it will be not useful to check the model as the measure of surprise will equal 0.

1.5 Relative Maximized and Expected Measures of Surprise

There are other methods to find measures of surprise. Instead of computing p-values, Berger (1980-85) suggested to compare their relative likelihoods. Once more we need the prior predictive distribution $m(x)$ and if $m(x_{\text{obs}})$ is small, then data are surprising. Two different likelihoods have been defined,

$$m^*(x_{\text{obs}}) = \frac{m(x_{\text{obs}})}{\sup_x m(x)}, \tag{1.13}$$

$$m^{**}(x_{\text{obs}}) = \frac{m(x_{\text{obs}})}{E^{m(x)}\{m(X)\}}. \tag{1.14}$$

We notice that (1.14) is the inverse of the index λ_1 given in (1.2) when applied to the prior predictive distribution $m(x)$ and then it has the same properties. On the contrary, as $c \to 0$, we have that (1.13) is the limiting case of the inverse of (1.3).

Measure of surprise m^* has a property that is related with the robust Bayes approach. This approach has a natural measure of surprise in the infimum of Bayes factors derived from Bayesian global robustness analyses. If we accept to approximate H_1 by defining π_1 as a prior belonging to a large class of priors, then the infimum of the Bayes factor in favour of H_0 corresponds to the natural measure of surprise. The model for H_1 is defined as $f(x \mid \theta, \xi)$ and the marginal prior distribution $\pi(\theta)$ is the same under both hypotheses. Then we have that $\pi(\theta, \xi) = \pi_1(\xi)\pi(\theta)$, where $\pi_1 \in \Gamma$ and Γ is the class of all priors π_1 for the alternative values ξ. The **lower bound on the Bayes factor of H_0 to H_1** is

$$\underline{B} = \inf_{\pi_1 \in \Gamma} \frac{\int f(x \mid \theta)\pi(\theta)d\theta}{\int\int f_1(x \mid \theta, \xi)\pi(\theta)\pi_1(\xi)d\theta d\xi} \tag{1.15}$$

and data x_{obs} resulting in small $\underline{B}$ would be considered surprising.

Let H_0 be simple, without considering θ. Then the infimum of (1.15) becomes

$$\underline{B} = \frac{f(x_{\text{obs}})}{\sup_\xi f_1(x_{\text{obs}} \mid \xi)}.$$

We have the same problem for these measures of surprise: there is no invariance under non-linear, one-to-one transformation. Furthermore, if the dimension or the number of observations n is large, then it will be difficult to explain these values.

In more than one Bayesian situation we have seen that taking supremum or expectations over large spaces is not a good idea. This is underlined principally when measure m^* is applied to data x that are independent under H_0. Then we have as $n \to \infty$

$$m^*(x) = \prod_{i=1}^{n} \frac{f(x_i)}{f(x_{\max})} \to 0, \quad \text{with probability 1,}$$

even when the data come from the correct model.

In order to reduce and remove the problem of non-invariance and the impact of high dimensions, we introduce a "natural" statistic T, whose purpose is to measure the "distance" between the observations and the null hypothesis and apply m^* and m^{**} to its predictive distribution. The choice of T has to be done carefully and as we have already seen, the most evident difficulty to overcome is the lack of invariance. Therefore, Bayarri and Berger (1997) suggested that it is better to look for an appropriate alternative hypothesis rather than to get a statistic T, so that we can carry out a Bayesian analysis.

1.6 Conditional Predictive Distribution

In the previous sections we have seen that two difficulties arise when we use the prior predictive distribution (1.5). The first concerns the use of an improper prior or not well-defined proper prior $\pi(\theta)$. The second refers to the impossibility of separating the surprise in the model and in the prior.

We notice that sometimes also the use of a statistic T does not give a solution to the problem.

An attractive solution is conditioning on an appropriate statistic U as proposed by Bayarri and Berger (1997) so that we will achieve all the advantages of the prior and posterior predictive p-values in the same procedure. The most important features are the following ones. First, these p-values are based on the prior predictive distribution $m(x)$, which has a natural Bayesian meaning. Second, if we choose the statistic U appropriately, the prior has a secondary role. Third, if $\pi(\theta)$ is proper, the prior can be also non-informative. Finally, the data are not used twice.

A **conditional predictive** $m(t \mid u)$ is obtained for the statistic T defined previously and is

$$m(t \mid u) = \int f(t \mid u, \theta)\pi(\theta \mid u)d\theta, \tag{1.16}$$

where $\pi(\theta \mid u) = f(u \mid \theta)\pi(\theta) / \int f(u \mid \theta)\pi(\theta)d\theta$.

Since an improper prior is used, we have to choose the statistic U so that $\pi(\theta \mid u)$ is proper, so that $m(t \mid u)$ will be also a proper distribution. If we compare the conditional predictive p-value to the posterior predictive p-value, the data are not used twice: the part of the data represented by U will be used to eliminate the nuisance parameter and the part represented by T will be used to measure the surprise.

The separation of the effects of the model inadequacy and prior inadequacy can also be reduced if we choose an appropriate U.

Once we get the conditional predictive distribution, we can use it in any of the surprise measures explained in the previous sections. The relative measures of surprise (1.13) and (1.14) become

$$m^*(t_{\text{obs}} \mid u_{\text{obs}}) = \frac{m(t_{\text{obs}} \mid u_{\text{obs}})}{\sup_t m(t \mid u_{\text{obs}})}, \tag{1.17}$$

$$m^{**}(t_{\text{obs}} \mid u_{\text{obs}}) = \frac{m(t_{\text{obs}} \mid u_{\text{obs}})}{E^{m(t|u_{\text{obs}})}\{m(T \mid u_{\text{obs}})\}}. \tag{1.18}$$

The **conditional predictive p-value** is

$$p_{\text{cond}} = \Pr^{m(\cdot|u_{\text{obs}})}\{T(X) \geq T(x_{\text{obs}})\}, \tag{1.19}$$

where $T(x_{\text{obs}}) = t_{\text{obs}}$.

In the next paragraphs we detail different choices for the statistic U.

1.6.1 Through one-to-one Transformation of X

Let (T, X^*) be a one-to-one transformation of X. Then we can take $U = X^*$, where $\dim U = n - \dim T$. This means taking "the rest" of the data concerning T for the statistic U. This is the easiest and the most evident choice because it is not difficult to implement. We obtain $m(t, u)$ from $m(x)$. Then we compute the measure of surprise (1.17) multiplying by the Jacobians,

$$m^*(t_{\text{obs}} \mid u_{\text{obs}}) = \frac{m(t_{\text{obs}} \mid u_{\text{obs}})}{\sup_t m(t \mid u_{\text{obs}})} = \frac{m(t_{\text{obs}}, u_{\text{obs}})}{\sup_t m(t, u_{\text{obs}})}, \tag{1.20}$$

so that $m(t \mid u)$ does not have to be derived. Since $m(t \mid u)$ is proper and the constants cancelled, we can always use this method even though $m(x)$ would usually be improper.

The **partial posterior predictive p-value** (Bayarri and Berger, 1999) is

$$p_{\text{part}} = \Pr^{m^*(\cdot)}\{T \geq t_{\text{obs}}\}, \tag{1.21}$$

where

$$\begin{aligned} m^*(t) &= \int f(t \mid \theta)\pi^*(\theta)d\theta, \\ \pi^*(\theta) &\propto f(x_{\text{obs}} \mid t_{\text{obs}}, \theta)\pi(\theta) \propto \frac{f(x_{\text{obs}} \mid \theta)\pi(\theta)}{f(x_{\text{obs}} \mid t_{\text{obs}})}. \end{aligned}$$

In this case the double use of the data is removed because the contribution of t_{obs} to the posterior is cancelled out before θ is eliminated by integration.

Some examples given by Bayarri and Berger (1997) show that there is an effect of "too much conditioning". This phenomenon could be reduced if we find a suitable "orthogonal" transformation so that we have independence between T and X^*. The choice of $U = X^*$ might be quite appropriate.

A natural choice of U is often a statistic of the same dimension as θ, because we must take the dimension of U bigger or equal to the dimension of θ in order that $\pi(\theta \mid u)$ be proper.

1.6.2 Ideal Choice of U

Having sufficient statistics (T, U) of low dimension and conditionally independent corresponds to the ideal situation. In this case we have

$$m(t \mid u) = \int f(t \mid \theta)\pi(\theta \mid u)d\theta.$$

The data are used twice: once with independent pieces of it in order to learn about the nuisance parameter and once to detect surprising features.

1.6.3 Asymptotic Independence

It could be difficult to have independence between T and U. For this reason we look for an U that is asymptotically independent of T under some regularity conditions. To be more precise, we choose U such that

$$\begin{pmatrix} T \\ U \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} m \\ \theta \end{pmatrix}, \Sigma\right),$$

where Σ is a block diagonal matrix. We can sometimes choose U as the MLE $\hat{\theta}$ of some linear transformations of $(T, \hat{\theta})$. Unfortunately this idea is not as good as hoped because once more U results in "too much conditioning".

1.6.4 Sufficiency for the Nuisance Parameter

The need to learn about the nuisance parameter θ is the reason for conditioning on some statistic U. One proposal is to choose U as a sufficient statistic for θ. In this case $f(x \mid u, \theta) = f(x \mid u)$ does not involve θ. Furthermore, we have that $m(t \mid u)$ is given by $f(t \mid u)$ and as θ is not involved, no prior is needed.

1.6.5 An Attractive Choice of U

Information in the data and in T are used to find a suitable conditioning statistic U to eliminate θ. The distribution $f(x \mid t, \theta)$ is very interesting because it removes the information provided by T from the likelihood for θ.

Taking U as a low-dimensional sufficient statistic of this conditional distribution is not always possible because sufficient statistics may not exist. On the contrary we choose an approximate sufficient statistic with the same dimension as θ, so that we are sure of its existence, and we can define it as follows,

$$U = \hat{\theta} = \text{ arg max } f(x \mid t, \theta) = \text{ arg max } \frac{f(x \mid \theta)}{f(t \mid \theta)} \quad \text{for } T(x) = t. \quad (1.22)$$

1.6.6 Computational Issues

Numerical computations are usually necessary to obtain surprise measures. In Bayesian analysis inference is based on samples which are generated from the target distribution via MCMC methods. We develop the computations for T and U having dimension 1.

If we do not know the conditional predictive distribution $m(t \mid u_{\text{obs}})$ but we have a simulated sample $x_1, \ldots, x_M$ of size M from $m(x \mid u_{\text{obs}})$, it is easy to calculate:

- p-values:

$$\Pr\{T(X) \geq t_{\text{obs}} \mid u_{\text{obs}}\} = \frac{\#\{T(x_i) : T(x_i) \geq t_{\text{obs}}\}}{M},$$

- relative maximized surprise:

$$m^*(t_{\text{obs}} \mid u_{\text{obs}}) = \frac{\#\{T(x_i) :\mid T(x_i) - t_{\text{obs}} \mid < \epsilon\}}{\max \#\{T(x_i) : T(x_i) \in (T(x_i) - \epsilon, T(x_i) + \epsilon)\}},$$

- relative expected surprise:

$$m^{**}(t_{\text{obs}} \mid u_{\text{obs}}) = \frac{\#\{T(x_i) :\mid T(x_i) - t_{\text{obs}} \mid < \epsilon\}}{\sum_{j=1}^{M} \#\{T(x_i) :\mid T(x_i) - T(x_j)/M \mid < \epsilon\}}.$$

These computations can also be applied when the measure of surprise is obtained from $m(x)$.

We simulate a sample of $m(x \mid u_{\text{obs}})$ using one of the following algorithms: the first based on a Gibbs scheme and the second based on the Metropolis–Hastings approach (Robert and Casella, 2005).

In order to use both of them, we need to know an explicit expression for U. The sample is generated from $m(x \mid \ \mid u - u_{\text{obs}} \mid < \delta)$ and not directly from $m(x \mid u_{\text{obs}})$. If δ is small, we will have that $m(x \mid \ \mid u - u_{\text{obs}} \mid < \delta)$ is an approximation to $m(x \mid u_{\text{obs}})$. Otherwise if δ is large, the computations will be faster and there will be less conditioning than that one provided by u_{obs}. Furthermore, if $\delta \to \infty$, we will have $m(x \mid \ \mid u - u_{\text{obs}} \mid < \delta) \to m(x)$, which corresponds to the prior predictive distributions.

We can rewrite $m(x \mid \ \mid u - u_{\text{obs}} \mid < \delta)$ as follows:

$$\begin{aligned} m(x \mid \ \mid u - u_{\text{obs}} \mid < \delta) &= \int f(x, \theta \mid \ \mid u - u_{\text{obs}} \mid < \delta) d\theta \\ &= \frac{\int f(x \mid \theta)\pi(\theta)\mathcal{I}_{\{\mid u - u_{\text{obs}} \mid < \delta\}} d\theta}{\Pr\{\mid u - u_{\text{obs}} \mid < \delta\}}, \end{aligned}$$

where the denominator is a constant and therefore is not relevant to both algorithms.

Gibbs Sampler

Gibbs Sampler chain is based on the following steps (Bayarri and Berger, 1999):

1. Generate $\theta \sim \pi(\theta \mid x)$.
   We notice that the generation comes from the posterior distribution.
2. Generate $X \sim f(x \mid \theta)\mathcal{I}_{\{\mid u - u_{\text{obs}} \mid < \delta\}}$.
3. After many iterations of Steps 1 and 2, estimate p by the fraction of the generated x's for which $T(x)$ is greater than $T(x_{\text{obs}})$. This means that the chain is built only for "surprise" evaluations.

Metropolis–Hastings Algorithm

This algorithm (Bayarri and Berger, 1999) generates a chain (x_j, θ_j) through the following steps. First of all, we define the proposal as

$$f(x \mid \theta)\pi(\theta \mid x_{\text{obs}})\mathcal{I}_{\{|u-u_{\text{obs}}|<\delta\}}. \tag{1.23}$$

Then, from (x_t, θ_t) at time t,

1. Generate a candidate (x^*, θ^*) from the proposal (1.23) by taking $\theta \sim \pi(\theta \mid x_{\text{obs}})$, simulating $x \sim f(x \mid \theta)$ and repeating this procedure until the distance between $u(x)$ and u_{obs} is less than δ. If $u(x)$ is not within δ of u_{obs}, a new θ has to be generated from $\pi(\theta \mid x_{\text{obs}})$.

2. Accept the candidate with probability

$$\alpha = \min\left\{1, \frac{\pi(\theta^*)}{\pi(\theta^* \mid x_{\text{obs}})}\frac{\pi(\theta_t \mid x_{\text{obs}})}{\pi(\theta_t)}\right\} = \min\left\{1, \frac{f(x_{\text{obs}} \mid \theta_t)}{f(x_{\text{obs}} \mid \theta^*)}\right\}.$$

3. After sufficiently many iterations of Steps 1 and 2, the estimate p by the fraction of the generated x_j in the chain for which $T(x_j)$ is greater than $T(x_{\text{obs}})$.

If U has an explicit form, we can easily implement both schemes. Otherwise, if U is defined as in (1.22), we need more computations. In the first case $f(t \mid \theta)$ is known and to obtain U we have a numerical maximisation to compute from x. The second case is more complicated because the closed form of $f(t \mid \theta)$ is not available. Therefore, we have to implement an algorithm which computes $u = u(x^*)$ for a given x^* and $t^* = T(x^*)$. Three steps are required (Bayarri and Berger, 1997):

1. Take a grid of θ values.

2. For each θ generate a sample x_i from $f(x \mid \theta)$ and compute

$$r(\theta) = \frac{f(x^* \mid \theta)}{\hat{f}(t^* \mid \theta)},$$

where $\hat{f}(t^* \mid \theta)$ is some estimate of the density $f(t^* \mid \theta)$. The crudest estimate is

$$\hat{f}(t^* \mid \theta) = \frac{\#\{T(x_i) :\mid T(x_i) - t \mid < \epsilon\}}{2M\epsilon},$$

thought we could use a more sophisticated kernel estimator.

3. Take u as the value of θ maximizing $r(\theta)$ over the grid.

We need only the values of u so that the distance between u and u_{obs} is less than δ. So, once we have computed the values u_{obs}, we only need a grid of values of θ such that $\mid \theta - u_{\text{obs}} \mid < \delta$ and we have to look if $\max r(\theta)$ occurs in this grid.

We have seen that measures of surprise based on likelihood ratios are more in accord with Bayesian reasoning than ones based on tail areas or p-values. But tail areas are easier to compute; they do not change under one-to-one transformations and they can be applied to discrepancy measures. Therefore it is advisable to compute tail area of the observed $T(x_{\text{obs}})$ in the predictive distribution $m(t \mid u)$. On the contrary p-values are analysed in the next section, as they are highly misleading measures of evidence against H_0.

1.7 Calibration of p-values

It is well-known that there are many difficulties to interpret p-values. For this reason, in this section we investigate the possibility of developing an adjustment to the p-value. A possibility is to calibrate the p-value such that it will be closer to an infimum of Bayes factors (see Equation (1.15)). The proposal for calibrating a p-value is to compute

$$\underline{B} = -ep \log p, \quad p < e^{-1}, \tag{1.24}$$

and interpret this as a lower bound on the Bayes factor of H_0 to H_1. For this purpose we need to consider alternative models to the null one.

Let $f(x)$ be the model under the null hypothesis and recall that for surprise purposes we usually define $f(x)$ like $m(t \mid u)$. As the alternative model is usually larger than the null model, it will be denoted as $f(x \mid \xi)$ while the null model will be $f(x) = f(x \mid \xi_0)$, where ξ and the fixed ξ_0 denote the parameters of the alternative and the null models respectively. Let the p-value be $p = p(x_{\text{obs}})$ (Bayarri and Berger, 1997), where

$$p(x) = \int_x^{\infty} f(z \mid \xi_0) dz. \tag{1.25}$$

Furthermore, we compute the measure of surprise

$$m(x) = \int f(x \mid \xi) \pi(\xi) d\xi \tag{1.26}$$

in order to obtain the Bayes factor in favour of ξ_0 given by

$$B = \frac{f(x_{\text{obs}} \mid \xi_0)}{m(x_{\text{obs}})}.$$

Let define the **hazard rate** or **failure rate** function of the null model,

$$h_0(x) = \frac{f(x)}{1 - F(x)} = \frac{f(x \mid \xi_0)}{\int_x^\infty f(y \mid \xi_0) dy}.$$

An attractive approach to compute the Bayes factor is suggested by Sellk *et al.* (2001) which consists in directly considering alternative distributions for p itself and the uniform distribution for the null hypothesis. This means that we have to test

$$H_0 : p \sim U(0,1) \text{ versus } H_1 : p \sim f_p(p \mid \xi).$$

This is equivalent to compute the infimum of the Bayes factor in favour of the null hypothesis. A possible class of alternatives for p is the class of $Be(\xi, 1)$ distributions, where $0 \leq \xi \leq 1$, so that the distributions are decreasing:

$$f(p \mid \xi) = \xi p^{\xi - 1} = \frac{\xi}{p^{1-\xi}}. \tag{1.27}$$

It is suitable to work with $Y = -\log p$ and its distributions under H_0 and H_1. By a simple computation, if $p \sim Be(\xi, 1)$, then we have

$$\Pr\{Y > y\} = \Pr\{p < e^{-y}\} = e^{-\xi y},$$

that is $Y \sim \exp(\xi)$. In both cases, the null hypothesis is obtained for $\xi = 1$. Therefore, the infimum of the Bayes factor over all priors for ξ is

$$\underline{B} = \begin{cases} \inf_{\text{all } \pi_1} \frac{f(y|\xi)}{\int f(y|\xi)\pi_1(\xi)d\xi} = \frac{\exp(y|1)}{\sup_\xi \exp(y|\xi)} = ye^{1-y}, & y \geq 1, \\ 1, & \text{otherwise.} \end{cases} \tag{1.28}$$

Substituting $p = e^{-y}$ in the lower bound (1.28) we have the calibrating p-value given in equation (1.24). This calibration assumes that the alternative models and the priors are such that the distribution of $Y = -\log p$ is exponential, that is it has a constant failure rate. In order to relax this assumption but at the same time to still require that the distribution of p should decrease sufficiently fast so that most of the mass will be close to 0, we need that the distribution of Y has a decreasing failure rate. This is equivalent to requiring that the distribution $Y - y \mid Y > y$ is stochastically

increasing with y. In a similar way, for $p = e^{-y}$, this requirement of decreasing failure rate is equivalent to say that the distribution of $p/p_0 \mid p < p_0$ is stochastically decreasing with p. This means that, for any fixed p_0 and ρ, the probability $\Pr\{p < \rho p_0 \mid p < p_0\}$ increases as p_0 goes to zero. This corresponds to the natural condition implying that the mass under the alternative is appropriately concentrated near zero.

We have to show that the Bayesian factor for p is still valid when we suppose that the distribution of Y has a decreasing failure rate. The failure rate function of the distribution of Y is defined as follows,

$$h_1(y) = \frac{f_1(y)}{\int_y^\infty f_1(z)dz}$$

and according to the alternative model f_1 has a decreasing failure rate. This function f_1 can be written as

$$f_1(y) = h_1(y)\exp\{-\int_0^y h_1(z)dz\} \geq h_1(y)\exp\{-yh_1(y)\}.$$

In this case the infimum of the Bayes factor of H_0 to H_1 is

$$\underline{B} = \begin{cases} \frac{e^{-y}}{f_1(y)} \geq \frac{e^{-y}}{h_1(y)\exp\{-yh_1(y)\}} \geq ye^{1-y}, & y \geq 1, \\ 1, & \text{otherwise.} \end{cases}$$

It is simple to verify the decreasing failure rate for the distribution of Y when the alternative model and the prior have been already assessed. First of all, we assume under H_0 that $X \sim f(x)$ and under H_1 that $X \sim m(x)$, where $m(x)$ corresponds to the Bayesian marginal or the predictive density defined in (1.5). Let F and M denote their probability distributions, respectively. Knowing that the p-value under H_0 is given by (1.25), we compute the survival function of $Y = -\log\{p(X)\}$ under H_1,

$$\Pr\{Y > y\} = \Pr\{p < e^{-y}\} = 1 - M\{F^{-1}(1 - e^{-y})\} \tag{1.29}$$

and its density has the following form,

$$f_1(y) = \frac{m\{F^{-1}(1 - e^{-y})\}}{e^y f\{F^{-1}(1 - e^{-y})\}}. \tag{1.30}$$

The hazard rate function of Y is given by dividing (1.30) by (1.29) and it is decreasing if and only if

$$\frac{m(x)}{1 - M(x)} \Big/ \frac{f(x)}{1 - F(x)} \tag{1.31}$$

is decreasing, which is equivalent to the ratio of the alternative hazard rate to the null one.

Chapter 2

Threshold Selection for the Generalized Pareto Distribution

Extreme value theory is a statistical discipline that allows to model and study the tail of distributions. Many different approaches exist like the generalized extreme value model, the threshold exceedance model and the point process model. Coles (2001) gives a detailed explanation of all these models.

In this section, we focus our interest on modelling observations above a certain threshold u. More precisely, we study different ways to estimate the threshold for a generalized Pareto distribution. One advantage of this approach is that more data can be considered as extreme events compared to the GEV model which takes only the maximum on each block.

2.1 Generalized Extreme Value Model

First of all we define the generalized extreme values distribution which are necessary when the generalized Pareto distribution will be introduced.

In order to develop the model for extreme value theory we need to know the distribution of

$$M_n = \max\{X_1, \ldots, X_n\},$$

where $X_1, \ldots, X_n$, is a sequence of independent random variables having a common distribution function F. These random variables represent values of a process measured on a regular time-scale, as daily mean temperature. Therefore M_n corresponds to the maximum of process over n time units of

observation. The distribution function of M_n is

$$\Pr\{M_n \leq z\} = F(z)^n,$$

where F is unknown. There are two approaches to estimate F. The first one is based on observed data, applying standard statistical techniques. The second one consists in finding approximate families of models for F^n, which can only be estimated on the basis of extreme data.

The behaviour of F^n as $n \to \infty$ is observed, but it is not sufficient: for any $z < z_+$, $F^n(z) \to 0$ as $n \to \infty$, so that the distribution of M_n degenerates to a point mass on z_+, where z_+ is the upper end-point of F (i.e. z_+ is the smallest value of z such that $F(z) = 1$). In order to overpass the above said degeneration we need a linear renormalization of the variable M_n as follows,

$$M_n^* = \frac{M_n - b_n}{a_n},$$

where $\{a_n > 0\}$ and $\{b_n\}$ are sequences of constants.

If appropriate $\{a_n\}$ and $\{b_n\}$ are chosen, the location and the scale of M_n^* will be stabilized. This avoids the problem of finding the limiting distribution of M_n. For this reason we look for limit distributions for M_n^*. The following definition gives the whole range of possible limit distributions for M_n^*.

Definition 2 (Jenkinson (1955))**.** The **generalized extreme value** (GEV) may be formulated into a single family of models that have distribution function of the form

$$G(z) = \exp\left[-\left\{1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right\}_+^{-1/\xi}\right], \tag{2.1}$$

where $-\infty < \mu < \infty$, $\sigma > 0$ and $-\infty < \xi < \infty$.

This model depends on three parameters: μ (location), σ (scale) and ξ (shape). The shape parameter determines the rate of tail decay, with

- $\xi > 0$ giving the heavy-tailed (Fréchet) case,
- $\xi = 0$ giving the light-tailed (Gumbel) case,
- $\xi < 0$ giving the short-tailed (negative Weibull) case.

Joining the original three families into a single family simplifies the statistical implementation and we obtain the following result.

Theorem 1. (Coles (2001), p. 48). *If there exist sequences of constants* $\{a_n > 0\}$ *and* $\{b_n\}$ *such that*

$$\Pr\{(M_n - b_n)/a_n \leq z\} \to G(z) \quad as\ n \to \infty \tag{2.2}$$

for a non-degenerate distribution function G*, then* G *is a member of the GEV family*

$$G(z) = \exp\left[-\left\{1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right\}_+^{-1/\xi}\right]$$

where $-\infty < \mu < \infty, \sigma > 0$ *and* $-\infty < \xi < \infty$.

2.2 Threshold Exceedance Model

As explained above, the generalized extreme value models are inefficient if other data on extremes are available. In addition, if an entire time series of observations is available, then it is better not to use this approach. For this reason, we consider the generalized Pareto distribution.

Theorem 2. (Coles (2001), p. 75). *Let* $\{X_i\}_{i\geq 1}$ *be a sequence of independent random variables with common distribution function* F*, and let*

$$M_n = \max\{X_1, \ldots, X_n\}.$$

Denote an arbitrary term in the X_i *sequence by* X*, and suppose that* F *satisfies Theorem 1, so that for large* n*,*

$$\Pr\{M_n \leq z\} \approx G(z),$$

where

$$G(z) = \exp\left[-\left\{1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right\}^{-1/\xi}\right]$$

for some $\mu, \sigma > 0$ *and* $-\infty < \xi < \infty$*. Then, for large enough* u*, the distribution function of* $(X - u)$*, conditional on* $X > u$*, is approximately*

$$H(x) = 1 - \left\{1 + \xi\left(\frac{x-u}{\tilde{\sigma}}\right)\right\}^{-1/\xi} \tag{2.3}$$

defined on $\left\{x : x - u > 0 \text{ and } 1 + \xi\left(\frac{x-u}{\tilde{\sigma}}\right) > 0\right\}$*, where*

$$\tilde{\sigma} = \sigma + \xi(u - \mu). \tag{2.4}$$

From equation (2.3) we define the generalized Pareto distribution.

Definition 3 (Behrens *et al.* (2004) and Embrechts *et al.* (1997))**.** A random quantity X follows a **generalized Pareto distribution** (GPD) with threshold u if its distribution function is

$$H(x \mid \tilde{\sigma}, \xi, u) = \begin{cases} 1 - \left\{1 + \xi \left(\frac{x-u}{\tilde{\sigma}}\right)\right\}^{-1/\xi}, & \text{if } \xi \neq 0, \\ 1 - \exp\left\{-\left(\frac{x-u}{\tilde{\sigma}}\right)\right\}, & \text{if } \xi = 0, \end{cases} \qquad x > u, \tag{2.5}$$

where $\tilde{\sigma} > 0$ and $-\infty < \xi < \infty$ are the scale and shape parameters, respectively. Equation (2.5) is valid when $x - u \geq 0$ for $\xi \geq 0$ and for $0 \leq x - u \leq -\tilde{\sigma}/\xi$ for $\xi < 0$. The data present heavy tailed behaviour when $\xi > 0$.

The parameters of threshold excesses are uniquely determined by those of the GEV distribution of block maxima. The parameter ξ is the same as that defined for the GEV distribution. Even if the block size n varies, it would not affect the generalized Pareto distribution, but only the values of the GEV parameters. This means that ξ is invariant to block size. Also the changes in μ and σ, which compensate each other, do not perturb the calculation of $\tilde{\sigma}$. There is a duality between the two distributions, then the shape parameter ξ is dominant in determining their qualitative behaviour.

- If $\xi < 0$ the distribution of excesses has an upper bound of $u - \tilde{\sigma}/\xi$;
- if $\xi > 0$ the distribution has no upper limit;
- if $\xi = 0$ the distribution is unbounded.

Data analysis for a generalized Pareto model is carried out in two steps. Firstly, the threshold u is chosen by using one of several existing procedures. Secondly, the other parameters are estimated assuming that u is known. A disadvantage of this method is that only the observations above the threshold are considered for the estimation of the other parameters. Namely, if we choose too low a threshold, then the data cannot be approximated by a GPD model. Therefore there is a bias. Otherwise, if the threshold is high, the data will be well approximated by a GPD model, but we do not have a lot of observations; this means that the variance is high.

2.3 Threshold Selection

In the next sections, three different approaches to select the threshold u are investigated. Only the first two methods will be applied in the simulation study carried out in Section 3.

2.3.1 Parameter Stability

This first procedure (Coles, 2001) bases the selection of the threshold on fitting the generalized Pareto distribution at a range of thresholds and looking for stability of parameter estimates. We notice that if the generalized Pareto distribution fits well for u_0, it also fits well for $u > u_0$. Both distributions have the same shape parameter. On the other hand, the scale parameter σ_u is defined as

$$\sigma_u = \sigma_{u_0} + \xi(u - u_0), \quad \xi \neq 0. \tag{2.6}$$

In order to simplify the estimation, the scale parameter can be reparametrized as follows,

$$\sigma^* = \sigma_u - \xi u,$$

which is constant with respect to u. Estimates of σ^* and ξ should be roughly constant above u, if u_0 has been correctly chosen. If they are not constant, they have to be stable after the value u_0.

A suggestion could be to plot $\hat{\sigma}^*$ and $\hat{\xi}$ against u with their confidence intervals and choose u_0 as the lowest value of u for which the estimates remain near-constant. To obtain the confidence intervals for $\hat{\xi}$ we use the variance-covariance matrix. On the other hand, the confidence intervals for $\hat{\sigma}^*$ require the delta method as $\hat{\sigma}^*$ depends on σ_u and ξ. The variance of $\hat{\sigma}^*$ is

$$\text{var}(\sigma^*) \approx \nabla\hat{\sigma}^{*T} V \nabla\hat{\sigma}^*,$$

where $\nabla\hat{\sigma}^{*T} = \left[\frac{\partial\hat{\sigma}^*}{\partial\sigma_u}, \frac{\partial\hat{\sigma}^*}{\partial\xi}\right] = [1, -u]$ and V is the variance-covariance matrix of $\hat{\sigma}^*$.

2.3.2 Mean Residual Life Plot

Coles (2001) suggested also another method, which is based on the mean of the generalized Pareto distribution. If Y is a random variable having a generalized Pareto distribution with parameters $\tilde{\sigma}$ and ξ, then the expected value of Y is

$$E(Y) \quad = \begin{cases} \frac{\tilde{\sigma}}{1-\xi}, & \xi < 1 \\ +\infty, & \xi \geq 1. \end{cases} \tag{2.7}$$

Consider the generalized Pareto distribution as a good model for the excesses of a threshold u_0 generated by a series $X_1, \ldots, X_n$, where X is any term. Applying (2.7) for $\xi < 1$, we have

$$E\left(X - u_0 \mid X > u_0\right) = \frac{\sigma_{u_0}}{1-\xi},$$

where σ_{u_0} corresponds to the scale parameter of u_0. If the generalized Pareto distribution is valid for excesses of the threshold u_0, it should be also valid for all $u > u_0$, choosing an adequate change of scale parameter σ_u. Therefore, by equation (2.4) and for $u > u_0$, we have

$$E\left(X - u \mid X > u\right) = \frac{\sigma_u}{1-\xi} = \frac{\sigma_{u_0} + \xi u}{1-\xi}. \tag{2.8}$$

This expectation is a linear function of u. This means that these estimates might change linearly with u, at level of u for which the generalized Pareto model is appropriate.

Let $X_{(1)}, \ldots, X_{(n_u)}$ be n_u observations that exceed u and let $x_{\max}$ be the largest of the X_i. Then the pair of points

$$\left\{ \left(u, \frac{1}{n_u} \sum_{i=1}^{n_u} (x_{(i)} - u) \right) : u < x_{\max} \right\}$$

corresponds to the **mean residual life plot**.

This plot has to be linear in u and confidence intervals can be added as it is based on the approximate normality of sample mean.

2.3.3 Bayes Estimation

In contrast with the previous procedures, Behrens *et al.* (2004) mentioned another way to select the threshold. The model contains uncertainty because a prior, possibly flat, for u is chosen. He proposed a model to fit data characterized by extremal events where the threshold is defined as another model parameter.

Let $X_1, \ldots, X_n$ be independent and identically distributed observations and u the threshold. Then we have that

$$(X_i \mid X_i \geq u) \sim H(\cdot \mid \tilde{\sigma}, \xi, u).$$

On the other hand, the observations below this threshold are distributed according to J, which can be estimated either parametrically or non-parametrically. In the parametric case, we often choose for the data below the threshold J like a gamma, Weibull or normal distribution. Otherwise, if J is estimated non parametrically, usually mixtures of these previous parametric forms are a convenient basis for J.

Suitable prior distributions are chosen for each parameter of the model. In particular Coles and Powell (1996)'s prior is used, that is the eliciting information. Unfortunately, analytical computations are impossible. For this reason, Markov Chain Monte Carlo methods are applied, in particular Metropolis–Hastings and Gibbs Sampler.

Model Definition

Assume that the data under the threshold u are distributed according to $J(\cdot \mid \boldsymbol{\eta})$, where $\boldsymbol{\eta}$ are the parameters of the distribution. Assume also that the data above the threshold u come from a generalized Pareto distribution. Then we can define the distribution for any X as follows,

$$F(x \mid \boldsymbol{\eta}, \tilde{\sigma}, \xi, u) = \begin{cases} J(x \mid \boldsymbol{\eta}), & x < u, \\ J(u \mid \boldsymbol{\eta}) + \{1 - J(u \mid \boldsymbol{\eta})\} H(x \mid \tilde{\sigma}, \xi, u), & x \geq u. \end{cases} \tag{2.9}$$

Let define two sets, $A = \{i : x_i < u\}$ and $B = \{i : x_i \geq u\}$. For a sample $\boldsymbol{x} = (x_1, \ldots, x_n)$ from F and $\boldsymbol{\theta} = (\boldsymbol{\eta}, \tilde{\sigma}, \xi, u)$ the parameter vector, then the likelihood function is

$$L(\boldsymbol{\theta}; \boldsymbol{x}) = \begin{cases} \prod_A j(x \mid \boldsymbol{\eta}) \prod_B \{1 - J(u \mid \boldsymbol{\eta})\} \left[\frac{1}{\tilde{\sigma}} \left\{ 1 + \xi \left(\frac{x_i - u}{\tilde{\sigma}} \right) \right\}_+^{-1/\xi - 1} \right], & \xi \neq 0, \\ \prod_A j(x \mid \boldsymbol{\eta}) \prod_B \{1 - J(u \mid \boldsymbol{\eta})\} \left[\frac{1}{\tilde{\sigma}} \exp \left\{ - \left(\frac{x_i - u}{\tilde{\sigma}} \right) \right\} \right], & \xi = 0. \end{cases} \tag{2.10}$$

Graphically, we can imagine to have a density function which has a discontinuity point in u. This jump represents the difficulty to estimate the threshold. This means that if we have a small jump, the estimation of u will be more difficult. On the contrary, if the jump is large, there is evidence of separation of the data, then the estimation will be easier.

Figure 3.1 of the simulation study shows a jump between the data distributed below (uniform data) and above (generalized Pareto data) the threshold u. This discontinuity is represented by the red line at the point $u = 5$.

Prior and Posterior distribution

The parameters in the model are $\boldsymbol{\theta} = (\boldsymbol{\eta}, \tilde{\sigma}, \xi, u)$. In the next paragraphs we describe in details the priors for the parameters above, on and below the threshold u.

Prior for parameters above the threshold

As it is not easy to express directly prior beliefs of GPD parameters, the elicitation of information is used (Coles and Powell (1996) and Coles and Tawn (1996)). Equation (2.5) is inverted and we get the $1 - p$ **quantile of the distribution**,

$$q = u + \frac{\tilde{\sigma}}{\xi}(p^{-\xi} - 1).$$

The value q corresponds to the return level associated with a return period of $1/p$ time units.

For the generalized Pareto parameters, the prior elicitation is carried out in term of (q_1, q_2, q_3) specifying the values of $p_1 > p_2 > p_3$. Hence, we order the parameters and $q_1 < q_2 < q_3$. Coles and Tawn (1996) proposed to work with the differences $d_i = q_i - q_{i-1}$, $i = 1, 2, 3$. In addition, they assume $q_0 = e_1$, where e_1 is the physical lower bound of the variable. The differences d_i are supposed to be gamma distributed with parameters (α_i, β_i) for $i = 1, 2, 3$. The prior distribution of each d_i is supposed to be independent to the others. Usually we use e_1 equal to zero.

The procedure to obtain the prior information is the following: first, the median and the 90% quantile (or any other) estimates for specific values of p are required. Second, we transform the elicited parameters to obtain the equivalent gamma parameters. Notice that neither d_i nor q_i depend on u for $i > 1$. Then, we have that $p(d_i \mid u)$ is approximated by $(d_i \mid u^*) \sim Ga(a_i(u^*), b_i(u^*))$, where u^* is the prior mean for u.

In this particular case, we do not consider the location parameter, but only the scale and shape parameters. For this reason, we need only two quantiles. The gamma distributions for the differences with known parameters are given by

$$d_1 = q_1 \sim Ga(a_1, b_1), \quad d_2 = q_2 - q_1 \sim Ga(a_2, b_2).$$

The marginal prior distribution for parameters $\tilde{\sigma}$ and ξ is

$$\begin{aligned}
\pi(\tilde{\sigma}, \xi) \quad \propto \quad & \left\{ u + \frac{\tilde{\sigma}}{\xi}(p_1^{-\xi} - 1) \right\}^{a_1 - 1} \exp\left[-b_1 \left\{ u + \frac{\tilde{\sigma}}{\xi}(p_1^{-\xi} - 1) \right\} \right] \\
\times \quad & \left\{ \frac{\tilde{\sigma}}{\xi}(p_2^{-\xi} - p_1^{-\xi}) \right\}^{a_2 - 1} \exp\left[-b_2 \left\{ \frac{\tilde{\sigma}}{\xi}(p_2^{-\xi} - p_1^{-\xi}) \right\} \right] \\
\times \quad & \left| -\frac{\tilde{\sigma}}{\xi^2} \left\{ (p_1 p_2)^{-\xi}(\log p_2 - \log p_1) - p_2^{-\xi} \log p_2 + p_1^{-\xi} \log p_1 \right\} \right|,
\end{aligned}$$

where a_1, a_2, b_1 and b_2 are hyperparameters obtained from the experts information, $\tilde{\sigma} > 0$ and $\xi \in \mathbb{R}$.

Prior for the threshold

Different alternatives to define a prior distribution for u exist. The most used are the continuous uniform prior, the discrete distribution or a truncated normal distribution with parameters (μ_u, σ_u^2), truncated from below at e_1 with density

$$\pi(u \mid \mu_u, \sigma_u^2, e_1) = \frac{1}{\sqrt{2\pi\sigma_u^2}} \frac{\exp\{-(u - \mu_u)^2 / 2\sigma_u^2\}}{\Phi[-(e_1 - \mu_u)/\sigma_u]} \tag{2.11}$$

with μ_u set at some high data percentile, σ_u^2 large enough to represent a fairly non informative prior (Behrens *et al.*, 2004), $e_1 = q_0$ and $e_1 < u < \infty$.

Prior for parameters below the threshold

According to the distribution chosen for the data below the threshold u, the prior for the parameters $\boldsymbol{\eta}$ could be modified. The most suitable choice for the prior would be a conjugate prior so that the problem has a simpler form analytically.

In this case, we assume that the data have a gamma distribution $j(x \mid \boldsymbol{\eta})$ with parameters $\boldsymbol{\eta} = (\alpha, \beta)$, where α is the shape and β the rate parameter. It is easier to reparametrize in terms of α and $\mu = \alpha/\beta$ to have a more natural interpretation. Moreover, we assume that the shape parameter α and the mean μ are independent to simplify the computations. Both parameters have a gamma distribution,

$$\alpha \sim Ga(a, b), \quad \mu \sim Ga(c, d),$$

where a, b, c and d are known hyperparameters. Then, the joint prior density function can be written as follows,

$$\pi(\boldsymbol{\eta}) = \frac{b^a}{\Gamma(a)} \alpha^{a-1} e^{-b\alpha} \frac{d^c}{\Gamma(c)} \left(\frac{\alpha}{\beta}\right)^{c-1} e^{-d\alpha/\beta} \left(\frac{\alpha}{\beta^2}\right),$$

where $a, b, c, d > 0$.

Posterior inference

We take the likelihood defined in equation (2.10) and the prior distributions given in the previous paragraphs to compute the posterior distribution given by applying Bayes Theorem. As the calculations are too much complicated to carry out analytically, we apply the Markov Chain Monte Carlo methods, in particular the Metropolis–Hastings algorithm.

Chapter 3

Simulation Study

In order to check the theoretical properties of measures of surprise, simulations are performed. In particular, we consider two cases, the prior predictive p-values (1.7) and the posterior predictive p-values (1.9). After that, we compare the results with two of the approaches explained in Section 2.3: the parameter stability plot (see Section 2.3.1) and the mean residual life plot (see Section 2.3.2).

Firstly, we look at the posterior predictive measures of surprise considering three different samples and after that we get on to the prior predictive measures of surprise.

3.1 Posterior Predictive p-values

Considering the posterior predictive p-values defined in (1.8),

$$m(\boldsymbol{x} \mid \boldsymbol{x}_{\text{obs}}) = \int f(\boldsymbol{x} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid \boldsymbol{x}_{\text{obs}}) d\boldsymbol{\theta},$$

we notice that $m(\boldsymbol{x} \mid \boldsymbol{x}_{\text{obs}})$ could be approximated by

$$m(\boldsymbol{x} \mid \boldsymbol{x}_{\text{obs}}) \approx \frac{1}{N} \sum_{i=1}^{N} f(\boldsymbol{x} \mid \boldsymbol{\theta}^{(i)}), \quad \boldsymbol{\theta}^{(i)} \sim \pi(\boldsymbol{\theta} \mid \boldsymbol{x}_{\text{obs}}). \tag{3.1}$$

Then, for the parameter $\boldsymbol{\theta} = (\tilde{\sigma}, \xi)$ of a generalized Pareto distribution, the likelihood of a set of independent observations $\boldsymbol{x} = (x_1, \ldots, x_n)$ can be written as

$$f(\boldsymbol{x} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i \mid \boldsymbol{\theta}), \tag{3.2}$$

where

$$f(x_i \mid \boldsymbol{\theta}) \quad = \quad \begin{cases} \frac{1}{\tilde{\sigma}} \left\{1 + \xi \left(\frac{x_i - u}{\tilde{\sigma}}\right)\right\}^{-1/\xi - 1}, & \text{if } \xi \neq 0, \\ \frac{1}{\tilde{\sigma}} \exp\left\{-\left(\frac{x_i - u}{\tilde{\sigma}}\right)\right\}, & \text{if } \xi = 0, \end{cases} \qquad x_i > u.$$

In order to compute the measure of surprise (3.1), the Metropolis–Hastings algorithm is implemented to draw the posterior distribution. Usually to simplify the calculations of the likelihood-ratio in the MCMC algorithm, computation is performed on the log-scale (i.e. difference of log likelihoods). This avoids evaluations of the likelihood being numerically rounded to zero.

However, in evaluating (3.1) the likelihood must be evaluated on its natural scale, and so is rounded accordingly to 0. It is difficult to get round this by using log likelihood computations as

$$\log\left(\int f(\boldsymbol{x} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid \boldsymbol{x}_{\text{obs}}) d\boldsymbol{\theta}\right) \neq \int \log f(\boldsymbol{x} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid \boldsymbol{x}_{\text{obs}}) d\boldsymbol{\theta}$$

or any other computation with $\log f(\boldsymbol{x} \mid \boldsymbol{\theta})$.

For this reason, the likelihood defined in equation (3.2) as the product of the $f(x_i \mid \boldsymbol{\theta})$ is replaced by the mean of the $f(x_i \mid \boldsymbol{\theta})$ so that higher likelihood values will produce a higher mean of the densities and then the surprise is more evident,

$$f(\boldsymbol{x} \mid \boldsymbol{\theta}) \approx \frac{1}{n} \sum_{i=1}^{n} f(x_i \mid \boldsymbol{\theta}). \tag{3.3}$$

While this approach is non-standard, this approximation gives credible results and we can see them in the simulation studies. Unfortunately, no information has been found on how to compute these marginal likelihoods numerically in the literature and to support this choice. Only algebraic computations could have been found in certain circumstances.

In addition, the prior distribution is given by the Jeffrey's prior (Castellanos and Cabras, 2007),

$$\pi(\boldsymbol{\theta}) \quad = \quad \frac{1}{\tilde{\sigma}} \frac{1}{1+\xi} \frac{1}{\sqrt{1+2\xi}} \quad \tilde{\sigma} > 0, \quad \xi > -0.5. \tag{3.4}$$

The procedure to approximate the integral consists of several steps. Firstly, the Metropolis–Hastings algorithm produces a chain of values of $\boldsymbol{\theta}$. These parameter values come from the posterior distribution $f(\boldsymbol{x} \mid \boldsymbol{\theta})$ using the Metropolis–Hastings algorithm. The prior of $\boldsymbol{\theta}$ is the Jeffrey's prior defined in equation (3.4) and the distribution $f(\boldsymbol{x} \mid \boldsymbol{\theta})$ is the mean of the

generalized Pareto distribution of each observation with scale parameter $\tilde{\sigma}$ and shape parameter ξ. The proposal densities for $\tilde{\sigma}$ and ξ are a log-normal density distribution and a normal density distribution, respectively.

In order to generate the sample, $R = 10000$ iterations have been carried out, where a burn in period of 1000 has been cut off. Furthermore, we consider for the analysis the chains consisting in every $10-$th observation. These values are used to evaluate the distribution $f(\boldsymbol{x} \mid \boldsymbol{\theta})$ and calculate approximatively $m(\boldsymbol{x} \mid \boldsymbol{x}_{\text{obs}})$. Finally, the probability

$$\Pr^{m(\cdot|\boldsymbol{x}_{\text{obs}})}\{m(\boldsymbol{X} \mid \boldsymbol{x}_{\text{obs}}) < m(\boldsymbol{x}_{\text{obs}} \mid \boldsymbol{x}_{\text{obs}})\}$$

is estimated by counting the number of times that $m(\boldsymbol{X} \mid \boldsymbol{x}_{\text{obs}})$ is less than $m(\boldsymbol{x}_{\text{obs}} \mid \boldsymbol{x}_{\text{obs}})$ divided by the total number of simulations. Then, the posterior predictive p-values can be written as

$$p_{\text{post}} = \Pr\{T(\boldsymbol{X}) \geq T(x_{\text{obs}}) \mid x_{\text{obs}}\},$$

where $T(\boldsymbol{X}) = m(\boldsymbol{X} \mid \boldsymbol{x}_{\text{obs}})$. The same study is carried out for three datasets which are created with different changepoints. Two samples are generated with a known changepoint location and the third one does not have a changepoint. Concerning the datasets with a changepoint, we have either some uniform or gamma data generated below it and some generalized Pareto data generated above it. The last dataset has a generalized Pareto distribution. The purpose of looking at these different datasets is the detection of the known changepoint, if it exists by using measures of surprise.

3.1.1 Uniform and Generalized Pareto Data

First of all we generate two different datasets. The first sample has a generalized Pareto distribution with parameters $\tilde{\sigma}$ equal to 1, ξ equal to 0.2 and u equal to 5. Its size is $n = 500$. The second sample ($n = 500$) has a uniform distribution on the interval $[0, 5]$. The histogram of the complete dataset (generalized Pareto and uniform data) is represented in Figure 3.1; the red line represents the threshold $u = 5$.

Before starting to compute the measures of surprise, we look at different plots. Figure 3.2 illustrates the traces of the sampled values of the parameters $\hat{\tilde{\sigma}}$ and $\hat{\xi}$ estimated by Metropolis–Hastings algorithm. The trace plots represent the behaviour of the parameters at each iteration for the new chain. The posterior means (red line in Figure 3.2) and their 95% central credibility intervals are displayed in Table 3.1. The marginal posterior densities are analysed too. Figure 3.3 shows the marginal posterior densities of each

parameter and the red lines correspond to the posterior mean. Finally, the correlogram of both parameters is displayed in Figure 3.4. This graph indicates that the chain has a stationary distribution and the observations are independent. For lags bigger than 2 the observed ACFs correspond to white noise.

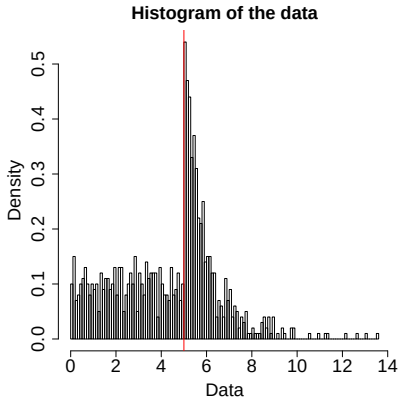

Figure 3.1: Histogram of the complete dataset. Data below the threshold $u = 5$ (red line) correspond to uniform data on the interval $[0, 5]$. Data above the threshold $u = 5$ (red line) have a generalized Pareto distribution with parameters $\tilde{\sigma} = 1$ and $\xi = 0.2$.

	Mean	2.5% quantile	97.5% quantile
$\tilde{\sigma}$	0.87	0.75	0.99
ξ	0.19	0.09	0.30

Table 3.1: Estimates of the parameters and their 95% central credibility intervals.

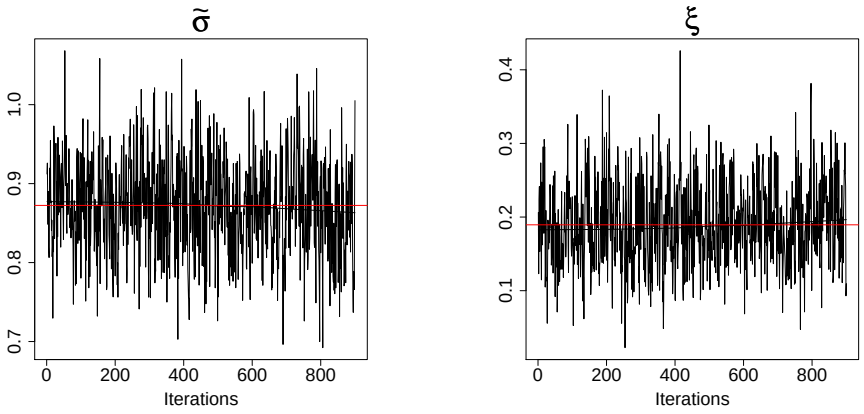

Figure 3.2: Trace plots of the parameters $\hat{\tilde{\sigma}}$ and $\hat{\xi}$ estimated by Metropolis–Hastings algorithm (10000 iterations have been carried out, a burn in period of length 1000 has been cut off and one every 10-th observation is considered). The red lines correspond to the posterior means for the sampled values $\tilde{\sigma}$ and ξ.

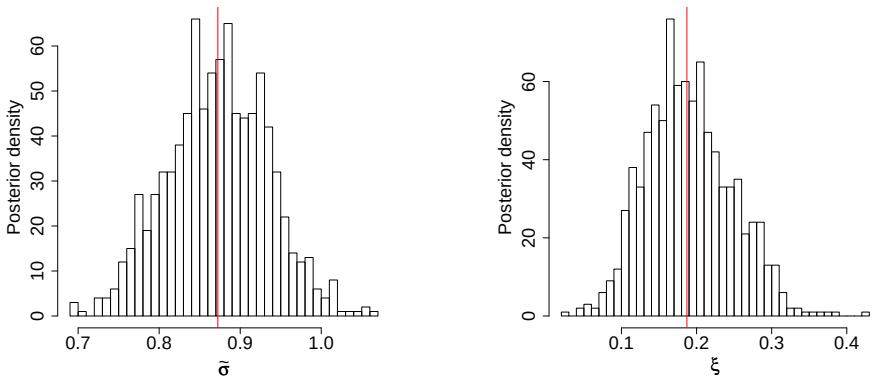

Figure 3.3: Marginal posterior density plots of the parameters $\hat{\tilde{\sigma}}$ and $\hat{\xi}$ estimated by Metropolis–Hastings algorithm (10000 iterations have been carried out, a burn in period of length 1000 has been cut off and one every 10-th observation is considered).

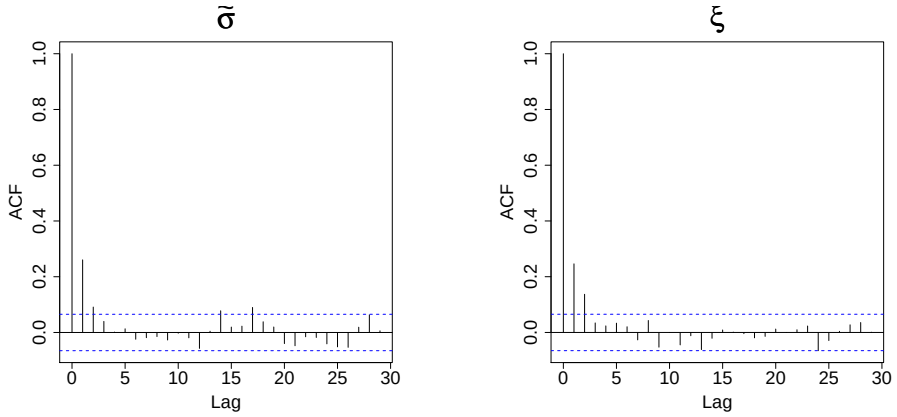

Figure 3.4: ACFs for the parameters $\hat{\tilde{\sigma}}$ and $\hat{\xi}$ estimated by Metropolis–Hastings algorithm (10000 iterations have been carried out, a burn in period of length 1000 has been cut off and one every 10-th observation is considered).

Further on, we analyse the measures of surprise and posterior predictive p-values for the thresholds from 2 to 9. Figure 3.5 shows the posterior predictive measures of surprise (left panel) and the posterior predictive p-values (right panel) for each threshold, where the vertical lines correspond to the 95% central credibility intervals. The true threshold (i.e $u = 5$) is highlighted by the red triangle.

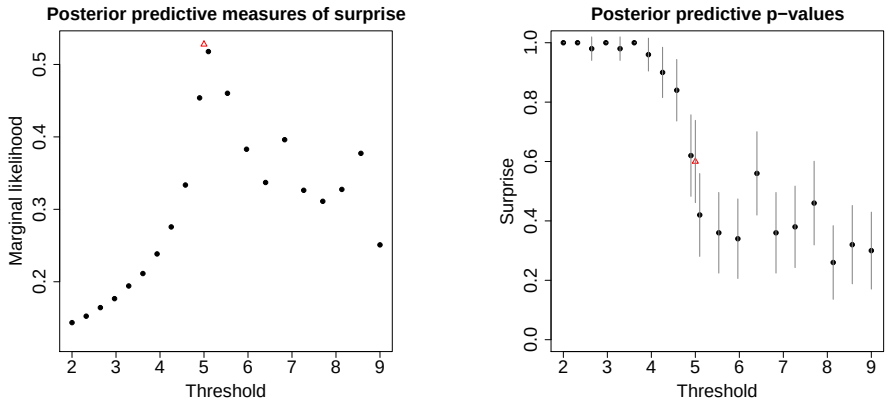

Figure 3.5: Plot of the measures of surprise $m(\boldsymbol{x} \mid \boldsymbol{x}_{\text{obs}})$ (left panel) and of the posterior predictive p-values with their 95% central credibility intervals (right panel) for the different thresholds u from 2 to 9 estimated by using the approximation given in equation (3.1). The red triangle corresponds to the proper threshold $u = 5$.

Both graphs indicate that the data below the true threshold do not have a generalized Pareto distribution, that is there is evidence of "surprise". The left plot shows the marginal likelihoods which increase from very small values, for a threshold far away from $u = 5$, to the highest value of $m(\boldsymbol{x} \mid \boldsymbol{x}_{\text{obs}}) = 0.53$ at u equal to 5. Then the marginal likelihood starts again to decrease. The way, how the estimated marginal likelihoods change, affects the results concerning the p-values as its graph indicates it. In fact, the right plot shows the p-values estimated around 1 below the true threshold and this means that the probability to have surprise is very high and therefore the model does not fit appropriately. The reason why the p-values are very big is due to the fact that having very small marginal likelihoods, the probability to obtain larger values of the marginal likelihoods, which are obtained from the simulated dataset, than the marginal likelihood of the dataset is slight. Thus, the probability to have "surprise" is very high. On the other hand, small p-values indicate a slight surprise, that is the generalized Pareto model fits appropriately to the data. Furthermore, in the p-values plot we notice that the probability jump from values around 1 to 0.6 for a threshold chosen just below the true one (i.e. $u = 4.9$). We remark that the dataset considering the threshold at 4.9 has just ten observations more than the dataset generated only from the generalized Pareto distribution. Then this jump of about 0.4 gives evidence of the appearance of a changepoint. Thus, in order to find

the suitable changepoint both graphs are necessary for the analysis.

Other tools to estimate the threshold are explained in Section 2 and we exploit two of them to check the goodness of the model.

First of all, we look at the threshold selection by using parameter stability (see Section 2.3.1) which consists in plotting the fitted GPD parameters at different thresholds in order to detect a good threshold for the dataset.

Figure 3.6 shows the estimated values of the scale and shape parameters at each threshold from 2 to 9. In addition, the vertical broken lines represent the 95% central credibility intervals. Both the left and the right panels indicate that the parameters are not stable below the threshold $u = 5$ and for u bigger than 5 the bands cover horizontal lines, suggesting stability. A similar result is not surprising as below the threshold $u = 5$ the data are uniform distributed.

Another approach to estimate the threshold u of a dataset is the use of the mean residual life plot which has been explained in Section 2.3.2. Figure 3.7 illustrates the mean residual life plots for every threshold (left panel) and for the thresholds chosen for the previous studies (from 2 to 9). Looking at the left plot, we observe a slight linearity of the mean exceedance from $u = 5$. This means that the GPD distribution does not fit appropriately before the true threshold. This linearity is shown more clearly in the right panel.

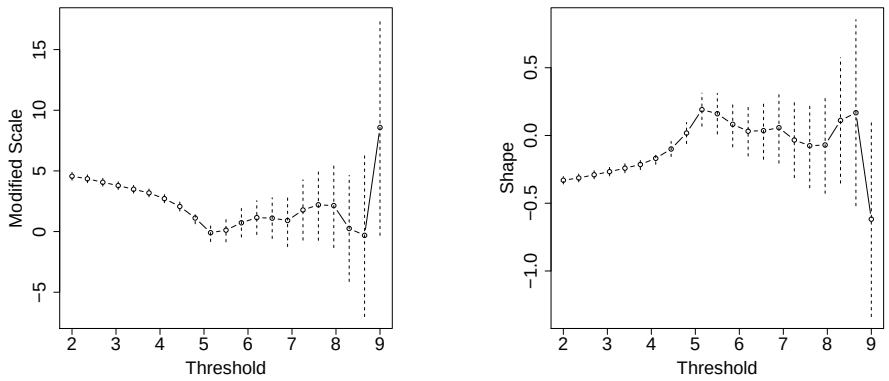

Figure 3.6: Parameter stability plots of $\hat{\tilde{\sigma}}$ (left panel) and $\hat{\xi}$ (right panel) against the thresholds. The vertical broken lines correspond to the 95% central credibility intervals.

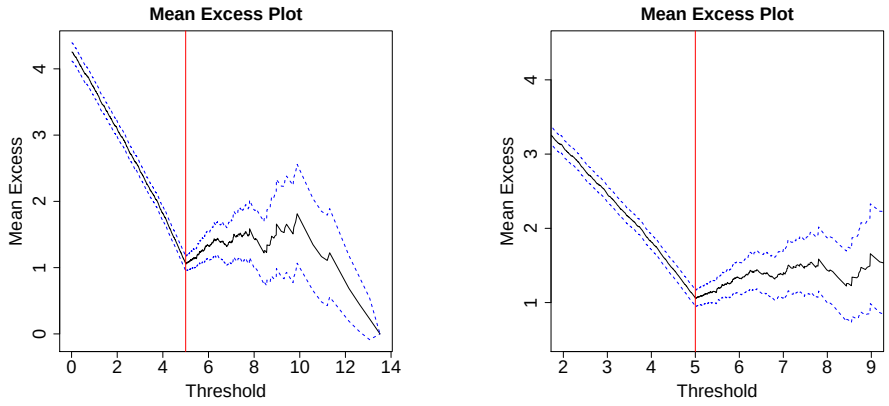

Figure 3.7: Mean residual life plot for every threshold (left panel) and mean residual life plot for the thresholds from 2 to 9 (right panel). The blue lines correspond to the 95% central credibility intervals and the red line corresponds to the threshold $u = 5$.

Once more, this result is coherent with the hypothesis considering that only the data above the threshold $u = 5$ have a generalized Pareto distribution. Furthermore, in the left panel we observe that for a big threshold (i.e. $u > 9$) the dataset becomes very small and thus the generalized Pareto model does not fit any more very well. In fact, there is a decreasing tendency of the mean exceedance instead of keeping a constant value.

3.1.2 Gamma and Generalized Pareto Data

A second simulation study is carried out on a sample generated by gamma and generalized Pareto data. The first sample has a generalized Pareto distribution with parameters $\tilde{\sigma}$ equal to 5, ξ equal to 0.1 and u equal to 4.5. Its size is $n = 500$. The second sample has a gamma distribution with shape parameter equal to 10 and rate parameter equal to 5. We take into account only the values which are less or equal than 4.5. Figure 3.8 represents the histogram of the complete dataset (gamma and generalized Pareto data) and the vertical red line corresponds to the changepoint ($u = 4.5$). The histogram let suggest that the data appear to come from a single continuous distribution with a bit of data removed at around u equal to 4.5.

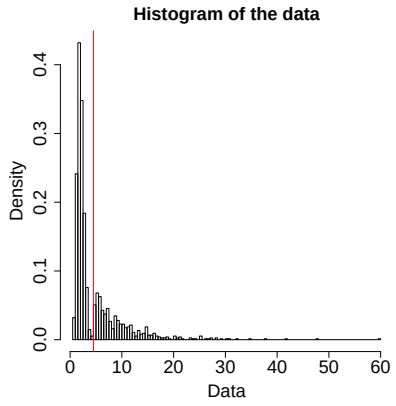

Figure 3.8: Histogram of the complete dataset. Data below the threshold $u = 4.5$ (red line) correspond to gamma data with shape parameter equal to 10 and rate parameter equal to 5. Data above the threshold $u = 4.5$ (red line) have a generalized Pareto distribution with parameters $\tilde{\sigma} = 5$ and $\xi = 0.1$.

A similar analysis about the outputs (posterior densities of the parameters) of the Metropolis–Hastings algorithm is carried out: the marginal posterior density plots, the trace plots and the independence of the chain are studied before looking at the measures of surprise and their p-values.

Figure 3.9 shows the posterior marginal likelihoods (left panel) and the posterior predictive p-values with their 95% central credibility intervals (right panel). These graphs highlight some interesting but at the same time unexpected results. The behaviour of both the marginal likelihoods and their p-values for the thresholds between 1 and 3 is not regular. First of all, the marginal likelihoods increase until u equal to 1.76 and after that it decreases until becoming infinitesimal for u equal to 3.27. Respectively, the p-values plot shows that for the highest marginal likelihood (at $u = 1.76$), the "surprise" to have a generalized Pareto model is very small. After that it increases until 1 (for u equal to 3.64).

No surprise means that the p-values is equal to zero, that is every dataset generated during the simulation is less likely than the observed one. Furthermore, as the observed marginal likelihood is very small, the probability to obtain bigger marginal likelihoods for the generated dataset is extremely difficult and then the surprise is estimated around zero.

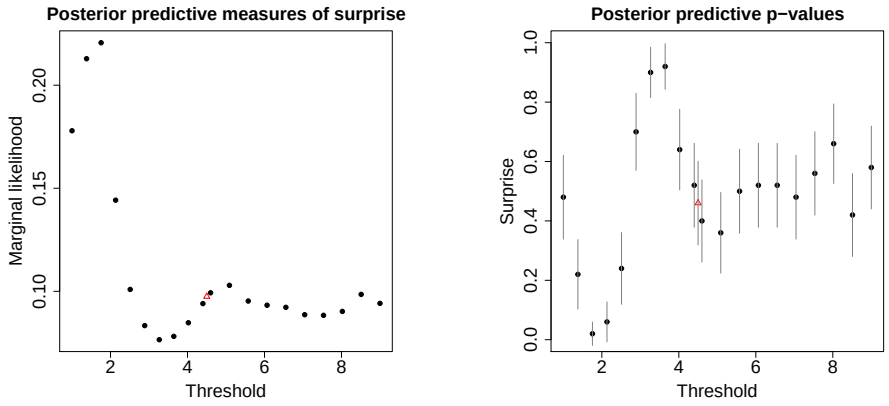

Figure 3.9: Plot of the measures of surprise $m(\boldsymbol{x} \mid \boldsymbol{x}_{\text{obs}})$ (left panel) and of the posterior predictive p-values with their 95% central credibility intervals (right panel) for the different thresholds u from 1 to 9 estimated by using the approximation given in equation (3.1). The red triangle corresponds to the proper threshold $u = 4.5$.

Observing the histogram (see Figure 3.8) we notice a drop just before the true threshold. Two different interpretations of this plot can be possible: the first one is that the data come from two separate datasets and the second one it that there is a unique dataset coming from a continuous distribution with some data missing. As shown in Figure 3.9 this duel interpretation affects the results of the posterior marginal likelihoods and the posterior predictive p-values. In fact, for low threshold (i.e. $u < 3$), when data are generated to estimate the p-values, consistently with the second interpretation, the drop is not taken into account because we have a lot of data and they seem to have a generalized Pareto distribution. Thus the "surprise" will be small for very low threshold values. On the other hand, when the threshold u is chosen just around 4 (or possibly < 1.76), the data will not be generalized Pareto distributed. Therefore instead of having a small surprise we have a surprise which increases as we move below the threshold $u = 4$ (or $u = 1.76$). This occurs in the first case ($u < 1.76$) because we have too much gamma data than generalized Pareto data and in the second case ($3 < u < 4$) because of the presence of the drop.

Then for the analysis we cannot consider the results for u less than 3. The reason why we cannot consider the threshold u equal to 1.76 as credible is because if data are really generalized Pareto distributed, they should be GPD for all thresholds above u. This is not true for $u = 1.76$ but it is

true for $u = 4.5$. Therefore, analysing both graphs we can conclude that approximately the right threshold is around 4.5.

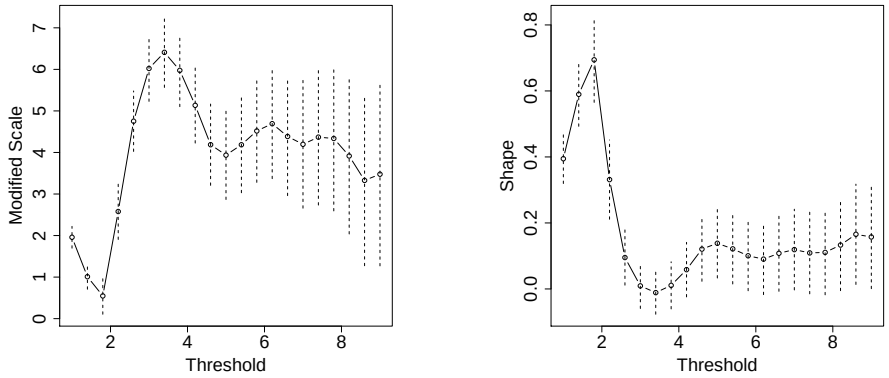

Figure 3.10: Parameter stability plots of $\hat{\tilde{\sigma}}$ (left panel) and $\hat{\xi}$ (right panel) against the thresholds. The vertical broken lines correspond to the 95% central credibility intervals.

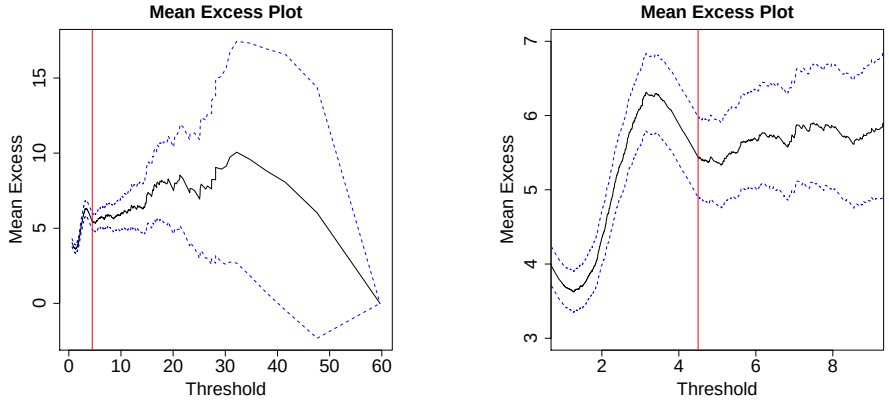

Figure 3.11: Mean residual life plot for every threshold (left panel) and mean residual life plot for the thresholds from 1 to 9 (right panel). The blue lines correspond to the 95% central credibility intervals and the red line corresponds to the threshold $u = 4.5$.

Looking at the outcomes displayed in the parameter stability plots (see Figure 3.10) and the mean residual life plots (see Figure 3.11), we have

that the generalized Pareto model fits appropriately for the threshold $u = 4.5$. Opposite of the p-values plot, both graphs do not indicate that for the threshold u equal to 1.76 the generalized Pareto distribution fits well. Looking more attentively at the left mean residual life plot we notice a similar behaviour as for the previous study for high thresholds: for u bigger than 15, the stability of the exceedence disappears slowly.

3.1.3 Generalized Pareto Data

A last simulation study is analysed: no changepoint exists and only generalized Pareto data are generated with parameters $\tilde{\sigma}$ equal to 1, ξ equal to 0.2 and u equal to 2. Its size is $n = 1000$. In Figure 3.12 the histogram of the sample is represented and the red line indicates the threshold at $u = 2$.

As for the previous studies, some different analysis are carried out on the chain of $\tilde{\sigma}$ and ξ produced by Metropolis–Hastings algorithm. After that the posterior marginal likelihoods and the posterior predictive p-values are estimated and the results are displayed in Figure 3.13.

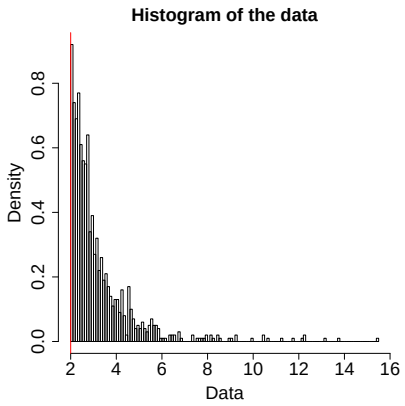

Figure 3.12: Histogram of the complete dataset. All data above the threshold $u = 2$ (red line) have a generalized Pareto distribution with parameters $\tilde{\sigma} = 1$ and $\xi = 0.2$.

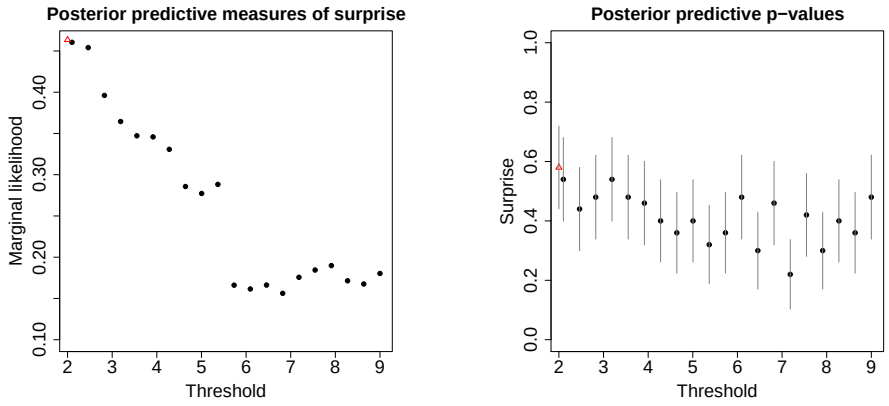

Figure 3.13: Plot of the measures of surprise $m(\boldsymbol{x} \mid \boldsymbol{x}_{\text{obs}})$ (left panel) and of the posterior predictive p-values with their 95% central credibility intervals (right panel) for the different thresholds u from 2 to 9 estimated by using the approximation given in equation (3.1). The red triangle corresponds to the proper threshold $u = 2$.

The left panel shows the posterior predictive measures of surprise and the right panel their respective p-values. Both graphs indicate the goodness of the model for u equal to 2. The marginal likelihood plot shows a decreasing behaviour of the measures of surprise. Already this first graph suggests to take into account u equal to 2 as the most appropriate threshold because its marginal likelihood is the highest. Furthermore, we look at the right panel, where the posterior predictive p-values are displayed. We notice that the "surprise" for each threshold is quite constant and not very high. This can be interpreted as that the model fits appropriately to the data.

As in the previous studies, we look at the usual tools to assess the appropriateness of the model: the parameter stability plots (see Figure 3.14) and the mean residual life plots (see Figure 3.15).

In Figure 3.13 we notice a "jump" in the marginal likelihood plot between $u = 5$ and $u = 6$. This odd behaviour is also represented in the Figures 3.14 and 3.15. It looks like that there are like two different generalized Pareto distributions, the first one before the "jump" and the second one after the "jump". In fact, looking at the parameter stability plots (see Figure 3.14) we observe that the parameters are stable but their values change according to the threshold (before or after the "jump"). Similarly, the mean residual life plot (see Figure 3.15) is linear in u giving evidence of the goodness of the generalized Pareto distribution but the straight line has also a jump when

the threshold is between 5 and 6. In other words, all these graphs indicate that there is another generalized Pareto distribution for a dataset having a threshold higher than $u = 5$.

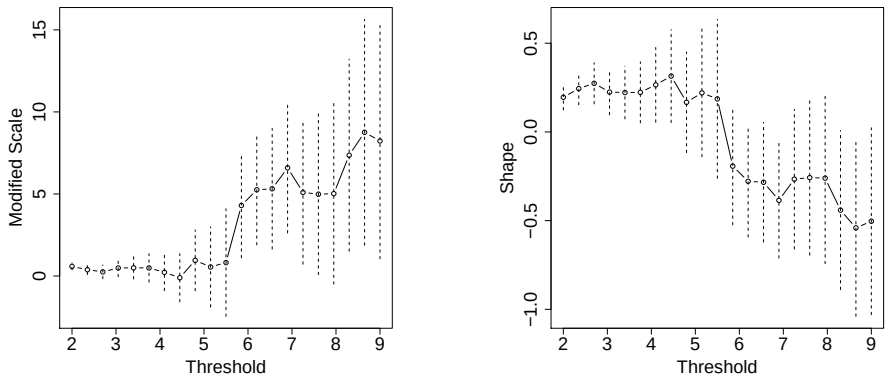

Figure 3.14: Parameter stability plots of $\hat{\tilde{\sigma}}$ (left panel) and $\hat{\xi}$ (right panel) against the thresholds. The vertical broken lines correspond to the 95% central credibility intervals.

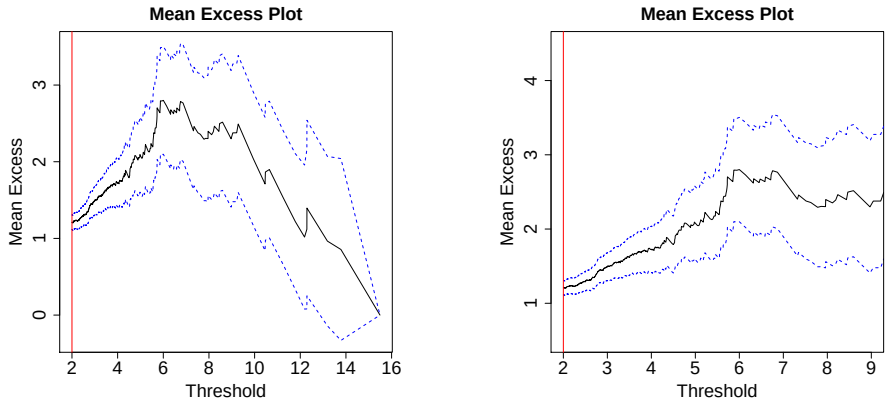

Figure 3.15: Mean residual life plot for every threshold (left panel) and mean residual life plot for the thresholds from 2 to 9 (right panel). The blue lines correspond to the 95% central credibility intervals and the red line corresponds to the threshold $u = 2$.

In this last study, when no changepoint exists, we can notice that the

use of the measure of surprise like a tool to estimate the right threshold is harder. The reason is that having only a generalized Pareto dataset, for each threshold that we choose from u equal to 2, the model will fit appropriately to the sample. In fact, for none of the thresholds the "surprise" is very high or even 1, which indicates the goodness of the model. However, there is a more accurate generalized Pareto distribution fitted to the data above the threshold $u = 5$.

3.2 Prior Predictive p-values

Let us now define the prior predictive distribution as in (1.5),

$$m(\boldsymbol{x}) = \int f(\boldsymbol{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta},$$

where, for the same reasons as in the previous section, we approximate the likelihood $f(\boldsymbol{x} \mid \boldsymbol{\theta})$ as the mean of the density of each observation x_i given in equation (3.3) and the prior distribution is defined as the Jeffrey's prior given in equation (3.4).

Computing marginal likelihoods is extremely difficult; therefore we need to estimate these quantities separately. Many different ways to do it exist, such as computing integrals by choosing conjugate $f(\boldsymbol{x} \mid \boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta})$ and making exact computations by hand. Other tools are numerical integration (e.g. Gaussian quadrature), analytical approximation and simulation. Difficulties, even if different, appear in each of these methods.

In this particular case the Laplace approximation approach is studied and the most important steps concerning this method to estimate the prior predictive distribution are explained in the next section.

The main drawback of analytical approximation is that the result is not very precise if compared to the result carried out by simulation. Nevertheless, simulations are sometimes hard to implement and need a lot of tuning.

3.2.1 Laplace Approximation

The one-dimensional integral is defined as follows,

$$I_n = \int_{-\infty}^{\infty} e^{-nh(v)}dv, \tag{3.5}$$

where $h(v)$ is a smooth convex function with minimum at $v = \tilde{v}$, at which point $dh(\tilde{v})/dv = 0$ and $d^2h(\tilde{v})/dv^2 > 0$ (Davison, 2003). Then we carry out

a Taylor series expansion close to $\tilde{v}$ obtaining the following approximation

$$I_n \doteq \left(\frac{2\pi}{nh_2}\right)^{1/2} e^{-nh(\tilde{v})}, \tag{3.6}$$

where $h_2 = d^2h(\tilde{v})/dv^2$.

In the multivariate case, the integral I_n is approximated as follows,

$$I_n \doteq \left(\frac{2\pi}{n}\right)^{p/2} \mid h_2 \mid^{-1/2} e^{-nh(\tilde{\boldsymbol{v}})}, \tag{3.7}$$

where $h(\boldsymbol{v})$ is again a smooth convex function but $\boldsymbol{v}$ is a vector of dimension p. Furthermore, we have that $\tilde{\boldsymbol{v}}$ solves the $p \times 1$ system of equations $\partial h(\boldsymbol{v})/\partial \boldsymbol{v} = 0$ and $\mid h_2 \mid$ is the determinant of the $p \times p$ matrix of second derivatives $\partial^2 h(\boldsymbol{v})/\partial \boldsymbol{v} \partial \boldsymbol{v}^T$ evaluated at $\boldsymbol{v} = \tilde{\boldsymbol{v}}$, at which point the matrix is positive definite (Davison, 2003).

In this case, we have that

$$h(\boldsymbol{\theta}) = -\log f(\boldsymbol{x} \mid \boldsymbol{\theta}) - \log \pi(\boldsymbol{\theta})$$

because $\exp\{-h(\boldsymbol{\theta})\} = f(\boldsymbol{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})$ and where $f(\boldsymbol{x} \mid \boldsymbol{\theta})$ is defined in (3.3) and $\pi(\boldsymbol{\theta})$ is defined in (3.4). Moreover p is equal to 2. Hence the approximation of the integral is

$$\log I_n \doteq \log(2\pi) - \log n - \frac{1}{2} \log \mid h_2 \mid -h(\boldsymbol{\theta}).$$

The following analysis using the Laplace approximation is based on the same dataset generated in Section 3.1.1 for the posterior predictive measures of surprise. The sample consists in uniform and generalized Pareto data with a threshold $u = 5$.

The main steps to approximate the integral are the following. Firstly, we estimate the parameters of $h(\boldsymbol{\theta})$ using the `R` function `optim` for the true threshold $u = 5$. In the second place, the measure of surprise $m(\boldsymbol{x}_{\text{obs}})$ is estimated and finally, similarly to the posterior distribution, we compute the probability

$$\Pr^{m(\cdot)}\{m(\boldsymbol{X}) < m(\boldsymbol{x}_{\text{obs}})\}$$

in order to obtain the prior predictive p-values

$$p_{\text{prior}} = \Pr^{m(\cdot)}\{T(\boldsymbol{X}) \geq T(\boldsymbol{x}_{\text{obs}})\},$$

where $T(\boldsymbol{X}) = m(\boldsymbol{X})$ Afterwards several thresholds are taken into account from 2 to 4.9 and from 5.1 to 9 and the parameters estimated by using once

more the `R` function `optim` for each threshold. Finally, their measures of surprise and their prior predictive p-values are estimated. The results are illustrated in Figure 3.16: for each threshold the measure of surprise and the prior predictive p-values are displayed on the left and on the right panel, respectively. The graph of the marginal likelihoods has a similar behaviour as the one corresponding to the posterior marginal likelihoods in Figure 3.5: it indicates that below the threshold $u = 5$, the generalized Pareto model does not fit appropriately to the data. On the contrary, the p-values plot does not show the "surprise" as well as the posterior predictive p-values plot: in fact, the values below the true threshold are not as surprising as expected. The main reason is that the Laplace approximation does not compute precisely the marginal likelihoods as this method approximates these measures having a normal distribution. Therefore the results carried out by using the Laplace approximation have a poorer quality compared to the results obtained by using the posterior predictive approximation given by equation (3.1).

The same parameter stability plots and mean residual life plots displayed in Section 3.1.1 give evidence of the goodness of a generalized Pareto model from u equal to 5.

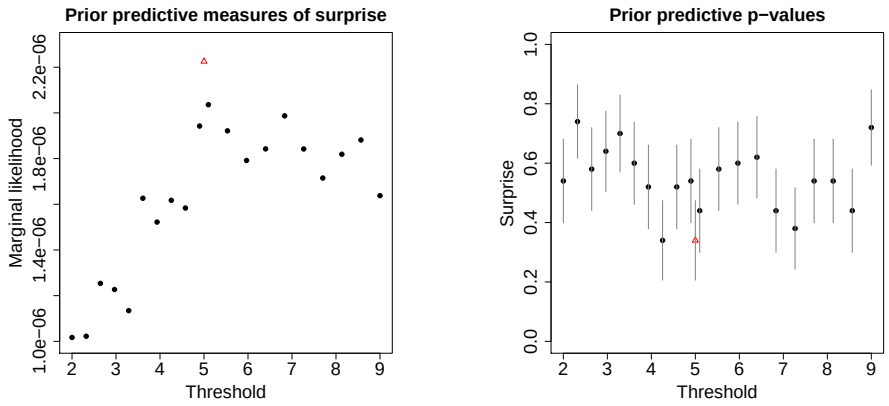

Figure 3.16: Plot of the measures of surprise $m(\boldsymbol{x})$ (left panel) and of the prior predictive p-values with their 95% central credibility intervals (right panel) for the different thresholds u from 2 to 9 estimated by using Laplace approximation. The red triangle corresponds to the proper threshold $u = 5$.

Conclusion

Our objective was to estimate the true threshold u for generalized Pareto models. A new tool, the measure of surprise, has been exploited. An important point to consider is that "surprise" exists only in the presence of uncertainty and is related to the expectations of the observer. In fact, it can only be defined in a relative, subjective, way.

Section 1 presents many different ways to measure the surprise. We decided to consider only the most natural ones, the prior and the posterior predictive measures of surprise. Unfortunately, several obstacles appear when we started coding the algorithms allowing the computations of both the marginal likelihoods and the predictive p-values.

The most important problem concerns the computation of the likelihood: in most of the cases the likelihood rounds everything to zero so that no credible measures of surprise have been obtained. Therefore to avoid this difficulty, we decided to replace the likelihood with the mean of the generalized Pareto density for each observation. Thanks to this alternative substitute, credible results for the surprise have been carried out using both the posterior and the prior predictive p-values, even if the posterior results are much more precise (see Sections 3.1 and 3.2.1).

Another complication during the computations is related to the importance sampling approach for the prior p-values. We introduced the importance sampling to estimate the marginal likelihoods as it is more precise than the Laplace approximation, which approximates the marginal likelihoods via a Gaussian distribution. We tried to use this method but the alternative likelihood substitute given in (3.3) cannot be used because the resulting hessian is not positive-definite (according to the function `optim` in `R`). In addition, even if we approximate the marginal likelihoods, considering that the logarithm of the marginal likelihoods is not equal to the integral of the logarithm of a density function (in this case $f(\boldsymbol{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})$), the results are not credible enough to allow identifications of the most appropriate threshold. For this reason, we have to look for an attractive and credible approximation

for the prior marginal likelihoods, which allows us to improve the results obtained by importance sampling. One possibility consists of choosing another prior density like a log-normal for $\tilde{\sigma}$ and a normal for ξ, in order to achieve a positive-definite hessian matrix, although the choice must be well considered. Another suggestion could be to investigate if any other kind of approximation of the marginal likelihoods exist instead of the mean of the densities or the logarithm.

In order to solve the difficulty related to the hessian, we could estimate the posterior densities of $\tilde{\sigma}$ and ξ by the Metropolis–Hastings algorithm so that we can derive a suitable importance sampling density. However, the required computation of these estimates of the marginal likelihood is very high. In fact, this procedure has to be repeated for each sample.

Other approaches to estimate the marginal likelihoods are proposed by Han and Carlin (2001). Unfortunately, because of the short time on our hands, we did not get further.

It is however important to underline that on the whole the surprise is a credible way of estimating the threshold. This means that, in principle, this approach works correctly and gives interesting results even if some problems with the computations are present and in order to avoid them some approximations have been taken without any particular demonstration.

The most important difference between the "surprise" method to select the threshold and the parameter stability plot or the mean residual life plot is that the first one is a Bayesian approach and the others are not. Let consider also the last method studied only theoretically, which is explained in Section 2.3.3, that is the mixture model where both data above and below the threshold are modelled. We have that it allows a Bayesian approach too but unlike the "surprise" approach, we have to specify a model, not necessarily correct, for the data below the threshold.

After having carried out all these studies, we noticed easily that the amount of computations to obtain the "surprise" is extremely large compared to the speed with that the mean residual life plot or the parameter stability plot are obtained. Then, this suggests us to look for other ways to implement the algorithms in order to reduce the time of the computations. One possibility is the use of sequential Monte Carlo methods or sequential importance sampling to get the posterior distributions. This approach is based on increasing slightly at each step the threshold u rather than performing MCMC independently on each threshold.

Finally, it should have been interesting to take into consideration also other measures of surprise, like the Kullback-Leibler distance between prior and posterior distribution.

Let $\{\Pr(M)\}_{M\in\mathcal{M}}$ be the prior distribution of a model M and $\Pr(D \mid M)$ be the associated likelihood function quantifying how likely any data observation D is under the assumption that the model M is correct. Then as surprise is based on the Bayes' theorem, we can transform prior belief distributions into posterior belief distributions (Itti and Baldi, 2006),

$$\forall M \in \mathcal{M}, \quad \Pr(M \mid D) = \frac{\Pr(D \mid M)}{Pr(D)}\Pr(M).$$

Under this hypothesis, if the posterior distribution is identical to the prior distribution, then no surprise exists for the new data D; this means that the observer's beliefs are unaffected. On the other hand, the new data D are considered as surprising if the posterior distribution $\Pr(M \mid D)$ is different from the prior distribution. This is the reason why the surprise is measured by computing the difference between the prior and the posterior distributions. This distance is better measured if we use the Kullback-Leibler divergence. Then, the surprise is defined by the average of the log-odd ratio,

$$S(D) = \int_{\mathcal{M}} \Pr(M \mid D) \log \frac{\Pr(M \mid D)}{\Pr(M)} dM,$$

taken with respect to the posterior distribution over the model space $\mathcal{M}$.

This measure of surprise may give more attractive results. During the implementation we have to pay attention at the model space because the integral takes into account all the models for the measurement D. In fact, we cannot use directly this formula but we have to convert the model space into a parameter space having a specific parametric family of distributions as a model (Ranganathan and Dellaert, 2009). In order to compute the expected surprise, Monte Carlo approximation of the integral is carried out, as for the prior and posterior measures. We generate N measurements $x \in D$ with the prior distribution $\Pr(M)$ and we take the average of these values,

$$E(x) = \frac{1}{N}\sum_{i=1}^{N} S(x_i).$$

A difficulty of this measure of surprise consists in the choice of the prior because the Kullback-Leibler divergence measures the distance between the prior and the posterior. Therefore, if the prior is uninformative, then all the data should be equally surprising from the point of view of the model. This means that we could always obtain the same distance between prior and posterior. This implies that no surprise exists and thus the model fits

appropriately to the data. Not necessarily it could be the case given that we have defined a non-informative prior.

Finally, we deduce that when we desire to test a hypothesis for which no alternative has been proposed, the measures of surprise are an appropriate tool, which give interesting results. As many different measures exist, it would be interesting to get a deeper insight into them and analyse more accurately the other measures too.

Acknowledgements:

I wish to thank Prof. Davison for giving to me the opportunity to go through this experience abroad and to carry out my project. Furthermore, I am grateful to him for his help and all advices given to me during these last months working on my thesis.

I am grateful to Mr. Sisson. I appreciate the interesting subject he found for me. I appreciate also his welcome in the University of New South Wales. Moreover, I wish to thank him for his helpfulness each time I needed.

Bibliography

Aitkin, M. (1991) Posterior Bayes Factors. *Journal of the Royal Statistical Society B* **53**, 111–142.

Bayarri, M. and Berger, J. (1999) Quantifying surprise in the data and model verification. *Bayesian Statistics* **6**, 53–82.

Bayarri, M. and Berger, J. O. (1997) Measures of Surprise in Bayesian Analysis. *Institute of Statistic and Decision Sciences* .

Bayarri, M. and Berger, J. O. (2000) P-values for composite null models. *Journal of the American Statistical Association* **95**, 1127–1142.

Bayarri, M. and Morales, J. (2003) Bayesian measures of surprise for outlier detection. *Journal of Statistical Planning and Inference* **111**, 3–22.

Behrens, C. N., Lopes, H. F. and Gamerman, D. (2004) Bayesian Analysis of Extreme Events with Threshold Estimation. *Statistical Modelling* **4**, 227–244.

Berger, J. (1980-85) *Statistical Decision Theory and Bayesian Analysis.* Second edition. New York: Springer-Verlag.

Box, G. (1980) Sampling and Bayes Inference in scientific modeling and robustness. *Journal of the Royal Statistical Society A* **143**, 383–430.

Castellanos, M. E. and Cabras, S. (2007) A default Bayesian procedure for the generalized Pareto distribution. *Journal of Statistical Planning and Inference* **137**, 473–483.

Coles, S. (2001) *An Introduction to Statistical Modeling of Extreme Values.* London: Springer.

Coles, S. G. and Powell, E. A. (1996) Bayesian methods in extreme value modelling: a review and a new developments. *International Statistical Review* **64**, 119–136.

Coles, S. G. and Tawn, J. A. (1996) A Bayesian analysis of extreme rainfall data. *Applied Statistics* **45**, 463–478.

Davison, A. C. (2003) *Statistical Models*. Cambridge: Cambridge University Press.

Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997) *Modelling Extremal Events for Insurance and Finance*. New York: Springer.

Evans, M. (1997) Bayesian inference procedures derived via the concept of relative surprise. *Communications in Statistics* **26**, 1125–1143.

Gelman, A., Meng, X. and Stern, H. (1996) Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica* **6**, 733–807.

Good, I. (1953) The appropriate mathematical tools for describing and measuring uncertainty. *Uncertainty and Business Decisions* pp. 19–34.

Good, I. (1956) The surprise index for the multivariate normal distribution. *Annals of Mathematical Statistics* **27**, 1130–1135.

Good, I. (1988) Surprise index. *Encyclopedia of Statistical Sciences* **7**, 104–109.

Guttam, I. (1967) The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society B* **29**, 83–100.

Han, C. and Carlin, B. P. (2001) Markov Chain Monte Carlo Methods for Computing Bayes Factors: A Comparative Review. *Journal of the American Statistical Association* **96**(455), 1122–1132.

Itti, L. and Baldi, P. (2006) Bayesian surprise attracts human attention. *Advances in Neural Information Processing System* pp. 1–8.

Jenkinson, A. F. (1955) The frequency distribution of the annual maximum (or minimum) values of meteorological events. *Quaterly Journal of the Royal Meteorological Society* **81**, 158–172.

Meng, X. (1994) Posterior Predictive p-values. *The Annals of Statistics* **22**, 1142–1160.

Ranganathan, A. and Dellaert, F. (2009) Bayesian Surprise and Landmark Detection. *International Center for Relativistic Astrophysics* pp. 2017–2023.

Renyi, A. (1961) On measures of entropy and information. *Proc. Fourth Berkeley Symposium Math. Stastist. Prob.* **1**, 547–562.

Robert, C. and Casella, G. (2005) *Monte Carlo Statistical Methods*. Second edition. New York: Springer.

Rubin, D. (1984) Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics* **12**, 1151–1172.

Sellk, T., Bayarri, M. and Berger, J. (2001) Calibration of p-values for Testing Precise Null Hypotheses. *Institute of Statistics and Decision Sciences* **55**(1), 62–71.

Weaver, W. (1948) Probability, Rarity, Interest and Surprise. *Scientific Monthly* **67**, 390–392.

Weaver, W. (1963) *Lady Luck: The Theory of Probability*. New York: Doubleday.