

Statistical Modeling of Extreme Events

A Thesis

submitted to

Indian Institute of Science Education and Research Pune

in partial fulfillment of the requirements for the

BS-MS Dual Degree Programme

by

Birbal Prasad



Indian Institute of Science Education and Research Pune
Dr. Homi Bhabha Road,
Pashan, Pune 411008, INDIA

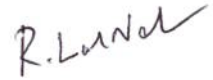
March, 2016

Supervisor: Dr. Laks Raghupathi

Birbal Prasad 2016

All rights reserved

This is to certify that this dissertation entitled " Statistical Modeling of Extreme Events" towards the partial fulfillment of the **BS-MS Dual Degree** programme at the **Indian Institute of Science Education and Research, Pune** represents original research carried out by **Birbal Prasad**, Indian Institute of Science Education and research under the supervision of **Dr. Laks Raghupathi**, Computational Researcher, **Shell India Markets Pvt. Ltd., Bangalore**, during the academic year **2015-2016**.



Dr. Laks Raghupathi

Committee:

Dr. Laks Raghupathi

Prof. Uttara V. Naik Nimbalkar

Declaration

I hereby declare that the matter embodied in the report entitled "Statistical Modeling of Extreme Events" are the results of the investigations carried out by me at Shell India Markets Pvt. Ltd., Bangalore, under the supervision of Dr. Laks Raghupathi, except where stated otherwise and the same has not been submitted elsewhere for any other degree.



Birbal Prasad

Acknowledgments

I would like to acknowledge the Computational Center of Excellence (CCOE) at Shell India Markets Pvt. Ltd. (SIMPL) for providing an internship opportunity without which this thesis would not have been possible. I thank Vianney Koelman and Bertwim van Beest from CCOE for the encouragement and support.

Thanks to Dr. Laks Raghupathi (CCOE, SIMPL), who initiated the project, introduced me to extreme events modeling and for supervising the project, whilst guiding my first attempts in this field. Also to Dr. David Randell and Dr. Philip Jonathan from the Statistics and Chemometrics group, Shell Global Solutions (UK) for their valuable inputs throughout the project. From sorting out the thesis plan to guiding me through all this while by teaching more about the statistical application of extreme value theory. Finally, to Prof. Uttara Naik Nimbalkar at IISER Pune for being my thesis committee member and providing me the local support and encouragement as well as for the valuable inputs for my project.

Birbal Prasad

Abstract

Extreme value (**EV**) analysis involves the estimation of the probability of events that are unusually large or small. EV methods have a wide range of application from modeling extreme wave heights and water levels in hydrology, structural engineering to share price return levels in finances. In case of univariate independent and identically distributed (i.i.d.) random variables, a number of statistical models do exist in literature. However, for dependent and non-stationary multivariate extremes the development of different statistical model remains an ongoing area of research.

Most of my reading and work has been motivated by the application of the EV analysis in designing oil and gas producing facilities, off-shore or on-shore for extreme ocean environments. It becomes essential to model covariate effects (wave directions, seasons etc.) for the data observed over the years in the oceans. We begin with the study of different existing models which incorporate these covariate effects such as conditional extremes model (Heffernan and Tawn, 2004) and Non-stationary conditional extremes (**NSCE**) model (Raghupathi et al. 2016). However, the application of the frequentist NSCE model seems to be computationally challenging and expensive. So, next we study a piece-wise model for a sample of peaks over threshold which is non-stationary with respect to multidimensional co-variates, estimated using a computationally efficient Bayesian inference. We then study the convergence diagnostics for the Markov Chain Monte Carlo (MCMC) procedure used in estimation of the desired Bayesian model by application on synthetic data. Most importantly, we study the problem of threshold estimation using the Bayesian inference in application for one covariate (direction).

Contents

1	Introduction to EV Theory	1
1.1	Univariate case	1
1.2	Multivariate Case	4
1.2.1	Common marginals	4
1.2.2	Types of model for multivariate extremes	5
	Extremal dependence model	5
	Conditional extremes model	7
1.3	Computational challenges and further study	13
2	Markov Chain Monte Carlo (MCMC) sampling	15
2.1	Markov chains	15
2.2	Metropolis-Hastings algorithm	16
2.2.1	Convergence of M-H algorithm	18
2.3	Gibbs' sampling	19
2.4	Convergence diagnostics	20
3	Threshold Estimation Using Bayesian Inference	23
3.1	Mixture model	24
3.2	Estimation of Non-exceedance probability	25
3.2.1	Application to synthetic 1D-case	26
	Convergence study of the MCMC chains	27
	Choice of beta priors	33
3.2.2	Discussion and Conclusions	39
3.3	Way Forward	46
	References	49

Chapter 1

Introduction to EV Theory

This chapter is an extensive literature survey of the main advances done in extreme value theory in last few decades. From univariate case to multivariate case this chapter introduces the basic concepts in extreme value theory and statistical inference on extremes. Moreover it also discusses some of the limitations of different extreme events modeling in univariate and multivariate cases. Though we state some of the main results but the proofs for the main results have been omitted as these are well documented elsewhere in literature.

1.1 Univariate case

Let X_1, X_2, \dots, X_n be a sequence of random variables, independent and have a common distribution function F . There are basically two approaches to model extreme events which lays down the foundation or represents the cornerstone of extreme value theory. The block maxima approach and the threshold models. The block maxima approach focuses on the statistical behavior of

$$M_n = \max\{X_1, X_2, \dots, X_n\} \quad (1.1)$$

In theory the asymptotic distribution of M_n can be described and stated by the extremal types theorem (**Fisher-Tippett-Gnedenko theorem**) [1] as follows

Theorem 1. *If \exists sequences of normalizing constants $\{a_n > 0\}$ and $\{b_n\}$ such that*

$$P\{(M_n - b_n)/a_n \leq z\} \rightarrow G(z) \quad (1.2)$$

as $n \rightarrow \infty$ and where G is a non-degenerate distribution function. Then, G belongs to one

of the following families

$$A : G(z) = \exp\{-\exp[-((z - b)/a)]\}, \quad -\infty < z < \infty; \quad (1.3)$$

$$B : G(z) = \begin{cases} 0, & z \leq b \\ \exp\{-((z - b)/a)^{-\alpha}\}, & z > b; \end{cases} \quad (1.4)$$

$$C : G(z) = \begin{cases} \exp\{-[((z - b)/a)^\alpha]\}, & z < b \\ 1, & z \geq b; \end{cases} \quad (1.5)$$

where $a > 0$ and b are both parameters. Also, $\alpha > 0$ in case of families B and C . These three classes of distribution are known as **Gumbel**(A), **Fréchet** (B) and **Weibull** (C) families. However, there is another result which provides a more complete generalization of the asymptotic distribution of M_n and can be stated formally as

Theorem 2. *If \exists sequences of normalizing constants $\{a_n > 0\}$ and $\{b_n\}$ such that*

$$P\{(M_n - b_n)/a_n \leq z\} \rightarrow G(z)$$

as $n \rightarrow \infty$ and where G is a non-degenerate distribution function. Then, G belongs to the following family

$$G(z) = \exp\{-[1 + \xi((z - \mu)/\sigma)]^{-1/\xi}\} \quad (1.6)$$

and is defined on $\{z : 1 + \xi(z - \mu)/\sigma > 0\}$, where $-\infty < \mu < \infty, \sigma > 0$ and $-\infty < \xi < \infty$.

This family of distributions is called as the **generalized extreme value (GEV)** family of distributions. In fact, the families of distributions described by equations 1.3, 1.4 and 1.5 as A, B and C belongs to this GEV family for different cases of parameterization of the GEV family. When $\xi > 0$ Eqn.(1.6) corresponds to the Fréchet and to Weibull when $\xi < 0$. Also, when the GEV family is interpreted as a limit as $\xi \rightarrow 0$ it leads to the Gumbel family. Motivated by Theorem 2, the GEV gives us a model for the distribution of block maxima. The parameter estimation in this model can be done using a number of proposed techniques but the one that stands out is the use of likelihood based techniques. It is well established that the usual regularity conditions are violated in cases like when $\xi \leq -0.5$ but this situation is rarely encountered in application. Hence, the maximum likelihood approach can be used to get the parameter estimates.

However, in many cases we don't have data in the form of block maxima and sometimes extremes are scarce. This has led to search for characterizations of extreme values when the data is not in the form of block maxima. In literature there are two widely known characterizations for this. Among these two, one is based on exceedances of a high threshold [2] whereas the other is based on the r largest order statistics behavior in a block [1], but for small values of r .

Let X_1, X_2, \dots be a sequence of i.i.d. random variables and define

$$M_n^{(k)} = k\text{th largest of } \{X_1, X_2, \dots, X_n\} \quad (1.7)$$

The next important result here identifies the limiting behavior of $M_n^{(k)}$, for a fixed k as $n \rightarrow \infty$.

Theorem 3. *If \exists a sequence of constants $\{a_n > 0\}$ and $\{b_n\}$ such that*

$$P\{(M_n - b_n)/a_n \leq z\} \rightarrow G(z)$$

as $n \rightarrow \infty$ and where G is a non-degenerate distribution function and is the GEV distribution function given by 1.6. Then, for a fixed k ,

$$P\{(M_n^{(k)} - b_n)/a_n \leq z\} \rightarrow G_k(z) \quad (1.8)$$

on $\{z : 1 + \xi(z - \mu)/\sigma > 0\}$, where

$$G_k(z) = \exp\{-\tau(z)\} \sum_{s=0}^{k-1} \frac{\tau(z)^s}{s!} \quad (1.9)$$

and

$$\tau(z) = [1 + \xi(\frac{z - \mu}{\sigma})]^{-1/\xi} \quad (1.10)$$

The parameters here are the ones of the limiting GEV distribution. But in case we have a complete vector

$$\mathbf{M}_n^{(r)} = (M_n^{(1)}, \dots, M_n^{(r)}) \quad (1.11)$$

which usually happens, there is another important result which gives us the joint density function of the limit distribution. We will omit that result here as it is not required but can be found in **Coles S. (2001)** [1].

The last result in this section gives us the main result in asymptotic model characterization

for extremes based on exceedances of a high threshold.

Theorem 4. *Let X_1, X_2, \dots be a sequence of i.i.d. random variables with a common distribution F and that Theorem 2 holds. Then, for a large enough threshold u , the distribution function of $X - u$, conditional on $X > u$, is defined by*

$$H(y) = 1 - \left(1 + \frac{\xi y}{\sigma + \xi(u - \mu)}\right)^{-1/\xi} \quad (1.12)$$

on $\{y : y > 0\}$ and $\{1 + \frac{\xi y}{\sigma + \xi(u - \mu)} > 0\}$

This family of distribution defined here by Eqn. 1.12 is known as the **generalized Pareto (GP)** family. Also, the parameters of this distribution are uniquely determined by the parameters of the associated generalized extreme value distribution of the block maxima.

1.2 Multivariate Case

In case of multivariate extremes, it becomes difficult to derive the asymptotic form of the distribution of threshold exceedances and others theoretically. However, **Beirlant et al. (2004)** [3] introduces us to multivariate extremes but it is the work of **Heffernan and Tawn (2004)** [4] which provides us a framework for multivariate extreme value modelling in a more flexible way. The conditional extremes model described by them can be easily implemented and extended and is the one useful in modelling covariate effects. For spatial extremes one can use methods related to max-stable processes. However, there are some unrealistic assumptions which goes into these methods related to max-stable processes. So, for a sample of values drawn from some multivariate distribution if we want to model extremes there are four different approaches in literature. Extremal dependence models, parametric models, max-stable models and the conditional extremes model. But, many of these models for the multivariate extremes are easily described and applied when all the variables follow a common marginal distribution.

1.2.1 Common marginals

Let X_1, X_2, \dots, X_p be p random variables and $\{x_{ij}\}_{i=1, j=1}^{n, p}$ be a sample of n observations on these p variables. To transform the marginals to a common marginal we use the idea of probability integral transform (**Jonathan et al. 2010**) [5] and the steps as described below:

- i. We model variable x_j marginally independently using an appropriate distribution F_j (e.g.: GP for threshold exceedances).
- ii. Evaluate the cumulative distribution function (CDF) $\hat{F}_j(x_{ij})$ for each observation x_{ij} of the variable x_j .
- iii. Find the value of the argument x_{ij}^* of the desired common marginal CDF $F^*(x_{ij}^*)$ such that

$$\hat{F}_j(x_{ij}) = F^*(x_{ij}^*) \quad (1.13)$$

for all i, j and thus establishing a transformed sample $\{x_{ij}\}_{i=1, j=1}^{n, p}$ having a common marginal distribution F^* .

The typical forms for F^* are the standard Gumbel CDF $F^*(x) = \exp(-\exp(-x))$ for $x \in (-\infty, \infty)$, and the standard Frechet CDF $F^*(x) = \exp(-\frac{1}{x})$ for $x > 0$.

1.2.2 Types of model for multivariate extremes

As described above we have four different approaches namely extremal dependence models, parametric models, conditional extremes model and the max-stable models. However our main focus remains on extremal dependence and conditional extremes model.

Extremal dependence model

As described in **Jonathan et al. (2013)** [6], for simplicity, at first, let us consider a bivariate random variable (X, Y) having a common marginal distribution function. Then, (X, Y) is said to be asymptotically dependent if

$$\lim_{x \rightarrow \infty} P(X > x | Y > x) > 0 \quad (1.14)$$

and asymptotically independent if

$$\lim_{x \rightarrow \infty} P(X > x | Y > x) = 0 \quad (1.15)$$

Now, for large x , it becomes important to look at quantities like the joint survivor function i.e.: $P(X > x, Y > x)$ and the conditional probability $P(X > x | Y > x)$. So, let us consider a bivariate random variable (X_F, Y_F) where X_F and Y_F have unit Frechet marginal

distributions, as

$$P(X_F \leq f) = e^{-1/f}, \quad f > 0. \quad (1.16)$$

In case X_F and Y_F are independent, they are also asymptotically independent as

$$P(X_F > f | Y_F > f) = P(X_F > f) \rightarrow 0 \quad \text{as } f \rightarrow \infty. \quad (1.17)$$

However, X_F and Y_F are asymptotically dependent in case $X_F = Y_F$, as

$$P(X_F > f | Y_F > f) = 1 > 0 \quad (1.18)$$

Using the theory of regular variation as described in **Bingham et al. (1987)** [7], one can assume that $P(X_F > f, Y_F > f)$ is regularly varying at ∞

$$\lim_{f \rightarrow \infty} \frac{P(X_F > sf, Y_F > sf)}{P(X_F > f, Y_F > f)} = s^{-1/\eta}. \quad (1.19)$$

Here, $-1/\eta, \eta \in (0, 1]$ is the index with which the regular variation is assumed for some fixed $s > 0$. Now, transforming the variables to Gumbel (X_G, Y_G) scale such that

$$P(X_G < g) = \exp(-e^{-(g)}) = P(X_F < e^g), \quad \text{for } g \in (-\infty, \infty) \quad (1.20)$$

we have, for $t > 0$ and large g

$$P(X_G > g + t, Y_G > g + t) = e^{-t/\eta} P(X_G > g, Y_G > g). \quad (1.21)$$

This is easy to see as Eqn. 1.19, for large values of f can be written as

$$P(X_F > sf, Y_F > sf) \approx s^{-1/\eta} P(X_F > f, Y_F > f). \quad (1.22)$$

Thus, the simple case discussed above suggests that the joint tail of a distribution is described using the coefficient of tail dependence η , and this is the what quantifies the extent of extremal dependence for the distribution. In general, for Gumbel margins and for large values (x_1, x_2, \dots, x_n) (from above and [8]), we have

$$P(X_1 > x_1 + t, X_2 > x_2 + t, \dots, X_n > x_n + t) \approx \exp\left(-\frac{t}{\eta}\right) P(X_1 > x_1, X_2 > x_2, \dots, X_n > x_n) \quad (1.23)$$

for fixed $t > 0$.

However, Eqn.1.23 can only be used if a suitable set $(X_1 > x_1, X_2 > x_2, \dots, X_n > x_n)$ exists for an extreme set of the form of $(X_1 > x_1 + t, X_2 > x_2 + t, \dots, X_n > x_n + t)$ and we have a method to estimate η . When $\eta = 1$ we say that the distribution is asymptotically dependent and independent otherwise. Using Ledford and Tawn,(1996) [9], η can be estimated directly from the sample. So, now the problem is of the threshold selection for sets $(X_1 > x_1, X_2 > x_2, \dots, X_n > x_n)$ and most of the methods for this are subjective. One also needs to be careful in case of asymptotic independence as wrong assumptions can be problematic.

Conditional extremes model

Directly fitting the parametric multivariate EV distributions has its own limitations. One limitation is that the exact distributional form is unknown. Also, since the samples may not be extreme in all their components and so the direct estimation by Eqn.1.23 for extremal dependence model (as described before) may not give us proper results. Thus there needs to be a different approach to model such extreme events. Though there are a number of approaches suggested in the literature but the conditional approach of Heffernan and Tawn, (2004) [4] stands out in providing a robust model and overcoming difficulties to some extent that the other conditional model have.

1. Conditional extremes (Heffernan and Tawn, 2004)

They present a semi-parametric approach. Based on asymptotic arguments, they derive a parametric equation for the form for one variable conditional on a large value of another. The choice of this statistical model is motivated by a range of theoretical results and are important to understand.

Assumption of limit representation

Let $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)$ be a random variable with Gumbel marginal distributions. Since, we are concerned about the conditional distribution, consider the conditional distribution $P(\mathbf{Y}_{-i} \leq \mathbf{y}_{-i} | Y_i = y_i)$. As $y_i \rightarrow \infty$, let us look at the limiting distribution of these conditional distributions. Now, as in univariate theory, the limiting distribution here also needs to be non-degenerate in all the margins. Hence, let $\mathbf{a}_i(y_i)$ and $\mathbf{b}_i(y_i)$ be vector of normalizing functions (constants in univariate case), defined from $\mathfrak{R} \rightarrow \mathfrak{R}^{n-1}$, for each given i , which can

be chosen such that $\forall \mathbf{z}_{|i}$ (fixed) and any sequence of y_i values as $y_i \rightarrow \infty$,

$$\lim_{y_i \rightarrow \infty} P[\mathbf{Y}_{-i} \leq \mathbf{a}_{|i}(y_i) + \mathbf{b}_{|i}(y_i)\mathbf{z}_{|i} | Y_i = y_i] = G_{|i}(\mathbf{z}_{|i}). \quad (1.24)$$

Thus, under assumption (1.24), we have, $Y_i - u_i$ and $\mathbf{Z}_{|i}$ to be independent in the limit conditionally on $Y - i > u_i$, as $u_i \rightarrow \infty$. Also, these variables $Y_i - u_i$ and $\mathbf{Z}_{|i}$ have limiting marginal distributions as exponential and $G_{|i}(\mathbf{z}_{|i})$ respectively (see [4]).

For the marginal and dependence characteristics of $G_{|i}(\mathbf{z}_{|i})$, let us define $G_{j|i}(\mathbf{z}_{j|i})$ to be the conditional distribution of

$$Z_{j|i} = \frac{Y_j - a_{j|i}(y_i)}{b_{j|i}(y_i)} \quad (1.25)$$

given $Y_i = y_i, y_i \rightarrow \infty, j \neq i$. Also, $a_{j|i}(y_i)$ and $b_{j|i}(y_i)$ are the component functions of $\mathbf{a}_{|i}(y_i)$ and $\mathbf{b}_{|i}(y_i)$. Hence, we have, $G_{j|i}$ to be the marginal distribution of $G_{|i}$ corresponding to Y_j . Moreover, for the elements of \mathbf{Y}_{-i} to be mutually conditionally independent given Y_i , the following condition needs to be satisfied

$$G_{|i}(\mathbf{z}_{|i}) = \prod_{j \neq i} G_{j|i}(z_{j|i}). \quad (1.26)$$

Choice of normalization functions

The main task now is how do we choose these normalization functions that has to be used. Heffernan and Tawn, 2004 [4] uses the two different ideas to choose the appropriate $\mathbf{a}_{|i}(y_i)$ and $\mathbf{b}_{|i}(y_i)$. First they identify the normalizing functions in terms of characteristics of the conditional distribution of $\mathbf{Y}_{-i} | Y_i$ and then make the observation that the normalizing functions as well as the limit distribution are not unique. Mathematically, if $\mathbf{a}_{|i}(y_i)$ and $\mathbf{b}_{|i}(y_i)$ give a non-degenerate limit distribution $G_{|i}(\mathbf{z}_{|i})$, then the normalizing functions

$$\mathbf{a}_{|i}^*(y_i) = \mathbf{a}_{|i}(y_i) + \mathbf{A}\mathbf{b}_{|i}(y_i); \mathbf{b}_{|i}^*(y_i) = \mathbf{B}\mathbf{b}_{|i}(y_i) \quad (1.27)$$

gives us the non-degenerate limit distribution as $G_{|i}(\mathbf{B}\mathbf{z}_{|i} + \mathbf{A})$. Here, \mathbf{A} and \mathbf{B} , with $\mathbf{B} > 0$ are arbitrary vector constants. Then using the idea from Leadbetter et. al(1983) [10] that this is a unique way when two different limits with no mass at infinity can arise and hence the class of distribution is unique up to type. Thus, the normalizing functions $\mathbf{a}_{|i}(y_i)$ and $\mathbf{b}_{|i}(y_i)$ can be identified up to the constants \mathbf{A} and \mathbf{B} . The next result in [4] forms the base in choice of the normalizing functions.

Theorem 5. *Suppose that the vector random variable \mathbf{Y} has an absolutely continuous joint*

density. If, for a given i , the vector functions $\mathbf{a}_{|i}(y_i)$ and $\mathbf{b}_{|i}(y_i) > 0$ satisfy the limiting property (1.24), then the components of these vector functions corresponding to the variable Y_j , for each $j \neq i$, satisfy, up to type, the following properties

$$\lim_{y_i \rightarrow \infty} [F_{j|i}(a_{j|i}(y_i)|y_i)] = p_{j|i} \quad (1.28)$$

where $p_{j|i}$ is a constant in the range $(0, 1)$. $F_{j|i}(a_{j|i}(y_i)|y_i)$ is the conditional distribution function and

$$b_{j|i}(y_i) = h_{j|i}(a_{j|i}(y_i)|y_i)^{-1}. \quad (1.29)$$

Here, $h_{j|i}$ is a conditional hazard function defined as

$$h_{j|i}(y_j|y_i) = \frac{f_{j|i}(y_j|y_i)}{1 - F_{j|i}(y_j|y_i)} \quad (1.30)$$

and $f_{j|i}(y_j|y_i)$ is the conditional density function of $Y_j|Y_i = y_i$ and $F_{j|i}(y_j|y_i)$ is the conditional distribution function of the same.

For proof refer Heffernan and Tawn, 2004(Appendix) [4]. It is then well established that the normalizing functions are all special cases of the parametric family [4]

$$\mathbf{a}_{|i}(y) = \mathbf{a}_{|i}y + I_{(\mathbf{a}_{|i}=0, \mathbf{b}_{|i}<0)}(\mathbf{c}_{|i} - \mathbf{d}_{|i}\log(y)); \mathbf{b}_{|i}(y) = y^{\mathbf{b}_{|i}} \quad (1.31)$$

where, $\mathbf{a}_{|i}, \mathbf{b}_{|i}, \mathbf{c}_{|i}, \mathbf{d}_{|i}$ are vector constants and I , an indicator function. The vector of constants, for all $j \neq i$ are

$$0 \leq a_{j|i} \leq 1; -\infty < b_{j|i} < 1; -\infty < c_{j|i} < \infty; 0 \leq d_{j|i} \leq 1 \quad (1.32)$$

Conditional dependence model

As described above, this is a semi-parametric model motivated by the findings of the discussed results in the sections assumptions of limit distributions and choice of normalization functions. The model describes the what happens to variable \mathbf{Y}_{-i} with large Y_i and we look at the model structure and its properties.

Let us suppose that for each $i = 1, 2, \dots, n$ there exists a high threshold u_{Y_i} such that we model

$$P[\mathbf{Y}_{-i} \leq \mathbf{a}_{|i}(y_i) + \mathbf{b}_{|i}(y_i)\mathbf{z}_{|i}|Y_i = y_i] = P[\mathbf{Z}_{|i} < \mathbf{z}_{|i}|Y_i = y_i] = G_{|i}(\mathbf{z}_{|i}) \quad \forall y_i > u_{Y_i}. \quad (1.33)$$

Here, $G_{|i}$ is the distribution function of the standardized residual $\mathbf{Z}_{|i}$. Also, that the standardized residual is independent of the random variable Y_i for all $Y_i > u_{Y_i}$. Now, to characterize the extremal dependence, we need to estimate the functions $\mathbf{a}_{|i}(y_i)$, $\mathbf{b}_{|i}(y_i)$ and $G_{|i}$. Using the parametric model described in Eqn.1.31, we can do the estimation of vector constants $\mathbf{a}_{|i}$, $\mathbf{b}_{|i}$, $\mathbf{c}_{|i}$ and $\mathbf{d}_{|i}$. For the estimation of the distribution function $G_{|i}$, Heffernan and Tawn, 2004 uses a non-parametric approach as the limiting condition described by 1.24 doesn't have restrictions on the structure of $G_{|i}$. The estimation is done using the empirical distribution of replicates of $\hat{\mathbf{Z}}_{|i}$ which is defined as

$$\hat{\mathbf{Z}}_{|i} = \frac{\mathbf{Y}_{-i} - \hat{\mathbf{a}}_{|i}(y_i)}{\hat{\mathbf{b}}_{|i}(y_i)} \quad (1.34)$$

for $Y_i = y_i > u_{Y_i}$ and where $\hat{\mathbf{a}}_{|i}$, $\hat{\mathbf{b}}_{|i}$ be the estimators of $\mathbf{a}_{|i}$ and $\mathbf{b}_{|i}$ respectively. Thus, we have a multivariate dependence model which is semi-parametric regression model and is of the form

$$\mathbf{Y}_{-i} = \mathbf{a}_{|i}(y_i) + \mathbf{b}_{|i}(y_i)\vec{Z}_{|i} \quad Y_i = y_i > u_{Y_i}. \quad (1.35)$$

Even though the model seems to be complete in itself under certain assumptions but the problem of threshold estimation (in this case u_{Y_i}) still remains the core of this model. In particular, it becomes even difficult to estimate the threshold in application. Though there are different techniques such as using quantile regression and then looking at the model fits but each one of these techniques have their own limitations. The next model studied in this chapter gives us a way to not only completely characterize conditional extremes but also addresses the problem of threshold estimation to some extent.

2. Non-stationary conditional extremes (NSCE) [11]

Since characterizing the joint structure of extremes of environmental variables is critical to understanding the ocean environments. Sometimes it becomes necessary to model covariates effects. For ex.: Let's say that we are trying to model wave heights (H_s) that occur during hurricanes and storms in the deep oceans and sea. Oil and gas producing facilities built offshore are very likely to get affected by these extreme events such as storms and hurricanes. Covariates such as directions, seasons, etc influences these extreme events, e.g.: wave heights are larger for storms in monsoon season in south China sea than in rest of the year. The NSCE model described by Jonathan et. al (2014) [11] and Raghupathi et al. (2016) [12] provides a complete characterization of the full joint non-stationary extremal

structure for any particular values of covariates. This model uses non-crossing quantile regression techniques for threshold estimation and incorporates the conditional extremes model of Heffernan and Tawn, 2004 [4].

Consider a set of random variables X_1, X_2, \dots, X_p and the respective multidimensional covariate vectors $\theta_1, \theta_2, \dots, \theta_p$.

Assumptions

- a. Marginal extreme value behavior of X_k can be explained adequately by θ_k alone.
- b. Pairwise extremal dependence of X_j and X_k can be explained adequately by $\theta_j \cup \theta_k$ of covariates.

Model for threshold exceedances

For each X_k , for a given fixed value \mathbf{t}^k of θ_k , we look for the distribution of threshold exceedances. Let $\psi_k^*(\mathbf{t}_k)$ be a pre-selected quantile threshold associated with non-exceedance probability (NEP) τ_k^* where NEP is defined as

$$P(X_k \leq \psi_k^*(\mathbf{t}_k) | \theta_k = \mathbf{t}_k) = \tau_k^* \quad (1.36)$$

Then, it is assumed that the threshold exceedances are GP distributed [11].

$$P(X_k > x_k | X_k > \psi_k^*(\mathbf{t}_k), \theta_k = \mathbf{t}_k) = (1 + \frac{\xi_k(\mathbf{t}_k)}{\zeta_k(\mathbf{t}_k)}(x_k - \psi_k^*(\mathbf{t}_k)))^{-\frac{1}{\xi_k^*(\mathbf{t}_k)}} \quad (1.37)$$

where $x_k > \psi_k^*(\mathbf{t}_k)$, $(1 + \frac{\xi_k(\mathbf{t}_k)}{\zeta_k(\mathbf{t}_k)}(x_k - \psi_k^*(\mathbf{t}_k))) > 0$ and $\zeta_k(\mathbf{t}_k) > 0$. Now, the estimates for the values of the parameter functions ξ_k and ζ_k at $[\mathbf{t}_k^i]_{i=1}^n$ of covariate values can be obtained by maximum likelihood estimation. In particular, by minimizing the negative log-likelihood [11], in this case

$$l_{GP,k} = \sum_{i=1}^n \log \zeta_k(\mathbf{t}_k^i) + \frac{1}{\xi_k(\mathbf{t}_k^i)} \log(1 + \frac{\xi_k(\mathbf{t}_k^i)}{\zeta_k(\mathbf{t}_k^i)}(x_k^i - \psi_k^*(\mathbf{t}_k^i))). \quad (1.38)$$

However, the basic question here is to ask about how do we choose a threshold. Jonathan et. al, 2014 [11], for the choice of thresholds uses a quantile regression (QR) model. In this model, for each X_k , the quantile threshold ψ_k corresponding to the quantile probability τ_k is estimated by minimizing the roughness penalized loss criterion, i.e.:

$$l_{QR,k}^* = [\tau \sum_{i, r_k^i \geq 0} |r_k^i| + (1 - \tau) \sum_{i, r_k^i < 0} |r_k^i|] + \lambda_{\psi_k} R_{\psi_k} \quad (1.39)$$

and the residuals $r_k^i = x_k^i - \psi_k(\mathbf{t}_k^i)$. Here, R_{ψ_k} and λ_{ψ_k} are the parameter roughness and roughness coefficient respectively. In this model, the value of λ_{ψ_k} is chosen such that it

maximizes the predictive performance of the QR model and this is achieved by using cross-validation techniques. Similarly, R_{ψ_k} , parameter roughness can be evaluated in closed form for efficient estimation.

Incorporating conditional extremes model

To incorporate the Heffernan and Tawn, 2004 model for conditional extremes, we need the random variables with standard Gumbel marginal distributions. Thus, we need to transform the marginals to standard Gumbel scale. In practice, this is done using the probability integral transform (Jonathan et al., 2010, 2014) [5] [11] and can be done for any choice of \mathbf{t}_k of θ_k . The dependence model for multidimensional responses X_1, X_2, \dots, X_p on Gumbel scale can be expressed in terms of the set of pairwise dependence models $X_j|X_k$ for $j = 1, 2, \dots, p$ and $k = 1, 2, \dots, p$. Now, as per the assumption (b) and using Heffernan and Tawn, 2004 asymptotic arguments, Jonathan et. al, 2014 [11] assumes the form of $X_j|X_k$, for large values of X_k to be

$$(X_j|X_k = x_k, \theta_j \cup \theta_k = \mathbf{t}_{jk}) = \alpha_{jk}(\mathbf{t}_{jk})x_k + x_k^{\beta_{jk}(\mathbf{t}_{jk})}Q_{jk}(\mathbf{t}_{jk}) \quad \text{for } x_k > \psi_k(\mathbf{t}_k). \quad (1.40)$$

Here, the threshold $\psi_k(\mathbf{t}_k)$ is a non-stationary threshold with respect to the covariate vector θ_k and Q_{jk} is a random variable drawn from an unknown distribution whose characteristics vary smoothly with $\theta_j \cup \theta_k$. α_{jk} and β_{jk} are parameter functions where $\alpha_{jk} \in [0, 1]$ $\beta_{jk} \in (-\infty, 1]$. Now, let

$$Z_{jk} = (Q_{jk} - \mu_{jk})/\sigma_{jk} \quad (1.41)$$

be a standardized variable such that it follows a common distribution G_{jk} , independent of covariates and where $\sigma_{jk} > 0$. Then, Eqn.1.40 can be written in terms of Z_{jk} and parameter estimation can be done. The parameters $\alpha_{jk}, \beta_{jk}, \mu_{jk}$ and σ_{jk} are estimated for the model described in the same way as in conditional extremes model of Heffernan and Tawn, 2004 [4]. The distribution function G_{jk} is then assumed to be a standard normal distribution [11] and the corresponding negative log-likelihood for a sample of pairs (x_j^i, x_k^i) is derived. The functional form as described by Jonathan et. al, 2014 for the negative log-likelihood is

$$l_{CE,jk} = \sum_{i, x_k^i > \psi_k^i} \log(\sigma_k^i(\mathbf{t}_{jk}^i)(x_k^i)^{\beta_{jk}(\mathbf{t}_{jk}^i)}) + \frac{(x_j^i - (\alpha_{jk}(\mathbf{t}_{jk}^i)x_k^i + \mu_k^i(\mathbf{t}_{jk}^i)(x_k^i)^{\beta_{jk}(\mathbf{t}_{jk}^i)}))^2}{2(\sigma_k^i(\mathbf{t}_{jk}^i)(x_k^i)^{\beta_{jk}(\mathbf{t}_{jk}^i)})^2} \quad (1.42)$$

for $x_k^i > \psi_k(\mathbf{t}_k^i)$.

For regulating the parameter roughness one uses the same approach as in case of model for threshold exceedances by penalizing the negative log likelihood. Also the parameter

roughness are evaluated similarly and the roughness coefficient are again estimated using the cross-validation method (Jonathan et al., 2014 [11]). Residuals are inspected to confirm the model fit and show if it is reasonable or not.

Model validation

Estimation of parameter uncertainties is carried out by using the non-parametric bootstrap procedure by drawing bootstrap samples from the original samples. Now, to validate the NSCE model, following procedure is followed:

1. An underlying true non-stationary bivariate extreme value model whose characteristics are known is specified.
2. Using the true model, one or more samples of data can be simulated.
3. Then, fit an NSCE model to the simulated data.
4. Comparison of the estimates of the marginal and conditional CDF's for arbitrary combinations of covariats and specified return periods by simulation under the true model and the fitted NSCE model.

1.3 Computational challenges and further study

In application, simulation studies using this NSCE model takes a lot of time and are computationally expensive [13]. This computation can be made faster by avoiding bootstrapping and cross-validation steps used in the algorithm for this model. Also, the threshold estimation in this model is done using the P-splines regression approach discussed in Bollaerts et. al, (2006) and Bollaerts, (2009). Even though this QR model can be formulated in terms of a linear programme [11], the threshold estimation using this approach becomes difficult in application. However, for application purposes we need a more computationally efficient model. The Bayesian non-stationary marginal extremes model described by Randell et. al, (2015) helps us in avoiding these difficulties to some extent. It incorporates specifying a prior information for estimating the non-exceedance probability (NEP) on which the threshold estimation is based. Thus we consider the Bayesian model described by Randell et. al, (2015) and try to address the problem of non-stationary threshold estimation in application. Chapter 2 covers the important aspect of sampling in Bayesian inference. It gives a brief description of the methods and approaches used as sampling techniques. Chapter 3 covers the study of threshold estimation using Bayesian inference for 1-dimensional marginal case.

Chapter 2

Markov Chain Monte Carlo (MCMC) sampling

As discussed later in the **chapter 3**, we see that the most important thing in Bayesian inference is the posterior distribution. Often, the posterior distribution cannot be obtained in closed form and hence the posterior of interest needs to be estimated by simulations and this is one of the main challenges of Bayesian analysis. Algorithms like Markov Chain Monte Carlo (**MCMC**) provide us a way to sample from and is widely used these days in Bayesian statistics to sample from the posterior distribution of interest. The approach of computing statistics of the concerned posterior with arbitrary precision given a large enough sample of simulated draws is called Monte Carlo simulation [14]. Here we discuss two most commonly used MCMC methods namely the Metropolis-Hastings (**MH**) algorithm and Gibbs sampling. These methods are based on some important results from Markov chains theory and can be used to sample when we don't have the full conditionals available to us for estimation or when we have closed form conditionals respectively. The notations used in this section are the same as described in Letham et al.(2012) [15]

2.1 Markov chains

Let $\theta^0, \theta^1, \dots$ be a sequence of random variables with $\theta^t \in \mathfrak{R}^d$. This sequence is a discrete state Markov chain if it satisfies the Markov property [15]

$$P(\theta^t | \theta^{t-1}, \dots, \theta^1) = P(\theta^t | \theta^{t-1}). \quad (2.1)$$

In other words, conditional on $\theta^0, \theta^1, \dots, \theta^{t-1}, \theta^t$, the next state only depends on the present state θ^{t-1} and not on all the past ones. Also, for a transition from θ^{t-1} to θ^t , let us define the transition kernel to be

$$K(\theta^{t-1}, \theta^t) = P(\theta^t | \theta^{t-1}) \quad (2.2)$$

and denote the unconditional probability distribution over states at a time t as $\pi^t(\theta)$. Now, the unconditional probability distribution $\pi(\cdot)$ on the state space is called as **invariant**(stationary) if the following equation is satisfied.

$$\pi K = \pi \quad (2.3)$$

Also, K is reversible with respect to π if

$$K(\theta^{t-1}, \theta^t)\pi(\theta^{t-1}) = K(\theta^t, \theta^{t-1})\pi(\theta_t) \quad (2.4)$$

$\forall \theta^{t-1}, \theta^t$. **Eqn.(2.4)** is what is referred to as the detailed balance equation in literature. The next result formally states a relationship between reversibility and invariant property of Markov chains.

Theorem 6. *If a distribution π of a Markov chain is reversible then it is invariant.*

However, for MCMC purposes we need all the Markov chains used to have a unique stationary distribution and limiting distribution. It is well studied in literature that not all Markov chains have a stationary distribution. A Markov chain can also have more than one stationary distribution and that not all the stationary distributions are also limiting distributions. Under conditions such as ergodic Markov chain, reversibility, etc., the sequence of state distributions will converge to a unique distribution $\pi(\theta)$, the stationary distribution [16].

2.2 Metropolis-Hastings algorithm

In principle, the target of MCMC is to construct a Markov chain whose invariant (stationary) distribution ($\pi(\theta)$) is the posterior distribution $P(\theta|y)$. In this algorithm, we start the chain say as θ^0 at time $t = 1$ and then specify a proposal distribution $J(\theta, \theta^*)$. Next, we propose new states from this proposal distribution and calculate the acceptance probability α based on which we accept or reject the proposed new state. The algorithms is as follows:

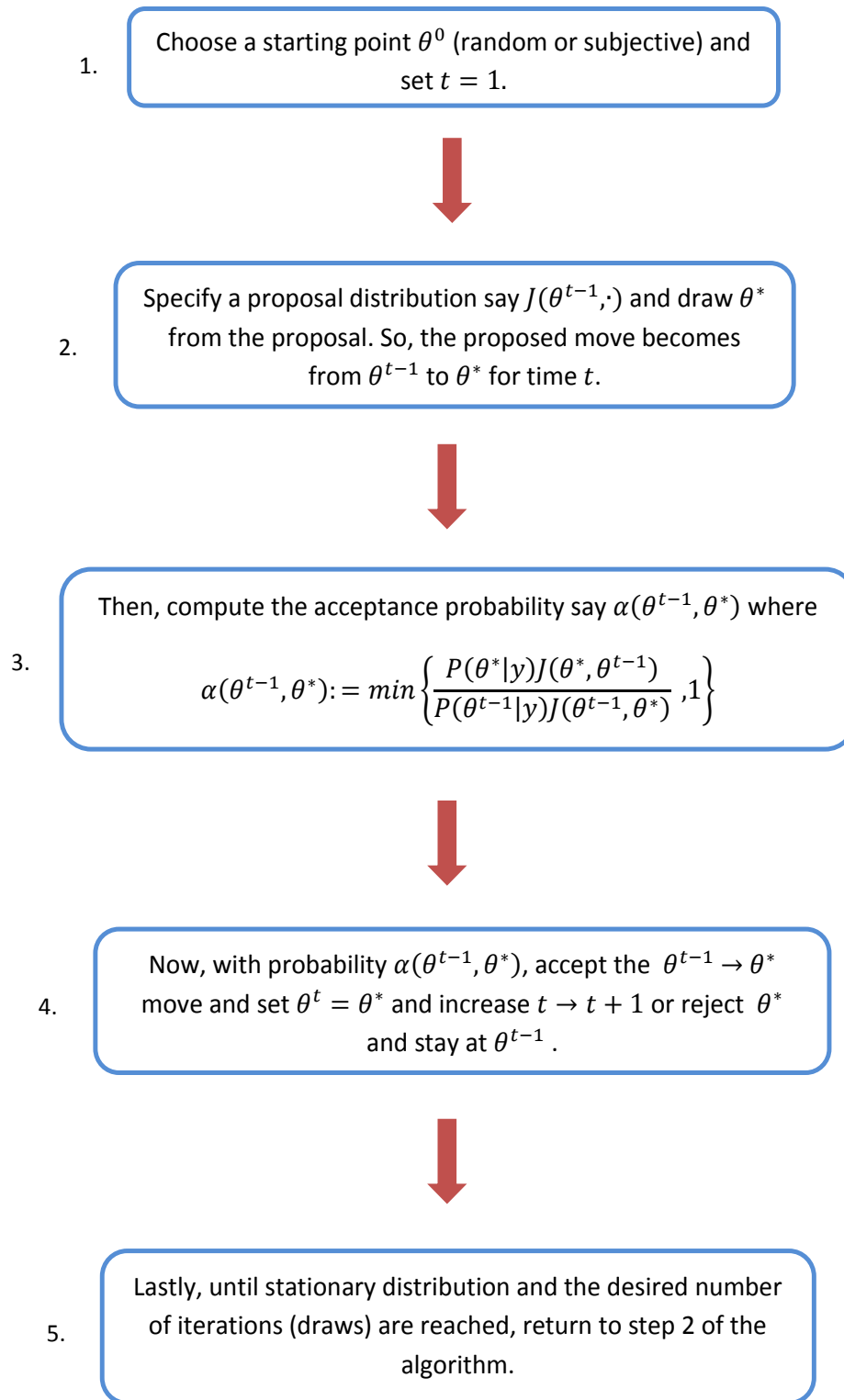


Figure 2.1: The Metropolis-Hastings algorithm and its flow step by step.

2.2.1 Convergence of M-H algorithm

The computation of the acceptance probability(α) in step 3 of the M-H algorithm requires the computation of ratios of the posterior probabilities. This step is done in MCMC without even worrying about the problematic normalization integral and is the key to MCMC methods. The other basic idea used in M-H algorithm is that if the chain is simulated long enough then eventually we will end up drawing from the posterior distribution. The next result formally states this as a theorem. For existence of the stationary distribution, the proposal distribution is specified such that there is a positive probability of reaching any state from any other state.

Theorem 7. *Let the proposal distribution $J(\theta, \theta^*)$ be such that the chain $\theta^0, \theta^1, \dots$ produced by M-H algorithm has a unique stationary distribution, then the stationary distribution $\pi(\cdot)$ is posterior distribution $P(\theta|y)$.*

Proof: Using Theorem 5 and Eqn.(2.4), it suffices to prove that if the posterior $P(\theta|y)$ satisfies Eqn.(2.4) then it is in fact a stationary distribution i.e.: to show that

$$K(\theta, \theta^*)P(\theta|y) = K(\theta^*, \theta)P(\theta^*|y), \forall \theta, \theta^*. \quad (2.5)$$

Here, the transition kernel $K(\cdot)$ is from the M-H algorithm and is

$$\begin{aligned} K(\theta, \theta^*) &= (P(\text{proposing } \theta^*) \times P(\text{accepting } \theta^* | \theta^* \text{ was proposed})) \\ &= J(\theta, \theta^*) \times \alpha(\theta, \theta^*). \end{aligned} \quad (2.6)$$

Now, let $\theta \rightarrow \theta^*$ be any transition of states and that $\alpha \leq 1$ for this transition (WLOG), i.e.:

$$\alpha(\theta, \theta^*) \leq 1 \Rightarrow \frac{J(\theta^*, \theta)P(\theta^*|y)}{J(\theta, \theta^*)P(\theta|y)} \leq 1 \quad (2.7)$$

and

$$\alpha(\theta^*, \theta) = 1. \quad (2.8)$$

So, the LHS of Eqn.(2.5) becomes

$$\begin{aligned}
K(\theta, \theta^*)P(\theta|y) &= J(\theta, \theta^*)\alpha(\theta, \theta^*)P(\theta|y) && [Substituting from Eqn.2.6] \\
&= J(\theta, \theta^*)\frac{J(\theta^*, \theta)P(\theta^*|y)}{J(\theta, \theta^*)P(\theta|y)}P(\theta|y) && [From Eqn.2.7] \\
&= J(\theta^*, \theta)P(\theta^*|y) && [On simplification] \\
&= J(\theta^*, \theta)P(\theta^*|y)\alpha(\theta^*, \theta) && [From Eqn.2.8] \\
&= K(\theta^*, \theta)P(\theta^*|y) \\
&= RHS of Eqn.2.5
\end{aligned} \tag{2.9}$$

Thus, we have the detailed balance equation satisfied by the posterior and hence the posterior is a stationary distribution.

2.3 Gibbs' sampling

This MCMC sampling technique is a special case M-H algorithm and is much faster. Gibbs' sampling is only used when we have full conditional distributions available. Let $\theta = [\theta_1, \theta_2, \dots, \theta_d] \in \mathbb{R}^d$ and that we can sample from the conditional distribution

$$P(\theta_j | \theta_1, \dots, \theta_{j-1}, \theta_{j+1}, \dots, \theta_d, y) \tag{2.10}$$

even if we cannot draw from the posterior $P(\theta|y)$ directly. Then each of the posterior variables $\theta_1, \theta_2, \dots, \theta_d$ is updated by Gibbs' sampler but one at a time. At each step of the algorithm, all the posterior variables are kept constant except for one of the j' 's. This one variable is then updated by drawing from its conditional posterior distribution as in Eqn.2.10. Similarly, the algorithm updates each of the variables subsequently in an iterative updating process. It is then if the algorithm draws long enough then eventually it simulate draws from the full posterior. In this algorithm the proposal distribution is the conditional posterior distribution. The algorithm as described below in Figure 2.2 is such that we accept every move and the probability of accepting any new proposed move is 1 [17]. This is why Gibbs' sampling is a faster algorithm as compared to ordinary Metropolis-Hastings algorithm. The update process for each component of θ^t can happen in any order and it is not necessary to follow 1, ..., d order.

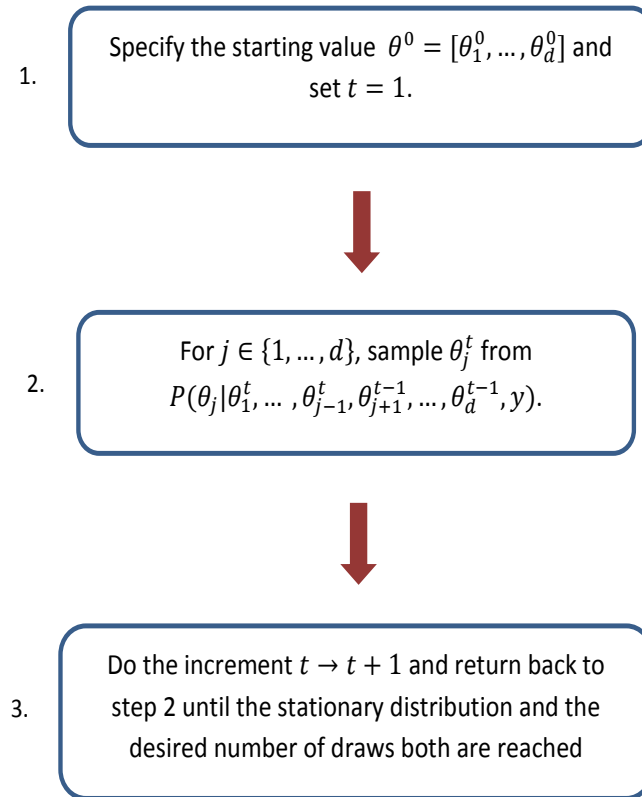


Figure 2.2: The Gibbs' Sampling algorithm.

2.4 Convergence diagnostics

For practical application, there are few issues which are encountered and hence one needs to look at different convergence diagnostics to assess convergence. The two important issues are **burn-in** and **slow mixing** of chains. It is often the case that the MCMC simulation is started at a random starting value (as is the case in chapter 3) in the parameter space. This sometimes leads to the problem that the starting point can indeed be far from the high density parts of the posterior. So, the values obtained from the simulation are not the representative of the true posterior in the early stages. So, we have a part of the chain (early stages) which is unlikely to be in the sample from the posterior. This is what is termed as burn-in. Often these samples are thrown away and are not used. However, in application one has to decide based on different convergence diagnostics and decide on how many samples are to be discarded or is to be termed as burn-in. The rest of the MCMC chain is what we call as the stationary part, i.e.: the part where the chain is assumed to have converged.

The other issue is the slow mixing of chains and this can happen because of the step size specified for the algorithm being too small or too large. A very small step size would mean a high acceptance rate and this leads to the samples (successive) moving very slowly. If the step size specified is too large then this leads to very low acceptance rate. This is because the proposals specified are more likely to be in the regions of very low probability density and in turn this leads to the chain moving too slowly around the space. It can also happen that if the starting value is chosen to be very far away from the expected true value then the chains can get stuck and not move at all with too small or too large step size values. Thus, the idea to look for an optimal step size such that the mixing is better and fast seems to be a valid task. To address these issues and to validate convergence of the Markov chains in MCMC there are a number of convergence diagnostics one can look at in Bayesian procedure [18] [19] [20] [21] [22] [23]. However, there are some of them commonly used and we have used some of them in our analysis later in chapter 3 and they are summarized as follows:

1. **Autocorrelation:-** Measures correlation (dependency) among the Markov chain samples. Correlations is measured for different lags and high correlation between long lags shows a poor mixing and poor convergence.
2. **Manhattan plots:-** This is more of a visual inspection diagnostics which shows if the chain is mixing well or if it has finished burning in (ex.: Fig.3.1 and 3.2). If the chain gets stuck in some parts of the parameter space or if the mean or variance of the chain change drastically with number of iterations then non-convergence is indicated.
3. **Effective sample size:-** This also is a measure of mixing of Markov chains as large difference between the effective sample size and the simulated sample size shows poor mixing.
4. **Gelman-Rubin test:-** This is a one-sided test based on the ratio test statistic. This test uses parallel chains with dispersed initial values to test whether they all converge to the same target distribution. Failure indicates the presence of multi-mode posterior distribution or the need to run a longer chain i.e.: the burn-in is yet to be completed.

In the next chapter, we study the threshold estimation problem using Bayesian inference. We apply the Metropolis-Hastings algorithm to sample from the posterior for the non-exceedance probability and study the convergence of it using some of the convergence diagnostics stated above.

Chapter 3

Threshold Estimation Using Bayesian Inference

Even though the NSCE model described by Raghupathi et al. (2016) [12] and Jonathan et. al (2014) [11] using the maximum likelihood estimation approach gives us a detailed characterization of non-stationary extreme events, it is computationally very expensive for higher dimensions. Also, the specification of the extreme value threshold is generally difficult in the model described by Raghupathi et al. (2016) and Jonathan et. al (2014) for characterizing storm peak significant wave height. Thus it is important to have a computationally efficient models in order to use them for real world applications.

Bayesian inference gives us an intuitive framework for environmental applications of the extreme value analysis. It allows incorporation of prior knowledge and a complete uncertainty quantification in a single step. In literature, there are many applications of Bayesian inference for extreme events models. Coles and Tawn (1996) [24] [25] and Coles and Tawn (2005) [26] [27] uses Bayesian analysis for extreme rainfall data and for improved flood risk assessment respectively. Beirlant et al. (2004) [3] considers the specification of priors for parameters of the extreme value model. Guedes-Soares (2001) [28] uses Bayesian inference to estimate the distributions of significant wave height. Bayesian inference has also been used in case of extremes of wild fires as discussed by Mendes et al. (2010) [29]. More recently, Davison et al.(2012) [30] [31] describes Bayesian hierarchical models for spatial extremes. There are other applications discussed by some other work in the literature.

In this chapter we study a piece-wise model described by Randell et al. (2015) [32] for sample of peaks over threshold which is non-stationary with respect to the multidimensional covariates and estimated using Bayesian inference. This model is a computationally efficient

in application and provides a detailed characterization of non-stationary extreme environments. Next, we study the problem of threshold estimation in application of this model to a directional analysis of storm peak significant wave height generated synthetic data. This is achieved by studying the convergence of the MCMC algorithm used to sample from the posterior probability distribution for the extreme value threshold non-exceedance probability and analyzing the negative log-likelihood.

3.1 Mixture model

As described in Randell et al. (2015) [32], theoretically and historically, the wave models derived gives us the idea that a suitably parameterized Weibull distribution provides us with a detailed description of the body of the distribution of wave heights, crest elevations and storms. On the other hand, asymptotic theory for extreme values suggest that EV model is required for describing largest threshold exceedances. Thus, the mixture model is described as a piece wise model.

Let x be the magnitude of H_s (wave heights), then for a response x , the non-exceedances of some threshold ψ are assumed to follow a three-parameter truncated Weibull distribution.

$$f_{TW}(x|\tau, \alpha, \gamma) = \frac{f_W(x|\alpha, \gamma)}{F_W(\psi|\alpha, \gamma, \tau)} \quad \text{for } x \in [\zeta, \psi] \quad (3.1)$$

where ζ is the non-stationary peak picking threshold used for storm peak identification before the model estimation is done. τ is the EV threshold non-exceedance probability (NEP) which is assumed to be stationary with respect to the covariates. Also,

$$f_W(x|\alpha, \gamma) = \frac{\gamma}{\alpha} \left(\frac{x - \zeta}{\alpha}\right)^{\gamma-1} \exp\left(-\left(\frac{x - \zeta}{\alpha}\right)^\gamma\right) \quad (3.2)$$

and

$$F_W(\psi|\alpha, \gamma, \tau) = 1 - \exp\left(-\left(\frac{\psi - \zeta}{\alpha}\right)^\gamma\right) = \tau \quad (3.3)$$

However, for a response x , the exceedances of some threshold ψ follows a GP distribution i.e.: x follows a GP distribution above some EV threshold ψ .

$$f_{GP}(x|\tau, \alpha, \gamma, \sigma, \xi) = \frac{1}{\sigma} \left(1 + \frac{\xi}{\sigma} (x - \psi)\right)^{-1/(\xi-1)} \quad (3.4)$$

Here, the threshold ψ is defined as a function of τ i.e.:

$$\zeta + \alpha(-\log(1 - \tau))^{-1/\gamma} \quad (3.5)$$

and is specified such that Eqn.3.3 holds.

Now, the basic work is the estimation of all the model parameters ($\rho, \alpha, \gamma, \sigma$ and ξ as well as the estimation of the threshold non-exceedance probability τ). Practically, we should expect these model parameters to vary smoothly with respect to the directional covariates (or higher dimensional covariates). In this model, this is achieved by expressing each of the model parameters in terms of an appropriate basis for the domain D of covariates. Then using the idea provided by Eilers (1998) [33] and Eilers et al. (2006) [34], the rest of the spline parameterization work is done. In particular, the whole problem of model parameter estimation reduces to estimation of appropriate sets of spline parameters for each of these model parameters. However, the estimation of the threshold NEP τ still remains a critical and elementary problem in application and rest of the model description would only go through if we estimate the threshold NEP. To estimate τ we use our prior belief and make the following prior specification.

The extreme threshold NEP τ is assumed to follow a beta distribution $B(\alpha_\tau, \beta_\tau)$, with fixed parameters. Here, α_τ and β_τ are the two positive shape parameters that control the shape of the Beta distribution. The choice of Beta distribution as the prior probability distribution of τ seems suitable because of the fact that Beta distribution is defined on the interval $[0, 1]$. The choice of Beta distribution parameters is done such that we have sufficient sample to estimate the other model GP parameters in a better way and ensure a reasonable EV tail fit. But, our focus here is to study if there is a range of parameter values for Beta or there is a particular parameter value for which this model would suitably work in application. The rest of our study is focused on addressing this issue.

3.2 Estimation of Non-exceedance probability

As in most of the cases, we don't have the posterior distribution of model parameters to be in a closed form. But, we do have different Markov Chain Monte Carlo (MCMC) algorithms available in literature to sample from full conditionals. Thus, the posterior inference is made with the help of MCMC sampling.

As described in **section 3.1**, we have the prior probability distribution for τ in the form of Beta distribution. At each iteration of the MCMC chain, the sampling is done from the full

conditional distribution of extreme threshold NEP τ . So, whenever the the full conditionals are available in closed form, Gibbs sampling (**Chapter 2**) is used and Metropolis-Hastings (MH)(**Chapter 2**) otherwise. In this case, the full conditional to be sampled from is

$$f(\tau|\mathbf{y}, \Omega \setminus \tau) \propto f(\mathbf{y}|\tau, \Omega \setminus \tau) \times (\tau) \quad (3.6)$$

but is not available in closed form. Hence, Metropolis-Hastings is used to sample from full conditionals in application.

3.2.1 Application to synthetic 1D-case

The model described in section 3.1 was applied to a synthetic data sample (Fig. 3.1) of around 2000 storm peaks (wave heights H_s) for directional covariate. The synthetic data

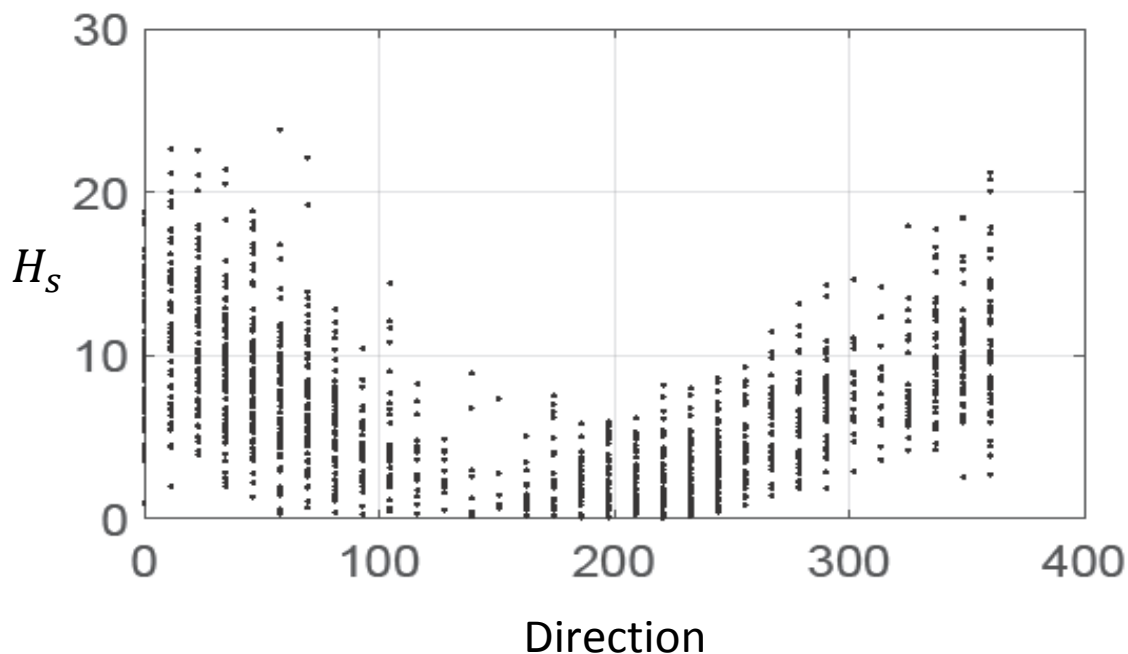


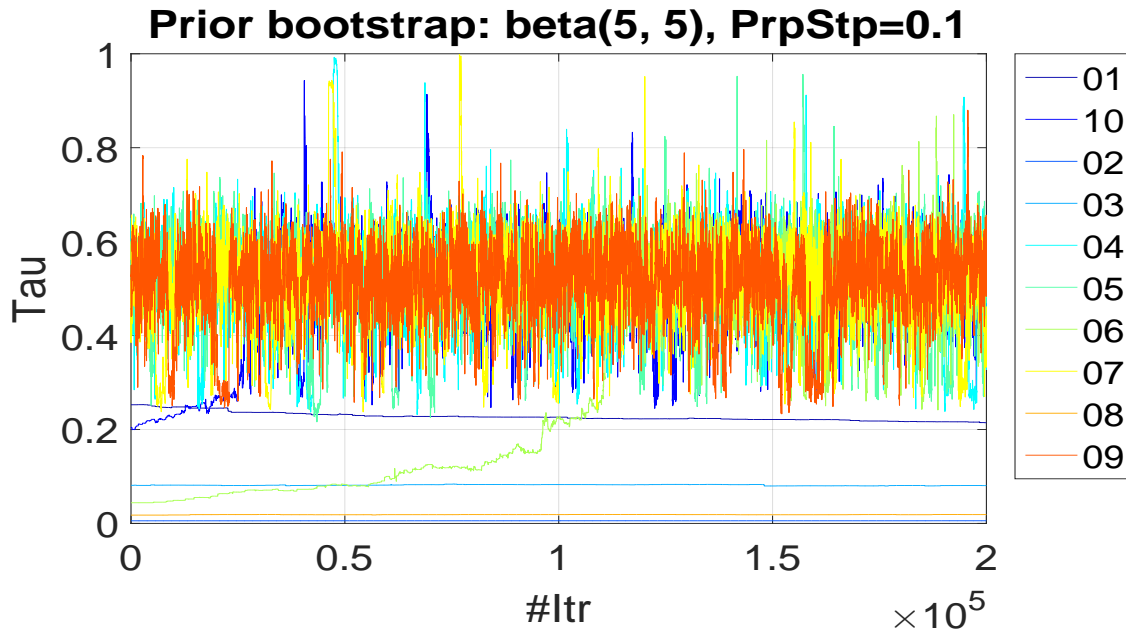
Figure 3.1: Storm peak significant wave height H_s (in meters) on direction. The direction is expressed in degrees (0, 360) clockwise with respect to north.

sample is chosen for application first because of the fact that the true NEP(τ) is known to be 0.6 and then it would be easier to validate the model. In this application, the posterior distribution of τ was estimated using MCMC, incorporating 20,000 burn-in iterations and

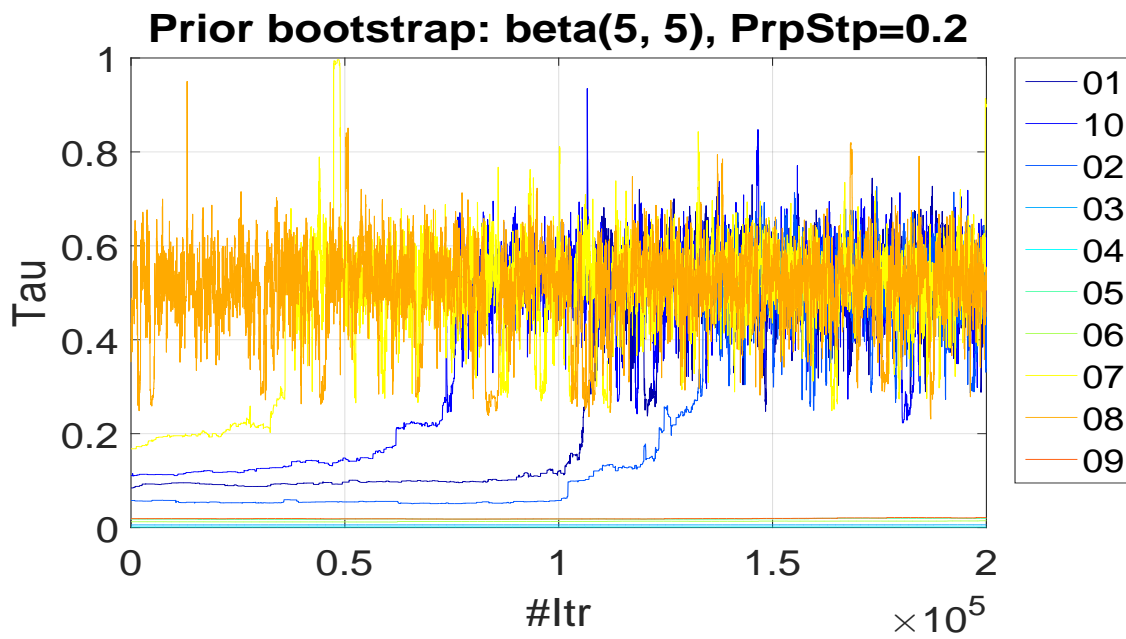
200,000 subsequent iterations. Then, we studied the convergence for the MCMC algorithm used here and later concluding the choice of beta priors for which the model works well.

Convergence study of the MCMC chains

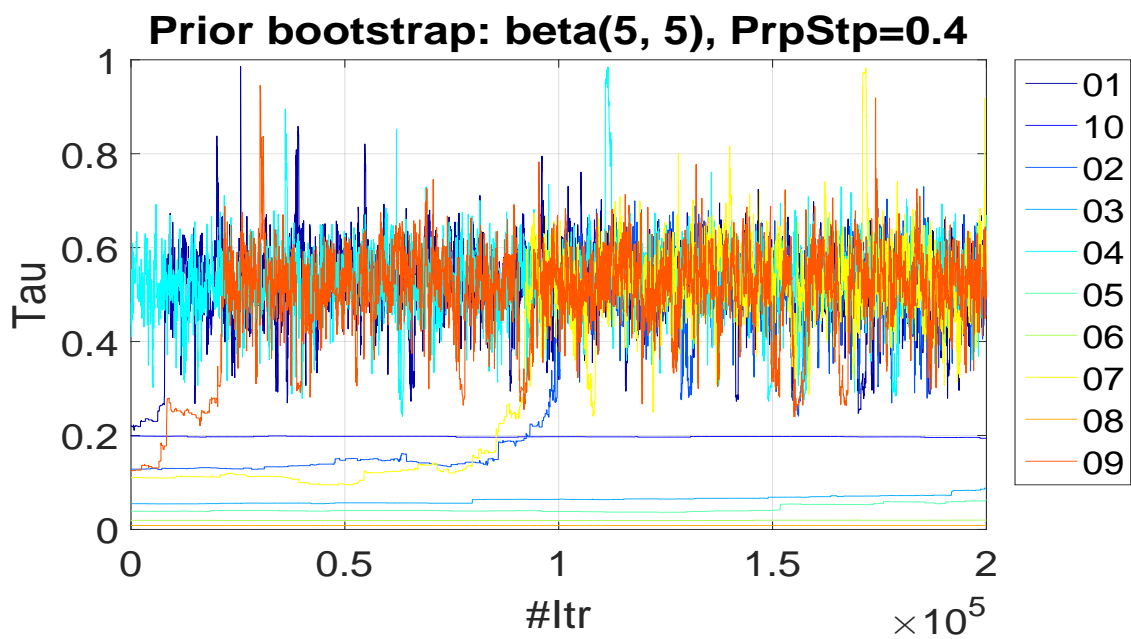
We looked at mixing of different MCMC chains for τ by plotting the Manhattan patterns for individual chains. The algorithm starts these chains randomly for each run by specifying a random value sampled from beta distribution. Visual inspection of these plots (**Figure 3.2**) and (**Figure 3.3**) for a relatively weaker beta prior $B(5, 5)$ and a stronger prior $B(55, 45)$ suggest good mixing for step sizes such as 0.5, 0.6 and 0.7, etc. A chain that shows good mixing traverses its posterior space quickly. This means that it can go from one remote region to another of the posterior relatively quicker. In case of $B(5, 5)$, for step sizes 0.1, 0.2 and 0.4 the chain mixing is good but for some chains are very slow or not at all. Similarly, in case of $B(55, 45)$, the mixing is not good for the step size 0.1 whereas the mixing improves with increasing step sizes of 0.3, 0.5 and 0.7. Figure 3.2 also suggests that the mixing improves for step sizes such as 0.5 and 0.6 and these seem to be the optimal step size for this MCMC algorithm for sampling. Moreover, the choice of number of iterations (200,000) is also justified as almost all the chains seem to have a constant mean and variance after 150,000 number of iterations, suggesting convergence of the chain.



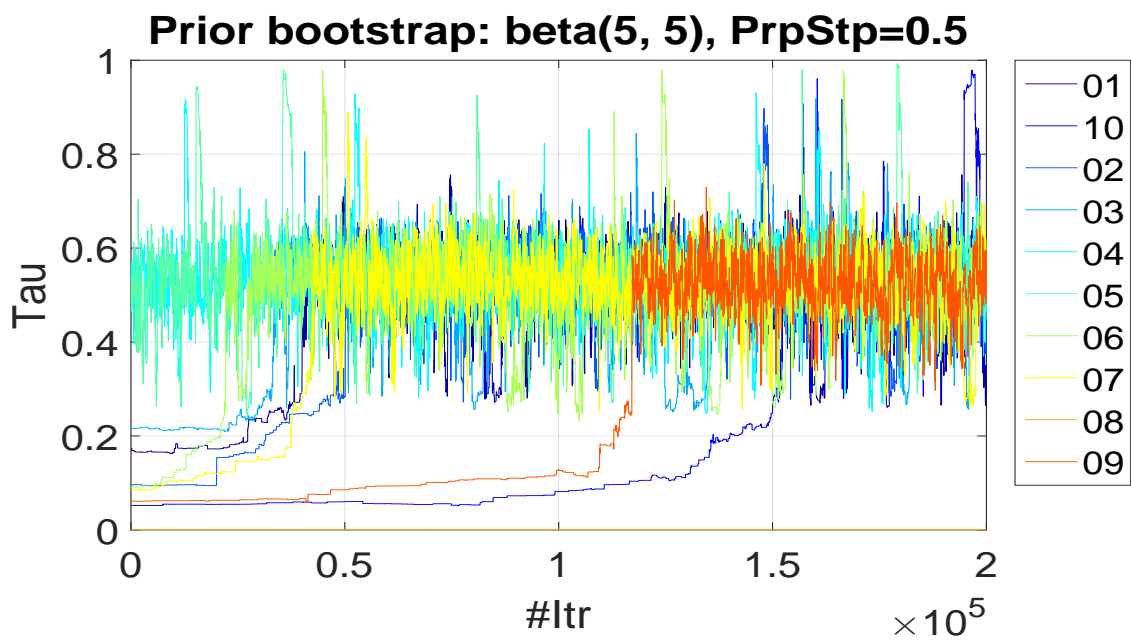
(i)



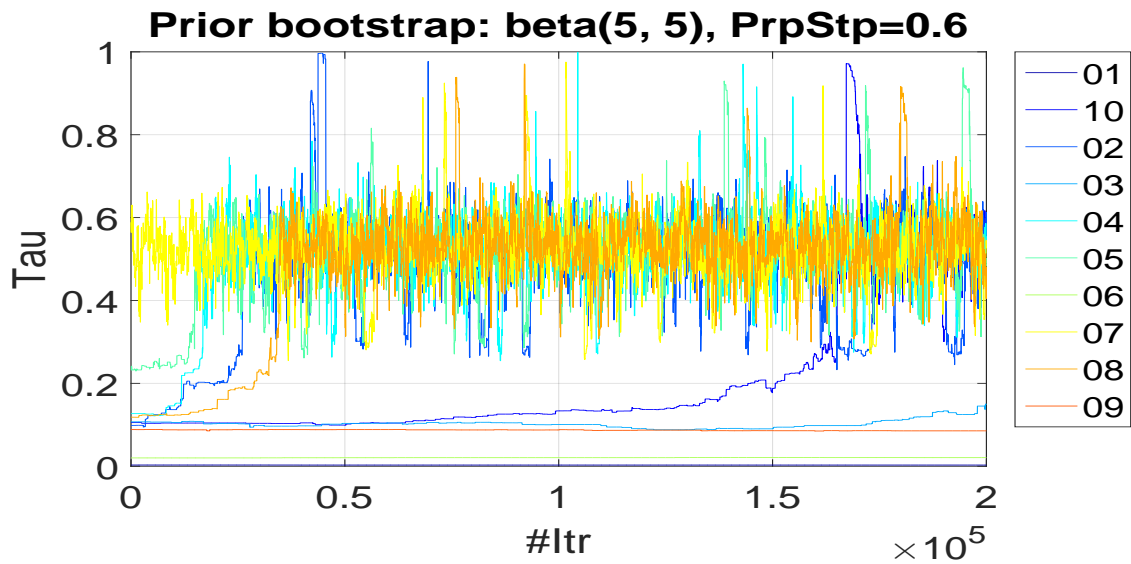
(ii)



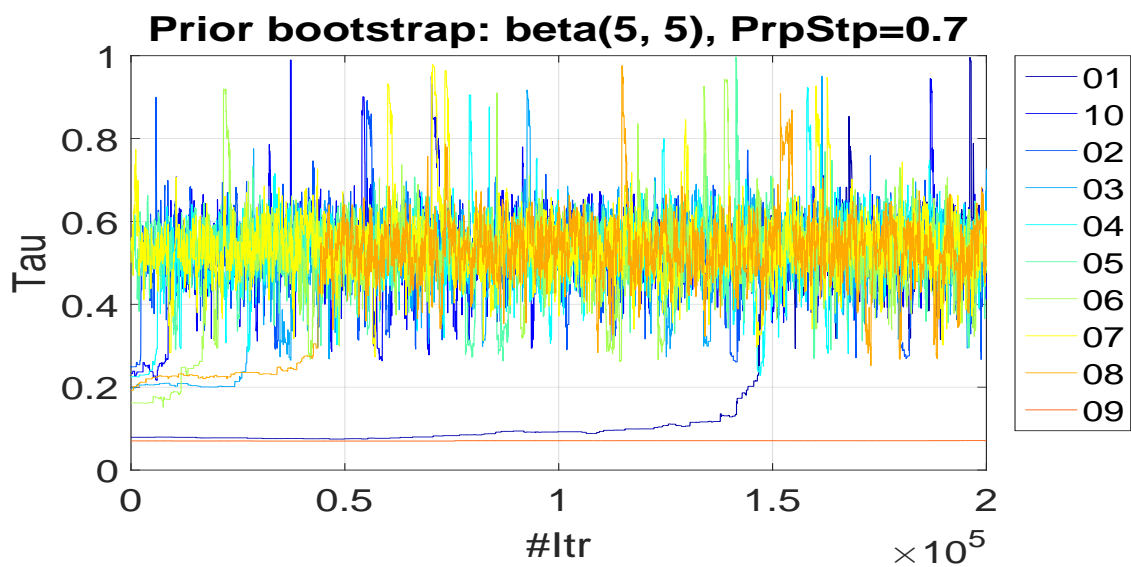
(iii)



(iv)

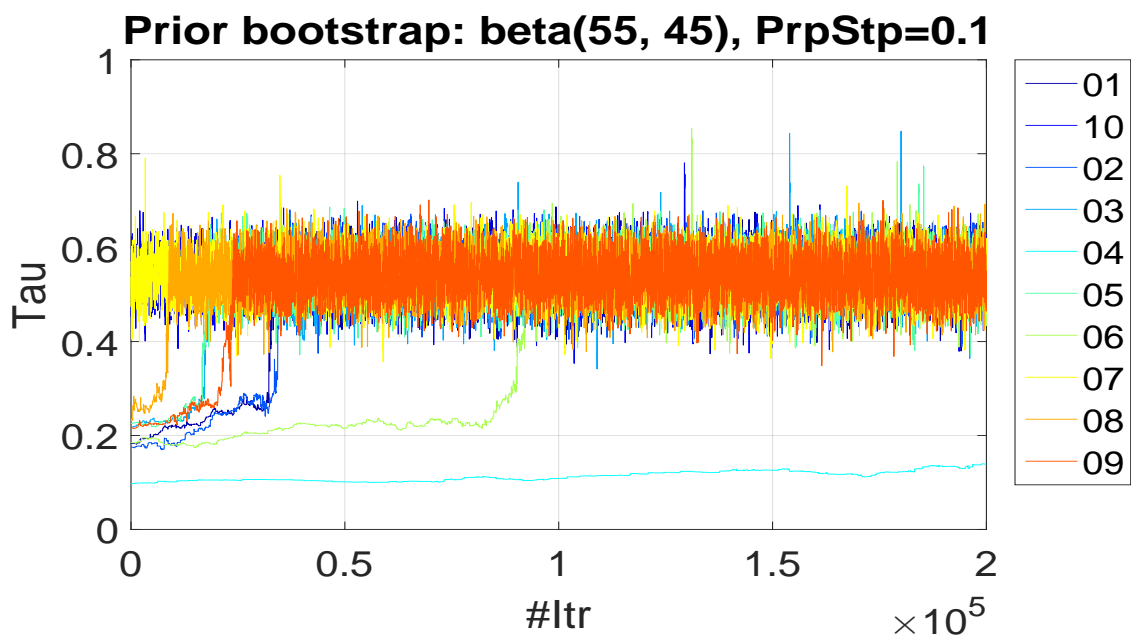


(v)

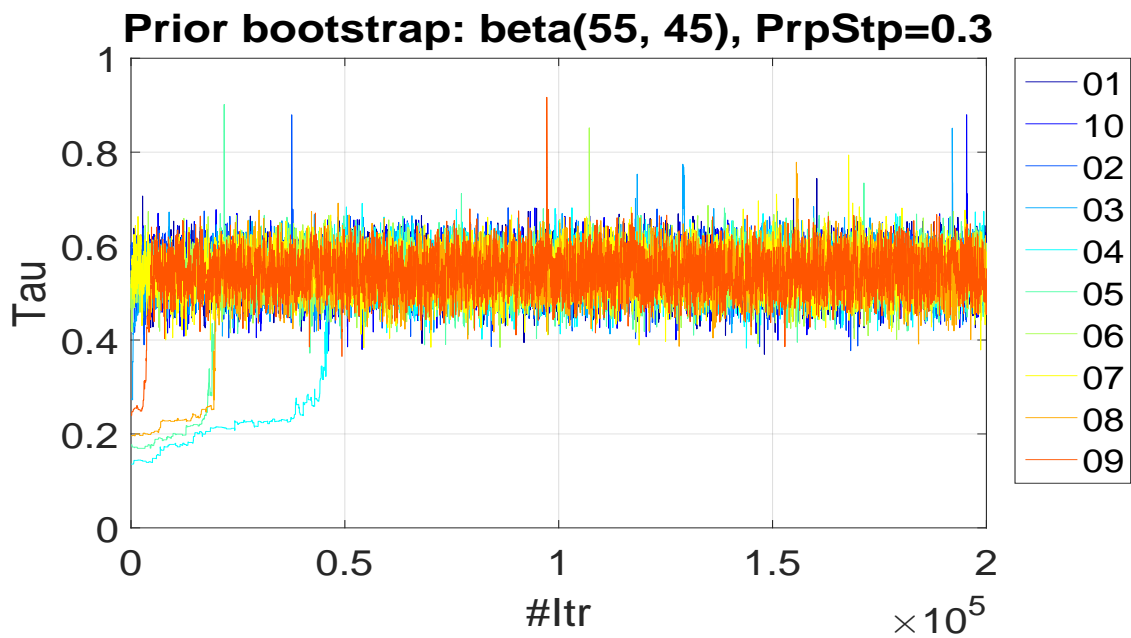


(vi)

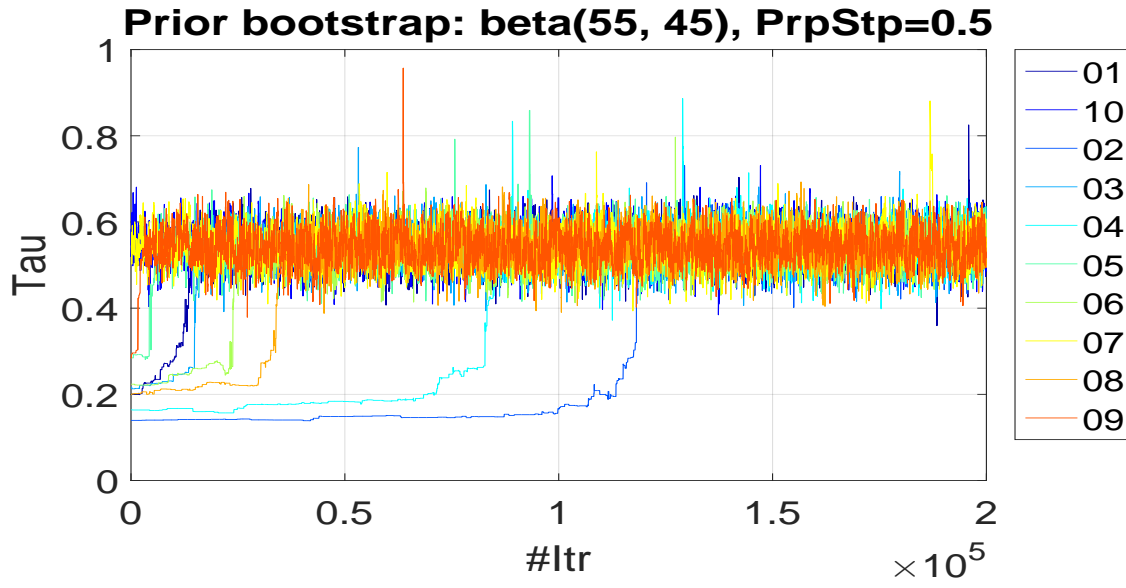
Figure 3.2: [(i)-(vi)] Manhattan pattern plots of τ values for $B(5, 5)$ prior showing variation in mixing of the MCMC chains for increasing values of step sizes (PrpStp at the top of the figures). Itr denotes the number of iterations of each MCMC run and Tau denotes the posterior values of τ . The different colors denotes a different bootstrap for a given step size. Plots suggest that step size 0.5 seems to perform the best in this case.



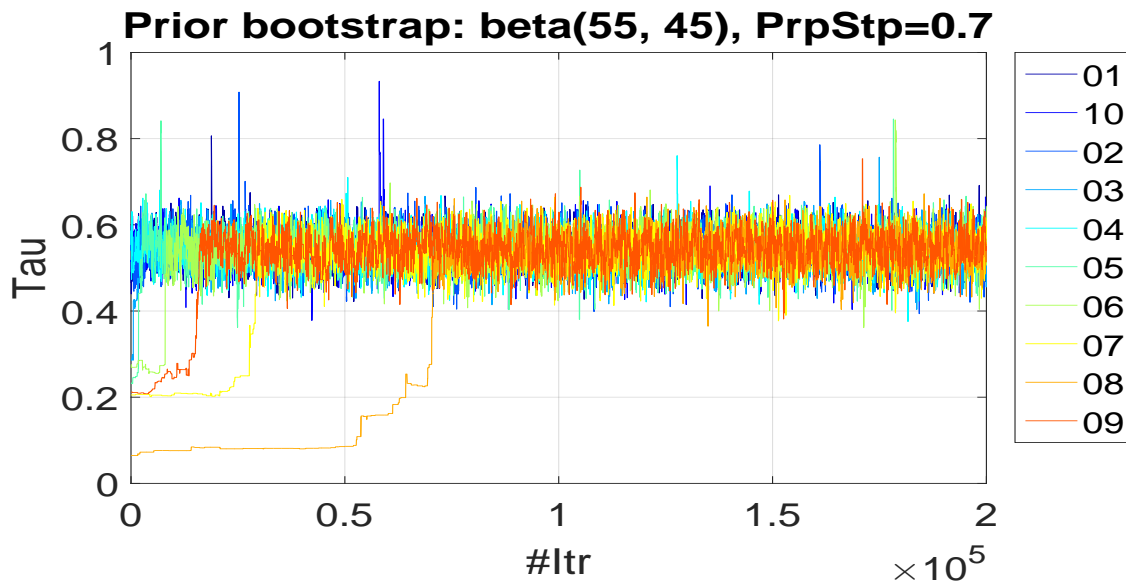
(i)



(ii)



(iii)



(iv)

Figure 3.3: [(i)-(iv)] Manhattan pattern plots of τ values for $B(55, 45)$ prior showing variation in mixing of the MCMC chains for increasing values of step sizes (PrpStp at the top of the figures). Itr denotes the number of iterations of each MCMC run and Tau denotes the posterior values of τ . The different colors denotes a different bootstrap for a given step size. Step size 0.5 also performs well in this case as suggested by the plot.

On the other hand, some chains seem to be stuck with very low values of τ and this could be because of the fact that the chains are started randomly at any point. If the starting point is randomly chosen to be very small or near zero then the condition of chains getting stuck and not moving at all might arise. However, chains starting from low values but not very near to zero eventually stabilizes and tries to reach the true value of 0.6 given sufficient number of iterations for step size values of 0.5, 0.6 and 0.7. We don't pick the step size to be 0.6 or 0.7 in this case because of the fact that we don't want the chains to traverse its posterior space very quickly or in some cases go out of the posterior space. In order to correctly assess the convergence, we also look at the auto-correlation between the draws of the Markov chain and is discussed in the next section. Now, with the choice of number of iterations and step size made we need to estimate the non-exceedance probability (NEP) τ such that the extreme threshold estimation is done. In particular, the choice of NEP then reduces to the choice of prior incorporated for τ . In this case, the choice of parameter values α_τ and β_τ for the beta prior based on which we get the posterior inference.

Choice of beta priors

The values for the beta parameters α_τ and β_τ for specifying prior beta probability distribution must be chosen such that the posterior distribution gives us the estimate of τ very close to the true known value. The other important question to answer is if there is one such value of beta parameters or a range of values of beta parameters for which the Bayesian inference would work. Here, since the posterior distribution of τ is estimated, we try to see if for a specified set of beta priors, the mode of the τ estimates approach the true value of 0.6 (Synthetic case) or not. In other words, whether the algorithm is trying to learn from the model or not. So, we looked at different diagnostics and statistics such as the tau trace plots, auto-correlation plots, sample negative log-likelihood estimated using Bayesian for a range of beta parameter values. The posterior distribution of sample negative log-likelihood is estimated for a range of pre-specified values of τ .

First, we start our analysis by looking the diagnostics and statistics described above for different, relatively strong beta priors. In this case, we choose $B(40, 60)$, $B(45, 55)$, $B(50, 50)$, $B(55, 45)$, $B(60, 40)$, **$B(65, 35)$** , **$B(70, 30)$** , $B(75, 25)$, $B(80, 20)$, $B(85, 15)$ and $B(90, 10)$ (**Figure 3.4**). Naturally, then comes in the second experiment where we have the specification of beta priors as a range of weak priors. In this case, we choose $B(4, 6)$, $B(4.5, 5.5)$, $B(5, 5)$, $B(5.5, 4.5)$, $B(6, 4)$, **$B(6.5, 3.5)$** , **$B(7, 3)$** , **$B(7.5, 2.5)$** , **$B(8, 2)$** , **$B(8.5, 1.5)$** and $B(9, 1)$ (**Figure 3.5**). We then draw our conclusions based on these two experiments.

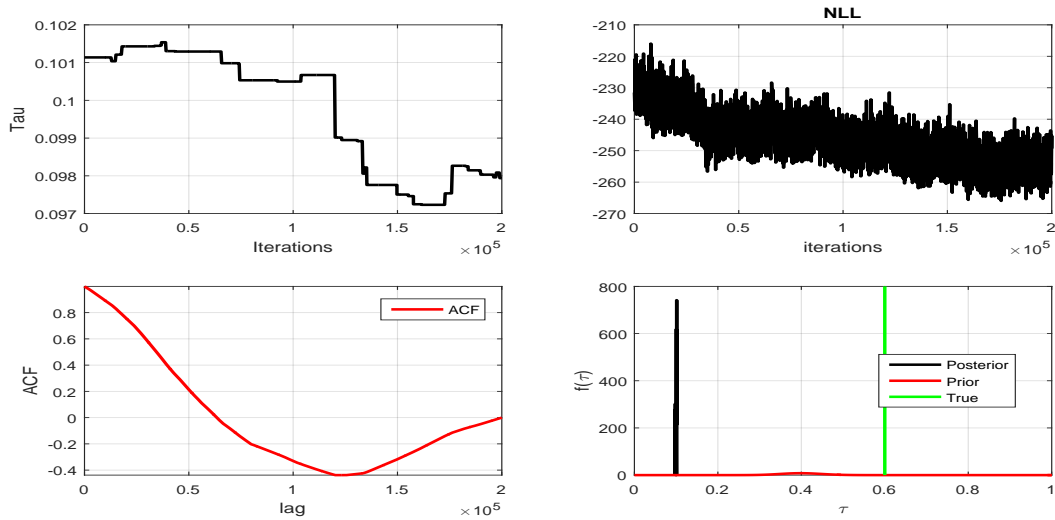
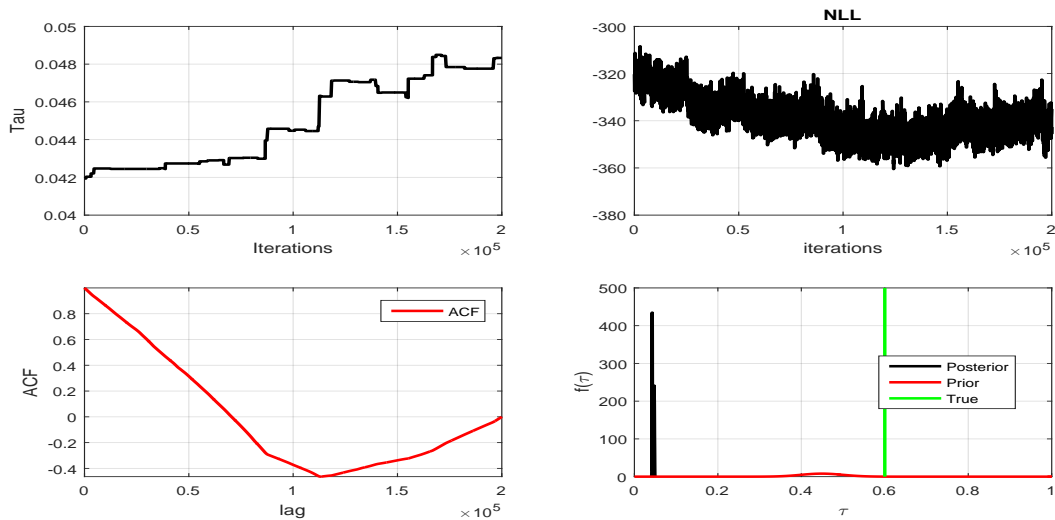
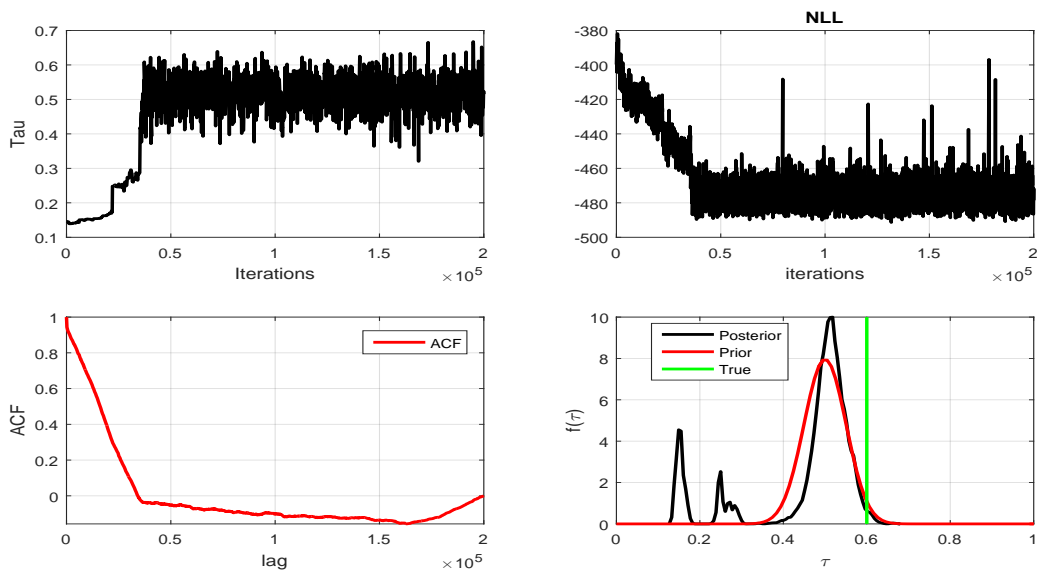


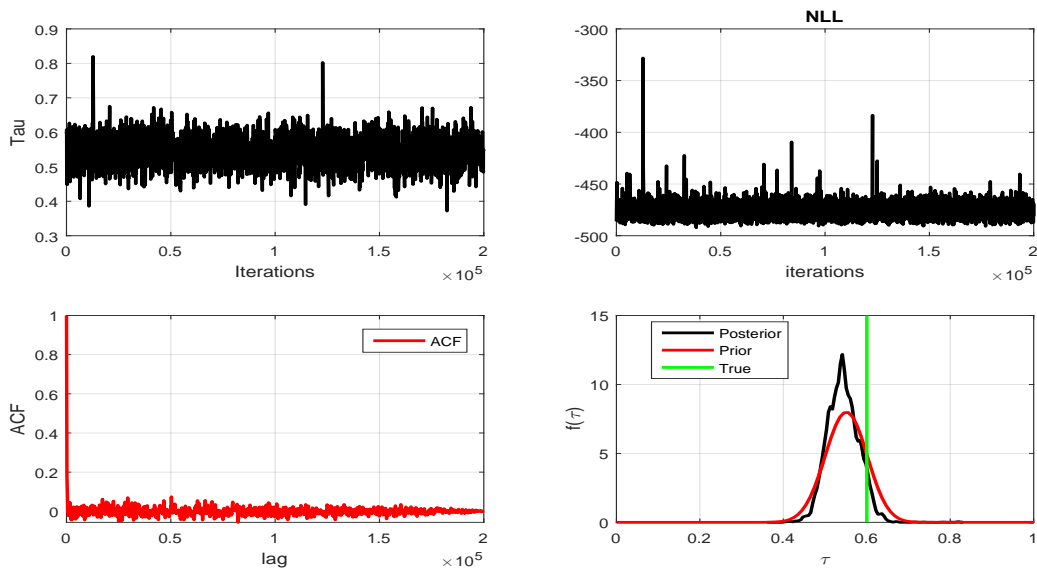
Figure 3.4: (a) The figure consists of Tau trace plot (top-left), Negative log-likelihood plot (top-right), the prior and posterior distributions plot (bottom-right) and the Auto-correlation plot (bottom-left) for $B(40, 60)$ prior. NLL denotes the negative log-likelihood and ACF denotes the auto-correlation function. Similar plots for other strong priors in (b),(c),(d),(e),(f),(g),(h),(i) and (j)



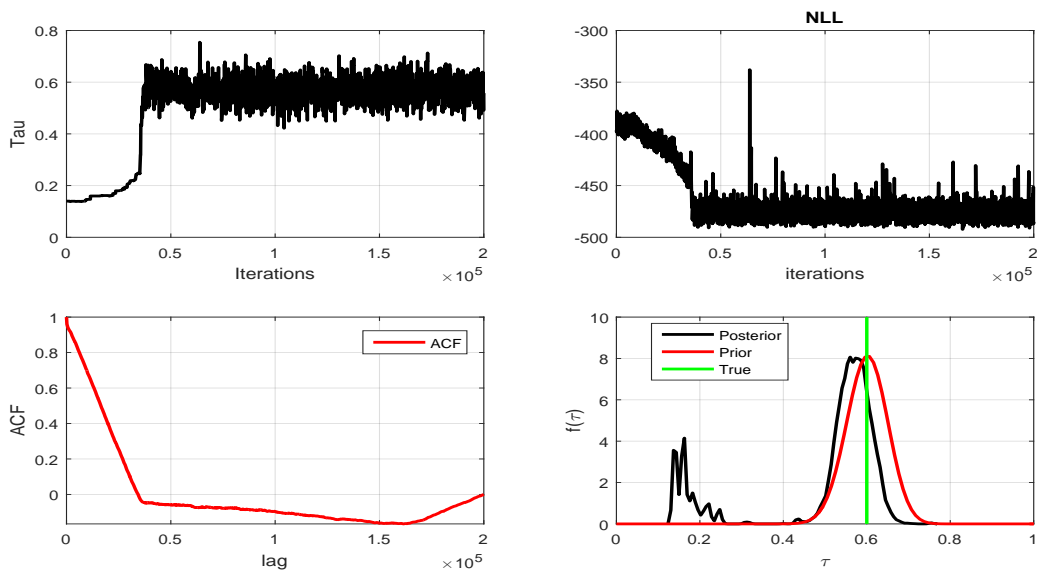
(b) For $B(45, 55)$ prior probability.



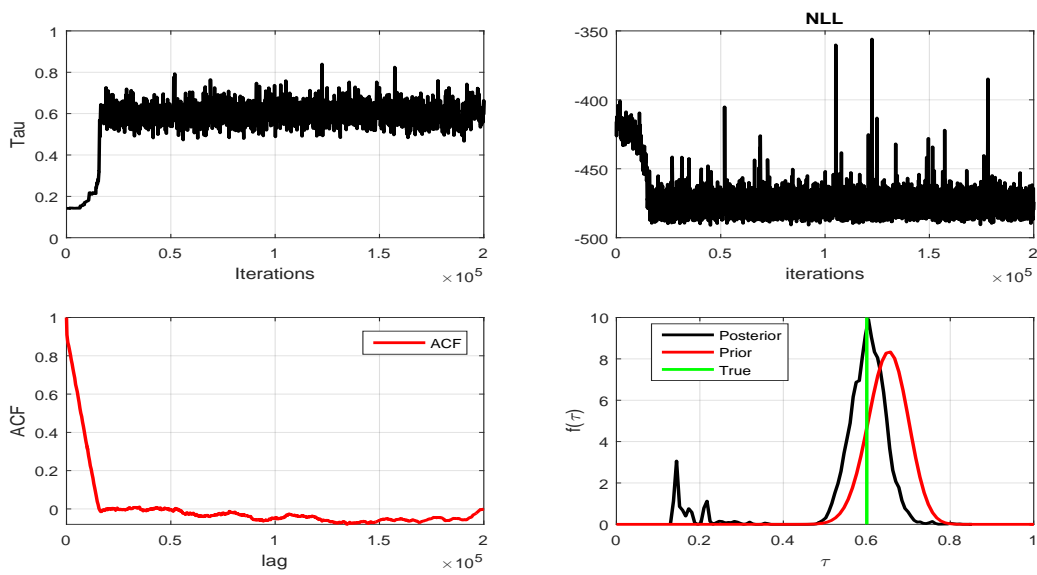
(c) For $B(50, 50)$ prior probability.



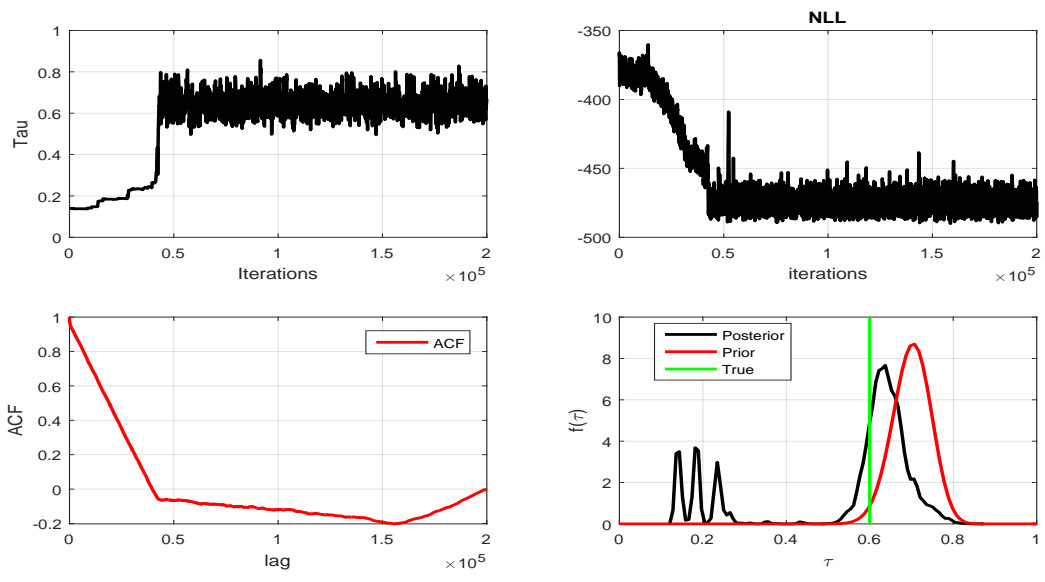
(d) For $B(55, 45)$ prior probability.



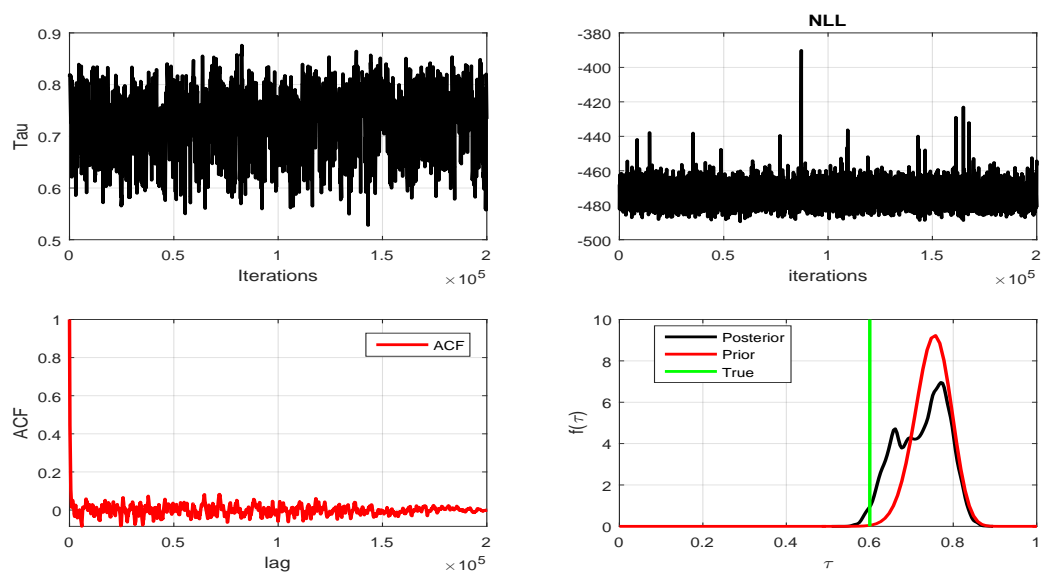
(e) For $B(60, 40)$ prior probability.



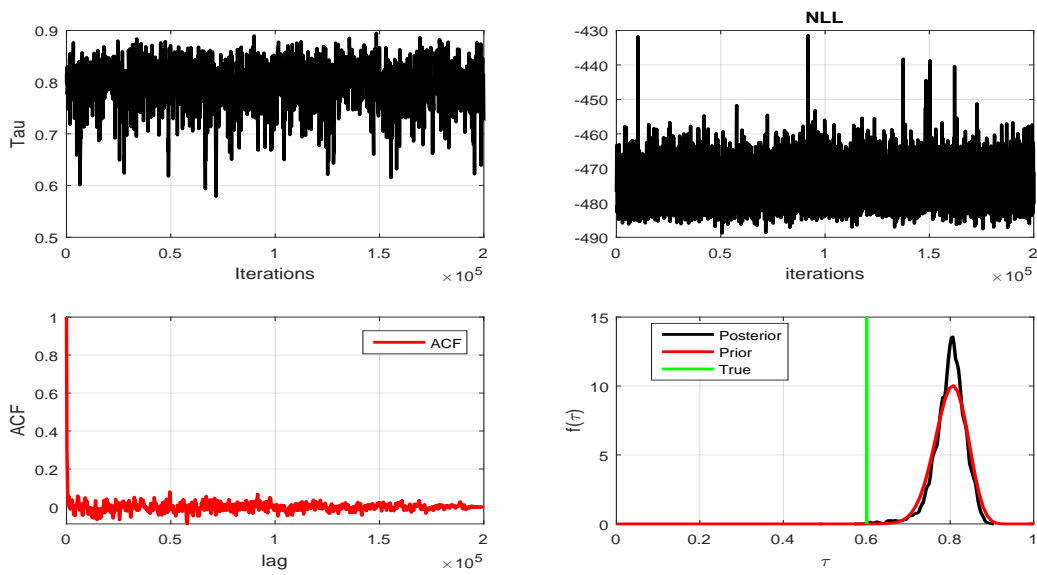
(f) For $B(65, 35)$ prior probability.



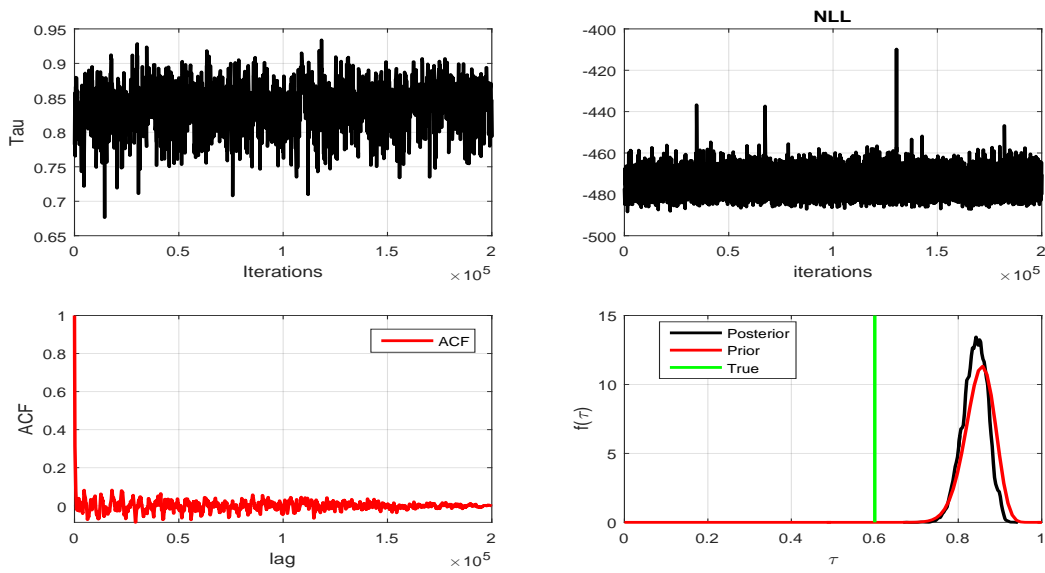
(g) For $B(70, 30)$ prior probability.



(h) For $B(75, 25)$ prior probability.



(i) For $B(80, 20)$ prior probability.



(j) For $B(85, 15)$ prior probability.

3.2.2 Discussion and Conclusions

As in **Figure 3.4 (a)** and **Figure 3.4 (b)**, we see that for $\alpha_\tau = 40$ and 45 and $\beta_\tau = 60$ and 55 , the τ -trace plots perform badly and the posterior has not moved at all. However, in case of $B(50, 50)$, $B(55, 45)$, $B(60, 40)$, $B(65, 35)$ and $B(70, 30)$, the τ -trace plot (where ever started) (**Figure 3.4 (c)**, **(d)**, **(e)**, **(f)** and **(g)**) approaches the true value of the synthetic data NEP, i.e.: 0.6 and the mode of the posterior distribution of τ moves closer and closer to the true value (green line in prior and posterior plot). This shows that for choice of these priors for τ the model works well and the algorithm does fine. On the contrary, if we specify even stronger priors such as $B(75, 25)$, $B(80, 20)$, $B(85, 25)$ and $B(90, 10)$ (**Figure 3.4 (h)**, **(i)** and **(j)**) the τ -trace plot moves away from the true value of 0.6 . This shows that specifying a too strong of a prior for τ lets the algorithm no degrees of freedom and that it does not learn from the model. In fact, the strong priors lead to bias and seems to dominate the algorithm. Thus, the idea to specify relatively weaker priors.

In case of weaker priors, **Figure 3.5** shows that the mode of the posterior distribution of τ is close to the true value of 0.6 except for $B(4, 6)$ (**Figure 3.5 (a)**). If we look at the τ -trace plots in **Figure 3.5**, all of them reach the true value. Some of them take longer than others but eventually the trace plots seem to be stationary around close to the true value. **Figure 3.5 (h)**, **(i)**, **(j)** gives us the best result in this case suggesting that a range of Beta priors $B(7.5, 2.5)$, $B(8, 2)$ and $B(8.5, 1.5)$ works well. Also, plotting the sample negative log-likelihood values (**Figure 3.6 (a)**) and (**Figure 3.6 (b)**), estimated using Bayesian inference assuming known pre-specified extreme value threshold NEP τ gives us another way of looking at the same. The pre-specified NEP is estimated using the mean of the beta prior. We have, for the synthetic data sample used here, there is evidence that any τ value ranging from 0.5 to 0.75 should be preferred and that they will provide somewhat better fit than others. Though it is not clear from **Figure 3.6** that τ values in the range above 0.75 should not be used directly. But as per the diagnostics discussed above in the section of choice of beta priors, it is evident that too strong a prior will not work. However, it must be kept in mind that in specifying the beta prior distribution, one must have sufficient sample to be able to estimate the GP tail parameters reasonably and keep τ sufficiently large such that fitting an extreme value model to the tail of the given data is somewhat reasonable. Now that we have an appropriate idea about the extreme threshold non-exceedance probability (τ) in this case, the threshold estimation becomes easy here as the threshold is a function of τ (**Eqn. 3.5**). The NEP range of 0.5 to 0.75 seems to be the optimal range here for the threshold estimation.

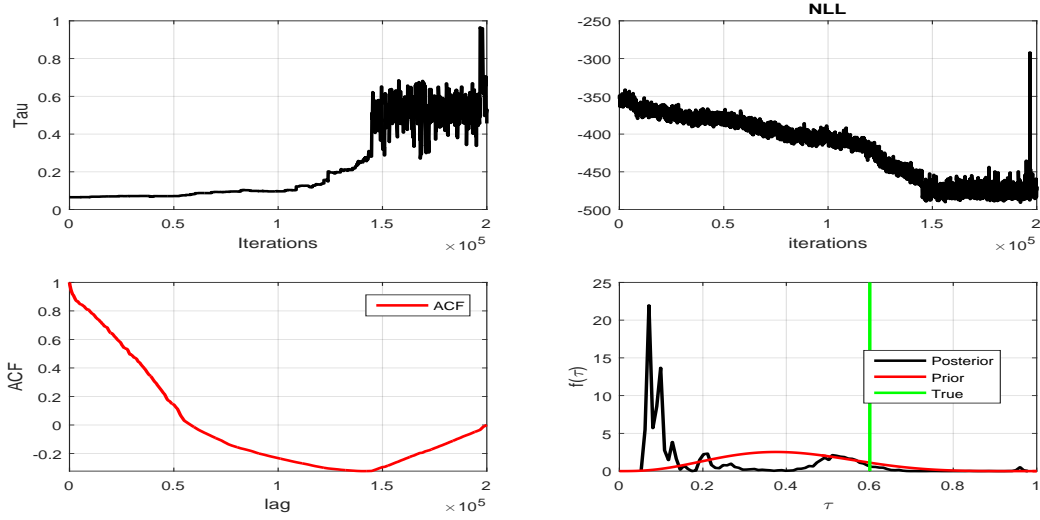
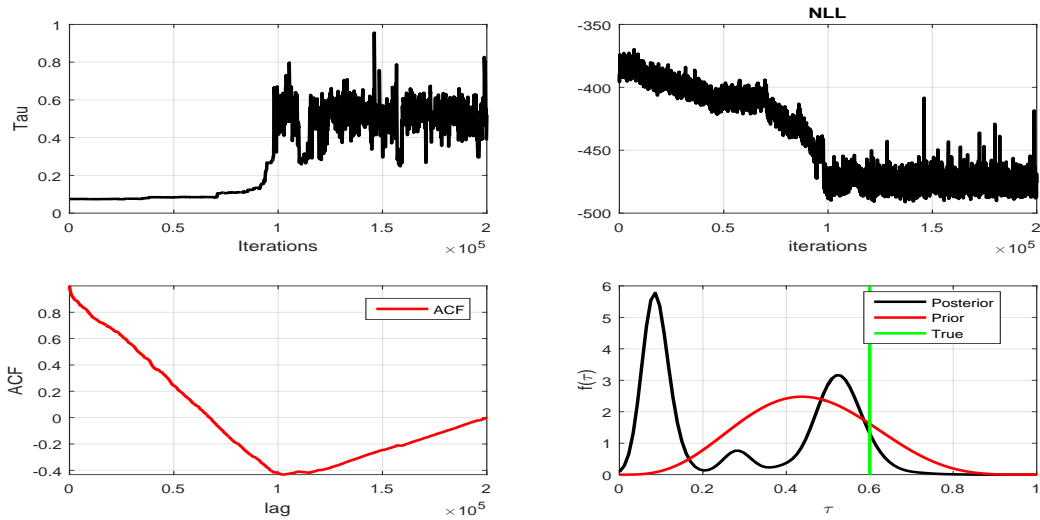
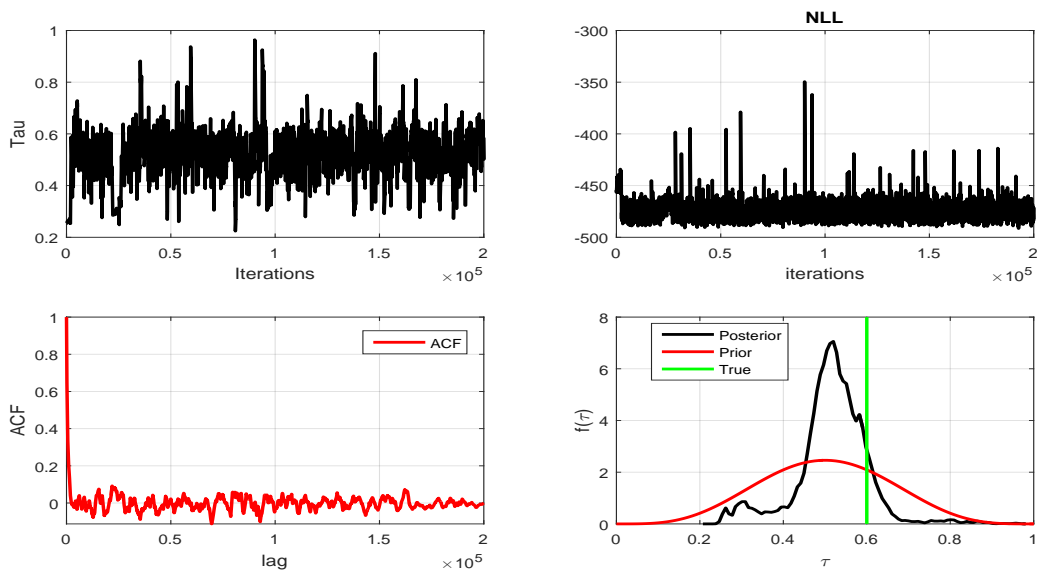


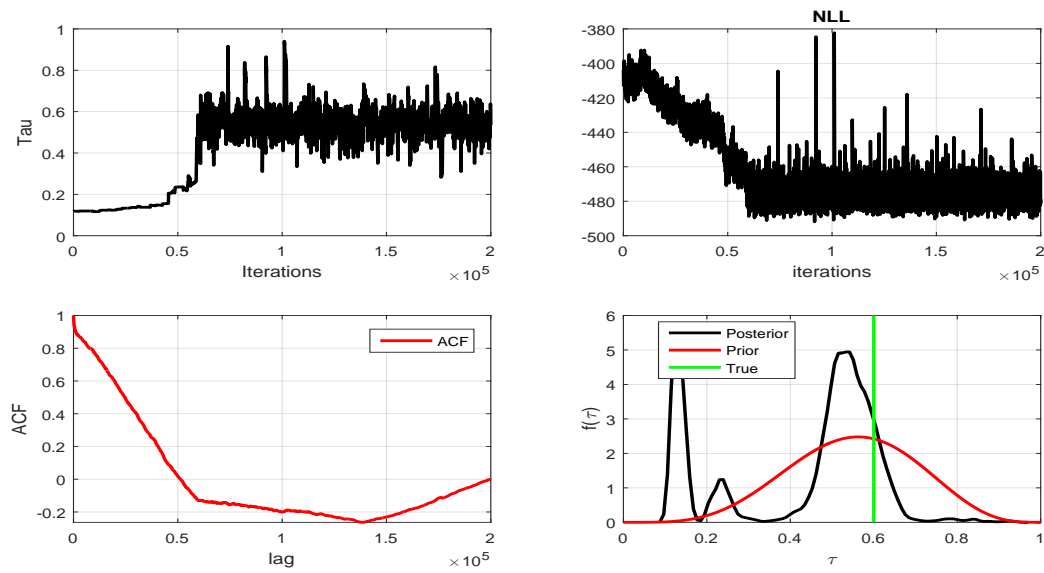
Figure 3.4: (a) The figure consists of (in clockwise sense) Tau trace plot, Negative log-likelihood plot, the prior and posterior distributions plot and the Auto-correlation plot for $B(4,6)$ prior. NLL denotes the negative log-likelihood and ACF denotes the auto-correlation function. Similar plots for other strong priors in (b),(c),(d),(e),(f),(g),(h),(i) and (j)



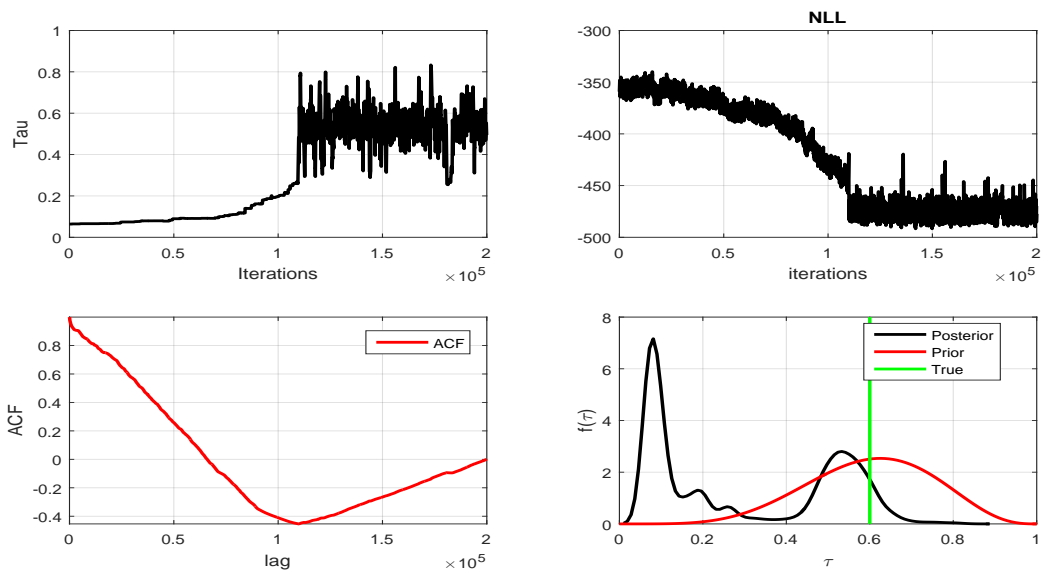
(b) For $B(4.5, 5.5)$ prior probability.



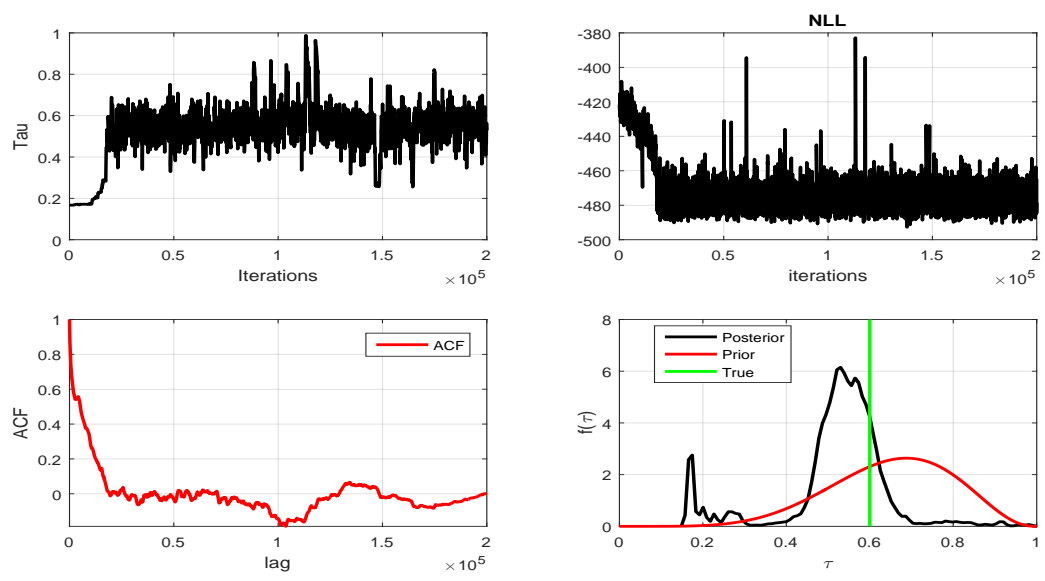
(c) For $B(5, 5)$ prior probability.



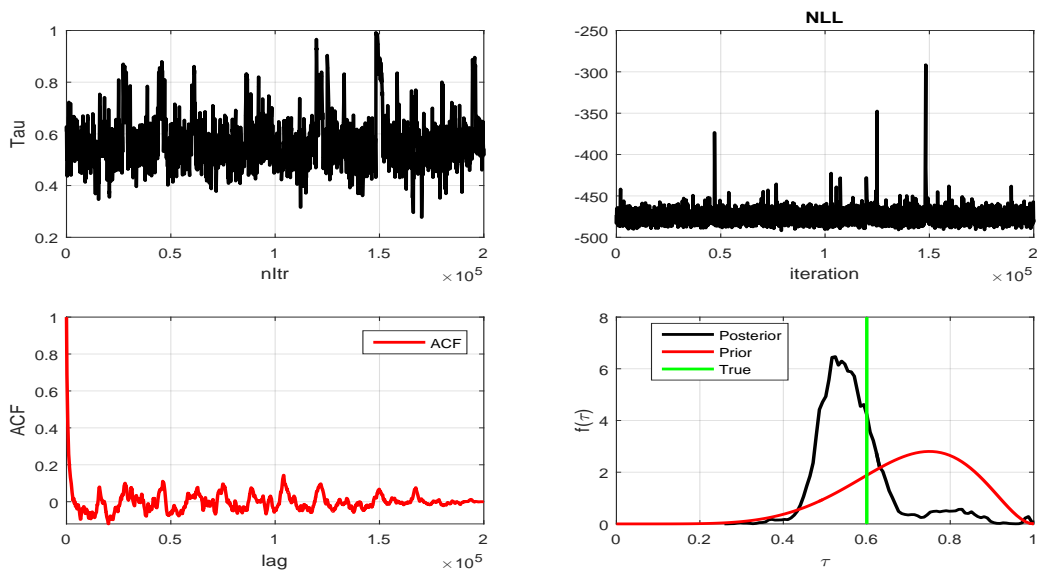
(d) For $B(5.5, 4.5)$ prior probability.



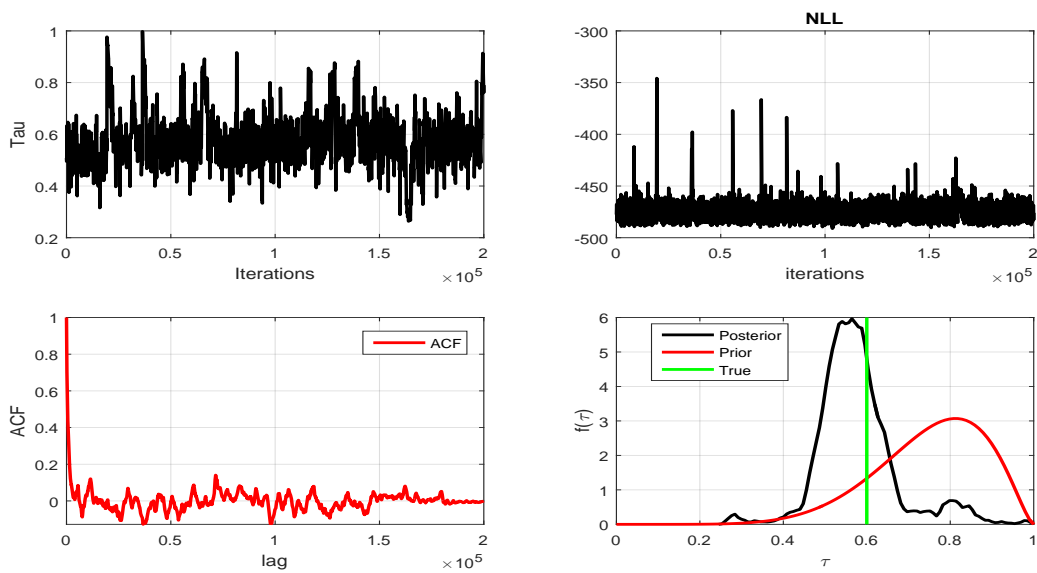
(e) For $B(6, 4)$ prior probability.



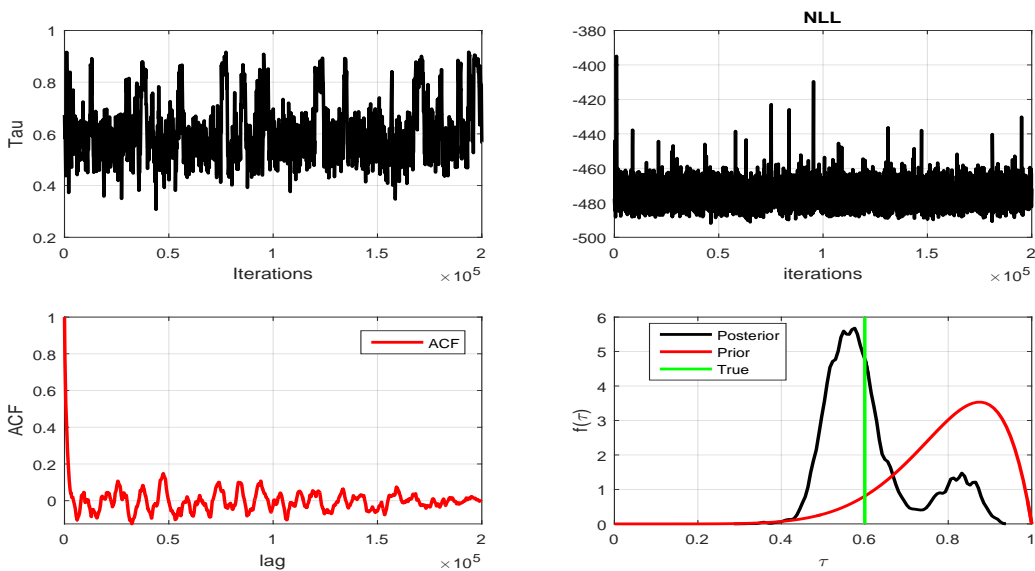
(f) For $B(6.5, 3.5)$ prior probability.



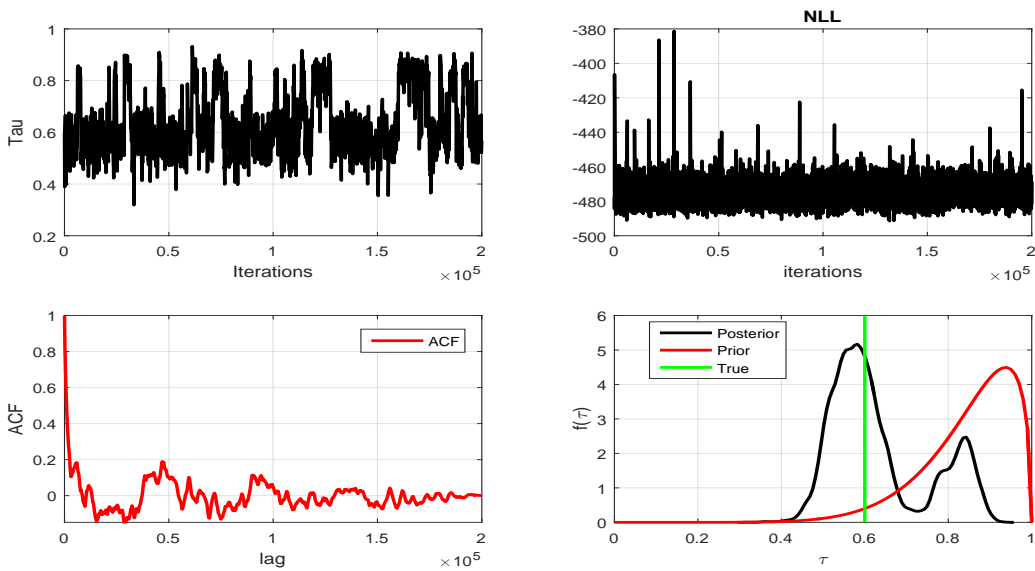
(g) For $B(7, 3)$ prior probability.



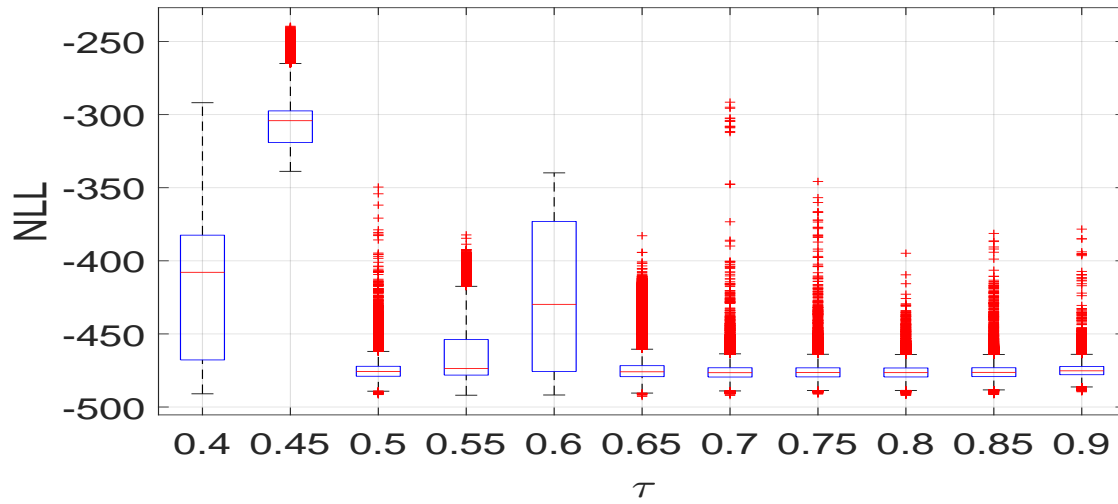
(h) For $B(7.5, 2.5)$ prior probability.



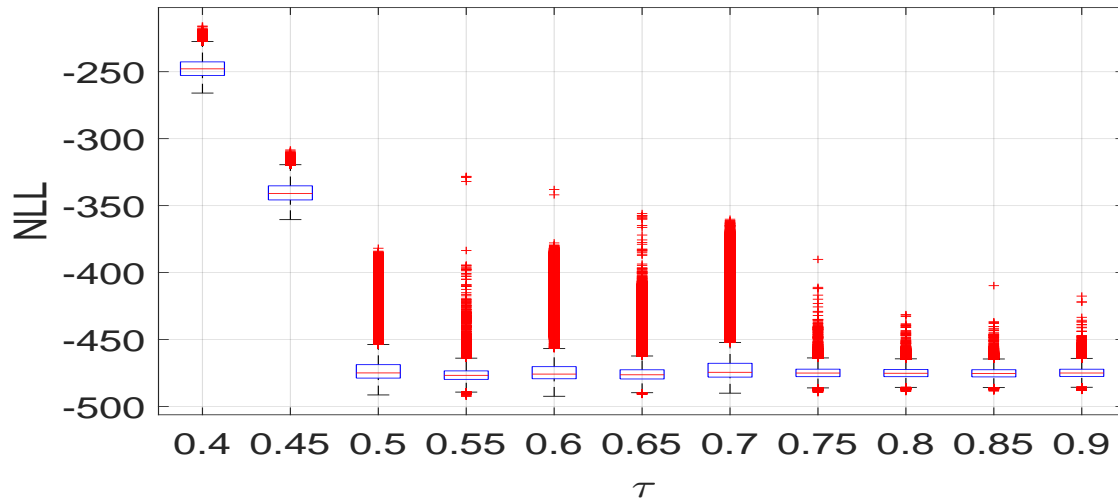
(i) For $B(8, 2)$ prior probability.



(j) For $B(8.5, 1.5)$ prior probability.



(a)



(b)

Figure 3.5: Box-plot of posterior distribution of sample negative log-likelihood, estimated using Bayesian inference assuming known pre-specified extreme threshold NEP $\tau = 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9$ for weaker priors (a) relatively strong beta priors (b). NLL represents the sample negative log-likelihood.

3.3 Way Forward

Currently, a number of improvements and extensions of this approach are in consideration. Since the Bayesian approach to threshold estimation doesn't incorporate techniques like cross-validation and a lot of bootstrapping, this Bayesian model for non-stationary marginal extremes is computationally less demanding than the frequentist maximum likelihood inference model. Also, to use P-splines regression in case of NSCE model for threshold estimation needs us to estimate different spline roughness penalties which is computationally more demanding. So, we plan to extend this idea of threshold estimation using Bayesian inference to model higher dimensional covariates. We currently are incorporating seasons as a new covariate and is trying to estimate extreme threshold when we have two different covariates. Later, this approach can be extended to n -dimensions in application depending upon the complexity and computational cost.

Bibliography

- [1] S. Coles, J. Bawa, L. Trenner, and P. Dorazio, *An introduction to statistical modeling of extreme values*, vol. 208. Springer, 2001.
- [2] A. C. Davison and R. L. Smith, “Models for exceedances over high thresholds,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 393–442, 1990.
- [3] J. Beirlant, Y. Goegebeur, J. Segers, J. Teugels, D. Waal, and C. Ferro, “Statistics of extremes: Theory and applications. 2004.”
- [4] J. E. Heffernan and J. A. Tawn, “A conditional approach for multivariate extreme values (with discussion),” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 66, no. 3, pp. 497–546, 2004.
- [5] P. Jonathan and K. Ewans, “A spatiodirectional model for extreme waves in the gulf of mexico,” *ASME. J. Offshore Mech. Arct. Eng.*, vol. 133, no. 1, pp. 011601–011601–9, 2010.
- [6] P. Jonathan and K. Ewans, “Statistical modelling of extreme ocean environments for marine design: a review,” *Ocean Engineering*, vol. 62, pp. 91–109, 2013.
- [7] N. H. Bingham, C. M. Goldie, and J. L. Teugels, “Regular variation (encyclopedia of mathematics and its applications),” 1987.
- [8] A. W. Ledford and J. A. Tawn, “Modelling dependence within joint tail regions,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 59, no. 2, pp. 475–499, 1997.
- [9] A. W. Ledford and J. A. Tawn, “Statistics for near independence in multivariate extreme values,” *Biometrika*, vol. 83, no. 1, pp. 169–187, 1996.
- [10] M. Leadbetter and G. Lindgren, “Rootz en, h.(1983). extremes and related properties of random sequences and processes,” 1983.
- [11] P. Jonathan, K. Ewans, and D. Randell, “Non-stationary conditional extremes of northern north sea storm characteristics,” *Environmetrics*, vol. 25, no. 3, pp. 172–188, 2014.

- [12] L. Raghupathi, D. Randell, P. Jonathan, and K. Ewans, “Non-stationary estimation of joint design criteria with a multivariate conditional extremes approach,” *Proceedings of The 35th International Conference on Ocean, Offshore and Arctic Engineering, OMAE*, jun 2016.
- [13] L. Raghupathi, D. Randell, K. Ewans, and P. Jonathan, “Fast computation of large scale marginal extremes with multi-dimensional covariates,” *Computational Statistics & Data Analysis*, vol. 95, pp. 243–258, mar 2016.
- [14] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian data analysis*, vol. 2. Taylor & Francis, 2014.
- [15] B. Letham and C. Rudin, “Probabilistic modeling and bayesian analysis,” feb 2012.
- [16] W. R. Gilks, *Markov chain monte carlo*. Wiley Online Library, 2005.
- [17] G. Casella and E. I. George, “Explaining the gibbs sampler,” *The American Statistician*, vol. 46, no. 3, pp. 167–174, 1992.
- [18] M. Cowels and B. Carlin, “Markov chain monte carlo convergence diagnostics: a comparative review,” *Journal of the American Statistical Association*, vol. 91, no. 434, pp. 883–904, 1996.
- [19] A. Gelman and D. B. Rubin, “Inference from iterative simulation using multiple sequences,” *Statistical science*, pp. 457–472, 1992.
- [20] J. Geweke, “Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments,” in *IN BAYESIAN STATISTICS*, pp. 169–193, University Press, 1992.
- [21] P. Heidelberger and P. D. Welch, “Simulation run length control in the presence of an initial transient,” *Operations Research*, vol. 31, no. 6, pp. 1109–1144, 1983.
- [22] A. E. Raftery, S. Lewis, *et al.*, “How many iterations in the gibbs sampler,” *Bayesian statistics*, vol. 4, no. 2, pp. 763–773, 1992.
- [23] R. E. Kass, B. P. Carlin, A. Gelman, and R. M. Neal, “Markov chain monte carlo in practice: a roundtable discussion,” *The American Statistician*, vol. 52, no. 2, pp. 93–100, 1998.
- [24] S. G. Coles and J. A. Tawn, “A bayesian analysis of extreme rainfall data,” *Applied statistics*, pp. 463–478, 1996.
- [25] S. G. Coles and J. A. Tawn, “Modelling extremes of the areal rainfall process,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 329–347, 1996.

- [26] S. Coles and J. Tawn, “Bayesian modelling of extreme surges on the uk east coast,” *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 363, no. 1831, pp. 1387–1406, 2005.
- [27] S. Coles and J. Tawn, “Seasonal effects of extreme surges,” *Stochastic Environmental Research and Risk Assessment*, vol. 19, no. 6, pp. 417–427, 2005.
- [28] C. G. Soares and M. Scotto, “Modelling uncertainty in long-term predictions of significant wave height,” *Ocean Engineering*, vol. 28, no. 3, pp. 329–342, 2001.
- [29] J. M. Mendes, P. C. de Zea Bermudez, J. Pereira, K. Turkman, and M. Vasconcelos, “Spatial extremes of wildfire sizes: Bayesian hierarchical models for extremes,” *Environmental and Ecological Statistics*, vol. 17, no. 1, pp. 1–28, 2010.
- [30] M. Ribatet, D. Cooley, and A. C. Davison, “Bayesian inference from composite likelihoods, with an application to spatial extremes,” *Statistica Sinica*, pp. 813–845, 2012.
- [31] A. C. Davison, S. Padoan, M. Ribatet, *et al.*, “Statistical modeling of spatial extremes,” *Statistical Science*, vol. 27, no. 2, pp. 161–186, 2012.
- [32] D. Randell, K. Turnbull, and P. Jonathan, “Bayesian inference for non-stationary marginal extremes,” *Environmetrics*, vol. submitted, 2015.
- [33] B. D. Marx and P. H. Eilers, “Direct generalized additive modeling with penalized likelihood,” *Computational Statistics & Data Analysis*, vol. 28, no. 2, pp. 193–209, 1998.
- [34] K. Bollaerts, P. H. Eilers, and I. Mechelen, “Simple and multiple p-splines regression with shape constraints,” *British Journal of Mathematical and Statistical Psychology*, vol. 59, no. 2, pp. 451–469, 2006.