Instituto Tecnológico Autónomo de México

Aprendizaje Bayesiano estadístico para modelos de colas pesadas

Tesis

que para obtener el título de

Licenciado en Matemáticas Aplicadas

Presenta

Daniel Salnikov Botello

Asesora

Dr. César Luis García García

pa: Dr. Abdellah Salhi

Ciudad de México 2021

«Con fundamento en los artículos 21 y 27 de la Ley Federal del Derecho de Autor y como titular de los derechos moral y patrimonial de la obra titulada "**Aprendizaje Bayesiano estadístico para modelos de colas pesadas**", otorgo de manera gratuita y permanente al Instituto Tecnológico Autónomo de México y a la Biblioteca Raúl Bailléres Jr., la autorización para que fijen la obra en cualquier medio, incluido el electrónico, y la divulguen entre sus usuarios, profesores, estudiantes o terceras personas, sin que pueda percibir por tal divulgación una contraprestación.»

Fecha

Daniel Salnikov Botello

A mis padres Lourdes y Leonardo
por su apoyo incondicional sin
importar las circunstancias
ni las distancias.

A mis abuelos Jaime y Helen
por todo lo que hicieron por mí.

Agradecimientos

Al Dr. Abdel Salhi, Dr. César Luis García por su apoyo y supervisión en la elaboración de esta tesis. A mis sinodales: Dr. Luis Enrique Nieto Barajas, Dr. Leonardo Rojas Nandayapa y el Lic. Carlos Samuel Pérez Pérez por leer y mejorar la calidad del trabajo. A mis profesores durante mi estancia en el ITAM, en particular al Dr. Carlos Bosch, Dr. Manuel Mendoza, Dr. Joao Morais y a mis comnpañeros por inculcar en mí el gusto por el estudio y la aplicación de las matemáticas y la estadística. Es imposible enlistar a toda la gente con la cual estoy agradecido por su participacion en mi etapa universitaria, por eso, aprovecho esta sección para agredecer a todas las personas con las que compartí experiencias durante mi tiempo en la universidad. Gracias a todas y a todos.

Resumen

El estudio de los modelos de colas pesadas es interesante porque utiliza argumentos asintóticos para estudiar el comportamiento y características de procesos que generan datos con colas pesadas. Las colas pesadas son aquellas que tienen una probabilidad mayor para eventos poco probables en comparación con los modelos clásicos de la estadística, por ejemplo, la distribución normal. Esto hace que el estudio este hecho en una forma muy diferente a los modelos clásicos porque la media y la varianza no proporcionan información útil con respecto al proceso que genera los datos, es más es posible que la variabilidad en las muestras y los datos sea tan alta que la media teórica no exista, y por ende no exista un patrón identificable a través de métodos estadísticos clásicos que están construidos sobre supuestos de concentración de probabilidad, como lo es la regresión lineal clásica, la cual asume una función lineal con errores cuya distribución es normal.

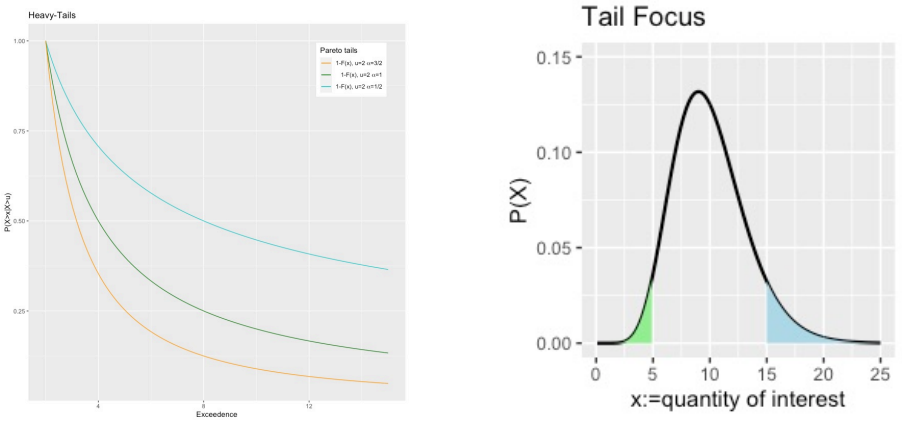

Figure 1. Cola pesada (izquierda) y cola ligera (derecha).

El objetivo es que el trabajo sea suficiente para que la lectora pueda comprender los fundamentos del análisis de colas pesadas y realizar un análisis de una colección de datos por medio de los algoritmos propuestos en este trabajo, para así poder aprender sobre el proceso que genera los datos. La tesis combina resultados teóricos con metodologías practicas. La mezcla entre teoría y práctica busca llenar una brecha que el autor ha encontrado en libros de texto básicos de probabilidad, teoría de la medida y estadística enfocados en teoría, y libros de texto, artículos y propuestas de análisis de datos y aprendizaje de maquina que solo están enfocados en una metodología práctica. Por esto, los primeros dos capítulos tienen un enfoque teórico; el capitulo tres es un puente que combina características de los modelos de colas pesadas y valores extremos con las bases necesarias para realizar un análisis computacional de datos que parecen tener una cola pesada. Los últimos dos capítulos tienen un enfoque computacional y consisten en una propuesta algorítmica para aprender

información relevante por medio de técnicas Bayesianas, y un experimento de simulación cuyo objetivo es poner a prueba los métodos propuestos en el capítulo anterior.

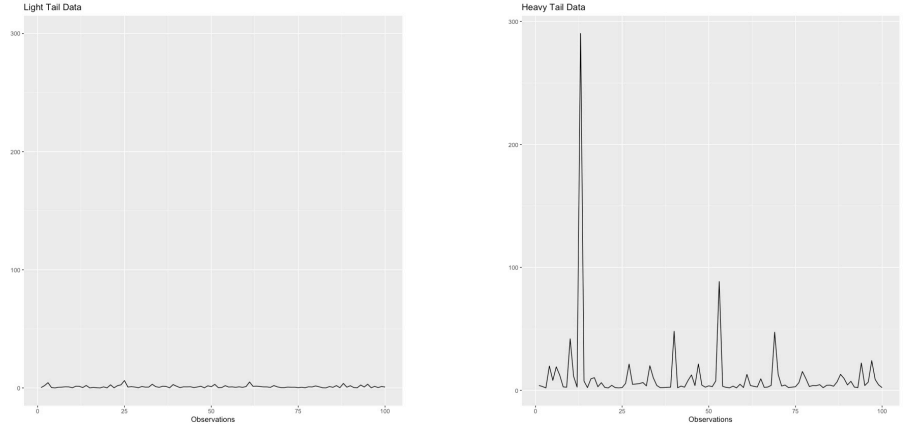

Figure 2. Serie de tiempo cola pesada (derecha) y cola ligera (izquierda).

El primer capítulo contiene conceptos de teoría de la medida y probabilidad necesarios para construir los modelos asintóticos para colas pesadas y valores extremos. Este capítulo contiene una demostración de porque una función no decreciente puede generalizarse a ser una medida sobre los reales que mide la longitud de un intervalo por medio de una función distinta a la distancia entre dos puntos en la recta real. Es decir, una función $g : X \to \mathbb{R}$ no decreciente y continua por la derecha puede extenderse para calcular la longitud entre dos puntos $a < b$ como $g(b) - g(a)$. Esta idea es muy poderosa porque permite estudiar propiedades asintóticas de los procesos que generan datos con colas pesadas. El segundo capítulo contiene la

demostración de que la única distribución no degenerada para variables aleatorias de una sola dimensión con colas pesadas es la distribución de Fréchet. Este resultado es esencial para estudiar modelos de colas pesadas porque sustenta la aproximación asintótica, incluso si el tamaño de la muestra no es muy grande. La demostración esta hecha con base en la construcción de Procesos de Puntos Poisson (Pickands, 1974) con base en borradores y ejercicios en [2, 1, 21], el objetivo es que la demostración sea accesible, pero al mismo tiempo mantenga el rigor de textos mas avanzados, como [2]. Finalmente, el capitulo cierra con una demostración del Teorema de Tipos de Convergencia de Valores Extremos para presentar la relación que hay entre medidas sobre los reales, modelos de colas pesadas y las distribuciones de valores extremos.

La idea fundamental es la diferencia que existe entre una función que decae como un polinomio (i.e., $f(x) \approx x^{-\alpha}$, $\alpha > 0$, x en la cola) y una que decae con la misma o mayor velocidad que una función exponencial (i.e., $f(x) \leq \exp(-\lambda x)$, $\lambda > 0$, x en la cola). Al decaer de forma más lenta el polinomio no concentra la probabilidad en una región, por ende, la gráfica suele estar caracterizada por una línea extendida en lugar de una campana; por esto, al estudiar estas funciones como posibles medidas para los modelos asintóticos es posible encontrar las distribuciones asintóticas y las diferencias que existen entre ellas. La más importante es que las colas pesadas no suelen presentar patrones porque la probabilidad (medida) esta repartida entre muchos valores sin tener una distribución que centre los momentos (una campana), por ejemplo, incluso una normal con una varianza muy alta concentra probabilidad porque la cola esta acotada por una función exponencial.

Tomar decisiones suele ser un proceso complicado, más aún cuando hay poca información y la decisión está rodeada de incertidumbre. Por esto, es complicado tomar decisiones que dependen de modelos de colas pesadas, ya que la volatilidad inherente al proceso que genera los datos afecta la credibilidad/confianza del tomador de decisiones. La inferencia Bayesiana reconoce la ignorancia y subjetividad del método que analiza los datos, por esto, todo análisis Bayesiano debe tomar en cuenta la incertidumbre del método propuesto; además, las técnicas Bayesianas están fundamentadas en teoría de decisión para así poder tomar la mejor decisión bajo algún criterio apropiado - por ejemplo, función de perdida cuadrática - incluso en un ambiente de incertidumbre. Con base en lo anterior, la tesis propone un método algorítmico para aprender información relevante de un proceso que genera una muestra con datos de cola pesada por medio de técnicas Bayesianas que reconocen la incertidumbre inherente tanto en el modelo como en el método inferencial.

Estos algoritmos están presentes en el capítulo cuatro. La metodología sigue el algoritmo clásico del análisis de colas pesadas: análisis exploratorio de datos, selección de un umbral, estimar los parámetros de la distribución asintótica y probar los resultados para estimar otras cantidades de interés, como funciones de un percentil.

El primer algoritmo propuesto es una prueba de hipótesis que compara si la muestra contiene más evidencia en favor de una cola pesada o de una cola que está acotada por una función exponencial, el segundo algoritmo es otra prueba de hipótesis que calcula la información que la muestra contiene en favor de alguno de los tres modelos clásicos de colas pesadas (Fréchet, Gumbel e Inverse-Weibull); el tercer algoritmo es una propuesta para buscar un umbral con el cual

los datos tienen compatibilidad por medio de una búsqueda estocástica; el cuarto algoritmo es una versión de *Metropolis Hastings* para colas pesadas con una estructura quasi-conjugada para simular la distribución posterior. El capitulo cuatro termina con una propuesta de pruebas empíricas de bondad de ajuste, estas pruebas proponen algoritmos para verificar si la muestra es compatible con el modelo y resaltar algún problema en la metodología y/o los datos.

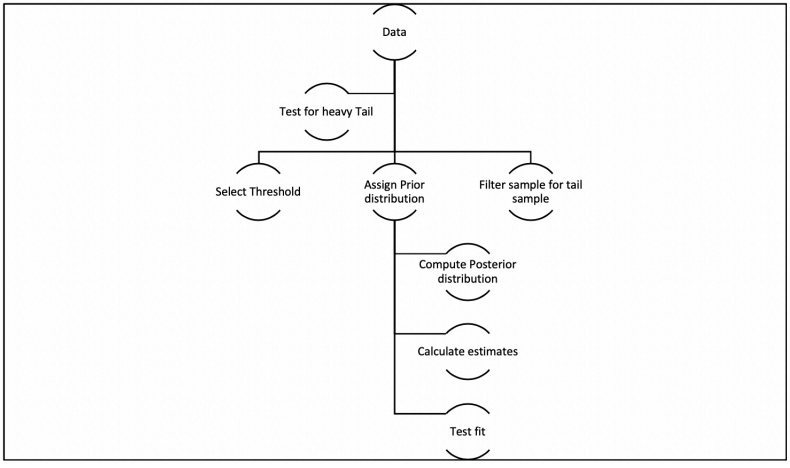

Figure 3. Diagrama del algoritmo de aprendizaje para modelos de colas pesadas.

El capítulo cinco pone a prueba la teoría y metodología de los capítulos anteriores en un experimento de simulación. El objetivo es que al conocer el modelo que genera los datos, los algoritmos de aprendizaje deben ser más o al menos igual de eficaces que al estudiar procesos para los cuales no es conocido el proceso que genera los datos. Este capítulo ilustra la forma en la que los algoritmos aprenden información importante de los datos con colas pesadas, al igual que la manera en la evalúan su desempeño. Inspirado por técnicas de

aprendizaje de maquina, el experimento aprende de los datos con el 80% de la muestra simulada y evalúa su desempeño con el otro 20% de la muestra. El resultado del experimento es positivo porque los algoritmos logran aprender las cantidades relevantes con precisión, los resultados son discutidos al final del capítulo cinco.

Finalmente, el ultimo capítulo (6) discute la relevancia y potencial de los modelos de colas pesadas en el nuevo contexto de análisis de datos, por ejemplo, la posible implementación de estos métodos de aprendizaje - los cuales son capaces de simular el proceso que genera los datos - para estresar y construir sistemas con mayor capacidad para enfrentar y administrar choques, como lo son excesos en la demanda de energía de un edificio. Espero que el lector y o la lectora considere que la mezcla de teoría y practica facilita la compresión de los resultados teóricos y su implementación por medio de los métodos estadísticos computacionales propuestos.

Abstract

This work lays out the basics in the study of one variable Heavy-Tail phenomena. These phenomena are characterised by processes in which large observations are likely. The paper is structured as a ladder that builds from abstract theory the essential limit results that enable Heavy-Tail analysis. The theoretical construction employs Point Processes - first due to Pickands in 1974. This construction highlights the importance of polynomial decay to understand events that do not behave as classic statistical models, such as the normal distribution. Heavy-Tail analysis techniques are different from classic statistical methodologies because they focus on the unusual, hence, the mean and variance are not part of the analysis. This thesis discusses modelling empirical data using Heavy-Tail techniques under the Bayesian Paradigm. Bayesian analysis is heavily influenced by decision theory, hence, it's useful to model and study data that are characterised by a lot of variability; furthermore, the computational advantages enable one to account for uncertainty as the algorithms study and learn relevant information about the process that generates Heavy-Tail data. Finally, the work tests the algorithms with a simulation experiment that validates the theory and techniques proposed in this work to estimate Heavy-Tail models; these results are discussed at the end.

Contents

List of Algorithms

List of Figures

Introduction

This work is a brief, yet precise collection of ideas that form the basis for Heavy-Tail analysis. The objective is that the reader is able to understand the theory behind Heavy-Tail models and to perform a basic Heavy-Tail analysis of empirical data. The thesis combines strong theoretical concepts with practical methodologies. This is an attempt to bridge the gap that I have seen between some statistical and mathematical books and computational practical ones focused solely on data analysis. It also exhibits the use of Bayesian techniques for Heavy-Tail analysis.

A fascinating thing about Heavy-Tail analysis is the way it uses, in a clever way, concepts from mathematical analysis to build asymptotic models that can extrapolate valuable information from scarce observations. This is particularly useful to study Extreme Values because, by definition, extreme values are rare; thus, there is a necessity for models that can extrapolate information from small and varied samples. Hence, it is desirable that the reader is familiar with notions from mathematical analysis.

Furthermore, the structure of the work permits the reader to jump ahead into the applications if he or she wishes to do that; on the contrary,

if the reader wants to focus on the theoretical results, it is possible to enjoy the work by just looking at the first sections. My intention is that those who wish to fully understand the construction of Heavy-Tail models are able to do so with the work in the first sections, or if the they prefer, take the lift to the statistical methodology section six proposes.

The first chapter presents the mathematical tools that give place to important results in Extreme Value Theory. It exhibits the key ideas from measure and probability theory used to construct Heavy-Tail models. These ideas are included to provide a connection between the mathematical theory behind a model's construction and the computational techniques needed to implement the model in a practical implementation. The prerequisite is some familiarity with real analysis and probability; nevertheless, a rigorous development of these ideas is beyond the scope of this work the reader should not be concerned with mathematical formality throughout her reading; nonetheless, if the reader desires to read these concepts in more detail he or she should consult [20, 21, 2, 5]. Usually, this chapter would be an appendix, however, the results in this chapter are crucial for the development of Heavy-tail analysis; thus, I think it is worthwhile to exhibit them rather than relegate them to the appendix.

Chapters two and three are the brain of this work. These chapters employ ideas from previous chapters to construct the Fréchet limit for Heavy-Tail extremes. This asymptotic result is crucial for Heavy-Tail analysis because it enables extrapolation from small samples. The limit is constructed with a Point Process that is aimed at highlighting the Power Law behaviours Heavy-Tail extremes have. A Power Law is a function that increases as a power of itself. In essence, these functions are proportional to the product of a polynomial, if c is a constant and

$y \in \mathbb{R}$ and f is a power law where $f(y), f(c), f(cy)$ are both defined, then for some constant $\alpha \in \mathbb{R}$:

$$f(cy) = c^{\alpha} f(y).$$

Furthermore, if for $x, y \in \mathbb{R}$, $f(x), f(y), f(xy)$ are properly defined, then:

$$f(xy) = x^{\alpha} f(y).$$

The function f behaves in such a way that resize values and applies f to them is proportional to multiplying the function by a constant elevated to a power, for instance, if the function measures energy consumption as a function of population size and the function f is power law, then an increase in population equivalent to c times the older population will result in an increase in energy consumption by a factor of c^{α}. Power Laws decay as polynomials for large values, therefore, power laws increase the probability of large deviations when compared with probability distributions characterised by exponential decay. This characteristic leads to the construction of the Fréchet distribution.

The fourth and fifth chapters discuss how to use the results from the previous chapters within a statistical/computational context. These chapters show possible likelihood functions for different Heavy-Tail models. Chapter four exhibits the relationships between different Extreme Value models, and how these relationships can be exploited within a statistical framework. Chapter five discusses a possible methodology to study Heavy-Tail data. This methodology proposes hypothesis tests, a threshold search and selection algorithm, a possible algorithm to simulate from the posterior distribution and

feasible goodness of fit tests.

The sixth chapter is the muscle of the dissertation. This chapter applies ideas from Bayesian Inference to Extreme Value models in an effort to produce reliable inferences. Chapter six applies the methodology from previous chapters propose in a simulation experiment, therefore, I think it is the muscle because all the theoretical work from the previous chapters is unleashed on Heavy-Tail models. The experiment studies four simulations that illustrate the behaviours of Heavy-Tail data; furthermore, chapter six analyses the efficacy given that the process that generates the data is completely known. The positive results are shown throughout chapter six, as well as a discussion about the limitations the algorithms in previous chapters have.

Finally, there is a brief discussion on the importance of Heavy-Tail analysis, as well as the limitations Heavy-Tail analysis and other statistical inference techniques have.

Chapter 1

Measure Theory and Probabilistic Foundations

This chapter lays out the theoretical foundations needed to establish the results present in this thesis. It exhibits ideas and results from Measure Theory that form the foundations of Probability Theory and Statistical Modelling. A complete rigorous development of these is in [5, 21]. The objective is that this work can serve as a bridge between purely theoretical works such as [21, 5] and ones which are more focused on applications, such as [23, 6, 1].

1.1 Measure Theoretic Foundations

Measure Theory is the mathematical backbone of mathematical Probability. The first idea to define and explain in probability is a way to quantify uncertainty. Probability is a measure of belief regarding possible outcomes (subsets) of a given collection (Event space).

A measure is a mathematical function that assigns a size to a set; it generalises the notion of length, area, volume to other concepts, such as probability.

These sets must belong to a measurable collection, in essence to a (σ-algebra= $\mathbb{A}$). Before stating these notions it is necessary to extend the real line because some events/objects might have infinite size, such as the area under some curves.

Definition (Extended Real Line [2, 20])**.** *The extended real line is the extension of the real line with the extended values* $\{-\infty, +\infty\}$.

$$\begin{aligned}\overline{\mathbb{R}} &= \mathbb{R} \cup \{-\infty, +\infty\},\\ \overline{\mathbb{R}^+} &= (0, +\infty) \cup \{+\infty\},\\ \overline{\mathbb{R}^-} &= (-\infty, 0) \cup \{-\infty\}.\end{aligned}$$

The operations are:

$$\begin{aligned}(\pm\infty) \pm (\pm\infty) &= x \pm (\pm\infty) = (\pm\infty), \quad \forall x \in \mathbb{R}.\\ (\pm\infty)(\pm\infty) &= +\infty \ ; \ (\pm\infty)(\mp\infty) = -\infty.\\ x(\pm\infty) &= \begin{cases} \pm\infty & if \ \ x > 0,\\ 0 & if \ \ x = 0,\\ \mp\infty & if \ \ x < 0.\end{cases}\end{aligned}$$

For any $x \in \mathbb{R}$ *the following expressions are not defined.*

$$"+\infty" - "+\infty". \quad "\frac{x}{+\infty}". \quad "\frac{x}{-\infty}". \quad "\frac{x}{0}".$$

Finally, for all $x \in \mathbb{R}$.

$$-\infty < x < +\infty.$$

The previous definition permits the notion of sets with infinite measure, for example, the area under the function $f(x) = \frac{1}{x}$ for $1 \leq x$ (i.e., $\int_1^{+\infty} \frac{1}{x} dx = +\infty$) is infinite, yet the function has a zero limit at infinity (i.e., $\lim_{x \to +\infty} \frac{1}{x} = 0$).

Definition (Sigma algebra [5, 20])**.** *A Sigma algebra* $\mathbb{A}$ *is a subset of the power set of A. In essence:*

$$\mathbb{A} \subseteq \wp(A).$$

Where $\wp(\cdot)$ *denotes the set of all possible subsets of set* $\cdot$*.* $\mathbb{A}$ *is closed under complement and countable unions, it includes A and* $\emptyset$ *(i.e.,* $A, \emptyset \in \mathbb{A}$ *) .*

If the set $\mathbb{A}$ is closed under finite union rather than countable unions, then it is an algebra rather than a σ-algebra [20].

Definition (Measure [21, 20])**.** *A measure is a non-negative function from a* σ*-algebra to the extended non-negative real line that assigns size to a subset of A.*

$$\mu : \mathbb{A} \longrightarrow \overline{\mathbb{R}^+}.$$

It satisfies the following properties.

$$\mu(\emptyset) = 0,$$
$$\mu(A) \geq 0 \quad \forall A \in \mathbb{A},$$
$$\textit{if } A = \bigcup_{n=1}^{\infty} A_n \textit{ , such that } A_i \cap A_j = \emptyset;$$
$$\textit{then } \mu(A) = \mu\big(\bigcup_{n=1}^{\infty} A_n\big) = \sum_{n=1}^{\infty} \mu(A_n).$$

The last equation is σ-additivity, it is the basis for studying limit results regarding measures.

Definition (σ-finite measure [20, 5])**.** *If the set A is measurable with a finite measure (i.e., $\mu(A) < +\infty$), then the measure is σ-finite if there exists a collection of disjoint sets (A_n), such that $\cup_{n=1}^{\infty} A_n = A$ and $\mu(A_n) < +\infty$ for all n. In essence, if it is possible to measure the universal set A as the union of disjoint sets with finite measure, then the measure is σ-finite.*

It might be the case that a measure is σ-finite but not finite, consider the length of the real line, clearly the length is infinite, but it is the countable union of intervals with finite length. For example:

$$\mathbb{R} = \bigcup_{n=-\infty}^{+\infty} (n-1, n].$$

Definition (Measure Space [20])**.** *A measure space $\big(A, \mathbb{A}, \mu\big)$ is a triple consisting of a set A, a collection of subsets of A and a measure μ, this set function assigns a non-negative number to each element in $\mathbb{A}$. If the σ-algebra is the smallest σ-algebra that contains certain subsets of A, then it is said that the σ-algebra is generated by the collection of those subsets, such as open intervals on the real line generate the Borel σ-algebra.*

Perhaps the most famous measure is Lebesgue measure, this measure extends the notion of length to other sets on the real line by approximating from outside such sets with intervals and assigning the length of the best approximating union of exterior intervals, more formally it is based on the extension theorem of measure theory.

Definition (Semi-Measure)**.** *μ is a semi measure if it is a function with the same properties a measure has, but it is defined over an algebra $\boldsymbol{A}$ rather than over a σ-algebra $\mathbb{A}$.*

Definition (Exterior Measure [20])**.** *ρ is an exterior measure if the set function.*

$$\rho : \wp(A) \longrightarrow \overline{\mathbb{R}^+}.$$

Satisfies the following properties.

$$\begin{aligned}
&\rho(\emptyset) = 0,\\
&\rho(B) \geq 0 \quad \forall B \subset A,\\
&\rho(B) \leq \rho(C) \;\; if \;\; B \subset C \subset A,\\
&\text{if } A = \bigcup_{n=1}^{\infty} A_n \quad \text{such that } A_i \cap A_j = \emptyset, \text{ then}\\
&\rho(A) = \rho\big(\bigcup_{n=1}^{\infty} A_n\big),\\
&\qquad = \sum_{n=1}^{\infty} \rho(A_n).
\end{aligned}$$

The exterior measure can be restricted to $A^\rho := \{B \subset A : B \text{ is } \rho \text{ measurable }\}$.

Definition (Restricted Measure [20])**.** *The measure ρ is a restricted measure if it is a measure μ restricted to a specific σ-algebra, such as A^ρ, rather than the whole σ-algebra where μ is defined. .*

$$\overline{\rho} = \rho|_{A^\rho} \longrightarrow \overline{\mathbb{R}^+}.$$

The previous concepts make it possible to extend a semi-measure to an exterior measure.

Lemma (Generated Exterior Measure [20])**.** *The exterior measure μ_*, generated by the semi-measure μ, is a function that extends μ to measure arbitrary sets with approximating sets from the algebra where μ is defined.*
This is done by approximating these sets from outside with sets from the algebra where μ is a semi-measure. In essence:

$$\mu_* := \inf\left\{ \sum_{n=1}^{\infty} \mu(A_n) : E \subset \cup_{n=1}^{\infty}, A_n \in \boldsymbol{A} \right\}$$

for all $E \subset A$. This function can measure all possible subsets of A.

$$\mu_* : \wp(A) \longrightarrow \overline{\mathbb{R}^+}.$$

The generated exterior measure is the same as the semi-measure when restricted to the algebra $\boldsymbol{A}$. The σ-algebra $\boldsymbol{A}^*$ is the σ-algebra generated by μ_* (i.e., it contains all possible subsets that this measure can approximate from the outside).

$$\mu_*|_{\boldsymbol{A}} = \mu.$$

The preceding definitions are necessary to state he following theorem. This theorem is the basis for building complicated probability distributions and measuring difficult sets. It also is an essential part of a key proof in section 4.

Theorem (Extension Theorem, K.Caratheodory - E. Hopf (1918) [20]). *Let $\rho : \wp(A) \to \overline{\mathbb{R}}$ be an exterior measure then:*

$$(i) \quad \boldsymbol{A}^{\rho} \text{ is a } \sigma\text{- algebra},$$

$$(ii) \quad \overline{\rho} = \rho|_{\boldsymbol{A}^{\rho}} : A^{\rho} \longrightarrow \overline{\mathbb{R}},$$

$\overline{\rho}$ is a measure. In the case that $\rho = \mu_*$ as an extension of a semi-measure μ, then the σ algebra generated by $\sigma(\boldsymbol{A})$ is a subset of $\mathbb{A}^*$.

The importance of this theorem is that it allows semi-measures to be defined on algebras rather than σ-algebras with the possibility to be extended to a σ-algebra. This is important because intervals aren't the only type of subsets in the real line. How is it proper to assign measure to a finite collection $\{a, b, c\} \subset \mathbb{R}$; the extension theorem provides a way of approximating each singleton with intervals, subsequently, it assigns the limit length as the size of the singleton. Furthermore, the following theorem states that the extended semi-measure is unique.

Theorem (H. Hahn (1921) [20]). *Let $\mu : \boldsymbol{A} \longrightarrow \overline{\mathbb{R}}$ be a semi-measure that is σ-finite, $S \subset \wp(A)$ a σ-algebra that is a subset of the power set of A and contains $\boldsymbol{A}^*$ the sigma algebra generated by the completion of μ and $\nu : S \longrightarrow \overline{\mathbb{R}}$ a measure, such that $\mu(A) = \nu(A)$ for all $A \in \boldsymbol{A}$, then $\overline{\mu}(B) = \nu(B)$ for all $B \in \boldsymbol{A}^*$.*

This theorem states that if a semi-measure can be extended then its extension is the unique measure over the σ-algebra generated by the extension. Let $\mathbf{A}$ be the algebra generated by finite countable unions of intervals in $\mathbb{R}$. In essence, for $m \in \mathbb{Z}^+$:

$$\mathbf{A} := \bigcup^{m} \{(-\infty, a], (a, b], (b, \infty)\}.$$

Then since singletons do not belong to $\mathbf{A}$, $\mathbf{A}$ is not a σ-algebra; not to worry, the Borel σ-algebra is generated by the algebra of open intervals, as well as the other interval types in $\mathbf{A}$.

Definition (Borel σ-algebra [21, 5])**.** *The Borel σ-algebra is the smallest σ-algebra generated by open intervals in $\mathbb{R}$. It is denoted by.*

$$B_{\mathbb{R}}.$$

This σ-algebra can be extended to be the smallest σ-algebra generated by the extended real line.

$$B_{\overline{\mathbb{R}}} = \sigma(\overline{\mathbb{R}}).$$

This algebra is essential to measure functions that map values to the real line, such as random variables. The concepts in this section could appear as daunting when first approached, however, understanding the relationship between measure theory and probability is what enables the study of complex processes, such as Heavy-Tail analysis - it would be unfeasible without results from measure theory, such as the ones present above. Nevertheless, a full development of these theorems is beyond the scope of this work. The theory is constructed rigorously in [21, 2, 8, 20].

This work includes a proof that permits extending a non-decreasing function to a measure over the real line, it is based on a proof and exercise in [20].

Definition (Generated Measure from g [20])**.** *Let g be a non-decreasing right continuous function; furthermore, let $B_{\overline{\mathbb{R}}}$ be the Borel σ-algebra generated by open intervals in $\overline{\mathbb{R}}$, then the*

Lebesgue-Stieljes semi-measure associated to g is:

$$g : \mathbb{R} \longrightarrow \mathbb{R}^+$$
$$\overline{\lambda_g} : B_{\overline{\mathbb{R}}} \longrightarrow \overline{\mathbb{R}^+}.$$

The exterior measure generated by the semi-measure over intervals:

$$\lambda_g^*(A) = \inf\{\lambda_g(I) : A \subset I = \bigcup_{i=1}^{\infty} I_i\}.$$

The generated measure restricted to the Borel- σ-algebra:

$$\overline{\lambda_g} := \lambda_g^*|_{B_{\overline{\mathbb{R}}}}.$$

The measure assigned to finite intervals:

$$\lambda_g\big((a,b]\big) := g(b) - g(a).$$

Finally, if $(+\infty)$, $(-\infty)$ are contained in the interval.

$$\lambda_g\big((-\infty,b]\big) = g(b) - \lim_{x\downarrow-\infty}(x)$$
$$\lambda_g\big((a,+\infty)\big) = \lim_{x\uparrow+\infty}(x) - g(a)$$
$$\lambda_g(\mathbb{R}) = \lim_{x\uparrow+\infty}(x) - \lim_{x\downarrow-\infty}(x)$$

Theorem (Lebesgue-Stieltjes measure over $\mathbb{R}$)**.** *This proof is an exercise in [20]. Let λ_g be the semi-measure generated by a non-decreasing right continuous function g. Then, the function $\overline{\lambda_g} : B_{\overline{\mathbb{R}}} \longrightarrow \overline{\mathbb{R}}$ is a Lebesgue-Stieljes measure for the measure space $\big(\overline{\mathbb{R}}, B_{\overline{\mathbb{R}}}, \overline{\lambda_g}\big)$; furthermore, by the Extension and H.Hahn theorems it is uniquely determined by g. The function g measures interval length as $g(b) - g(a)$.*

Proof. First it is necessary to prove that λ_g is a semi measure over the algebra of sets $\mathbf{A} = \{(a, b] \subset \mathbb{R}\}$. Clearly the measure of the empty set is zero because the empty set does not contain an interval.

$$\lambda_g(\emptyset) = 0.$$

The semi-measure is the length the function g assigns to intervals.

$$\lambda_g\big((a, b]\big) = g(b) - g(a).$$

Note that if $a < b$, since g is a non-decreasing function, $g(b) \geq g(a)$; therefore, $\lambda_g\big((a, b]\big) \geq 0$ for all intervals in $\mathbf{A}$, also, if $I = (-\infty, a]$ then there exists x_0, $-\infty < x_0 < a$, such that,

$$\lambda_g\big((-\infty, a]\big) = \lambda_g\big((-\infty, x_0]\big) + \lambda_g\big((x_0, a]\big).$$

Since, g is non-decreasing it is true that $\forall x \in \mathbb{R}$ the limits behave as follows: $g(x) \geq \lim_{x\downarrow-\infty} g(x)$ and $g(x) \leq \lim_{x\uparrow+\infty} g(x)$ so,

$$0 \leq g(a) - g(x_0) \leq g(a) - \lim_{x\downarrow-\infty} g(x),$$

analogously for $y_0 > b$,

$$0 \leq g(y_0) - g(b) \leq \lim_{x\uparrow+\infty} g(x) - g(b)$$

Therefore $\lambda_g \geq 0$ for all intervals in $\mathbb{R}$. Now we study countable unions. Case (I):

$$I = (a, b] = \bigcup_{i=1}^{\infty} (a_i, b_i]$$

disjoint.

Without loss of generality there is a finite subcollection of a_i, b_i, such that

$$a \leq a_{i_1} < b_{i_1} \leq a_{i_2} \cdots \leq b.$$

Therefore,

$$g(a) \leq g(a_{i_1}) < g(b_{i_1}) \leq a_{i_2} \cdots \leq g(b_{i_m}) \leq g(b),$$

Thus,

$$g(b) - g(a) = g(b) + \sum_{j=1}^{m} g(b_{i_j}) - g(a_{i_j}) +$$

$$\sum_{j=1}^{m-1} g(a_{i_{j+1}}) - g(b_{i_j}) + g(a_{i_1}) - g(b_{i_m}) - g(a)$$

$$\geq \sum_{j=1}^{m} g(b_{i_j}) - g(a_{i_j}).$$

The sum covers all sub-parts, g is non-decreasing so the total length must be greater than or equal to the sum of the lengths within. Therefore, $g(b) - g(a)$ is an upper bound for the sums over the interval $(a, b]$.

$$\therefore \sum_{i=1}^{\infty} g(b_n) - g(a_n) \leq g(b) - g(a).$$

Let $\epsilon \in (0, g(b) - g(a))$ since g is right continuous for all $b_n \in \mathbb{R}$ and $\forall \epsilon > 0$, it is possible to fix $\delta_n > 0$ such that

$$g(b_n + \delta_n) - g(b_n) < \frac{\epsilon}{2^n}.$$

Also, there exists a family $I_n = (a_n, b_n + \delta_n)$ of open covers for the compact set $[a + \epsilon, b]$ such that for a finite m,

$$[a + \epsilon, b] \subset \bigcup_{n \in I_G}^{m} I_{n_i}.$$

Where $G = \{I_n\}_1^\infty$ is an open cover, such that there exists a finite cover; at least one finite index cover exists because the interval is compact. It is also possible to select and rearrange a subcollection to get,

$$a_{n_1} < a + \epsilon < a_{n_2} < b_{n1} + \delta_{n_1} < a_{n_3}$$

$$< \cdots < b_{n_{m-1}} + \delta_{n_{m-1}} \leq b < b_{n_m} + \delta_{n_m}$$

And since g is non-decreasing and right continuous,

$$g(a_{n_1}) < g(a + \epsilon) < g(a_{n_2}) < g(b_{n1} + \delta_{n_1}) < g(a_{n_3})$$

$$< \cdots < g(b_{n_{m-1}} + \delta_{n_{m-1}}) \leq g(b) < g(b_{n_m} + \delta_{n_m})$$

Recall that because g is right continuous

$$g(b_{n_j} + \delta_{n_j}) < \frac{\epsilon}{2^{n_j}} + g(b_{n_j}).$$

So rearranging the sum leads to,

$$\begin{aligned}
g(b) - g(a+\epsilon) &\leq g(b_{n_m} + \delta_{n_m}) - g(a_{n_1}) \\
&= g(b_{n_1} + \delta_{n_1}) - g(a_{n_1}) + \cdots \\
&+ \sum_{j=2}^{m} g(b_{n_j} + \delta_{n_j}) - g(b_{n_{j-1}} + \delta_{n_{j-1}}) \\
&< g(b_{n_1}) + \frac{\epsilon}{2^{n_1}} - g(a_{n_1}) + \sum_{j=2}^{m} g(b_{n_j}) + \frac{\epsilon}{2^{n_j}} - g(a_{n_j}) \\
&= \sum_{j=1}^{m} g(b_{n_j}) - g(a_j) + \frac{\epsilon}{2^{n_j}} \\
&\leq \sum_{j=1}^{\infty} g(b_{n_j}) - g(a_j) + \frac{\epsilon}{2^{n_j}} \\
&\leq \sum_{j=1}^{\infty} g(b_{n_j}) - g(a_j) + \sum_{j=1}^{\infty} \frac{\epsilon}{2^j} \\
&= \sum_{j=1}^{\infty} g(b_{n_j}) - g(a_j) + \epsilon \frac{1/2}{1/2},
\end{aligned}$$

$$\therefore g(b) - g(a+\epsilon) < \sum_{j=1}^{\infty} g(b_{n_j}) - g(a_j) + \epsilon.$$

Therefore $\forall \epsilon \in (0, g(b) - g(a))$ there exist I_n and δ_n such that

$$g(b) - g(a+\epsilon) < \sum_{j=1}^{\infty} g(b_{n_j}) - g(a_j) + \epsilon.$$

Let $\epsilon \to 0$ and by the right continuity of g it is so that $g(a+\epsilon) \to g(a)$, this results in,

$$g(b) - g(a) \leq \sum_{j=1}^{\infty} g(b_{n_j}) - g(a_j).$$

Thus, it must be that,

$$g(b) - g(a) = \sum_{n=1}^{\infty} g(b_n) - g(a_n).$$

Case (II): $I = (a, +\infty)$ at least one extreme point of the interval is unbounded. To shorten notation consider.

$$g(\pm\infty) = \lim_{x \to \pm\infty} g(x),$$

Therefore, $\lambda_g(I) = g(+\infty) - g(a)$. $\mathbb{R} = I = \cup I_n$ disjoint unbounded sets that satisfy:

$$\begin{aligned}
\sum_{n=1}^{\infty} \lambda_g(I_n) &= \sum_{n=2}^{\infty} g(b_n) - g(a_n) - g(b_{n-1}) + g(a_{n-1}) \\
&\leq g(\sup\{b_n\}) - g(\inf\{a_n\}) \\
&= g(+\infty) - g(-\infty) \\
&= \lambda_g(I).
\end{aligned}$$

Suppose $I = (a, +\infty)$, I_n is bounded for all $(a_n, b_n]$; $a < b$, a fixed, set

$$\tilde{b_n} = \min\{b, b_n\}$$

$$(a, b] = (a, b] \cap I = \bigcup_{n=1}^{\infty} (a_n, b_n].$$

Then because the $(a_n, b_n]$ are disjoint sets, by case (I)

$$\begin{aligned}\lambda_g((a,b]) &= \sum_{a_n<b} g(\tilde{b_n}) - g(a_n) \\ &\leq \sum_{n=1}^{\infty} g(b_n) - g(a_n).\end{aligned}$$

Then it must be that

$$\therefore \lim_{b\to+\infty} g(b) - g(a) = \sum_{n=1}^{\infty} g(b_n) - g(a_n).$$

Analogously for $I = (-\infty, b]$ with the I_n bounded for all $(a_n, b_n]$; for $a < b$, fixed b, set

$$\tilde{a_n} = \max\{a, a_n\}$$

$$(a,b] = (a,b] \cap I = \bigcup_{n=1}^{\infty} (a_n, b_n].$$

Then because the $(a_n, b_n]$ are disjoint sets, by case (I)

$$\begin{aligned}\lambda_g((a,b]) &= \sum_{a<b_n} g(b_n) - g(\tilde{a_n}) \\ &\leq \sum_{n=1}^{\infty} g(b_n) - g(a_n) \\ \therefore g(b) - \lim_{a\to-\infty} g(a) &= \sum_{n=1}^{\infty} g(b_n) - g(a_n)\end{aligned}$$

Case (III): $I \in \mathbf{A}$ then is I the countable union of sets considered in the previous cases; $I = \cup^m I_j$, $m < +\infty$, such that

$$\begin{aligned}\lambda_g(I) &= \lambda_g(\bigcup^m I_j) \\ &= \sum_{j=1}^{m} \lambda_g(I_j).\end{aligned}$$

For each $I = \bigcup^\infty I_n$ it is possible to find the collection

$$\begin{aligned}\bigcup^m I_j &= \bigcup_{j=1}^{m} \bigcup_{n}^{\infty} I_{n_j} \\ \lambda_g(\bigcup^m I_j) &= \sum_{j=1}^{m} \sum_{n=1}^{\infty} \lambda_g(I_{n_j}) \\ &= \lambda_g(\bigcup I_{n_j}) \\ \therefore \lambda_g(I) &= \lambda_g(\bigcup I_{n_j}).\end{aligned}$$

Finally, since $\mathbb{R} = \cup_{n=-\infty}^{+\infty}(n-1, n]$ it is true that:

$$\begin{aligned}\lambda_g(\mathbb{R}) &= \lambda_g(\bigcup_{n=-\infty}^{+\infty} (n-1, n]) \\ &= \sum_{n=-\infty}^{+\infty} g(n) - g(n-1).\end{aligned}$$

A telescopic sum, leads to,

$$\begin{aligned} &= \lim_{x\uparrow+\infty} g(x) - \lim_{x\downarrow-\infty} g(x) \\ &= \sum_{n=1}^{\infty} g(b_n) - g(a_n); \\ b_n \to +\infty \;\; a_n \to -\infty. \end{aligned}$$

Therefore, λ_g is σ-additive, σ-finite and it satisfies the definition of a semi-measure over $\mathbf{A}$; therefore, by the extension theorem, λ_g^* is a complete measure and $\overline{\lambda_g}$ is a measure over the Borel σ-algebra $B_{\overline{\mathbb{R}}}$. Furthermore if g is bounded the measure is finite,

$$g(-\infty) \leq g(+\infty) < +\infty,$$

then,

$$\overline{\lambda_g}(\mathbb{R}) = g(+\infty) - g(-\infty) < +\infty.$$

Is a finite measure. Also, if g is continuous then the set:

$$x_0 = \bigcap_{n=1}^{\infty} (x_0 - \frac{1}{n}, x_0].$$

Has the following measure,

$$\begin{aligned}
\lambda_g(\{x_0\}) &= \lambda_g(\bigcap_{n=1}^{\infty}(x_0 - \frac{1}{n}, x_0]), \\
&= g(x_0) - \lim_{n\to+\infty} g(x_0 - \frac{1}{n}), \\
&= g(x_0) - g(x_0), \\
&= 0. \\
\therefore \lambda_g(a,b)) &= \lambda_g((a,b] - \{b\}), \\
&= \lambda_g((a,b]) - \lambda_g(\{b\}), \\
&= \lambda_g((a,b]), \\
&= g(b) - g(a).
\end{aligned}$$

■

This proves that this is a generalisation of Lebesgue measure by means of a non-decreasing function, one such function could be:

$$g : \mathbb{R} \longrightarrow \mathbb{R}^+$$

$$g(x) = \begin{cases} 0 & \text{if } x \leq \zeta \\ K - x^{-\alpha} & \text{if } x > \zeta \end{cases}$$

The function g is called the distribution of λ_g. It does not have to be continuous, nonetheless, it has to be non-decreasing; thus, the plot will either resemble a smooth hill climb or a step ladder.

Figure 1 illustrates the behaviour a distribution function g could have.

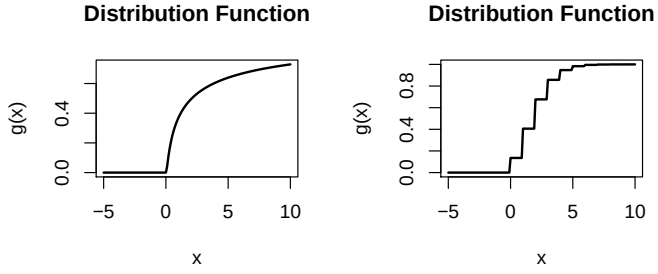

Figure 1.1. Distribution Function Plots: Smooth (left), Step (right).

There are two more important theorems from measure theory that are needed to obtain probabilistic results. To measure functions over arbitrary sets it is necessary to define the following functions that simplify and approximate the measures of interest over the set B. Before stating the theorems it is necessary to sate the following definitions.

Definition (Indicator Function)**.** *The indicator function,*

$$\chi_B(x) = \begin{cases} 1 & \text{if } x \in B \\ 0 & \text{if } x \notin B \end{cases}$$

Definition (Simple Function)**.** *A simple function s over a partition of A is the sum of non-negative coefficients for different subsets of A. It is simple because it is a sum, nonetheless, these functions can approximate any non-negative measurable function.*

$$s = \sum_{j=1}^{n} \beta_j \chi_{B_j}.$$

These are the basis needed to define the integral of a measurable function over a set $B \subseteq A$.

Definition (Measurable Function [20, 21])**.** *A function $f : A \to \mathbb{R}$ is Borel measurable if and only if the pre-image $f^{-1}((-\infty, x])$ belongs to the σ-algebra $\mathbb{A}$. In essence.*

$$f^{-1}((-\infty, x]) \in \mathbb{A}, \quad \forall x \in \mathbb{R}.$$

This enables one to measure sets of a measurable space $(A, \mathbb{A}, \mu)$ with sets from the real line.

Lemma (Non-negative measurable function approximation. [2, 20])**.** *Let (A, S) be a measurable space and $f : A \longrightarrow \mathbb{R}$ a non-negative S measurable function, then there exists a sequence (s_n) of S simple functions such that,*

(i) $0 \leq s_n \leq s_{n+1} \leq f$,
(ii) $\lim_{n\to\infty} s_n(x) \to f(x) \quad x \in A$,
(iii) if f is bounded then $s_n \to f$ uniformly in A.

Definition (Null set)**.** *A subset $N(A)$ of A is a null set with respect to the measure μ if $\mu(N(A)) = 0$.*

Definition (Almost Everywhere Convergence)**.** *The sequence of functions (f_n) over the measure space $(A, \mathbb{A}, \mu)$ converges almost everywhere with respect to the measure μ to the function f if for all $a \notin N(A)$, it is true that:*

$$f_n(a) \to f(a), \quad a \in A - N(A).$$

Theorem (Monotone Convergence Theorem [5, 20])**.** *Let (f_n) be a monotone non-decreasing sequence of measurable functions over $(A, \mathbb{A}, \mu)$. If $f_n(a) \uparrow f(a)$ for $a \in A - N(A)$, then the limit is interchangeable with the integral. In essence for $B \subseteq A$.*

$$\int_B f d\mu = \lim_{n \uparrow +\infty} \int_B f_n d\mu.$$

A way to study sets and their properties is to study integrals of functions over those sets, these integrals provide information about how the function behaves in such sets. The integral of measurable function f is the size of the function with respect to the measure μ over a set $B \subseteq A$. If it helps, think of the simple functions as rectangles that approximate f. An intuitive definition that is valid because of the Monotone Convergence Theorem and the approximation of measurable functions by simple non-decreasing functions is the following one.

Definition (Measurable Function Integral [20])**.**

$$\int_B f d\mu = \lim_{n\uparrow\infty} \int_B s_n d\mu.$$

Where the $s_n = \beta_j \chi_{B_j}$ *are simple functions and their integrals are,*

$$\int_B s_n d\mu = \sum_{j=1}^{n} \beta_j \mu(B_j).$$

This enables the study of more complicated sequences of functions. The next theorem is a handy tool that allows the interchange of the integral and limit under lax restrictions in comparison to uniform convergence.

Theorem (Dominated Convergence Theorem H. Lebesgue, 1910)**.** *Let* (A, S, μ) *be a measure space and* (f_n) *a sequence of measurable functions (i.e.,* $\int_A |f_n| d\mu < +\infty$ *) such that* $f_n(x) \to f(x)$ *a.s with respect to* μ*. The value of the functions may only differ on sets with measure zero with respect to* μ*. Furthermore, assume there exists a function* g*, such that* $\int_A g d\mu < +\infty$ *and* $|f_n| \leq g$ *a.s. with respect to* μ *for all* $n \in \mathbb{N}$*. Then,*

$$(i) \int_A |f| d\mu < +\infty,$$

$$(ii) \int_B f d\mu = \lim_{n\to\infty} \int_B f_n d\mu.$$

For all $B \subseteq A$*.*

This theorem is a powerful tool, it makes it possible to establish the convergence of a sequence of integrals.

1.2 Convergence Criteria for Random Variables

Sequences of random variables are sequences of measurable functions; hence, measure theory is the tool used to study sequences of random variables (X_n), $n \in \mathbb{N}$.

Definition. *The four most common types of convergence for random variables are [2, 5, 21]:*
Almost sure convergence (the random variables converge point-wise with probability one):

$$P\big(\lim_{n\to\infty} X_n = X\big) = 1$$

Convergence in probability (The probability of the sequence being far from the limit goes to zero) if for all $\epsilon > 0$:

$$\lim_{n\to\infty} P\big(|X - X_n| > \epsilon\big) = 0$$

Convergence in p-mean (The p-norm expected distance between the random variables goes to zero):

$$\int_{\mathbb{X}} |X_n - X|^p dP \to 0$$

Convergence in law (The distribution functions converge point-wise at points of continuity):

$$\lim_{n\to\infty} P(X_n \leq x) = \lim_{n\to\infty} F_n(x) = F(x) = P(X \leq x)$$

Points of continuity are those in which the distribution function does not jump (i.e., if $\delta > 0$ and $\delta \to 0$, then $F(x \pm \delta) \to F(x)$), otherwise, x could be a point at which F jumps (ladder function).

Definition (Weak Convergence [5, 2]). *If f is continuous, non-negative and bounded (i.e., $f \in C_b^+(\mathbb{X})$). Then the sequence (μ_n) converges weakly to μ, written ($X_n \Rightarrow X$), where $\ell(X_n) = \mu_n$ and $\ell(X) = \mu$ if and only if*

$$E(f(X_n)) = \int_{\mathbb{X}} f d\mu_n \to \int_{\mathbb{X}} f d\mu = E(f(X)), \ \ \forall f \in C_b^+(\mathbb{X}).$$

Definition (Vague Convergence [2]). *The sequence ν_n converges vaguely to ν in $M(\sigma(\mathbb{X}))$ (the space of measures over the σ-algebra), if and only if, $\nu_n(\mathbb{X}) < +\infty$ for all n, $\nu(\mathbb{X}) < +\infty$, and*

$$\nu_n(f) = \int_{\mathbb{X}} f d\nu_n \to \int_{\mathbb{X}} f d\nu = \nu(f), \ \ \forall f \in C_K^+(\mathbb{X}).$$

Weak convergence is more general than convergence in law because it establishes convergence in metric spaces, whereas, convergence in law is limited to the real line; nevertheless, there are important equivalences between distinct convergence types, such as the one the following theorem presents. Before it is useful to study the following result.

Proposition. *Assume that the random variable t is uniform, $t \sim Unif([0,1])$, then $X(t) = \inf\{x : F(x) \geq t\}$ has distribution function F. In essence:*

$$P(X(t) \leq x) = F(x).$$

Proof. By definition,

$$P(X(t) \leq x) = P(X(t) \in (-\infty, x])$$

Recall that F is right-continuous, thus, it obtains its minimum, as well as monotone increasing; hence,

$$\begin{aligned} &\inf\{x : F(x) \geq t\} = \min\{x : F(x) \geq t\}, \\ &0 \leq F(x) \leq 1 \quad \forall x \in \mathbb{R}. \end{aligned}$$

Therefore it is true that,

$$X(t) \leq x \iff t \leq F(x)$$

Hence,

$$\begin{aligned} &P(X(t) \leq x) = P(t \leq F(x)) \\ &= F(x) \\ &\therefore X \sim F(x). \end{aligned}$$

■

Thus, it is possible to construct a random variable from a cumulative distribution function by use of standard uniform random variables.

Theorem (Skorohod's Theorem). *Let (μ_n) be a sequence of Borel probability measures (i.e., $\int_{\mathbb{R}} d\mu_n = \mu_n(\mathbb{R}) = 1$.) Such that for every $x \in \mathbb{R}$ the numbers,*

$$\mu_n((-\infty, x] \to \mu((-\infty, x]) \quad \forall x \ \ s.t \ \ \mu(\{x\}) = 0.$$

The sequence converges at points of continuity. Then, there exists a sequence of random variables (X_n) with $\ell(X_n) = \mu_n$ that converges almost surely (strongest convergence) to the random variable X with law $\ell(X) = \mu$.

The following proof is an adaptation/outline completion of the proof presented in [5, 2].

Proof. Skorohod's Theorem is an essential result with respect to weak convergence. This proof is built from the ones in [5, 2]. Suppose,

$$\mu_n((-\infty, x] \to \mu((-\infty, x]), \quad \forall x \ \ s.t \ \ \mu(\{x\}) = 0.$$

Define the cumulative distribution functions as,

$$F_n(x) = \mu_n((-\infty, x]) \ \ ; \ \ F(x) = \mu((-\infty, x]),$$

the random variables $X(a) = \inf\{x : F(x) \geq a\}$, and $X_n(a) = \inf\{x : F_n(x) \geq a\}$; furthermore, let $\Omega = [0, 1]$, $\sigma(\Omega) = B_{[0,1]}$ and $\mathbb{P} = \lambda$ (Lebesgue measure) length. Then because of the proposition above $\omega \in \Omega$ is a standard uniform random variable (i.e., $F(\omega) = \omega$.). By the proposition stated before this theorem, the laws of X_n and X above are well defined, such that, $\ell(X_n) = \mu_n$ and $\ell(X) = \mu$. Since $F_n(x) \to F(x)$ it is natural to study if $X_n(\omega) \to X(\omega)$ at most points.

By hypothesis for $a, b, s, t \in [0, 1]$,

$$F_n(x) < a \Rightarrow X_n(a) \geq x,$$
$$F_n(x) \geq b \Rightarrow X_n(b) \leq x,$$
$$F(x) < t \Rightarrow X(t) \geq x,$$
$$F(x) \geq s \Rightarrow X(s) \leq x.$$

Assume that $X(\omega) = y$ is continuous at ω, furthermore, suppose that for some $\epsilon > 0$, $F(x - \epsilon) \geq \omega$. Then setting $x = y - \epsilon$ and $s = \omega$ above leads to,

$$F(y - \epsilon) \geq \omega \Rightarrow X(\omega) \leq y - \epsilon = X(\omega) - \epsilon \ \ddagger.$$

A contradiction. Therefore, $F(y - \epsilon) < \omega$.
Suppose now that $F(y + \epsilon) \leq \omega$, then setting $t = \omega + \delta$ and $x = y + \epsilon$ above results in,

$$F(y + \epsilon) \leq \omega < \omega + \delta, \quad \forall \delta > 0.$$

This means that for the conditions above we get,

$$F(y + \epsilon) < \omega + \delta \ \ \forall \delta > 0 \Rightarrow X(\omega + \delta) \geq x \quad \forall \delta > 0,$$

$$\Rightarrow X(\omega + \delta) \geq X(\omega) + \epsilon \ \ \forall \delta > 0 \ \ \ddagger.$$

This contradicts the continuity of X at ω. Therefore for all $\epsilon > 0$,

$$F(y - \epsilon) < \omega < F(y + \epsilon).$$

For a fixed ϵ it is possible to select $\tilde{\epsilon}$ such that $0 < \tilde{\epsilon} < \frac{\epsilon}{2}$, and,

$$\mu(\{y - \tilde{\epsilon}\}) = 0 = \mu(\{y + \tilde{\epsilon}\}).$$

By hypothesis,

$$F_n(y - \tilde{\epsilon}) \to F(y - \tilde{\epsilon}),$$

$$F_n(y + \tilde{\epsilon}) \to F(y + \tilde{\epsilon}),$$

Therefore for sufficiently large $n \geq N(\epsilon, \tilde{\epsilon})$ the limit results in.

$$F_n(y - \tilde{\epsilon}) < \omega < F_n(y + \tilde{\epsilon}).$$

This means that for $n \geq N(\epsilon, \tilde{\epsilon})$ setting $x = y - \tilde{\epsilon}$ and $a = \omega$ above results in,

$$\begin{aligned} F_n(y - \tilde{\epsilon}) < \omega &\Rightarrow y - \tilde{\epsilon} \leq X_n(\omega), \\ \omega < F_n(y + \tilde{\epsilon}) &\Rightarrow X_n(\omega) \leq y + \tilde{\epsilon}. \end{aligned}$$

Combining these results with the ones for $X(\omega)$ for $n \geq N(\epsilon, \tilde{\epsilon})$ leads to,

$$\begin{aligned} y - \tilde{\epsilon} \leq & X_n(\omega) \leq y + \tilde{\epsilon} \\ y - \tilde{\epsilon} \leq & X(\omega) \leq y + \tilde{\epsilon}. \end{aligned}$$

By rearranging,

$$\begin{aligned} y - \tilde{\epsilon} - y - \tilde{\epsilon} \leq & X_n(\omega) - X(\omega) \leq y + \tilde{\epsilon} - y + \tilde{\epsilon}, \\ & |X_n(\omega) - X(\omega)| \leq 2\tilde{\epsilon} < \epsilon. \end{aligned}$$

For sufficiently large $n \geq N(\epsilon, \tilde{\epsilon})$ at points of continuity of (X_n) and X, $X_n(\omega) \to X(\omega)$. The sequence (X_n) and X are non-decreasing measurable functions defined for $\omega \in [0,1]$; thus, these functions are monotone functions of a real variable mapped to the real numbers; hence, by Froda's Theorem [20] they can at most have a countable number of discontinuities, since countable sets have no length the discontinuity sets have Lebesgue measure zero. This means that for $\omega \in D(X_n)$ (the set of discontinuities) the probability that either (X_n) and/or X is discontinuous is,

$$\begin{aligned} P(D(X_n) \cup D(X)) &= P(\omega \in D(X_n) \cup D(X)), \\ &= \mathbb{P}(\omega \in \{\omega_i\}_1^n), \\ &= \lambda(\{\omega_i\}_1^n) \;\; n \in \mathbb{N}, \cup\{+\infty\}, \\ &= 0. \end{aligned}$$

For all points of continuity it is true that $X_n(\omega) \to X(\omega)$, and $P(D(X_n) \cup D(X)) = 0$. Since the event that the entire sequence converges almost surely only happens if convergence occurs at all points of continuity, convergence at a single point of continuity is a subset (event), of almost sure convergence; thus,

$$P(X_n \to X) \geq P(X_n(\omega) \to X(\omega) \;\; and \; \omega \notin D(X_n) \cup D(X)) = 1.$$

Therefore,

$$P(X_n \to X) = 1.$$

■

The theorem above is surprising because weak convergence does not imply almost sure convergence; nevertheless, Skorohod's theorem

exhibits that there exists a way in which it's possible to build a sequence (X_n), as well as a random variable X defined over a special probability space, such that convergence in distribution for said sequence within said probability space does imply almost sure convergence. If it helps, think of this as the possibility to find a sport in which it's possible to find an athlete for whom expertise at the high school level guarantees mastery at the professional level. It shouldn't be possible but it is so for unique cases. This helps understand convergence criteria for random variables and the relationships between, transformations, cumulative distribution functions and point-wise limits. This theorem is useful for proving the continuous mapping theorem.

Theorem (Continuous Mapping Theorem)**.** *Let X_n, X be random variables with respective laws $\ell(X_n) = \mu_n$, $\ell(X) = \mu$, such that.*

$$\mu_n((-\infty, x] \to \mu((-\infty, x]) \ \forall x \ \ s.t \ \ \mu(\{x\}) = 0.$$

Then if $g(X_n)$ is a bounded measurable function such that, $\mu(D(g)) = 0$, the set of discontinuities of g has probability zero, then g preserves weak convergence:

$$\int g d\mu_n \to \int g d\mu.$$

Furthermore if the function g is continuous,

$$g(X_n) \Rightarrow g(X).$$

Remark. *If the function g is the identity $g(x) = x$, the original random variables converge weakly, $X_n \Rightarrow X$. Continuous functions preserve weak convergence of random variables, as well as convergence of real number sequences.*

Proof. Another outline completion from [2].
By Skorohod's Theorem; if $\mu_n((-\infty, x] \to \mu((-\infty, x])$ and $\mu(\{x\}) = 0$. At all points of continuity, then there exist random variables X_n, X defined over some probability space, such that $\ell(X_n) = \mu_n$, $\ell(X) = \mu$ and $P(X_n \to X) = 1$ (i.e., the random variables converge point wise almost everywhere). Let g be a bounded measurable function, such that $\mu(D(g)) = 0$, thus, $\mu(D^c(g)) = 1$ (All continuity regions have probability); if g is continuous since $X_n(\omega) \to X(\omega)$ the continuity of g leads to

$$g(X_n(\omega)) \to g(X(\omega)).$$

At all points of continuity, since $g(X_n) \to g(X)$ only happens if the sequence converges point-wise almost everywhere, convergence at a point of continuity is a subset of this event; therefore,

$$\begin{aligned} P(g(X_n) \to g(X)) &\geq P(g(X_n(\omega)) \to g(X(\omega)) \cdots \\ &\quad \textit{and } X_n(\omega), X(\omega) \notin D(g)), \\ &= 1, \\ \therefore\ P(g(X_n) \to g(X)) &= 1. \end{aligned}$$

Given that g is bounded by the Dominated Convergence Theorem,

$$E(g(X_n)) = \int_{\mathbb{X}} g d\mu_n \to \int_{\mathbb{X}} g d\mu = E(g(X)).$$

Since all bounded continuous functions are bounded measurable functions whose set of discontinuities have no measure (probability) for all $f \in C_K^+(\mathbb{X})$,

$$\int_{\mathbb{X}} f d\mu_n \to \int_{\mathbb{X}} f d\mu.$$

Therefore for any sequence (Y_n) such that, $\ell(Y_n) = \mu_n$, $\ell(Y) = \mu$ and $\forall f \in C_K^+(\mathbb{Y})$ if the function g is continuous, then the composition is continuous (i.e., $f \circ g \in C_K^+(\mathbb{Y})$); furthermore, because of the previous results,

$$\int_{\mathbb{Y}} f(g(Y_n))d\mu_n \to \int_{\mathbb{Y}} f(g(Y))d\mu.$$

Hence, $g(Y_n) \Rightarrow g(Y)$. ■

This shows that weak convergence is a handy tool. It turns out that convergence in law and weak convergence are equivalent.

Lemma (Weak convergence is equivalent to convergence in law)**.** *Let $(X_n) \in \mathbb{R}$ be a sequence of random variables, then $X_n \Rightarrow X$ (converge weakly), if and only if, $P(X_n \leq x) \to P(X \leq x)$ the sequence also converges in distribution.*

Proof. This proof is a more detailed presentation of the one in [5]. Suppose $X_n \Rightarrow X$, then $\forall f \in C_K^+(\mathbb{X})$,

$$\int_{\mathbb{X}} f d\mu_n \to \int_{\mathbb{X}} f d\mu$$

For x fix $\epsilon > 0$ and define the bounded continuous function:

$$f(t) = \begin{cases} 1 & \text{if } t \leq x \\ -\frac{1}{\epsilon}t + 1 + \frac{1}{\epsilon} & \text{if } t \in (x, x+\epsilon] \\ 0 & \text{if } t > x + \epsilon \end{cases}.$$

Therefore, $\chi_{(-\infty,x]}(t) \leq f(t) \leq \chi_{(-\infty,x+\epsilon]}(t)$; hence,

$$\begin{aligned}\limsup_{n\to\infty} \mu_n((-\infty,x]) &\leq \limsup_{n\to\infty} \int f d\mu_n, \\ &= \int f d\mu \leq \mu((-\infty,x+\epsilon]), \\ \therefore \limsup_{n\to\infty} \mu_n((-\infty,x]) &\leq \mu((-\infty,x+\epsilon]).\end{aligned}$$

Next define the bounded continuous function:

$$g(t) = \begin{cases} 1 & \text{if } t \leq x-\epsilon \\ \frac{1}{\epsilon}(x-t) & \text{if } t \in (x-\epsilon, x] \\ 0 & \text{if } > \epsilon \end{cases}.$$

Then, $\chi_{(-\infty,x-\epsilon]}(t) \leq g(t) \leq \chi_{(-\infty,x]}(t)$, hence,

$$\begin{aligned}\mu((-\infty,x-\epsilon]) &\leq \int g d\mu, \\ &= \liminf_{n\to\infty} \int f d\mu_n \leq \liminf_{n\to\infty} \mu_n((-\infty,x]), \\ \therefore \mu((-\infty,x-\epsilon]) &\leq \liminf_{n\to\infty} \mu_n((-\infty,x]).\end{aligned}$$

Therefore for all $\epsilon > 0$ and points x of continuity (i.e, $\mu(\{x\}) = 0.$) the equation results in,

$$\begin{aligned}\mu((-\infty,x)) &\leq \liminf_{n\to\infty} \mu_n((-\infty,x]), \\ \limsup_{n\to\infty} \mu_n((-\infty,x]) &\leq \mu((-\infty,x]).\end{aligned}$$

Given that $\mu(\{x\}) = 0$. There is no probability jump at x, thus,

$\mu((-\infty, x)) = \mu((-\infty, x])$; hence, it must be that

$$\lim_{n\to\infty} \mu_n((-\infty, x]) = \mu((-\infty, x]).$$

Therefore, the random variables converge in law.
Suppose the random variables converge in law, (μ_n), μ such that,

$$\mu_n((-\infty, x] \to \mu((-\infty, x]) \quad \forall x \ \ s.t \ \ \mu(\{x\}) = 0.$$

Then by the Continuous Mapping Theorem if $\ell(X_n) = \mu_n$, and $\ell(X) = \mu$, leads to:

$$X_n \Rightarrow X.$$

■

Therefore, $X_n \Rightarrow X \iff F_n(x) \to F(x)$; therefore, the probability law of the sequence is well approximated by the law of the limit even though the actual value of the random variables might differ. Also, once $X_n \Rightarrow X$, then any continuous transformation will also converge weakly, such as the exponential function.

$$(X_n \Rightarrow X) \Rightarrow (\exp(X_n) \Rightarrow \exp(X)).$$

Remark. *Clearly if $\mu_n \Rightarrow \mu$ at points of continuity. It is true that*

$$F_n(x) = \mu_n((-\infty, x] \to \mu((-\infty, x]) = F(x).$$

Then the tail function (Probability that X is as big as a surplus) $S(x) = 1 - F(x)$ converges at points of continuity too.

$$S_n(x) = \mu_n((-\infty, x]^c) \to \mu((-\infty, x]^c) = S(x).$$

Because the function S does not affect the sequential limit.

To study probabilities of sequences of random variables it is possible to study the distribution of the limit random variables if the sequence converges in some type to a random variable. This is particularly useful for extreme values because extreme value distributions are difficult to compute, as well as inherent scarcity extreme value samples are characterised by. However, it's possible to study the limit distribution s for extreme values and Heavy-Tails, this makes it feasible to compute and handle the challenging functions involved, as well as generalise the techniques to more random variables. The following theorems proof is extensive and beyond the scope of this work; nonetheless, it is key to establishing results in the coming sections. A rigorous proof of this result can be found in [5, 21]

Definition (Characteristic Function [21])**.** *Let X be a random variable, its Characteristic Function, $t \in \mathbb{R}$, is:*

$$\phi_X(t) = \int_{\mathbb{X}} e^{ixt} d\mu_n(x) = E(\exp(itX)).$$

Definition (Laplace Transform [21])**.** *Let X be a non-negative random variable, its Laplace Transform for $s > 0$ is:*

$$\varphi_X(s) = \int_{\mathbb{X}} e^{-sx} d\mu(x) = E(\exp(-sX)).$$

Theorem (Levy Continuity Theorem [21])**.** *Case (I): Let (X_n) be a sequence of random variables then if the sequence of characteristic functions converge point wise and are continuous at 0, then the random variables converge in distribution, hence, the random variables converge weakly.*

In essence, if the sequence of random variables (X_n) and the random variable X satisfy:

$$\begin{aligned}
\phi_n(t) &\to \phi(t) \ \ \forall t \in \mathbb{R}, \\
\int_{\mathbb{X}} e^{ixt} d\mu(x) &= \phi(t) \ \ \forall t \in \mathbb{R}, \\
\int_{\mathbb{X}} e^{ixt} d\mu_n(x) &= \phi_n(t) \ \ \forall t \in \mathbb{R}, \\
\lim_{t \to 0} \phi(t) &= \phi(0).
\end{aligned}$$

Then the sequence (X_n) converges weakly to the random variable X, probabilistically they're quite similar if not identical. Written $X_n \Rightarrow X$.

Remark. *Given that the moment generating function and Laplace transforms can be obtained from the characteristic function,*

$$\begin{aligned}
M_X(t) &= \phi(-it) = E(\exp(-i^2 tX)), \\
\varphi_X(t) &= \phi(it) = E(\exp(-tX)), \quad t > 0.
\end{aligned}$$

If either one of these functions is continuous in a neighbourhood of 0 and the functions converge point wise within that neighbourhood, then by the Levy Continuity Theorem the sequence (X_n) converges weakly to the random variable X. This is only valid if the moment generating function and Laplace transform are finite in a neighbourhood of zero.

The preceding theorem results from the uniqueness of Fourier and Laplace transforms. It is extremely useful in proving limit results for random variables, such as the Central Limit Theorem. As the previous results in this chapter, this theorem will allow the derivation of limit distributions for extreme values with a Heavy-Tail.

Chapter 2

Heavy-Tail Models Probabilistic Construction

2.1 Extreme Values and Heavy-Tails

This section introduces important concepts that help understand how to study and analyse Heavy-Tail models. There is a mixture between definitions and concepts presented in an effort to clarify the particular behaviour that is characteristic of heavy-Tail models and the connection these models have with extreme values. Given that extreme values are thought of as unlikely events, it is necessary to establish some concepts that make the study of extreme events and Heavy-Tails feasible.

Definition (Upper Bound of a Random variable [1, 2])**.** *The largest value x_+ the random variable can attain without reaching cumulative probability 1. If the random variable is unbounded then,*

$$x_+ := \sup\{F(x) = 1\} = +\infty.$$

Therefore, a random variable X is unbounded if and only if the support of X is unbounded.

Definition (Tail of a Random Variable [1, 25])**.** *The tail (survival) of X is the function,*

$$S(x) = 1 - F(x) = P(X \in (x, +\infty]).$$

This function is the probability X has of surpassing a large value x, if X is unbounded, the tail probability ($P(X > x)$) is greater than zero for all positive real numbers. If X has a proper distribution function, then this function goes to zero for large values of X because $P(X \leq +\infty) = 1 \Rightarrow S(+\infty) = 0$. The interesting thing about the tail function is: how fast does it approach zero for large values of X. The answer is that it either approaches zero exponentially fast, power law fast or some decay in between, such as the Weibull with decay parameter equal to one - slower than a exponential tail but faster than a power law [25, 2]. Heavy-Tail models are partly inspired by power laws because if the tails decays as a polynomial it resembles a Power Law [23, 22].

Definition (Regular Varying Function [2, 22])**.** *A measurable function f is a regular varying function if, for $t > 0$, and for some $\rho \in \mathbb{R}$.*

$$\lim_{x \to +\infty} \frac{f(xt)}{f(x)} = t^{\rho}.$$

Slowly varying if,

$$\lim_{x \to +\infty} \frac{f(xt)}{f(x)} = 1.$$

Where the parameter ρ is the regularly varying index; usually, if the function f satisfies these limits, then it is said that f belongs to the family of ρ regularly varying functions, written as $f \in RV_{\rho}$.

These functions behave somewhat nicely at infinity because they resemble a Power Law. Heavy-Tail data is characterised by large deviations from the mean, since large deviations are more probable when probability isn't concentrated (the expected value does not exist) The tail S is exponentially bounded if for large values of x there exists a constant $\lambda > 0$, such that $S(x) \leq e^{-\lambda x}$ [16, 9]. Heavy-Tails are characteristic of random variables with infinite moments [5, 23, 8]. In fact, if the tail is bounded by an exponential tail, then it might be worthwhile to study the moment generating function.

Definition (Moment Generating Function [6])**.**

$$E(\exp(tX)) = \int_{\mathbb{X}} e^{tx} dF(x) = M_X(t).$$

If the moment generating function is continuous at zero, and bounded for all $t \in (-h, h)$ for some $h > 0$, then all of the moments of X are finite, hence, in that case X does not have a Heavy-Tail. Therefore, Heavy-Tails are not bounded by an exponential tail, thus, large deviations have a larger probability because the tail decays slower than the moment expectations grow.

Definition (Heavy-Tail [2, 1])**.** *A random variable X has a Heavy-Tail if and only if its tail is a regular varying function for some $\alpha > 0$, written as $S \in RV_{-\alpha}$.*

$$\lim_{x \to +\infty} \frac{S(tx)}{S(x)} = t^{-\alpha}.$$

If X has a Heavy-Tail, then for all $t \in (-h, h)$ for some $h > 0$.

$$\int_{-\infty}^{+\infty} e^{tx} dF(x) = +\infty.$$

Proof. If $S \in RV_{-\alpha}$ then for large values of x the function is approximately a power law, $S(tx) \approx x^{-\alpha} S(t)$. Recall that the exponential grows much faster than any polynomial (i.e, $e^x >> x^\rho$) [21]. Thus, the tail does not decay fast enough to balance the exponential. Suppose $S(x) \approx x^{-\alpha}$. Then,

$$\begin{aligned} \lim_{x \to +\infty} e^{tx} S(x) &\to \lim_{x \to +\infty} e^{tx} x^{-\alpha} L(x), \\ &= +\infty \ \ \forall \alpha > 0. \end{aligned}$$

Furthermore by Karamata's theorem [21, 2]; for a sufficiently large s,

$$\begin{aligned} \int_{-\infty}^{+\infty} e^{tx} dF(x) &\geq \int_{s}^{+\infty} e^{tx} dF(x), \\ &\to \int_{s}^{+\infty} e^{tx} x^{-\alpha-1} L(x) dx, \\ &\to L(s) \int_{s}^{+\infty} e^{tx} x^{-\alpha-1} dx, \\ &= +\infty. \end{aligned}$$

Therefore, the moment generating function does not exist if the tail decays slower than an exponential function. Above, L is a slowly varying function. ■

Remark. *In fact for any $k > \alpha$ the k moment of X does not exist [23].*

$$\int_0^{+\infty} x^k f(x)dx \sim \int_0^{+\infty} x^k x^{-\alpha-1}dx = \begin{cases} E(X^k) & if \;\; k < \alpha \\ +\infty & if \;\; k \geq \alpha. \end{cases}.$$

Theorem (Moment Generating Function Existence and continuity imply Finite Moments [21, 7])**.**

$$If, \;\; M_X(t) = \int_{-\infty}^{+\infty} e^{tx} dF(x) < +\infty.$$

For all $t \in (-h, h)$ for some $h > 0$ and $M_X(t)$ is continuous at $t = 0$, then, all moments of X exist; thus,

$$\int_{-\infty}^{+\infty} x^k dF(x) < +\infty.$$

For all $k \in \mathbb{Z}^+$.

Proof. This proof is a completion of an outline in [5]. Suppose that the moment generating function exists, in essence that,

$$\int_{-\infty}^{+\infty} e^{tx} dF(x) < +\infty.$$

For $t \in (-h, h)$ for some $h > 0$ the moment generation function $M_X(t)$ of X exists. Thus,

$$M_X(t) = E(\exp(tX)) = \int_{-\infty}^{+\infty} e^{tx} dF(x).$$

Is finite in a neighbourhood of zero and continuous at zero.

$$\begin{aligned} E(\exp(tX)) &= \int_0^{+\infty} e^{tx} dF(x), \\ &= \int_{\mathbb{X}} e^{xt} d\mu. \end{aligned}$$

Consider the sequence,

$$Y_n = \sum_{k=1}^{n} \frac{t^k X^k}{k!}.$$

Then because the exponential is continuous in $\mathbb{R}$.

$$\begin{aligned} \lim_{n\to\infty} Y_n(\omega) &= \lim_{n\to\infty} \sum_{i=1}^{n} \frac{t^i X^i(\omega)}{i!}, \\ &= \exp(tX(\omega)). \end{aligned}$$

For points $\omega \in \Omega$ *in a non-measure zero set.*

$$\therefore\ P(Y_n \to \exp(tX)) = 1.$$

Define the random variable $Z = \exp(-tX) + \exp(tX)$.

$$\begin{aligned} |Y_n| &= \left| \sum_{k=1}^{n} \frac{t^k X^k}{k!} \right|, \\ &\leq \sum_{k=1}^{n} \frac{|t^k||X^k|}{k!}, \\ &\leq \exp(|tX|), \\ |tX| &= tX \ \textit{or,}\ |tX| = -tX. \end{aligned}$$

The exponential is non-negative; thus,

$$\exp(|tX|) \leq \exp(-tX) + \exp(tX) = Z.$$

Therefore, $Y_n \to \exp(tX)$ and $|Y_n| \leq Z$ for all n; thus, by the Dominated Convergence Theorem,

$$\int_{\mathbb{R}} Y_n d\mu \to \int_{\mathbb{R}} \exp(tX) d\mu < +\infty.$$

This results in,

$$\int_{\mathbb{R}} \sum_{k=1}^{n} \frac{t^k X^k}{k!} d\mu = \sum_{k=1}^{n} \int_{\mathbb{R}} \frac{t^k X^k}{k!} d\mu < +\infty.$$

Therefore, $E(|X^k|) = \int_{\mathbb{X}} |X|^k d\mu < +\infty$ for all $k \in \mathbb{Z}^+$. ■

For random variables with Heavy-Tails, no moment for $k > \alpha$ exists. Thus, if the moment generating function exists for some $t > 0$ the random variable does not have a Heavy-Tail [22]. To better understand the concept and why Heavy-Tails are the result of tails that are not bounded by an exponential tail consider two random variables X, Y such that the tail of X, $S_X \in RV_{-\alpha}$ and that the tail of Y is bounded by an exponential tail (i.e., $S_Y(x) \leq e^{-\alpha x}$.) for sufficiently large values of x. The tail odds are,

$$\text{odds}(\alpha) = \frac{e^{-\alpha x}}{x^{-\alpha}}.$$

Since $e^{\alpha x} >> x^{\alpha}$ for large x and for all $\alpha > 0$. The odds are $X^{-\alpha} >> e^{-\alpha x}$. Therefore, the probabilities of observing a large deviation are

much higher for Heavy-Tail random variables. Not all moments of a Heavy-Tail random variable exist; thus, it is not wise to study centres of mass because the random variable does not concentrate near a point, as it does when the tail is exponentially bounded [23, 1, 2].

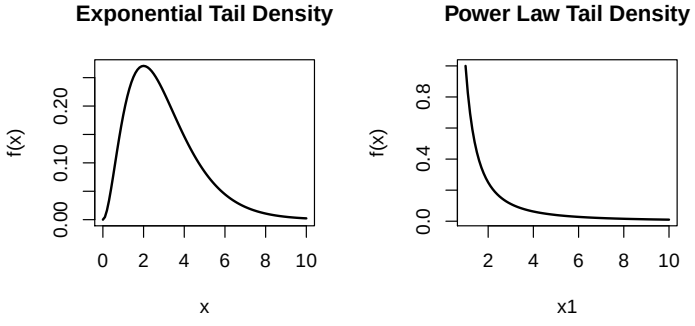

Figure 2.1. Exponential Tail and Heavy-Tail. (x1 is typo).

Finally, the smaller the value of α is the heavier the tail is because the probability of large deviations decays slower for small values of α. For large, $x > 1$, the tail can be written as,

$$x^{-\alpha} = \exp(-\alpha \ln(x)).$$

This exponential function is monotone decreasing; thus, if $\alpha_1 < \alpha_2$, then $x^{-\alpha_1} > x^{-\alpha_2}$; the result is a heavier tail. The key idea of this section is that large deviations are much more probable for Heavy-Tails; therefore, observations do not centre around a particular value.

To better look at this recall that if the random variable X has a Heavy-Tail, then not all of its moments are finite; whereas, if the tail is exponential, then $S(x) \leq e^{-\lambda x}$; thus,

$$\begin{aligned}\int_0^{+\infty} e^{tx} S(x)dx &\leq \int_0^{+\infty} e^{tx} e^{-\lambda x} dx, \\ &= \int_0^{+\infty} e^{-x(\lambda - t)} dx, \;\; t \in (-\lambda, \lambda), \\ &< +\infty.\end{aligned}$$

Therefore, the moments are finite and the random variable centres; on the contrary, if the tail is bounded by a Pareto like tail,

$$\begin{aligned}\int_0^{+\infty} e^{tx} S(x)dx &\leq \int_0^{+\infty} e^{tx} x^{-\alpha} dx, \;\; \forall \alpha > 0, \;\; t, \\ &= +\infty.\end{aligned}$$

This comparison shows that the study of Heavy-Tail data should focus on the tail index α, rather than on values where the random variable might centre, such as the mean and variance; therefore, study of Heavy-Tail data differs from classical statistical analysis that focuses on centres of mass of the random variable [25, 1, 23].

2.2 Point Process and Poisson Point Process

Point Processes are useful random measures that count the random number of events within a region of interest. These objects are random measures because they count the random number of independent points in a given region; thus, the argument of the point process is not a number or event but a region; also, once the region and number of possible events are fixed the Point Process counts the number of successes in a given region [2, 1]. Formally, a point process is a measure over a measurable space that counts the random number of points within a measurable region (set) of that space, for example, a ball in the plane could be the argument, and points within the ball the measure assigned to the ball, if point are randomly sampled, then the region is randomly measured.

Definition (Dirac measure [20])**.** *The Dirac measure for a subset A of the Event Space $\mathbb{X}$ is the function:*

$$\delta_{X_i}(A) = \begin{cases} 1 & \text{if } X_i \in A, \\ 0 & \text{if } X_i \notin A. \end{cases}$$

Note that this is not a function of the Random Variable, it is a function of the set A. This is to avoid confusion with the case in which A is fixed and the function evaluates if X_i is in A - written $\delta_A(X_i)$.

Definition (Point Process [2, 23])**.** *The Point Process is the random measure that counts the random number of points from the collection $\{X_i\}_1^n$ of independent identically distributed random variables that fell in some region of the support of X ($X_i \in A \subset \mathbb{X}$). It is the function.*

$$N_n(\cdot) = \sum_{i=1}^{n} \delta_{X_i}(\cdot).$$

Proposition. *It is important to note that neither n and/or A need to be fixed for the Point Process to be well defined; nonetheless, once both of these are fixed, the Point Process resembles a Binomial random variable [1, 6] - note that the Point Process is not a Binomial random variable, however, once A, n are fixed the Point Process probabilities can be computed as follows.*

Proof. For fixed n and A it is true that $\sum_{i=1}^{n} \delta_{X_i}(A) \in \{0, 1, \cdots, n\}$ where $P(\delta_{X_i}(A) = 1) = P(X_i \in A)$; also, the X_i are independent so the δ are independent because if one X_i falls in A it does not alter the probability of the other X_j, $j \neq i$, of falling in A. Therefore, the number of points the process counts are independent trials with the same probability of success,

$$\begin{aligned} P\big(N_n(A) = k\big) &= P\big(\sum_{i=1}^{n} \delta_{X_i}(A) = k\big), \\ &= \binom{n}{k} \big(P(X_i \in A)\big)^k \big(1 - P(X_i \in A)\big)^{n-k}, \\ k &\in \{0, 1, \cdots, n\}. \end{aligned}$$

Therefore, the Point Process for a fixed number of points within a fixed region has probabilistic characteristics of a Binomial random variable; in essence, $N_n|A, n \sim Binomial(n, P(X_i \in A)$. ■

The Point Process generalises the binomial distribution to count arbitrary successful trials in arbitrary regions. The Poisson Point Process generalises the Poisson distribution to count the number of points in a region with a measure of intensity for the given region [23, 1].

Definition (Intensity Point Measure)**.** *The measure ν is a point intensity measure for the set A if ν is a measure over a measure space that contains A, and $\nu(A)$ measures the intensity with which points might fall in A.*

If it helps, think of $\nu(A)$ as the intensity with which points fall in the set A; the larger $\nu(A)$ is the more intensely points fall in A. It is not a probability but a measure of intensity, for example, the intensity of cars being on the road could be infinite because all cars will be on some road at some point in time.

Definition (Poisson Point Process [2, 25])**.** *A Poisson Point Process $N(\cdot)$ is a random measure with an intensity measure ν, such that disjoint regions have independent counts; in essence, $N(A_j)$ is independent of $N(A_i)$ if $A_i A_j = \emptyset$. The probabilities for a given region $A \in \sigma(\mathbb{X})$ and $k \in \mathbb{Z}_+$ are:*

$$P\big(N(A) = k\big) = \begin{cases} \frac{(\nu(A))^k}{k!} \exp(-\nu(A)) & \text{if } \nu(A) < +\infty \\ 0 \quad \text{if } \nu(A) = +\infty \end{cases}$$

Therefore, if ν is finite in A, $N(A)$ is a Poisson random variable with parameter $\nu(A)$. There are some regularity conditions that relate the Poisson distribution as the limit of a Binomial distribution when the number of events goes to infinity and the the probability of success times the number of events converges to a finite limit (i.e., $\lim_{n\to\infty} np = \lambda$). This can be proven by the use of characteristic functions. An alternative of the characteristic function is the Laplace functional of a random measure [2, 21].

Definition (Laplace Functional [2, 21]). *The Laplace functional of the random measure N, such that there exists at least one region with a finite measure (i.e., $\nu(\cdot) < +\infty$); with regards to a non-negative measurable function f is,*

$$\Psi_N(f) = E\big(\exp(-N(f)\big).$$

Where,

$$\begin{aligned} N(f) &= \int_{\mathbb{X}} f dN, \\ &= \sum_{\{X_i \in (\cdot)\}} f(X_i), \\ &= \sum_{i=1}^{N(\cdot)} f(X_i) \geq 0. \end{aligned}$$

The accumulated value of the function by the random number of points that fell in the region.

It is possible to generalise the Continuity Theorem.

Definition (Continuous function Metric [2, 20]).

$$d(f,g) = \frac{1}{2}|N_n(f) - N_n(g)| \geq 0.$$

It is possible to measure distance between continuous functions as the distance between finite integrals, $\forall f \in C_K^+(\mathbb{X})$.

Theorem (Continuity Theorem for Laplace Functionals [2])**.** *Let $(N_n) \in M_+$ be a sequence of measures and N a measure in M_+. If the Laplace functionals converge point wise and are continuous at the zero fucntional, then the random measures converge weakly. In essence:*

$$\Psi_{N_n}(f) \to \Psi_N(f), \ \forall f \in C_K^+(\mathbb{X}) \iff N_n \Rightarrow N.$$

Proof. This proof is a completion and adaptation of the outline in [2]. Suppose,

$$N_n \Rightarrow N.$$

Since $N_n(f) \in \mathbb{R}_0^+$ for all f in $C_K^+(\mathbb{X})$, the composition of the exponential with these real numbers is a continuous function for each f; thus, $\exp(-N_n(f))$ is a continuous function of $N_n(f)$; hence, by the continuous mapping theorem,

$$\exp(-N_n(f)) \Rightarrow \exp(-N(f)).$$

Furthermore, $0 \leq \exp(-N_n(f)) \leq 1$ for all $f \in C_K^+(\mathbb{X})$, therefore, by the Bounded Convergence Theorem,

$$\int_{M_+} \exp(-N_n(f))dP \to \int_{M_+} \exp(-N(f))dP.$$

For all f in $C_K^+(\mathbb{X})$.

$$\therefore \Psi_{N_n}(f) \to \Psi_N(f) \ \forall f \in C_K^+(\mathbb{X}).$$

Now suppose

$$\Psi_{N_n}(f) \to \Psi_N(f) \ \forall f \in C_K^+(\mathbb{X}).$$

Thus, it is sufficient to show that the Laplace transforms of the non-negative generated random variables converge for any given family $\{f\} \in C_K^+(\mathbb{X})$. Consider an arbitrary family of size m and define the random non-negative vectors,

$$\underline{Y_m^n} = \big(N_n(f_1), N_n(f_2), \cdots, N_n(f_m)\big)^t,$$
$$\underline{Y_m} = \big(N(f_1), N(f_2), \cdots, N(f_m)\big)^t.$$

Then for $s = (s_1, \cdots, s_m)^t \geq 0$ the Laplace transform is:

$$\begin{aligned}
\varphi_n(s) &= E\big(\exp(-s^T \underline{Y_m}^n)\big), \\
&= E\big(\exp(-\sum_{i=1}^m s_i N_n(f_i))\big), \\
&= E\big(\exp(-\sum_{i=1}^m s_i \int_{\mathbb{X}} f_i dN_n)\big), \\
&= E\big(\exp(-\int_{\mathbb{X}} \sum_{i=1}^m s_i f_i dN_n)\big), \\
&= E\big(\exp(-N_n(\sum_{i=1}^m s_i f_i))\big), \\
&= \Psi_{N_n}(\sum_{i=1}^m s_i f_i).
\end{aligned}$$

Since all $s_i \geq 0$, then $g = \sum_{i=1}^m s_i f_i \in C_K^+(\mathbb{X})$.

Combining the previous arguments.

$$
\begin{aligned}
&\Psi_{N_n}(\sum_{i=1}^{m} s_i f_i) \to \Psi_N(\sum_{i=1}^{m} s_i f_i),\\
&= E\big(\exp(-\int_{\mathbb{X}} \sum_{i=1}^{m} s_i f_i dN)\big),\\
&= E\big(\exp(-\sum_{i=1}^{m} s_i \int_{\mathbb{X}} f_i dN)\big),\\
&= E\big(\exp(-\sum_{i=1}^{m} s_i N(f_i))\big),\\
&= E\big(\exp(-s^T \underline{Y_m})\big),\\
&= \varphi(s),\\
&\therefore \varphi_n(s) \to \varphi(s).
\end{aligned}
$$

Therefore, the Laplace transforms converge for all $s \geq 0$; hence, by the Levy Continuity Theorem for any given family $N_n(\{f\}) \Rightarrow N(\{f\})$. Therefore, $N_n \Rightarrow N$. [2]. ■

The remarkable result of the theorem above is that it does not matter what the number of points observed could be or in which region is the measure concentrated, as long as the Laplace Functional converges for all continuous bounded non-negative functions; the random variables generated by a fixed number of points and a fixed region will converge weakly. Therefore, $N_n \Rightarrow N$ in regions of interest. This property is useful to study collections of random variables $\{X\}$ as events within a region, such as where do extreme observations concentrate. As a final note on Point Processes and the Poisson Point Process, the thesis computes the Poisson Point Process Laplace

Functional. This will, hopefully, clarify the concept and show the benefits and setbacks of this approach.

Proposition. *Let N be a Poisson Point Process with intensity measure ν such that for the set $A \neq \emptyset$, it's true that $P(N(A) = 0) = \exp(-\nu(A))$. Then,*

$$\Psi_N(f) = \exp\big(-\int_{\mathbb{X}}(1 - e^{-f})d\nu\big) \;\; \forall f \in B_b^+(\mathbb{X}).$$

Where $B_b^+(\mathbb{X})$ is the set of non-negative bounded Borel measurable functions whose set of discontinuities have measure zero with respect to ν (i.e., $\nu(D(f)) = 0$) [25]. Recall that all continuous functions are measurable.

Proof. This proof adapts and fills the outline in [2].
Let $f = \beta\chi_A(x)$ and $\beta > 0$ be the scaled indicator function of the set $A \subseteq \mathbb{X}$. Then,

$$N(f) = \int_{\mathbb{X}} dN = \int_A \beta dN = \beta N(A).$$

Thus, $N(f) = \beta N(A)$; for the fixed set A, the Poisson Point Process is a Poisson random variable, $N(A) \sim Poisson(\nu(A))$.

Hence, the Laplace Functional of the indicator function is the Laplace transform of $N(A) \sim Poisson(\nu(A))$.

$$\begin{aligned} E\big(\exp(-N(f))\big) &= E\big(\exp(-\beta N(A))\big), \\ &= \sum_{n=0}^{\infty} e^{-\beta n} \frac{(\nu(A))^n}{n!} e^{-\nu(A)}, \\ &= e^{-\nu(A)} \sum_{n=0}^{\infty} \frac{(e^{-\beta}\nu(A))^n}{n!}, \\ &= e^{-\nu(A)} \exp(e^{-\beta}\nu(A)), \\ &= \exp(-\nu(A)(1-e^{-\beta})). \end{aligned}$$

Recall that,

$$\begin{aligned} \int_{\mathbb{X}} (1-e^{-f}) d\nu &= \int_{\mathbb{X}} (1-e^{-\beta}) \chi_A(x) d\nu, \\ &= \int_A (1-e^{-\beta}) d\nu, \\ &= \nu(A) - \nu(A)e^{-\beta}, \\ &= \nu(A)(1-e^{-\beta}); \\ \therefore \exp(-\nu(A)(1-e^{-\beta})) &= \exp\bigg(-\int_{\mathbb{X}} (1-e^{-f}) \bigg). \end{aligned}$$

Let $\{f_n\} \subset B^+$, such that $f_k = \beta_k \chi_{A_k}$ where the, $A_i A_j = \emptyset$ if $i \neq j$, are disjoint and $A_k \subset \mathbb{X}$. Therefore, the Poisson random variables for each set A_k are independent.

Let $f = \sum_{k=1}^{n} f_k$ be a sum of $\sigma(\mathbb{X})$ simple functions.

$$
\begin{aligned}
\Psi_N(f) &= E\big(\exp(-N(f))\big), \\
&= E\big(\exp(-N(\sum_{i=1}^{n} f_k))\big), \\
&= E\big(\exp(-\sum_{i=1}^{n} \int_{\mathbb{X}} f_k dN)\big), \\
&= E\big(\exp(-\sum_{i=1}^{n} N(f_k))\big), \\
&= E\big(\prod_{i=1}^{n} \exp(-N(f_k))\big), \\
&= \prod_{i=1}^{n} E\big(\exp(-\beta_k N(A_k))\big) \ \textit{by independence.}, \\
&= \prod_{i=1}^{n} \exp\Big(-\int_{\mathbb{X}} 1 - e^{-f_k} d\nu\Big), \\
&= \exp\Big(\sum_{i=1}^{n} -\int_{\mathbb{X}} 1 - e^{-f_k} d\nu\Big), \\
&= \exp\Big(-\int_{\mathbb{X}} \sum_{i=1}^{n} 1 - e^{-f_k} d\nu\Big).
\end{aligned}
$$

The A_k are disjoint; hence, if one of the indicators is not zero, all the other ones are zero ($f_k = \beta_k \chi_{A_k}$). For the set A_i this results in:

$$\begin{aligned}
\sum_{i=1}^{n} 1 - e^{-f_i} &= n - (n-1) - e^{\beta_i}, \\
&= 1 - e^{-\beta_i}, \\
&= 1 - e^{-f_i(x)}, \\
&= 1 - \exp\big(-\sum_{i=1}^{n} f_k(x)\big), \\
&= 1 - e^{-f(x)}.
\end{aligned}$$

For each set A_k, once A_k is fixed the indicator function results in:

$$\sum_{i=1}^{n} (1 - e^{-\sum_{i=1}^{n} f_k}) \chi_{A_k} = 1 - e^{-f}.$$

Then the integral of the sum of the f_k,

$$\begin{aligned}
\int_{\mathbb{X}} \sum_{i=1}^{n} 1 - e^{-f_k} d\nu &= \int_{\mathbb{X}} \sum_{i=1}^{n} (1 - e^{-\beta_k}) \chi_{A_k} d\nu, \\
&= \sum_{i=1}^{n} \int_{A_k} 1 - e^{-f_k} d\nu, \\
&= \sum_{i=1}^{n} \int_{A_k} 1 - e^{-\sum_{i=1}^{n} f_k} d\nu, \\
&= \sum_{i=1}^{n} \int_{\mathbb{X}} 1 - e^{-\sum_{i=1}^{n} f_k} \chi_{A_k} d\nu, \\
&= \int_{\mathbb{X}} \sum_{i=1}^{n} 1 - e^{-\sum_{i=1}^{n} f_k} \chi_{A_k} d\nu, \\
&= \int_{\mathbb{X}} 1 - e^{-f} d\nu, \\
\therefore \int_{\mathbb{X}} \sum 1 - e^{-f_k} d\nu &= \int_{\mathbb{X}} 1 - e^{-f} d\nu.
\end{aligned}$$

Substituting this into the previous equation results in:

$$\exp\left(- \int_{\mathbb{X}} \sum 1 - e^{-f_k} d\nu \right) = \exp\left(- \int_{\mathbb{X}} 1 - e^{-f} d\nu \right).$$

Finally, for all $f \in B^+$ there exists a monotone sequence $(f_n) \subset B^+$ such that $f_n = \sum_{i=1}^{m} \beta_i \chi_i$ for disjoint A_i for all n and $f_n(x) \uparrow f(x)$ for all $x \in \mathbb{X}$ outside of a zero set.

It is possible to manipulate the sequence as follows,

$$\begin{aligned} & f_n \leq f \ \ \forall n, \\ & \Rightarrow -f_n \geq -f, \\ & \Rightarrow e^{-f_n} \geq e^{-f}, \\ & \Rightarrow 1 - e^{-f_n} \leq 1 - e^{-f}. \end{aligned}$$

$1 - e^{-x}$ is a continuous continuous functions, hence, it preserves sequential limits; so $1 - e^{-f_n(x)} \uparrow 1 - e^{-f(x)}$. Then, by the Monotone Convergence Theorem,

$$\int_{\mathbb{X}} 1 - e^{-f_n} d\nu \uparrow \int_{\mathbb{X}} 1 - e^{-f} d\nu.$$

Note that the expectation is with respect to the function $0 \leq \exp(-N(f)) \leq 1$ for all $f \in B^+$ and $x \in \mathbb{X}$; therefore, by the Dominated Convergence Theorem,

$$E(\exp(-N(f_n))) = \int_{M_+} \exp(-N(f_n)) dP$$

$$\rightarrow \int_{M_+} \exp(-N(f)) dP = E(\exp(-N(f))).$$

Also, by the Monotone Convergence Theorem if $f_n(x) \rightarrow f(x)$, then the previous result holds and given that the exponential function preserves sequential limits,

$$\exp\left(- \int_{\mathbb{X}} 1 - e^{-f_n} d\nu \right) \uparrow \exp\left(- \int_{\mathbb{X}} 1 - e^{-f} d\nu \right).$$

Combining the previous two results,

$$E(\exp(-N(f))) = \exp\left(-\int_{\mathbb{X}} 1 - e^{-f} d\nu \right).$$

As claimed for all f in B^+. ■

The Laplace Functional is more flexible than the Laplace Transform because it can be calculated with test functions in arbitrary regions; in contrast, the Laplace Transform can only be calculated for fixed regions that define a Poisson random variable [5, 2].

2.3 Fréchet Limit for Heavy-Tail Extremes

The following definitions are useful to study the intensity and number of extreme observations.

Definition (Mean Extreme Measure [2, 1])**.** *Let X be a non-negative random variable with a Heavy-Tail. Then the Point Process that counts large deviations $X > z$ has mean measure ν_n, such that:*

$$\nu_n(A_z) = E(N_n(A_z)) = nP(X > z).$$

Definition (Extreme Intensity Measure [2, 1])**.** *Let X be a non-negative random variable with a Heavy-Tail. Then the Poisson Point Process that counts large deviations $X > z$ has intensity measure ν, such that:*

$$\nu(A_z) = \lim_{z \to +\infty} \frac{S(xz)}{S(z)} = x^{-\alpha}.$$

It's sensible to take a moment to understand the difference between these two measures. The Mean Extreme Measure counts the number of

expected extreme events within a region A_z, in essence, it is the mean count of points that are bigger than z if points are sampled frequently; on the other hand, the intensity measure is the asymptotic limit for the tail probability as a function of the value z. One counts the expected above a certain value z, the other one is the limit for tail probabilities with respect to the value z. The next Theorem is the tool that makes it possible to extrapolate information from scarce observations.

Theorem (Extremal Types Theorem for Heavy-Tails [1])**.** *Let $\{X_i\}_1^n$ be a sequence of unbounded from above independent identically distributed random variables, such that the tail of X is regularly varying at infinity, in essence that $S(x) \in RV_{-\alpha}$ for $\alpha > 0$. Then either there exist sequences $a_n, b_n > 0$ such that for $X^{(n)} = \max\{X_i\}$ the distribution converges to a Fréchet distribution:*

$$\lim_{n\to+\infty} P\left(\frac{X^{(n)} - a_n}{b_n} \leq x\right) = \exp(-x^{-\alpha}).$$

Or the stabilised Maxima degenerate,

$$P\left(\frac{X^{(n)} - a_n}{b_n} \leq x\right) \to \delta_x(\{x_0\}) \;\; n \to +\infty.$$

The proof of the previous theorem presented here is built upon fragments and combinations of derivations, proofs and outlines in [1, 2, 5, 23].

Proof. Let $\{X_i\}_1^n$ be a sequence of non-negative unbounded independent identically distributed random variables, such that the tail of X is regularly varying at infinity; $S(x) \in RV_{-\alpha}$ for $\alpha > 0$.

$$\lim_{t\to+\infty} \frac{S(xt)}{S(t)} = x^{-\alpha}.$$

Set $b_n := \inf\{z \in \mathbb{R}^+ : S(z) \geq \frac{1}{n}\}$, $n \in \mathbb{N}$; thus, given that S is a non-increasing function $b_n \leq S^{\leftarrow}(\frac{1}{n})$, also, $S(x) \to 0$ as $x \to +\infty$; therefore, b_n is well defined and since the X_i are unbounded; $b_n \leq b_{n+1} \to +\infty$, the sequence is unbounded; otherwise, there would be a n, such that the upper bound $x_+ = \sup\{x : F(x) < 1\} < +\infty$, however,

$$\frac{1}{n+1} < \frac{1}{n} \to 0 \quad n \uparrow +\infty.$$

Thus, $b_n \to x_+$ but the X_i are unbounded; hence, $x_+ = +\infty$, so $b_n \to +\infty$.
Set $n(t) = \inf\{m \in \mathbb{Z}^+ : t < b_m\}$; therefore, $t \leq b_{n(t)}$ for all n, t; also, there exists a $s > 0$, such that

$$t \leq b_{n(t)} < t + s.$$

Thus, for $x > 0$ the following inequalities are valid.

$$xt \leq xb_{n(t)} \leq x(t + s).$$

Recall that $S(x)$ is a non-increasing function so,

$$S(x(t + s)) \leq S(xb_{n(t)}) \leq S(xt).$$

$$S(t + s) \leq S(b_{n(t)}) \leq S(t).$$

Therefore:

$$\frac{S(x(t+s))}{S(t)} \leq \frac{S(x(t+s))}{S(b_{n(t)})},$$

$$\frac{S(x(t+s))}{S(b_{n(t)})} \leq \frac{S(xb_{n(t)})}{S(b_{n(t)})},$$

$$\frac{S(xb_{n(t)})}{S(b_{n(t)})} \leq \frac{S(xt)}{S(b_{n(t)})},$$

$$\frac{S(xt)}{S(b_{n(t)})} \leq \frac{S(xt)}{S(t+s)}.$$

Putting it all together:

$$\frac{S(x(t+s))}{S(t)} \leq \frac{S(xb_{n(t)})}{S(b_{n(t)})} \leq \frac{S(xt)}{S(t+s)}.$$

Let $t \to +\infty$, then $n(t) \to +\infty$ and since $t \leq b_{n(t)}$ this results in $b_{n(t)} \to +\infty$; also, for large enough t the addition of a fixed $s > 0$ for each t becomes negligible, since

$$\lim_{t \to +\infty} \frac{t+s}{t} = 1.$$

Also $S(x)$ is a regularly varying function; thus, the following limits are valid.

$$\lim_{t \to +\infty} \frac{S(x(t+s))}{S(t)} = \frac{S(xt)}{S(t)} = x^{-\alpha}.$$

$$\lim_{t \to +\infty} \frac{S(xt)}{S(t+s)} = \frac{S(xt)}{S(t)} = x^{-\alpha}.$$

By the limit squeeze theorem,

$$\lim_{t \to +\infty} \frac{S(xb_{n(t)})}{S(b_{n(t)})} = \lim_{n \to +\infty} \frac{S(xb_n)}{S(b_n)} = x^{-\alpha}.$$

Note that $S(b_n) \geq \frac{1}{n}$, so $\forall \epsilon > 0$ there exists N_ϵ such that if $n \geq N_\epsilon$ then

$$|S(b_n) - \frac{1}{n}| < \frac{\epsilon}{2}.$$

Therefore, for large x, b_n the next inequalities are valid:

$$S(b_{n+1}) \leq \frac{1}{n} \leq S(b_n),$$

$$\frac{S(xb_n)}{S(b_{n+1})} \leq \frac{S(xb_n)}{1/n} \leq \frac{S(xb_n)}{S(b_n)}.$$

These limits result in:

$$\lim_{t\to+\infty} \frac{S(xb_n)}{S(b_{n+1})} = \lim_{n\to+\infty} \frac{S(xb_n)}{S(b_n)} = x^{-\alpha}.$$

$$\therefore \lim_{n\to+\infty} nS(xb_n) = \lim_{n\to+\infty} \frac{S(xb_n)}{1/n} = x^{-\alpha}.$$

The previous result asserts that the $\big(b_n\big)$ act as a stabilising sequence for large deviations, this sequence prevents $S(x)$ from becoming zero as X increases. Recall that,

$$S(xb_n) = P\big(X > xb_n\big),$$

$$\therefore \lim_{n\to+\infty} nP\left(\frac{X}{b_n} > x\right) = x^{-\alpha}.$$

For large x and $\alpha > 0$. The equation above is a limit for a sequence of measures, recall that the sequence $\nu_n \to \nu$ vaguely iff $\forall f \in C_K^+(\mathbb{X})$ the integrals $\int\limits_{\mathbb{X}} f d\nu_n \to \int\limits_{\mathbb{X}} f d\nu$ converge. Given $\{X_i\}_1^n$, such that $S(x) \sim x^{-\alpha}$ is a regularly varying function; the previous limit asserts that,

$$\lim_{n\to+\infty} nP\left(\frac{X}{b_n} \in (x, +\infty]\right) = x^{-\alpha}.$$

Therefore, on the basis of Lebesgue-Stieljes Measure, a natural non-decreasing function assigned to the measure ν is g - where $K = \zeta^{-\alpha}$ for $\zeta > 0$.

$$g : \mathbb{R} \longrightarrow \mathbb{R}^{+},$$

$$g(x) = \begin{cases} 0 & \text{if } x \leq \zeta, \\ K - x^{-\alpha} & \text{if } x > \zeta. \end{cases}$$

$$\nu := \lambda_g : B_{\mathbb{R}} \longrightarrow \mathbb{R}^{+},$$

$$\nu((a, b]) \mapsto K - b^{-\alpha} - K + a^{-\alpha} = a^{-\alpha} - b^{-\alpha}.$$

K is a big enough constant, such that for all $x \in \mathbb{X}$ it is true that $x^{-\alpha} < K$, also $\nu = 0$ for any subset of $(-\infty, \zeta)$. These measures focus on large deviations, values far from zero. By the Lebesgue-Stieljes Measure Theorem, ν is a measure valid for large values of X. The ν_n are measures because they are the product of a positive number and a measure.

$$\nu_n : B_{\mathbb{R}} \longrightarrow \mathbb{R}^{+},$$

$$\nu_n((x, +\infty]) \mapsto nP\left(\frac{X}{b_n} \in (x, +\infty]\right).$$

Given that, $nS(b_n) \to x^{-\alpha}$ the next step will prove that $\nu_n \to \nu$ vaguely. Without loss of generality there exists a $\zeta \in \mathbb{R}^{+}$ such that for every $x > \zeta > 0$ the following happens:

$$\nu\left((\zeta, +\infty]\right) = \zeta^{-\alpha},$$

$$\nu\left((x, +\infty]\right) = x^{-\alpha}.$$

The previous limit result and the fact that $\nu_n > 0$ for $(\zeta, +\infty]$ leads to,

$$\frac{\nu_n\Big((x, +\infty]\Big)}{\nu_n\Big((\zeta, +\infty]\Big)} \to \frac{\nu\Big((x, +\infty]\Big)}{\nu\Big((\zeta, +\infty]\Big)}.$$

Where $\zeta^{-\alpha} > x^{-\alpha}$; thus, it is possible to define the probability measure for $x \in (\zeta, +\infty)$ as,

$$\begin{aligned} P_n\Big(X \in (x, +\infty]\Big) &= \frac{nP\Big(\frac{X}{b_n} > x\Big)}{nP\Big(\frac{X}{b_n} > \zeta\Big)}, \\ &= \frac{\nu_n\Big((x, +\infty]\Big)}{\nu_n\Big((\zeta, +\infty]\Big)}. \end{aligned}$$

Therefore $0 \leq P\Big(X \in (x, +\infty]\Big) \leq 1$ for random variables $X \in \mathbb{X} = (\zeta, +\infty]$. Furthermore, the previous results lead to,

$$\begin{aligned} \lim_{n\to+\infty} F_n(x) &= \lim_{n\to+\infty} P_n\Big(X \in (\zeta, x]\Big), \\ &= P\Big(X \in (\zeta, x]\Big), \\ &= F(x), \\ &= 1 - \frac{x^{-\alpha}}{\zeta^{-\alpha}}. \end{aligned}$$

Because the distribution functions converge point-wise at continuity points, the random variables converge weakly; therefore, the expected values of continuous non-negative functions with a compact support

converge; meaning,

$$E\big(f|P_n\big) = \int_{\mathbb{X}} f dF_n \to \int_{\mathbb{X}} f dF = E\ \big(f|P\big).$$

These implies that,

$$\begin{aligned}
&\frac{\int_{\mathbb{X}} f d\nu_n}{\int_{\mathbb{X}} d\nu_n} \to \frac{\int_{\mathbb{X}} f d\nu}{\int_{\mathbb{X}} d\nu}, \\
\Rightarrow &\frac{1}{\int_{\mathbb{X}} d\nu_n} \int_{\mathbb{X}} f d\nu_n \to \frac{1}{\int_{\mathbb{X}} d\nu} \int_{\mathbb{X}} f d\nu, \\
\Rightarrow &\frac{1}{\nu_n(\mathbb{X})} \int_{\mathbb{X}} f d\nu_n \to \frac{1}{\nu(\mathbb{X})} \int_{\mathbb{X}} f d\nu.
\end{aligned}$$

Since $\frac{1}{\nu_n(\mathbb{X})} \to \frac{1}{\nu(\mathbb{X})}$ then it must be true that

$$\int_{\mathbb{X}} f d\nu_n \to \int_{\mathbb{X}} f d\nu.$$

For $\mathbb{X} = (\zeta, +\infty]$, $\zeta > 0$ and $\forall f \in C_K^+(\mathbb{X})$; given that the measure is intended for large positive values this is valid for any $\zeta > 0$, such that large deviations of X belong to $(\zeta, +\infty]$. Therefore, it is always possible to find ζ such that:

$$\int_{\mathbb{X}} f d\nu_n \to \int_{\mathbb{X}} f d\nu,$$

$$\therefore \nu_n \xrightarrow{\text{v}} \nu.$$

This results proves that regular variation implies vague convergence of the measures ν_n [2]; this measure can be thought of as the measure of intensity for large deviations. Given the sequence $\{X_i\}_1^n$ of non-negative

unbounded independent identically distributed random variables with a Heavy-Tail (i.e., $S(x) \in RV_{-\alpha}$ for $\alpha > 0$) . It is possible to define the Point Process N_n as follows:

$$N_n(A_z) = \sum_{i=1}^{n} \delta_{X_i}(A_z).$$

Where,

$$A_z := \left\{ \# \frac{X_i}{b_n} > z : \frac{X_i}{b_n} \in (z, +\infty], \;\; i = 1, \cdots, n \right\}.$$

Meaning that X_i is in A_z iff $\frac{X_i}{b_n} > z$. Therefore, the Point Process counts the number of stabilised extreme events larger than the threshold z. It is a random measure that counts large deviations of X. For a fixed $z > 0$ and a number of observations n the region is determined, and the random measure behaves as the random variable:

$$N_n(A_z) \sim Binomial(n, p).$$

In which $p = S(xb_n)$. In this case the expected number of points in A_z is $np = \nu_n\Big((x, +\infty]\Big)$. This is an intensity measure for the given set, it is possible to test, by means of the Laplace Functional, different sets for valid thresholds $\{z\}$, such that $\nu_n \xrightarrow{\text{v}} \nu$. Recall that the Laplace Functional of the Poisson Point Process N is:

$$\Psi_N(f) = \exp\left(- \int_{\mathbb{X}} 1 - e^{-f} d\nu \right) [2].$$

The measure ν is the intensity measure of the Poisson Point Process. Now for the Point Process that counts the number of stabilised large

deviations the Laplace functional is:

$$\begin{aligned}\Psi_{N_n}(f) &= E\Big(\exp(-N_n(f))\Big),\\ &= E\Big(\exp\big(-\int_{\mathbb{X}} f dN_n\big)\Big),\\ &= E\Big(\exp\big(-\sum_{i=1}^{n} f\big)\Big),\\ &= E\Big(\prod_{i=1}^{n}\exp(-f(X_i))\Big).\end{aligned}$$

By independence of the X_i

$$\begin{aligned}&= \prod_{i=1}^{n} E\Big(\exp(-f(X_i))\Big),\\ &= \Big(E\Big(\exp(-f(X_i))\Big)\Big)^n.\end{aligned}$$

Recall that

$$\begin{aligned}\int_{\mathbb{X}} dP_n &= 1,\\ E\Big(\exp(-f(X_i))\Big) &= \Big(\int_{\mathbb{X}} e^{-f} n\frac{dP\big(X_n/b_n > x\big)}{n}\Big),\\ &= \Big(\int_{\mathbb{X}} e^{-f}\frac{d\nu_n}{n}\Big).\end{aligned}$$

Substitute these into the equations above to get:

$$\begin{aligned}
&= \left(-\int_{\mathbb{X}} e^{-f} dP_n \right)^n, \\
&= \left(-\int_{\mathbb{X}} e^{-f} \frac{n}{n} dP_n \right)^n, \\
&= \left(-\int_{\mathbb{X}} e^{-f} \frac{d\nu_n}{n} \right)^n, \\
\left(-\int_{\mathbb{X}} e^{-f} \frac{d\nu_n}{n} \right)^n &= \left(1 - \frac{1}{n} \int_{\mathbb{X}} 1 - e^{-f} d\nu_n \right)^n.
\end{aligned}$$

From the previous result $\nu_n \xrightarrow{\text{v}} \nu$; thus, for all non-negative continuous functions with compact support (i.e., $f \in C_K^+(\mathbb{X})$); results in that if $f \in C_K^+(\mathbb{X})$, then the function,

$$g = 1 - e^{-f}.$$

Is continuous because it is the composition of two continuous functions over $\mathbb{X} \subset \mathbb{R}^+$; also, $0 \leq g = 1 - e^{-f} \leq 1$; thus, the integrals of these continuous functions are bounded and by Vague Convergence,

$$\int_{\mathbb{X}} 1 - e^{-f} d\nu_n \to \int_{\mathbb{X}} 1 - e^{-f} d\nu.$$

Therefore, $\forall f \in C_K^+(\mathbb{X})$ it is true that

$$\left(1 - \frac{1}{n} \int_{\mathbb{X}} 1 - e^{-f} d\nu_n \right)^n \to \left(1 - \frac{1}{n} \int_{\mathbb{X}} 1 - e^{-f} d\nu \right)^n.$$

Then because continuous functions preserve sequential limits.

$$\begin{aligned}\lim_{n\to+\infty}\left(1-\frac{1}{n}\int_{\mathbb{X}}1-e^{-f}d\nu_n\right)^n &= \lim_{n\to+\infty}\left(1-\frac{1}{n}\int_{\mathbb{X}}1-e^{-f}d\nu\right)^n,\\ &= \exp\left(-\int_{\mathbb{X}}1-e^{-f}d\nu_n\right),\\ &= \Psi_N(f).\end{aligned}$$

Consequentially, by the Continuity Theorem the Laplace Functional of the Extreme Point Process converges at points of continuity to the Laplace Functional of a Poisson Point Process with intensity measure equal to the variation of the tail at infinity. Therefore, the Point Process converges weakly to the Poisson Point Process - $N_n \Rightarrow N$. This means that probabilistically as $n \to +\infty$ both processes behave similarly. Thus, to compute probabilities of one is analogous to computing probabilities of the other one.

This remarkable result means that for any valid threshold z this convergence still holds, given that the X_i are unbounded as long as $z \geq x_- = \inf\{x : F(x) > 0\}$, the Point Process will converge weakly to the Poisson Point Process because the Laplace functionals converge, in particular for $f = \delta_{X_i}(A_z)$.

$$P\big(N_n(\cdot) \leq k\big) \to P\big(N_n(\cdot) \leq k\big).$$

For all valid sets. To further illustrate this convergence for a fixed $x \geq z$, the random variable $N_n(A_z) \sim Binomial(n, p)$ is well defined; therefore, its probability mass function tends to the Poisson Point

Process probability mass function, the limit is as follows:

$$P\big(N_n(A_z)=k\big)\to P\big(N(A_z)=k\big),$$

This means that

$$\begin{aligned}
P\big(N_n(A_z)=k\big) &= \binom{n}{k}(S(b_nx))^k(1-S(b_nx))^{n-k},\\
&= \frac{n(n-1)\cdots(n-k+1)}{k!}(S(b_nx))^k(1-S(b_nx))^{n-k},\\
&= \frac{n^k}{k!}\big(1-\frac{1}{n}\big)\big(1-\frac{2}{n}\big)\cdots\\
&\big(1-\frac{k+1}{n}\big)(S(b_nx))^k(1-S(b_nx))^{n-k},\\
&= \frac{(nS(b_nx))^k}{x!}\big(1-\frac{nS(b_nx)}{n}\big)\cdots\\
&(1-S(b_nx))^{-k}\prod^{k+1}\big(1-\frac{i}{n}\big),
\end{aligned}$$

By the previous results this results in

$$\exp(-x^{-\alpha})\frac{x^{-k\alpha}}{k!}=P\big(N(A_z)=k\big).$$

Finally, for the sequence of $\{X_i\}$, such that there exists a sequence a_n that makes the X_i non-negative and a sequence b_n that stabilises large deviations; the events $\frac{X^{(n)}-a_n}{b_n}\leq z$ and $N_n(A_z)=0$ are equivalent because if the largest observation is smaller than z then all observations are smaller. (i.e., $X^{(1)}\leq X^{(2)}\leq\cdots\leq X^{(n)}\leq z$). Also, no such observations would fall in the region $(z,+\infty)$ meaning that $N_n(A_z)=0$, therefore, for $\frac{X_i^n-a_n}{b_n}=Y^{(n)}$ the probabilities tend as follows:

$$\begin{aligned}
P\big(Y^{(n)}\leq z\big) &= P\big(N_n(A_z)=0\big),\\
&\to P\big(N(A_z)=0\big),\\
&= \exp(-z^{-\alpha}).
\end{aligned}$$

Thus, the limit right tail distribution for all unbounded from above random variables with a Heavy-Tail (the tail is a regularly varying function) of stabilised extremes (i.e., the limit is not degenerate) is the Fréchet distribution.

$$P(Z \leq z) = \exp(-z^{-\alpha}).\ \alpha, z > 0.$$

■

This result is a consequence of the Power Law like behaviour the tail of X exhibits for large deviations. A natural question is: is it possible to approximate the conditional tail of Extremes? It turns out that this is possible and it is the basis of the Peaks Over Threshold method, first due to Pickands (1974) [1, 5, 14].

Theorem (A Heavy-Tail can be approximated with a Power Law [23, 1])**.** *For large enough X and a large enough threshold $u > 0$ the conditional distribution that $X > x|X > u$ for a Heavy-Tail random variable is a Pareto distribution.*

Proof. The proof below is a completion of the outline in [1, 23]. Consider the limit:

$$nP\left(\frac{X}{b_n}\right) \to x^{-\alpha}.$$

It can be extended further, since $\forall \epsilon > 0$ there exists a N_ϵ such that for big enough $x > 0$ if $n > N_\epsilon$ then:

$$\begin{aligned}
&\left|nP\left(\frac{X}{b_n} > x\right) - x^{-\alpha}\right| < \epsilon, \\
\Rightarrow &\left|n\left(P\left(\frac{X}{b_n} > x\right) - \frac{x^{-\alpha}}{n}\right)\right| < \epsilon, \\
\Rightarrow &n\left|P\left(\frac{X}{b_n} > x\right) - \frac{x^{-\alpha}}{n}\right| < \epsilon, \\
\Rightarrow &\left|P\left(\frac{X}{b_n} > x\right) - \frac{x^{-\alpha}}{n}\right| < \frac{\epsilon}{n} < \epsilon, \\
&\therefore P\left(X > b_n x\right) \to \frac{x^{-\alpha}}{n}.
\end{aligned}$$

Thus, the conditional distribution of events above the threshold u - $0 < u < x$ - where u stabilises large values, as b_n does, has the following distribution:

$$\begin{aligned}
P\left(X > b_n x | X > b_n u\right) &= \frac{P\left(\frac{X}{b_n} > x, X > u\right)}{P\left(\frac{X}{b_n} > u\right)}, \\
&= \frac{P\left(X > b_n x\right)}{P\left(X > b_n u\right)}, \\
&\to \frac{\frac{x^{-\alpha}}{n}}{\frac{u^{-\alpha}}{n}}, \\
&= \left(\frac{x}{u}\right)^{-\alpha}.
\end{aligned}$$

The tail of a Pareto distribution for values greater than or equal to the threshold u.

$$\therefore \lim_{n \to +\infty} P\left(\frac{X}{b_n} > x | X > u b_n\right) = \left(\frac{x}{u}\right)^{-\alpha}$$

■

This result states that if a random variable has a Heavy-Tail, then, the probabilities of exceeding a value above a threshold of interest behave asypmtotically as a Pareto Distribution; therefore, the tail of these distributions can be approximated with a Pareto Distribution [2, 25, 1]. The tail can also be measured as the excess Y above u, where $Y = X - u | X > u$; thus, the distribution of $y > 0$ is:

$$F(y) = 1 - \left(1 + \frac{y}{u}\right)^{-\alpha}.$$

Recall that stabilising constants are necessary to study the existence of these limits, thus, notation at the end of the section doesn't account for the translation and scaling constants. The following sections and chapters study these constants with more detail.

2.4 Extremal Types Theorem

There is a connection between Heavy-Tail models and Extreme Value Theory because both study the distributions of large deviations. The extremes Point Process characterisation is first due to Pickands (1974) [1]. The equation that gave birth to Extreme Value Theory is the limit distribution of the Maxima from a sample of independent identically

distributed random variables. It is not hard to derive this equation [23].

$$\begin{aligned} P(X^{(n)} = \max X_{i1}^{n} \leq x) &= P(X_1 \leq x, X_2 \leq x, \cdots, X_n \leq x), \\ &= P(X_1 \leq x)P(X_2 \leq x)\cdots P(X_n \leq x), \\ &= \prod_{i=1}^{n} F(x) \quad \textit{Identically distributed.} \\ &= F^n(x). \end{aligned}$$

Note that $0 \leq F(x) \leq 1$ for all $x \in \mathbb{R}$; thus,

$$\lim_{n \to +\infty} F^n(x) = \begin{cases} 0 & if \;\; F(x) < 1, \\ 1 & if \;\; F(x) = 1. \end{cases}$$

Therefore, it appears that the Maxima distribution is degenerate. But what if there are conditions that turn the degenerate limit into a non-degenerate limit. The natural question, which led to Extreme Value Theory, is ¿ Do sequences $(a_n) \in \mathbb{R}$, $(b_n > 0)$ exist, such that the stabilised Maxima converge to a non-degenerate limit?

$$F^n(b_n x + a_n) \to G(x).$$

Where G is non-degenerate. This equation is known as the max-stable postulate. The first thing to notice is that if G is max-stable, meaning that there exist sequences $(a_n) \in \mathbb{R}$, $(b_n > 0)$ such that for every $n \in \{2, 3, \cdots\}$, then

$$G^n(b_n x + a_n) = G(x).$$

Then it is clear that if G is max-stable, it solves the max stable postulate [1].

$$\lim_{n \to +\infty} G^n(b_n x + a_n) = \lim_{n \to +\infty} G(x) = G(x).$$

The next inquiry is: which distributions functions are max-stable and are these solutions unique. The Extremal Types Theorem answers these questions.

Theorem (Extremal Types Theorem (Fisher, Tippet Gnedenko 1948))**.** *Let* $\{X_i\}_1^n$ *be a sequence of independent identically distributed random variables, then if there exist sequences* $(a_n) \in \mathbb{R}$, $(b_n > 0)$, *such that the distribution of* $Y^{(n)} = \frac{X^{(n)} - a_n}{b_n}$ *converges to a non degenerate limit, then the limit is one of the following distribution functions. In essence if*

$$F^n(b_n x + a_n) \to G(x).$$

Then the distribution is one of the following three.

$$\begin{aligned}
&\textit{If} \quad nS(b_n x + a_n) \to x^{-\alpha} \ \textit{, then,} \\
&\quad G(x) = G_F(x) = \exp\big(-x^{-\alpha}\big)\chi_{\mathbb{R}^+}(x). \\
&\textit{If} \quad nS(b_n x + a_n) \to e^{-x} \ \textit{, then,} \\
&\quad G(x) = G_G(x) = \exp\big(-e^{-x}\big)\chi_{\mathbb{R}}(x). \\
&\textit{If} \quad nS(b_n x + a_n) \to (-x)^{\alpha} \ \textit{, then,} \\
&\quad G(x) = G_W(x) = \exp\big(-(-x)^{\alpha}\big)\chi_{\mathbb{R}^-}(x).
\end{aligned}$$

Where $\alpha > 0$ *is the tail index.*

The proof requires use of the max-stable theorem. The proof of the Max-stable Theorem is long, and departs from the main ideas discussed in this section; thus, to avoid interrupting the flow of ideas the proof of the Max Stable Theorem can be found in [21, 16].

Theorem (Max Stable Theorem [1, 23])**.** *A distribution* G *is max-stable if and only if it is one of the three Extreme Value distributions.*

The Max-Stable Theorem states that the only possible solutions to the max-stable postulate are the three Extreme Value distributions; nevertheless, it does not say if there exist random variables X, such that the stabilised extremes of X converge to one of the three Extreme Value distributions [1]. The max-Stable postulate does not provide information about the domain of attraction, it shows that the only solutions to the functional equation are the extreme value distributions. The Point Process characterisation of extremes clarifies the different types of convergence for extreme values (i.e., the limit tail for different random variables) [22]. The result is the convergence to types theorem [23, 25]. One possible proof, built on concepts and outlines in [1, 2, 21], follows.

Proof. Suppose that the random variable satisfies one of the three following limits.

$$\begin{aligned} &nS(b_n x + a_n) \to x^{-\alpha} \quad x > 0, \\ &nS(b_n x + a_n) \to e^{-x}, \\ &nS(b_n x + a_n) \to (-x)^{\alpha} \quad x < 0. \end{aligned}$$

The first case is the Fréchet distribution; thus, since extremes of Heavy-Tails converge to a Fréchet distribution (section 2.3 proves the result) the first case is that proof. Also, from section 3.3 follows that if the mean measure ν_n converges vaguely to the intensity measure ν, the Point Process converges weakly to the Poisson Point Process with intensity measure ν; furthermore, by hypothesis the mean measures converge, in essence:

$$\nu_n((x, +\infty]) \to \nu((x, +\infty]) \ ; \ \nu(\{x\}) = 0.$$

Then it follows from subsection 2.3 that,

$$\nu_n((x,+\infty]) \to \nu((x,+\infty]), \quad \nu(\{x\}) = 0,$$
$$\nu_n \to \nu \quad \textit{vaguely.}$$

The focus is on large deviations, thus, the lower bound is not a concern. Let $\zeta <= 0$ be a small value that is smaller than a possible upper bound for X (i.e., ζ is far from the right tail). Consider the function g_1 that resembles exponential tails,

$$g_1(x) = \begin{cases} 0 \quad if \;\; x \leq 0, \\ 1 - e^{-x} \quad if \;\; x > 0. \end{cases}$$

Then g_1 resembles a continuous distribution function because $0 \leq g_1 \leq 1$, also $g_1'(x) = e^{-x} > 0$; hence, $g_1' \geq 0$ so g_1 is monotone increasing. Then by the Lebesgue-Stieljes Measure theorem.

$$\nu = \overline{\lambda}_{g_1}.$$

Is a finite measure over $B_{\overline{\mathbb{R}}}$.

$$\begin{aligned} \lim_{x \uparrow +\infty} g_1(x) &= 1, \\ \lim_{x \downarrow -\infty} g_1(x) &= 0, \\ \nu((x,+\infty]) &= 1 - 1 + e^{-x}, \\ &= e^{-x}. \end{aligned}$$

The intensity measure of Exponential Tails. Therefore, the probability of large deviations is exponentially bounded. Define the Point Process

for Exponential Tail random variables.

$$N_n^1(A_z) = \{Number\ of\ \ \frac{X_i - a_n}{b_n} > z\},$$
$$= \sum_{i=1}^{n} \delta_{X_i}(A_z).$$

The event $N_n^1(A_z) = 0$ is equivalent to the event $P(\frac{X_i - a_n}{b_n} \leq z)$. A consequence of hypothesis (I) is that the Laplace functionals of the Point Processes converge for continuous non-negative bounded functions.

$$\Psi_{N_n^1}(f) \to \Psi_{N^1}(f).$$

Therefore, $N_n^1 \Rightarrow N^1$, where N^1 is a Poisson Point Process with intensity measure $\nu_1 = \overline{\lambda}_{g_1}$; hence,

$$P\left(\frac{X_i - a_n}{b_n} \leq z\right) = P(N_n^1(A_z) = 0),$$
$$\to P(N^1 = 0),$$
$$= \exp\big(-\nu_1(A_z)\big),$$
$$= \exp\big(-e^{-x}\big).$$

Therefore, the limit for Exponential tail random variables is a Gumbel distribution.
If $x_+ = \sup\{x : F(x) = 1\} = x_m < +\infty$, then X has an upper bound x_m (i.e., $P(X \leq x_m) = 1$). The extremes might centre close to x_m, or stay closer to the mean of X. For $M = |x_m|^\alpha$ and $\alpha > 0$; consider the

function g_2,

$$g_2(x) = \begin{cases} 0 \quad if \;\; x \leq -|x_m|, \\ M - (-x)^{\alpha} \quad if \;\; -|x_m| < x \leq 0, \\ M \quad if \;\; x > 0. \end{cases}$$

Then g_2 resembles a continuous distribution function because $0 \leq g_2 \leq M < +\infty$, also $g_2'(x) = \alpha(-x)^{\alpha-1}\chi_{(-|x_m|,0)}(x) > 0$,and the other two derivatives are zero. Hence, $g_2' \geq 0$ so g_2 is monotone increasing. Then, by the Lebesgue-Stieljes Measure theorem.

$$\nu = \overline{\lambda}_{g_2}.$$

Is a finite measure over $B_{\overline{\mathbb{R}}}$.

$$\lim_{x \uparrow +\infty} g_2(x) = M,$$
$$\lim_{x \downarrow -\infty} g_2(x) = 0.$$

The interesting values are $-|x_m| < x \leq 0$.

$$\begin{aligned} \nu((x, +\infty]) &= M - M + (-x)^{\alpha}, \\ &= (-x)^{\alpha}. \end{aligned}$$

The intensity measure of bounded from above random variables that have a Power Law like tail.
In a similar fashion define the Point Process that counts extremes of random variables that are bounded from above, but do not have an exponentially bounded tail.

The Y_i are independent identically distributed random variables with an upper bound and a tail that is not exponentially bounded.

$$\begin{aligned} N_n^2(A_z) &= \{Number\ of\ \ \frac{Y_i - a_n}{b_n} > z\}, \\ &= \sum_{i=1}^{n} \delta_{X_i}(A_z). \end{aligned}$$

The event $N_n^2(A_z) = 0$ is equivalent to the event $P(\frac{Y_i - a_n}{b_n} \leq z)$. A consequence of hypothesis (II) is that the Laplace functionals of the Point processes converge for bounded non-negative continuous functions.

$$\Psi_{N_n^2}(f) \to \Psi_{N^2}(f).$$

Therefore, $N_n^2 \Rightarrow N^2$, where N^2 is a Poisson Point Process with intensity measure $\nu_2 = \overline{\lambda}_{g2}$; hence,

$$\begin{aligned} P\left(\frac{Y_i - a_n}{b_n} \leq z\right) &= P(N_n^2(A_z) = 0), \\ &\to P(N^2 = 0), \\ &= \exp\big(-\nu_2(A_z)\big), \\ &= \exp\big(-(-x)^{\alpha}\big). \end{aligned}$$

For $x < 0$, since $X \leq x_m < +\infty$ it is always possible to make $X \leq 0$. Thus the limit for random variables with an upper bound and a tail that is not exponentially bounded is the Inverse-Weibull distribution.

■

Notice that the even though the random variable are bounded, the tail decreases slower than an exponential tail; thus, extremes will cluster near the upper bound because the Power law behaviour near

the upper bound x_m results in large deviations having higher probabilities of occurrence. The Continuity Theorem does not only imply convergence of random variables [5, 21], it also implies that these limits are unique.

Therefore, the only possible non-degenerate limit for Heavy-Tail random variables is the Fréchet distribution; additionally, the only possible limits for random variables with an Exponential Tail is the Gumbel family; furthermore, the only possible limit for random variables with an upper bound but with a Power Law like tail is the Inverse-Weibull family.

This work focuses on Heavy-tails, thus, it is important to compare it with the limit for unbounded random variables with an Exponential Tail. Furthermore, since these are the only possible solutions to the Max-Stable postulate the only Extreme Values with a non-degenerate limit are random variables with a Heavy-Tail, or, random variables with an Exponential Tail, or, random variables with an upper bound and a tail that is not Exponential.

Finally, notice that the intensity measure of Heavy-Tail random variables is a Power Law and that the intensity measure of Exponential Tails is an exponential function. The exponent of the Fréchet distribution is a Power Law and the exponent of the Gumbel distribution is an Exponential Tail.

Chapter 3

Statistical Modelling for Heavy-Tail Models

3.1 Extreme Value Models

Because of the Extremal Types Theorem [1]; the distribution for Maxima of Heavy-Tail data can be approximated with the Fréchet distribution:

$$G_F(x) = \exp\left(-\left(\frac{x-\gamma}{\theta}\right)^{-\alpha}\right),$$

$$F^n(b_n x + a_n) \to G_F(x).$$

With $\gamma = 0$ and $\theta = 1$, this random variable stabilises because the sequences, (a_n) and (b_n) relocate and scale $X^{(n)} = \max\{X_i\}$; therefore, it is valid to study if there is another Fréchet distribution that

incorporates the scale and location parameters.

$$P(X^{(n)} \leq x) \to G_F\left(\frac{x - a_n}{b_n}\right),$$
$$\to \tilde{G}_F(x).$$

Where $\tilde{G}_F$ is a Fréchet distribution with $\gamma \in \mathbb{R}$ and $\theta \in (0, +\infty)$. These location and scale parameters stabilise extremes. Therefore, it is necessary to estimate γ, θ to fully account for uncertainty in approximating the Heavy-Tail [23, 22]. To clarify this consider the following limit.

Proposition. *Let $X \sim$ Fréchet(α, θ, γ); if $a_n = \gamma$ and $b_n = G_F^{-1}(1 - \frac{1}{n})$, then $\frac{X^{(n)} - a_n}{b_n}$ converges weakly to a random variable with a standard Fréchet distribution.*

Proof. Let $b_n = G_F^{-1}(1 - \frac{1}{n}) - \gamma$ and $a_n = \gamma$, then for $q \in (0, 1)$,

$$\begin{aligned} q &= \exp\left(-\left(\frac{x - \gamma}{\theta}\right)^{-\alpha}\right), \\ &\iff -\ln(q) = \left(\frac{x - \gamma}{\theta}\right)^{-\alpha}, \\ &\iff (-\ln(q))^{-\frac{1}{\alpha}} = \frac{x - \gamma}{\theta}, \\ &\iff \theta(-\ln(q))^{-\frac{1}{\alpha}} + \gamma = x. \end{aligned}$$

Hence, $b_n = \theta(-\ln(1-n^{-1}))^{-\frac{1}{\alpha}}$, $a_n = \gamma$.

$$\begin{aligned}
&\lim_{n\to+\infty} G^n(b_n x + a_n) = \lim_{n\to+\infty} \exp\left(- n\left(\frac{b_n x + a_n - \gamma}{\theta}\right)^{-\alpha}\right),\\
&= \lim_{n\to+\infty} \exp\left(- n\left(\frac{\theta(-\ln(1-n^{-1}))^{-\frac{1}{\alpha}} x + \gamma - \gamma}{\theta}\right)^{-\alpha}\right),\\
&= \lim_{n\to+\infty} \exp\left(- n\left((-\ln(1-n^{-1}))^{-\frac{1}{\alpha}} x\right)^{-\alpha}\right),\\
&= \lim_{n\to+\infty} \exp\left(n\ln(1-n^{-1})(x)^{-\alpha}\right).
\end{aligned}$$

Therefore it is necessary to solve.

$$\begin{aligned}
&\lim_{n\to+\infty} n\ln(1-n^{-1}) = \lim_{n\to+\infty} \frac{\ln(1-n^{-1})}{n^{-1}},\\
&\textit{Indeterminate kind, not a number } "\frac{0}{0}".
\end{aligned}$$

If the limit exists it is the same limit as for continuous y.

$$\lim_{y\to+\infty} \frac{\ln(1-y^{-1})}{y^{-1}}.$$

Using L'hopital's rule.

$$\begin{aligned}
&\frac{d}{dy}\ln(1-y^{-1}) = \frac{1}{1-y^{-1}} y^{-2},\\
&\frac{d}{dy} y^{-1} = -y^{-2};\\
&\lim_{y\to+\infty} \frac{\frac{1}{1-y^{-1}} y^{-2}}{-y^{-2}} = \lim_{y\to+\infty} -\frac{y^{-2}}{y^{-2}} \frac{1}{1-y^{-1}},\\
&= \lim_{y\to+\infty} -\frac{1}{1-y^{-1}},\\
&= -1.
\end{aligned}$$

The exponential is a continuous function; thus, it preserves limits.

$$\therefore \lim_{n\to+\infty} \exp\big(n\ln(1-n^{-1})(x)^{-\alpha}\big) = \exp\big(-x^{-\alpha}\big).$$

Therefore,

$$\lim_{n\to+\infty} G^n(b_n x + a_n) = \exp\big(-x^{-\alpha}\big).$$

■

This max-stable property of Extreme Value distribution allows one to estimate the sequences (b_n), (a_n) as location and scale parameters. The scale and location parameters probably aren't the sequences' limits; nonetheless, they serve the same purpose of stabilising extremes. Hence, they enable statistical modelling of Extreme Values [1, 23]. It's interesting to estimate these quantities because they provide important information about extreme value quantiles and functions.

If the extremes come from a distribution with an exponential tail, then the limit is the Gumbel distribution [16],

$$F^n(b_n x + a_n) \to G_G(x).$$

$$G_G(x) = \exp(-e^{-x}).$$

And by an analogous argument for the location and scale parameters,

$$P(X^{(n)} \le x) \approx \exp\left(-\exp\left(\frac{x-\gamma}{\theta}\right)\right).$$

This distribution has no shape parameter for the tail because it decays exponentially; instead, the scale parameter θ is the parameter that affects decay. If the random variables are bounded with a Power Law

like tail, then the limit distribution of extremes is the Inverse-Weibull distribution,

$$G_W(x) = \exp\Big(- \Big(- \frac{x-\gamma}{\theta} \Big)^{\alpha} \Big).$$

This distribution has most mass close to the upper bound γ; nevertheless, the shape parameter α indicates that events far from the upper bound have a higher probability than they would have if the left tail or right tail decay exponentially. This work focuses on unbounded random variables with Heavy-Tails; thus, the Inverse-Weibull distribution will not be of much use.

The Extremal Types Theorem is the basis to approximate the distribution of Maxima and extreme values. The three distributions can be grouped as the Generalised Extreme Value Distribution (GEV), whose support is $\mathbb{X} := \{x : 1 + \xi x > 0\}$ for $\xi \in \mathbb{R}$ [12].

$$\begin{aligned} &\mathrm{GEV}(x) = \exp\big(- (1+\xi x)^{-\frac{1}{\xi}} \big) \quad \textit{if } \xi \neq 0, \\ &\mathrm{GEV}(x) = \exp(-e^{-x}) \quad \textit{if } \xi = 0. \end{aligned}$$

Remark. *If $\xi > 0$ then the random variables are unbounded and the GEV becomes a Fréchet distribution with tail index $\frac{1}{\xi} = \alpha > 0$ and support $\mathbb{X} = (-\frac{1}{\xi}, \infty)$.*

$$GEV(x) = \exp\big(- (1+\xi x)^{-\alpha} \big).$$

If $\xi < 0$ then the random variables are bounded from above and the GEV becomes an Inverse-Weibull distribution with tail index $-\frac{1}{\xi} = \alpha > 0$ and support $\mathbb{X} = (-\infty, -\frac{1}{\xi})$; hence, extremes tend to concentrate close to the upper bound.

$$GEV(x) = \exp\big(- (1+\xi x)^{\alpha} \big).$$

If $\xi = 0$ then the random variables have an exponential tail and the GEV becomes a Gumbel distribution with no tail index and support $\mathbb{X} = (-\infty, +\infty)$; hence, the moment generating function exists and extremes should be close to the expected values of the limit Gumbel distribution.

$$GEV(x) = \exp\big(-e^{-x}\big).$$

[1, 21, 22].

Proposition. *The definition makes the distribution continuous for $\xi \in \mathbb{R}$. In essence:*

$$\lim_{\xi \to 0} GEV(x) = \exp(-e^{-x}).$$

Proof. It is necessary to calculate the two directional limits.

$$\lim_{\xi \to 0^+} G(x) = \lim_{\xi \to 0^+} \exp\big(-(1+\xi x)^{-\frac{1}{\xi}}\big).$$

Look at the limit inside the exponential.

$$\begin{aligned} &\xi \to 0^+ \Rightarrow \exists n \in \mathbb{N} \ \ s.t. \\ &\frac{1}{n+1} < \xi < \frac{1}{n} \ \ ; \ \ n < \frac{1}{\xi} < n+1. \end{aligned}$$

Thus,

$$\begin{aligned} \left(1 + \frac{x}{n+1}\right)^n \leq (1+\xi x)^{\frac{1}{\xi}} \leq \left(1 + \frac{x}{n}\right)^{n+1}, \\ \therefore \lim_{\xi \to +\infty} (1+\xi x)^{\frac{1}{\xi}} = e^x. \end{aligned}$$

By continuity and the fact that $e^x > 0$, this results in,

$$\lim_{\xi \to 0^+} G(x) = \exp\left(\big(\lim_{\xi \to 0^+} (1+\xi x)^{\frac{1}{\xi}}\big)^{-1}\right) = \exp(-e^{-x}).$$

Similarly if $\xi \to 0^-$, then , there exists a $n \in \mathbb{N}$, such that

$$n < -\frac{1}{\xi} < n+1 \;\; ; \;\; -\frac{1}{n} < \xi < -\frac{1}{n+1},$$
$$\left(1 - \frac{x}{n}\right)^{n+1} \leq (1+\xi x)^{-\frac{1}{\xi}} \leq \left(1 - \frac{x}{n+1}\right)^{n},$$
$$\therefore \lim_{\xi \to 0^-} (1+\xi x)^{-\frac{1}{\xi}} = e^{-x},$$
$$\therefore \lim_{\xi \to 0^-} G(x) = \exp(-e^{-x}).$$

Therefore, $\lim_{\xi \to 0} G(x) = \exp(-e^{-x})$. ■

As the tails thin, becomes less heavy, both extreme value distributions converge to the Gumbel distribution (Exponential tail) as a limit of the other types of decay [9, 25]. The bigger ξ is the smaller α is. The parameter α is the tail index; it controls how fast the tail decays. The tail distribution can be studied by calculating the following limit - that gives rise to the Generalised Pareto Distribution [23, 26]. Below is an outline of the proof in [1]. Suppose that for a distribution function F, threshold u, stabilising constant ξ and random variable X the extreme value cumulative distribution function F has the following limit as $n \to +\infty$.

$$F^n(x) \to \exp(-(1+\xi x)^{-\frac{1}{\xi}}).$$

Therefore, the limit exists because of the Extremal Types Theorem and by continuity of $\ln(x)$,

$$-n \ln(F(x)) \to (1+\xi x)^{-\frac{1}{\xi}}.$$

A Taylor series expansion for $s \to 0^+$ *exhibits that,*

$$s \to -\ln(1-s) \;\; ; \;\; s \to s \to 0^+,$$
$$\textit{For large } x \; F(x) = 1 - S(x) \to 1 \;\; \textit{as} \;\; S(x) \to 0^+.$$

Thus, substituting these equations,

$$-\ln(F(x)) \to \frac{(1+\xi x)^{-\frac{1}{\xi}}}{n},$$
$$-\ln(F(x)) \to 1 - F(x) \to S(x),$$
$$\therefore S(x) \to \frac{(1+\xi x)^{-\frac{1}{\xi}}}{n}.$$

For a sufficient sample size and a big enough threshold u,

$$P(X > x | X > u) = \frac{P(X > x)}{P(X > u)},$$
$$= \frac{S(x)}{S(u)} \to \frac{\frac{(1+\xi x)^{-\frac{1}{\xi}}}{n}}{\frac{(1+\xi u)^{-\frac{1}{\xi}}}{n}},$$
$$= \left(\frac{(1+\xi x)}{(1+\xi u)}\right)^{-\frac{1}{\xi}}.$$

Set $\tilde{\xi} = \frac{\xi}{1+\xi u}$ and substitute it into the equation to get,

$$\frac{1+\xi(x-u+u)}{1+\xi u} = 1 + \frac{\xi(x-u)}{1+\xi u} = 1 + \tilde{\xi}(x-u).$$

Therefore the approximate Pareto Tail becomes,

$$\text{GPD}(x) = \left(1 + \tilde{\xi}(x-u)\right)^{-\frac{1}{\xi}}.$$

This is the kernel of the Generalised Pareto distribution. If $\xi = 0$, then extremes have an exponential tail and the proper approximation is:

$$\mathrm{GPD}(x) = e^{-x}.$$

In the case of the Fréchet distribution the Generalised Pareto distribution is a Pareto tail,

$$\begin{aligned} P(X > x | X > u) &\to \left(\frac{\frac{x-\gamma}{\theta}}{\frac{u-\gamma}{\theta}} \right)^{-\alpha}, \\ &= \left(\frac{x-\gamma}{u-\gamma} \right)^{-\alpha}, \\ &= \left(1 + \frac{x-u}{u-\gamma} \right)^{-\alpha}. \end{aligned}$$

Therefore if X has a Heavy-Tail, the distribution of exceedances above the threshold u is approximately a Pareto distribution [9, 6]. On the other hand, if X has an exponential tail.

$$\begin{aligned} P(X > x | X > u) &\to \frac{\exp(-\left(\frac{x-\gamma}{\theta}\right)}{-\left(\frac{u-\gamma}{\theta}\right)}, \\ &= \exp\big(-\frac{1}{\theta}(x - \gamma - u + \gamma)\big), \\ &= \exp\big(-\frac{1}{\theta}(x - u)\big). \end{aligned}$$

The loss of the parameter γ is the loss of memory property of exponential random variables, since X has an exponential tail for large enough u the tail can be approximated by an exponential distribution. These statistical models - based on the Limit results from the previous section - are the basic tools employed to study Extreme Values [23, 14, 10].

3.2 Poisson Point Process Likelihood Function Construction

The Poisson Point Process Likelihood has the advantage that it incorporates the rate at which the data exceed the threshold u and Pareto like behaviour of extreme observations; in contrast, the Fréchet and GEV likelihoods can only be fitted with independent block Maxima that dispose of other valuable extreme observations; while the Pareto Tail method, oppositely, ignores the rate at which an observation might exceed the threshold u [25].

It should also be noted that the parameters of the Poisson Point Process likelihood correspond to the parameters of the extreme value distribution and hold for the Pareto tail approximation. For the fixed threshold u (select u based on previous information experience and/or necessity, as well as by use of statistical methods present in following sections), define the region $A_u := \{\#\frac{X_i - a_n}{b_n} \geq u\}$, in essence, it is the set that counts the number of stabilised observations that are larger than u in n_p periods; the periods could be years, days, months or any other time measurement [1].

The larger the period is the more extreme large deviations are because more LARGE deviations can occur (i.e., the 10 year Maxima has to be at least as big as the annual Maxima). On the basis of the Fréchet limit for extremes with a Heavy-Tail. It is possible to model the data as a Poisson Point Process with parameters $\underline{\theta} = (\alpha > 0, \gamma \in \mathbb{R}, \theta > 0)$ and adjusted intensity measure [23, 2]:

$$\nu((x, +\infty]) = \left(\frac{x - \gamma}{\theta}\right)^{-\alpha}.$$

The likelihood function is the joint probability of the observed sample seen as a function of the parameters, where $N(A_u) \sim Poisson(\nu(A_u))$, and the joint probability is the probability of observing those extreme observations at that time and in that point sample. $N(A_u)$ is the number of extreme observations above u. Therefore $N(A_u) = m$ means that there are m extremes above u [3, 19].

Definition (Extremes Sample [1]). *Let $\underline{X_{(n)}}$ be a random sample, the Extremes above the threshold u random sample is the sub-sample where all observations are at least as large as u. $\underline{X_m^{(u)}}$.*

Furthermore, the block Maxima random sample is the sample one gets by selecting the Maximum observation every k observations. The block (period) has size $k \leq n$. $\underline{X^{(n)}_{(m=\frac{n}{k})}}$.

The likelihood construction that follows is a completion of outlines in [1, 6]. The observed region can be separated as follows for period size p with a total of n_p periods that serve as the extreme per period,

$$B = A_u - \bigcup_{i=1}^{N(A_u)} I_i.$$

The intervals are the singleton observations.

$$I_i = (X_i - \frac{1}{n}, X_i],$$
$$\therefore \bigcap_{n=1}^{\infty} (X_i - \frac{1}{n}, X_i] = \{X_i\}.$$

Unless $X_i = X_j$, the intervals are disjoint, however, if $X_i = X_j$ it is possible to remove one from the sample; in the probabilistically impossible case that the whole sample is the same value fitting a stochastic model would make no sense. Thus, it is possible to proceed

with a sample of distinct observations, therefore, the intervals are disjoint. The event $N(A_u) = m$ is equivalent to the event $A_u - \cup_{i=1}^{N(A_u)} I_i = 0, I_i = 1$ for all $i \in \{1, \cdots, m\}$. These regions are disjoint, hence, they are independent Poisson random variables [6].

$$P(N(A_u) = m) = P\bigg(A_u - \bigcup_{i=1}^{N(A_u)} I_i = 0, I_1 = 1, \cdots, I_m = 1\bigg),$$
$$= P\bigg(A_u - \bigcup_{i=1}^{N(A_u)} I_i = 0\bigg) \prod_{i=1}^{m} P(I_i = 1).$$

The observations in different periods are independent, therefore, the probability of the same event in different periods (no observation above u) is the product of the probability for one period. Since the X_i are identically distributed, the probability for each period is the same. Putting this into the equation (n_p is the period size: number of samples of size p that constitute the whole sample):

$$P\bigg(A_u - \bigcup_{i=1}^{N(A_u)} I_i = 0\bigg) = \big(\exp(-\nu(A_u - \bigcup_{i=1}^{N(A_u)} I_i)\big)^{n_p},$$
$$= \exp\bigg(-n_p\nu(A_u - \bigcup_{i=1}^{N(A_u)} I_i)\bigg).$$

Where,

$$\begin{aligned}\nu(A_u - \bigcup_{i=1}^{N(A_u)} I_i) &= \nu(A_u) - \sum \nu(\bigcap_{n=1}^{\infty} I_i),\\ &= \nu(A_u) - \sum \nu(\{X_i\}),\\ &= \nu(A_u) - 0,\\ &= (\frac{u-\gamma}{\theta})^{-\alpha},\\ \therefore P\Big(A_u - \bigcup_{i=1}^{N(A_u)} I_i = 0\Big) &= \exp\Big(-n_p\Big(\frac{u-\gamma}{\theta}\Big)^{-\alpha}\Big).\end{aligned}$$

Exponential decay is much faster than linear decay (i.e, $e^{-n} << \frac{1}{n}$ as n gets bigger.), thus, the contribution from the exponential term is much less than the Pareto tail term contribution.

$$\begin{aligned}P\Big(\bigcap_{n=1}^{\infty} I_i = 1\Big) &= \exp(-\nu(\bigcap_{n=1}^{\infty} I_i))(\nu(\bigcap_{n=1}^{\infty} I_i)),\\ &\to \exp(0)\nu(\bigcap_{n=1}^{\infty} I_i).\end{aligned}$$

The measure of the singleton is zero, nonetheless, by the Radon-Nykodym Theorem [20, 5, 21]; $\nu = \overline{\lambda_g}$ can be computed as an integral,

$$\nu(I_i) = \int_{x_i - \frac{1}{n}}^{x_i} g'(x)dx.$$

The rate of change of the intensity as the interval gets shorter (likelihood clustered at X_i) is,

$$\frac{g(x_i) - g(x_i - \frac{1}{n})}{1/n} \to g'(x).$$

As $n \to \infty$ this contribution goes to the derivative of g. Therefore, the likelihood of observation X_i is the value $g'(x)$.

$$\begin{aligned} g'(x) &= \frac{d}{dx} g(x), \\ &= \frac{d}{dx} \; K - \big(\frac{x-\gamma}{\theta}\big)^{-\alpha}, \\ &= \frac{\alpha}{\theta}\left(\frac{x-\gamma}{\theta}\right)^{-(\alpha+1)}. \end{aligned}$$

A Pareto tail for Heavy-Tail extremes, once the observation is fixed $X_i = x_i > u$. Combining these two results, the likelihood function is:

$$L(\underline{X^{(u)}_{(m)}}|\underline{\theta}) = \exp\left(- n_p \left(\frac{u-\gamma}{\theta}\right)^{-\alpha}\right) \prod_{i=1}^{N(A_u)=m} \frac{\alpha}{\theta}\left(\frac{x_i-\gamma}{\theta}\right)^{-(\alpha+1)}.$$

This likelihood has the information provided by the peaks over threshold method, furthermore, it incorporates the rate at which observations are smaller than u. Therefore, it employs more observations than fitting block Maxima; however, it also accounts for observations that were not over the threshold [1, 23]. Recall that it's also possible to study extreme values via the Pareto tail for values larger than the threshold u, and/or via the Generalised Extreme Value distribution; if one wishes to use those models, the likelihood functions would be as follows. The likelihood function of the Pareto Tail above u for $Y = X - u$ is,

$$L(\underline{Y^{(u)}_{(m)}}) = \prod_{i=1}^{N(A_u)=m} \frac{\alpha}{u}\left(1 + \frac{y_i}{u}\right)^{-\alpha-1}.$$

If it is desirable to fit the Generalised Pareto distribution the likelihood

function , where $\alpha = \frac{1}{\xi}$ $\xi \neq 0$, is:

$$L(\underline{X_{(m)}^{(u)}}|\underline{\theta}) = \frac{1}{\theta^m} \prod_{i=1}^{N(A_u)=m} \left(1 + \xi \frac{y_i - \gamma}{\theta}\right)^{-\frac{1}{\xi}-1}.$$

These likelihoods share a similar form (Pareto kernels), thus, inferences made with each of them should be similar; nonetheless, the Poisson Point Process likelihood uses more information and depends on the threshold, parameters and number of periods observed, in fact only the Point Process likelihood includes the parameter n_p because it's the only one that measures each extreme as a step from the Poisson Point Process with period size p [1, 25, 10]. This thesis proposes a Bayesian Computational algorithm to fit the Poisson Point Process likelihood.

3.3 Relationships Between Extreme Value Models

The three Extreme Value Models are related through transformations, also, some transformations do not affect some parameters of each distribution. These relationships are useful to understand the differences between the models and what techniques might be appropriate to study Extreme Values [25, 6].

Proposition. *Let* $X \sim Frechet(\alpha)$ *then,*

$$Y = \ln(X) \sim Gumbel(\alpha = \frac{1}{\theta}),$$
$$Z = -X^{-1} \sim \textit{Inv-Weibull}(\alpha).$$

Proof. First, $X > 0$; thus, $\ln(X) \in \mathbb{R}$. The logarithm is a continuous monotone non-decreasing transformation. Hence,

$$\begin{aligned} F_Y(y) &= P(Y \leq y), \\ &= P(\ln(X) \leq y), \\ &= P(X \leq e^y), \\ &= F_X(e^y), \\ &= \exp(-(e^y)^{-\alpha}), \\ &= \exp(-e^{-\alpha x}), \\ &= \exp(-e^{\frac{x}{\theta}}). \end{aligned}$$

A Gumbel distribution that is valid for all values in the transformed support $y \in \mathbb{R}$.
The other transformation results in. $X > 0$; thus, $-X^{-1} < 0$ note that this transformation bounds the random variable.

$$\begin{aligned} F_Z(z) &= P(Z \leq z), \\ &= P(-X^{-1} \leq z), \\ &= P(X^{-1} \geq -z), \\ &= P(X \leq -z^{-1}), \\ &= F_X(-z^{-1}), \\ &= \exp(-(-z^{-1})^{-\alpha}), \\ &= \exp(-(-z)^{\alpha}). \end{aligned}$$

An Inverse-Weibull distribution. The support is $z \in (-\infty, 0)$. ■

The quantile functions, for $q \in (0,1)$, are.

$$\begin{aligned} G_F(x) = q &\iff \theta(-\ln(q))^{-\frac{1}{\alpha}} + \gamma = x. \\ G_G(x) = q &\iff -\theta \ln(-\ln(q)) + \gamma = x. \\ GEV(x) = q &\iff \theta\left(\frac{-\ln(q)+1}{\xi}\right)^{-\xi} + \gamma = x. \end{aligned}$$

If $\xi = 0$ the quantile function is the quantile (percentile) function of the Gumbel distribution. An important quantile is the following.

Definition (m return level)**.** *The $m \in \mathbb{N}$ return level of the random variable X is the number x, such that X exceeds x with a probability of $\frac{1}{m}$. In essence:*

$$G(x) = 1 - \frac{1}{m} = P(X \leq x).$$

Ideally, X should be larger than x every m observations.

An interesting property is:

Proposition. *If X has a Pareto Tail above a threshold u, then the log-transform makes the tail of X exponential [10, 12].*

Proof. For simplicity let $Y = \ln(X)$ and $X \sim Pareto(\alpha, u)$, where $u = 1$.

$$\begin{aligned} P(Y \leq y) &= P(\ln(X) \leq y), \\ &= P(X \leq e^y), \\ &= F_X(e^y), \\ &= 1 - \left(e^y\right)^{-\alpha}), \\ &= 1 - e^{-\alpha y}. \end{aligned}$$

The distribution of an exponential random variable. ■

This section shows that the logarithm transform thins tails, nevertheless, this transformation should be done with caution. For starters it can only be done to non-negative random variables; furthermore, the tail index is lost when one log transforms the data. Therefore, one must be cautious when log transforming data, even though, this transformation stabilises the data it takes away the information about the Heavy-Tail. On that basis most analysis separate Large deviations from the rest of the sample and study the tail with Heavy-Tail models [24, 6, 14]. Otherwise, it is possible to lose valuable information when estimating quantities and functions, such as the m return level.

3.4 Exploratory Data Analysis for Heavy-Tail Data

Heavy-Tail data makes itself apparent with some Exploratory Data Analysis (EDA) techniques. The first thing to do is to study if the sample includes large deviations. Evidence of a Heavy-Tail could be that the mean is much bigger than the median, also, the ratio of Maxima divided by the sum of the sample could show that there is a LARGE deviation if it is not close to zero [9]. These are one of many ways to study the presence of a Heavy-Tail.

Definition (Mean Life Plot [14, 1])**.** *The Mean Life Plot quantifies the average value of excess for different thresholds. Where m_u is the number of observations at least as large as u.*

$$MLP(u) = \frac{1}{m_u} x_i - u\chi_{(u,+\infty)}(x_i).$$

The tail index should remain constant after a big enough threshold is surpassed, a range of plausible thresholds is the region where the Mean Life Plot grows linearly [1, 12]. Since the log transform of Heavy-Tail data is an exponential random variable a QQ-plot of the log data and an exponential distribution should be close to linear functions, for thresholds above a big enough threshold [15, 7]. The QQ-plot compares theoretical quantiles with quantiles of the empirical distribution. If the plot looks linear, then the empirical distribution is reasonably adjusted by the proposed model.

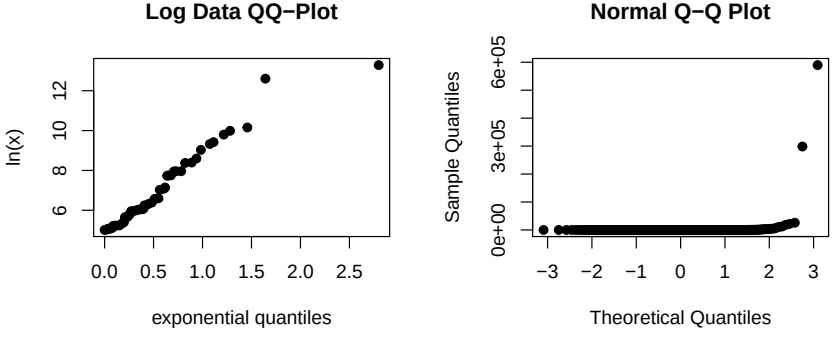

Figure 3.1. QQ-Plot of log data vs Normal QQ-Plot of the data.

This figure shows that a normal distribution has no correspondence with empirical quantiles that evidence a Heavy-Tail, but the plot on the left is close to a linear function, this is evidence that the logarithm of Heavy-Tail data should resemble a linear plot when compared with an exponential distribution's quantiles.

Definition (Empirical Distribution [21])**.** *The Empirical Distribution*

counts the number of observations smaller than x.

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \chi_{-(\infty,x]}(X_i).$$

This distribution gives a rough estimate of possible quantiles. The 75% percentile or any bigger percentile can be used as a threshold. It is important to remember that Heavy-Tail analysis requires extrapolation from the sample. A consequence of this is that the Empirical Distribution will probably be very different for independent samples because a sample may or may not contain a LARGE deviation; thus, the Empirical Distribution should be used only to look for a threshold not to fit the data or measure Goodness of Fit. The variability between distinct samples is much higher than it is for Exponential Tail data [9, 4]. This inherent volatility in Heavy-Tail samples makes bootstrapping challenging because the sample one has might be missing valuable information, thus, the bootstraps will not resemble future sampling. In essence, every Heavy-Tail sample tends to differ from other samples much more than Exponential Tail samples do [2].

These Exploratory Data Analysis (EDA) techniques help diagnose if the data have a Heavy-Tail. The presence of one LARGE deviation is evidence of a Heavy-Tail, furthermore, plenty LARGE deviations are decisive evidence of a Heavy-Tail. The other techniques (Mean Life Plot, QQ-plot, Empirical percentiles) help the analyst look for a possible threshold [19, 14, 10].

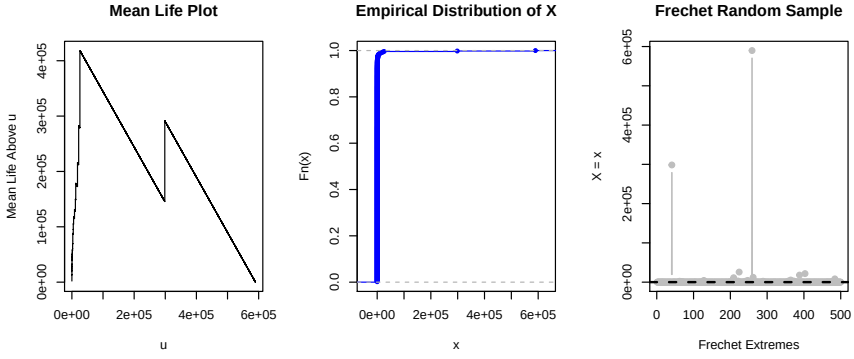

Figure 3.2. Exploratory Data Analysis example.

Chapter 4

Bayesian Inference for Heavy-Tail Models

4.1 Methodology Proposal

Heavy-Tail analysis is interesting because the models fitted are asymptotic. By definition extremes are rare, thus, study of Extreme Values requires the ability to extrapolate information from small sample sizes. Furthermore, the remarkable Convergence to Types Theorem shows that if X is a Heavy-Tail random variable with a non-degenerate limit, then the limit is a Fréchet distribution, furthermore, if X has an Exponential Tail (i.e., the tail of X is exponentially bounded) and a non-degenerate limit, then its limit Extreme Value distribution is a Gumbel distribution [22, 2, 19]. These results highlight the scope both the Fréchet and Gumbel families have, to study extreme values its necessary to study these two families, as well as the Inverse-Weibull family for bounded random variables.

Proper assessment of the tail of X can be done by fitting the Generalised Extreme Value distribution; it includes uncertainty about the type of tail X has, nonetheless, if there is evidence that X has a Heavy-Tail, fitting the Generalised Extreme Value distribution instead of the Fréchet distribution produces estimates with more variance; also, the GEV takes into consideration models that do not fit the data well [10, 9], such as taking into account the Gumbel likelihood for heavy-Tailm data, and or fitting the GEV to bounded data. Nevertheless, fitting extreme values is tricky, thus, all models should be under consideration, as well as a study of the data-generating process.

For example, there is substantial evidence that network traffic extremes have a Heavy-Tail; therefore, fitting the Generalised Extreme Value distribution instead of the Fréchet will produce less reliable estimates [2].

If there is evidence that the process that generated X has a Heavy-Tail, then for large enough thresholds the tail will resemble a Pareto distribution. It is remarkable that the Pareto tail approximation does not depend on sample size but rather on the Power Law like behaviour of the tail after a certain threshold has been surpassed. Hence, fitting the Generalised Pareto distribution, that includes the possibility of an Exponential Tail, is not as proper as fitting the Pareto Tail that does not resemble exponential decay [19, 9, 24].

This is important because the sample might not include a LARGE deviation that evidences a Heavy-Tail, nonetheless, Bayesian methods build on prior knowledge and given that there is both theoretical and empirical evidence that a lot of processes/phenomena, such as wage distributions, temperature fluctuations, financial returns, population

distributions (Zipfs law), just to name a few, have to be described with a Heavy-Tail; therefore, the analyst must take into consideration this even before exploring the sample [3, 25].

Recall that just one LARGE enough deviation is evidence that the process does not have an Exponential Tail; otherwise, estimates will communicate that events that take place on a yearly basis, such as market swings, should only happen once in a lifetime. The flexibility of the Generalised Extreme Value distribution permits exploration of data that fail to provide sufficient evidence of a Heavy-Tail.

Once the model limitations are taken into consideration Bayesian inference for extreme values follows the classic Bayesian algorithm. The algorithm will assign a prior distribution to the relevant quantities involved in the data-generating process, in this case the chosen distribution function's parameters; subsequently, the algorithm studies the posterior distribution of these quantities to update its beliefs about the data-generating process. if it helps, think of these as the learning procedure to update beliefs about a data-generating process, as well as to account for uncertainty in model selection and estimation.

The first thing to do is to test the data for a Heavy-Tail. This is done through exploratory data analysis techniques, such as the Mean Life plot and QQ-plots that study the empirical distribution. If there appears to be evidence of a Heavy-Tail one continues to select a threshold and/or perform hypothesis tests that either assert the Heavy-Tail, or provide evidence that the tail might not be compatible with a Heavy-Tail model. If there is reasonable evidence of a Heavy-Tail, then one selects a threshold u to fit one of the Heavy-Tail models [23].

The choice of threshold is a topic of debate, nevertheless, Mean Life plots provide an idea of the smallest possible threshold; also, prior information and concerns should be taken into account when selecting a threshold. For example, if the problem is to estimate the distribution of insurance claims above a known catastrophic threshold u_0, then the analyst should study if the threshold is large enough for a Pareto approximation, and if that is the case, then fit the Pareto distribution for values above the critical threshold u_0 [10, 1].

Finally, once the analyst selects a threshold u the next step is to fit one of the following models: Fréchet distribution for block Maxima, Generalised Extreme Value distribution for block Maxima, Pareto tail model for extremes over the threshold u, Generalised Pareto tail model for extremes over the threshold u and/or Poisson Point Process with a Power Law intensity measure to extremes larger than u to the data. After model fitting is done the analyst should look at the results and test the fit if possible [19, 16].

4.2 Hypothesis Tests for Heavy-Tails

An important difference between Heavy-Tail data and Exponential Tail data is that Heavy-Tails decay slower than an exponential function [2]. On this basis, a test can look for evidence in favour or against exponential decay. Consider the following heuristic to compare exponential decay with a Heavy-Tail.

If the data have a light tail, such that it's sensible that for some $\lambda > 0$ and large x the tail $S(x) \leq e^{-\lambda x}$ - the tail decays as fast or faster than an exponential tail. Then, consider the sample comes from an Exponential Tail distribution.

$X \sim Exp(\lambda)$ with a gamma prior (conjugate prior) for λ [3]. The model is,

$$\begin{aligned} P(\lambda|\underline{X_{(n)}}) &\propto \lambda^n \exp(-\lambda \sum_{i=1}^{n} X_i)\lambda^{a_0-1}e^{-\lambda b_0}, \\ &= \lambda^{n+a_0-1} \exp(-\lambda(b_0 + \sum_{i=1}^{n} X_i)). \end{aligned}$$

Hence, the posterior of λ is an updated gamma distribution. Then, for a sample of size n the n level return should approximately be:

$$\begin{aligned} P(X \geq x) &= \frac{1}{n}, \\ \iff e^{-\lambda x} &= \frac{1}{n}, \\ \Rightarrow x &= \frac{\ln(n)}{\lambda}. \end{aligned}$$

This a heuristic; thus, once the observed sample is fixed $\underline{X_{(n)}} = \underline{x_{(n)}}$ it is possible to substitute λ for its Bayes estimator (under this exponential-gamma conjugate model and squared error loss function),

$$\hat{\lambda}_{BS} = \frac{n + a_0}{b_0 + \sum_{i=1}^{n} x_i}.$$

If the value $\frac{\ln(n)}{\hat{\lambda}_{BS}}$ is much smaller than the largest value in the sample, then it is likely that the data do not have an Exponential Tail. This heuristic motivates, for data assumed to be unbounded, the following hypothesis test:

$\mathbf{H}_0$: the data have a Heavy-Tail vs $\mathbf{H}_1$: the data have an Exponential Tail.

This test will compare polynomial decay and exponential decay by

comparing the Fréchet model with the Gumbel model for the observed sample.

$\mathbf{M}_1$: Gumbel model for extremes, this model is compatible with the hypothesis $\mathbf{H}_0$ that the data do not have a Heavy-Tail.

$\mathbf{M}_0$: Fréchet model for extremes, this model is compatible with the hypothesis $\mathbf{H}_1$ that the data do have a Heavy-Tail.

For both posterior probabilities the normalising constant will cancel when computing the posterior odds (Bayes Factor: the posterior odds with respect to the models, the larger it is, the more likely $\mathbf{H}_0$ is). Thus, it is not wise to spend computational time calculating it.

$$P(\underline{X_{(n)}}) = \pi_0(\mathbf{M}_0) \int_A L(x_{(n)}|\alpha)\pi_0(\alpha|\mathbf{M}_0)d\alpha + \cdots$$
$$+\pi_0(\mathbf{M}_1) \int_{\xi=0} L(x_{(n)}|\xi = 0)\pi_0(\xi|\mathbf{M}_1)d\xi.$$

Under $\mathbf{M}_1$ the tail has no index and the only possible value is $\xi = 0$; therefore, the prior distribution is collapsed at zero,

$$\pi_0(\xi) = \delta_\xi(\{0\}).$$

Hence, there is no unknown parameter and the likelihood function for the Maxima of the sample is,

$$L_1(X^{(n)}|\xi = 0) = \exp(-(e^{-x} + x)).$$

Therefore the posterior probability of the model is,

$$\begin{aligned} P(\mathbf{H}_1|X^{(n)}) &= \int_{\xi=0} L_1(X^{(n)}|\xi=0)\pi_0(\xi|\mathbf{M}_1)\pi_0(\mathbf{M}_1)d\xi \\ &= \pi_0(\mathbf{M}_1)L_1(X^{(n)}|\xi=0). \end{aligned}$$

Under $\mathbf{H}_1$ the data do have a tail index $\alpha > 0$; thus, possible priors $\pi_0(\alpha|\mathbf{H}_0)$ are a Gamma, Exponential, Pareto or any other Borel probability measure for positive reals. The likelihood function for the sample maxima is,

$$L_0(X^{(n)}|\alpha) = \alpha x^{-(\alpha+1)} \exp(-x^{-\alpha}).$$

The posterior probability of $\mathbf{H}_0$ (Heavy-Tail) is:

$$\begin{aligned} P(\mathbf{H}_0|X^{(n)}) &= \frac{1}{P(\underline{X_{(n)}})} \int_\Theta L_0(X^{(n)}|\alpha)\pi_0(\alpha|\mathbf{M}_0)\pi_0(\mathbf{M}_0)d\alpha, \\ &= \frac{\pi_0(\mathbf{M}_0)}{P(\underline{X_{(n)}})} \int_\Theta L_0(X^{(n)}|\alpha)\pi_0(\alpha|\mathbf{H}_0)d\alpha, \\ &= \frac{\pi_0(\mathbf{M}_0)}{P(\underline{X_{(n)}})} E\big(L_0(X^{(n)}|\alpha) \ \ |\alpha \sim \pi_0(\alpha|\mathbf{H}_0)\big). \end{aligned}$$

The Bayes Factor for this presentation is:

$$BF(X^{(n)}) = \frac{P(\mathbf{H}_1|X^{(n)})}{P(\mathbf{H}_0|X^{(n)})},$$

The posterior odds for each hypothesis given the sample maxima;

$$= \frac{\frac{\pi_0(\mathbf{M}_0)}{P(\underline{X_{(n)}})} L_1(X^{(n)}|\xi = 0)}{\frac{\pi_0(\mathbf{M}_0)}{P(\underline{X_{(n)}})} E\big(L_0(X^{(n)}|\alpha) \ \ |\alpha \sim \pi_0(\alpha|\mathbf{H}_0)\big)},$$

Posterior odds evaluated via the posterior probabilities under each hypothesis and corresponding model assumption;

$$= \frac{\pi_0(\mathbf{M}_1)\exp(-(e^{-x}+x))}{\pi_0(\mathbf{M}_0) E\big(L_0(X^{(n)}|\alpha) \ \ |\alpha \sim \pi_0(\alpha|\mathbf{H}_0)\big)},$$

The final computation required to evaluate the Bayes factor and conclude the test.

The expected value can be calculated by use of Monte Carlo simulation as,

$$E\big(L_0(X^{(n)}|\alpha) \ \ |\alpha \sim \pi_0(\alpha|\mathbf{H}_0)\big) = \int_\Theta \alpha x^{-(\alpha+1)} e^{-x^{-\alpha}} \pi_0(\alpha|\mathbf{H}_0) d\alpha,$$
$$\approx \frac{1}{m}\sum_{t=1}^{m} L_0(\alpha^{(t)}, x^{(n)}).$$

Samples from $\alpha \sim \pi_0(\alpha|\mathbf{H}_0)$.

Because the ergodic theorem for Markov chain Monte Carlo (MCMC) implies,

$$\lim_{m\to\infty} \frac{1}{m}\sum_{t=1}^{m} L_0(\alpha^{(t)}, x^{(n)}) = \int_\Theta \alpha x^{-(\alpha+1)} e^{-x^{-\alpha}} \pi_0(\alpha|\mathbf{H}_0) d\alpha.$$

The prior odds affect how overwhelming the evidence for or against the null hypothesis has to be to reject. A rule of thumb is that a $BF > 10$ is strong evidence against the null hypothesis and that $BF > 100$ is decisive evidence against the null hypothesis [18]. For this test, suppose the posterior odds without considering prior probabilities of each model are equal to one (likelihoods ratios is equal to one), then if the prior odds of a thin tail are 20 times more likely than the prior odds of a Heavy-Tail the posterior evidence of a Heavy-Tail has to be at least as 20 times the posterior evidence of an Exponential Tail for the test to even consider failing to reject the null hypothesis. Also, the choice of $\pi_0(\alpha|\mathbf{H}_0)$ affects the posterior odds. The sensitivity of the test to prior information is a consequence of Lindley's paradox [18, 19, 28]; nevertheless, the test should be carried out in conjunction with other techniques to study if the data have a Heavy-Tail. A closer look at the test can help better understand the ratio. The heuristic to better understand the ratio resembles the likelihood ratio from frequentist statistics because it plugs in the value of $\hat{\alpha}_{BS}$. Given that under $\mathbf{H}_0$ the data have a Heavy-Tail the tail of the distribution is approximately a Pareto distribution above a valid threshold; thus, the conjugate prior is a gamma distribution and the bayes estimator is:

$$\hat{\alpha}_{BS} = \frac{m + a_0}{b_0 + \sum_{i=1}^{m} \ln(\frac{x_i}{u})}.$$

Where $\pi_0(\alpha) = Gamma(a_0, b_0)$; the sample size above the fixed threshold u is m and u is a threshold selected by exploratory data analysis techniques [1].

Plug in the estimate into the likelihood ratio to get,

$$BF = \frac{\pi_0(\mathbf{M}_1)\exp\big(-(e^{-x}+x)\big)}{\pi_0(\mathbf{M}_0)\big(\hat{\alpha}x^{-(\hat{\alpha}+1)}\exp(-x^{-\hat{\alpha}})},$$
$$\propto \exp\big(-(x+e^{-x}-x^{-\hat{\alpha}}\big)\frac{x^{(\hat{\alpha}+1)}}{\hat{\alpha}}.$$

Remark. *Note that the Bayes factor above is not proper because it does not sample different values of α to account for uncertainty, instead this Bayes factor plugs in an estimate to evaluate the odds ratio.*

Note that the Bayes factor above is not proper because it does not sample different values of α to account for uncertainty, instead this Bayes factor plugs in an estimate to evaluate the odds ratio. Then for large values of x this ratio will go to zero; also, the smaller $\hat{\alpha}$ is the faster this ratio will approach zero. A closer look reveals that for large values of x both $e^{-x} \to 0$ and $x^{-\hat{\alpha}} \to 0$. Thus,

$$\exp\big(-(x+e^{-x}-x^{-\hat{\alpha}}\big) \approx e^{-x}.$$

Putting these into the ratio one gets,

$$BF \to \frac{x^{(\hat{\alpha}+1)}}{\hat{\alpha}}e^{-x}.$$

The kernel of a gamma distribution for x. This ratio will only be large for values close to the mean and mode of this gamma distribution. Also, the ratio is a competition between exponential decay and polynomial decay because it compares a Power Law tail with an Exponential Tail. Therefore, for large values of x this ratio goes to zero. This agrees with the fact that large deviations of Heavy-Tails have much higher probabilities.

Remark.

$$\lim_{x\to\infty} \frac{x^{(\hat{\alpha}+1)}}{\hat{\alpha}} e^{-x} = 0.$$

For all $\hat{\alpha}$, also, the smaller $\hat{\alpha}$ is the faster the ratio goes to zero. This agrees with the fact that the smaller α is the Heavier the tail of X is.

The ratio above will be big if the biggest observation is not that big. This is evidence that the tail might not be Heavy; nonetheless, it is not conclusive evidence that the tail is Exponential or not Heavy. The prior odds can affect how fast this ratio reaches zero by either making it faster or smaller.

$$\frac{\pi_0(\mathbf{H}_1)}{\pi_0(\mathbf{H}_0)} = M, \ \textit{prior odds.}$$

Then the resulting Bayes factor needs to be much smaller than M to provide evidence that might be sufficient to fail to reject the null hypothesis; while, if M is small the Bayes factor will reach zero faster because M makes it smaller. Finally, if the value of x is small the ratio can be both huge and small; nonetheless, if x is small this means that the biggest observation is close to small values. This could be evidence that the data might be bounded and the test is inconclusive for $x < 1$, where x is the largest observation.

The second test takes advantage of the Generalised Extreme Value distribution (GEV). This distribution incorporates uncertainty about the tail index $\alpha = \frac{1}{\xi}$ into estimation. To perform this test one needs the sample of independent block Maxima from different periods. The size of the sample is the number of periods observed , n_p, and it consists of

block Maxima. Recall that the GEV is the distribution function,

$$\begin{aligned}
&\text{GEV}(x) = \exp\big(-(1+\xi x)^{-\frac{1}{\xi}}\big) \quad \textit{if } \xi \neq 0,\\
&\text{GEV}(x) = \exp(-e^{-x}) \quad \textit{if } \xi = 0.
\end{aligned}$$

One proceeds to obtain the posterior distribution of ξ.

$$P(\xi|\underline{X_{(n_p)}}) \propto P(\underline{X_{(n_p)}}|\xi)\pi_0(\xi).$$

The flexibility of the GEV is that $\xi \in \mathbb{R}$, thus, the prior distribution could be a normal distribution, t-student, or a mixture with different modes. The likelihood is only defined for Maxima that fall in:

$$\mathbb{X} = \{x : 1 + \xi x > 0\}.$$

The likelihood function for a fixed sample of size n_p and $\xi \neq 0$ is:

$$\begin{aligned}
&L(\underline{X_{(n_p)}}|\xi) = P(\underline{X_{(n_p)}} = \underline{x_{(n_p)}}|\xi),\\
&= \prod_{i=1}^{n_p} f(x_i|\xi),\\
&= \prod_{i=1}^{n_p} \exp\big(-(1+\xi x_i)^{-\frac{1}{\xi}}\big)\frac{\xi}{\xi}(1+\xi x_i)^{-(\frac{1}{\xi}+1)} \exp\big(-(1+\xi x_i)^{-\frac{1}{\xi}}\big),\\
&= \prod_{i=1}^{n_p} \exp\big(-((1+\xi x_i)^{-\frac{1}{\xi}} + (\frac{1}{\xi}+1)\ln(1+\xi x_i))\big),\\
&= \exp\left(-\sum_{i=1}^{n_p}(1+\xi x_i)^{-\frac{1}{\xi}} - (\frac{1}{\xi}+1)\sum_{i=1}^{n_p}\ln(1+\xi x_i)\right).
\end{aligned}$$

A rule of thumb is that if $|\xi| < \epsilon_0 = \frac{1}{2}$ then the likelihood should be replaced by the likelihood of the Gumbel distribution (limit $\xi \to 0$) [1],

in that case the likelihood function is:

$$L(\underline{X_{(n_p)}}|\xi) = \exp\left(- \left(\sum_{i=1}^{n_p} e^{-x_i} + x_i \right) \right).$$

This likelihood has a complex form and no apparent conjugate, therefore, simulation from the posterior distribution is done through numerical techniques. Given that $\xi \in \mathbb{R}$ and that the likelihood has a complex form a reasonable option is a Hamiltonian Monte Carlo (HMC) [19]. This algorithm takes an idea from physics to use a coefficient (momenta/momentum) to push the simulated ξ to regions with more posterior density (i.e., it optimises the random walk behaviour). The momenta vector acts as the gradient that pushes the values towards regions with more density. Furthermore, HMC permits the use of a flat improper prior $\pi_0(\xi) = 1$ that does not contribute to the posterior distribution. This allows the data to centre the posterior distribution of ξ just with information from the likelihood [19].

The gradient of the log-likelihood is:

$$
\begin{aligned}
\nabla(\xi) &= \Big(-\sum_{i=1}^{n_p} (1+\xi x_i)^{-\frac{1}{\xi}} - (\frac{1}{\xi}+1) \sum_{i=1}^{n_p} \ln(1+\xi x_i) \Big)', \\
&= -\Big(\sum_{i=1}^{m} \frac{1}{\xi} \ln(1+\xi x_i) + \ln(1+\xi x_i) + \exp(-\xi^{-1} \sum_{i=1}^{m} \ln(1+\xi x_i)) \Big)', \\
&= -\Big(\sum_{i=1}^{m} -\xi^{-2} \ln(1+\xi x_i) + \frac{\xi^{-1} x_i}{1+\xi x_i} + \frac{x_i}{1+\xi x_i} - \cdots \\
&\sum_{i=1}^{m} \xi^{-2} \ln(1+\xi x_i) - \frac{\xi^{-1} x_i}{1+\xi x_i} \exp\big(-\xi^{-1} \sum_{i=1}^{m} \ln(1+\xi x_i) \big) \Big), \\
&= -\Big(\sum_{i=1}^{m} -\xi^{-2} \ln(1+\xi x_i) + \frac{(\xi^{-1}+1) x_i}{1+\xi x_i} - \cdots \\
&\sum_{i=1}^{m} \big(\xi^{-2} \ln(1+\xi x_i) - \frac{\xi^{-1} x_i}{1+\xi x_i} \big) \exp\big(-\xi^{-1} \sum_{i=1}^{m} \ln(1+\xi x_i) \big) \Big), \\
&= \xi^{-2} \sum_{i=1}^{m} \ln(1+\xi x_i) - (\xi^{-1}+1) \sum_{i=1}^{m} \frac{x_i}{1+\xi x_i} + \cdots \\
&\sum_{i=1}^{m} \big(\xi^{-2} \ln(1+\xi x_i) - \xi^{-1} \frac{x_i}{1+\xi x_i} \big) \exp\big(-\xi^{-1} \sum_{i=1}^{m} \ln(1+\xi x_i) \big), \\
&= \xi^{-2} \sum_{i=1}^{m} \ln(1+\xi x_i) \big(1 + \exp\big(-\xi^{-1} \sum_{i=1}^{m} \ln(1+\xi x_i) \big) \big) \cdots \\
&- \xi^{-1} \sum_{i=1}^{m} \frac{x_i}{1+\xi x_i} \big(1 + \exp\big(-\xi^{-1} \sum_{i=1}^{m} \ln(1+\xi x_i) \big) \big) - \sum_{i=1}^{m} \frac{x_i}{1+\xi x_i}.
\end{aligned}
$$

It is necessary to add the derivative of the log-prior.

$$\frac{d}{d\xi} \ln(\pi_0(\xi)) = \psi_0(\xi).$$

The sum of these two is the derivative (one dimension gradient) of the log-posterior. Note that if $\pi_0(\xi) = 1$, then $\psi_0(\xi) = 0$; thus, in this case the gradient does not push ξ in any direction the prior suggests. Finally if $|\xi| < \epsilon_0$, the log-posterior has derivative:

$$\frac{d}{d\xi} \ln(\pi_0(\xi)) - \sum_{i=1}^{m} e^{-x_i} + x_i = \psi_0(\xi).$$

There is no change because all mass (probability) is at $\xi = 0$.

Algorithm 1 Hamiltonian Monte Carlo

Starting value $\xi^{(t)}$, s, c, L
for $t = 1, \cdots, T$ **do**

$$a \sim N(0, s)$$
$$\xi^* = \xi^{(t-1)}$$
$$a^* = a + \frac{c}{2}\nabla(\xi^*).$$

for $l = 1, \cdots, L$ **do**
Walk L **leapfrog steps:**

$$\xi_{(0)} = \xi^*$$
$$a_{(0)} = a^*$$
$$\xi_{(l)} = \xi_{(l-1)} + c a_{(l-1)}$$
$$a_{(l)} = a_{(l-1)} + c\nabla(\xi_{(l-1)}).$$

end

$$a^* = a_{((L)}$$
$$\xi^* = \xi_{((L)}$$
$$a^* = a^* - \frac{c}{2}\nabla(\xi^*)$$
$$R = \frac{P(\xi^*|\underline{X_{(n_p)}}) \exp(-\frac{a^{*2}}{2})}{P(\xi^{(t-1)}|\underline{X_{(n_p)}}) \exp(-\frac{a^2}{2})}$$

Accept or reject:
sample $U \sim Unif(0, 1)$

if $U \leq R$ **then**
$\xi^{(t)} = \xi^*$
else
$\xi^{(t)} = \xi^{(t-1)}$
end
end

Another way of carrying out this test is to employ a slice sampler to sample from the posterior distribution. The slice sampler model has the advantage of known conditional posterior distributions, hence, it is a Gibbs sampler [3]. The Slice Sampler would be:

$$\begin{aligned} f(\xi, t) &= \chi\big(0 \leq t < f(\xi|\underline{X_{n_p}})\big), \\ f(\xi) &= \int_{\mathbb{T}} \chi\big(0 \leq t < f(\xi|\underline{X_{n_p}})\big)dt, \\ &= \int_0^{f(\xi|\underline{X_{n_p}})} dt, \\ &= f(\xi|\underline{X_{n_p}}). \end{aligned}$$

Thus the conditional distributions are.

$$\begin{aligned} t \ | \ \xi &\sim Unif\big([0, f(\xi|\underline{X_{n_p}})]\big), \\ \xi \ | \ t &\sim Unif\big(\Xi\big), \\ &\textit{Where,} \quad \Xi = \{\xi : f(\xi|\underline{X_{n_p}}) \geq t\}. \end{aligned}$$

The chain moves from t to ξ and after a burn in period the chain samples from the posterior distribution. After sampling from the posterior distribution or if its possible solving for its exact form. The test proceeds to compute a credible interval for ξ.

Definition (Credible Interval [3])**.**

$$I_\eta|\underline{X_{(n_p)}} := P(\xi \in I_\eta|\underline{X_{(n_p)}}) = 1 - \eta.$$

To do this it is possible to look at a histogram of the simulations

and a summary of the empirical distribution of these simulations. The simulations that this tools uses are the ones obtained after the burn-in period. The smaller η is the more credibility (probability) there is that $\xi \in I_\eta$. Thus, if I_η does not include 0 or does not concentrate probability in a neighbourhood of 0, the test provides evidence that the tail is not exponential [19]. Therefore, given the sample $\underline{X_{(n_p)}}$ an Exponential Tail is less compatible than a Heavy-Tail model or a bounded Inverse-Weibull tail model. The likelihood will only be greater than zero for values of ξ that agree with the data.

Thus, if the sample $\underline{X_{(n_p)}}$ includes large deviations, negative values of ξ will have little to no likelihood; while, if the sample $\underline{X_{(n_p)}}$ includes small values, the likelihood will not increase for large values of ξ. Recall that, if $\xi > 0$ then the data have a Heavy-Tail and the GEV becomes the Fréchet distribution with a Pareto Tail for extremes above a large enough threshold u [23]. Thus, for $\frac{1}{\xi} = \alpha > 0$.

$$G_F(x) = \exp(-x^{-\alpha}),$$
$$P(X > x | X > u) \to \left(\frac{x}{u}\right)^{-\alpha}.$$

If $\xi = 0$ the data have an Exponential Tail. The limit model is a Gumbel distribution, thus, large deviations are much less likely and the parameter that affects the decay is the scale $\theta > 0$ [23].

$$G_G(x) = \exp(-e^{-x}),$$
$$P(X > x | X > u) \to \exp\big(-\frac{1}{\theta}(x-u)\big).$$

If $\xi < 0$ the limit is the Inverse-Weibull and the data have an upper bound with Power Law like behaviour, hence, extremes tend to

concentrate near the upper bound. For $\alpha = -\frac{1}{\xi} > 0$ and the upper bound $\gamma \in \mathbb{R}$ the limit distributions are:

$$G_W(x) = \exp(-(-x+\gamma)^{\alpha}),$$
$$P(X > x | X > u) \to \left(1 - \frac{x-u}{\gamma-u}\right)^{\alpha}.$$

These tests looks at the regions where the likelihood concentrates the posterior probability of ξ. Thus, if the values do not concentrate near zero, there is some evidence that the Exponential Tail is not compatible with the data; nevertheless, this test is for looking at incompatibility not at compatibility. The test might reject an Exponential Tail, when in fact the Exponential Tail is the correct model. That is the inherent difficulty of Hypothesis tests. These tests measure incompatibility not compatibility and are only able to compare the proposed models. One advantage of Extreme Value models is that there are only three possible models to compare, hence, the credibility interval test is able to compare the three possible models by looking at the posterior distribution [24, 19].

4.3 Threshold Search and Selection

The choice of Threshold has been and continues to be a topic of discussion [16, 12, 14]. The consensus is that exploratory data analysis techniques, such as the Mean Life Plot, QQ-plots, and probability plots are efficient at suggesting a threshold [23, 25]. The threshold selection is important because if the Threshold u is too small the Extreme Value model will not approximate as well as it would for a larger threshold; nevertheless, if the threshold u is too big the sample

size will probably be too small to produce robust inferences [1].

It is also important to remember that statistical inference does not happen in a vacuum, therefore, as with the other components of the model the threshold selection should take into account prior knowledge and information. This is helpful in applications where a specific threshold is of interest. For example, if the problem at hand is to estimate the distribution of Rain-Fall levels above a point u_0 that may cause damage to some structures. Another example is in Reinsurance where losses above a threshold u_0 are thought to be catastrophic and require reinsurance coverage [13, 15].

These excesses could also be good, some examples are off the charts financial returns, excess energy production and website traffic, just to name a few.

The first step to select a threshold is to fix a range of possible values, based on prior knowledge, and to produce a Mean Life Plot, QQ-plots and other exploratory data analysis heuristics that suggest the threshold is in a set of values. This is well paired with Bayesian Inference in its use of prior beliefs about the data-generating process, but this is not exclusive to the Bayesian approach, it is valid to select a range of possible thresholds and proceed with Frequentist estimation techniques; however, the algorithm this section proposes exploits Bayesian techniques, as well as possible ways to choose a threshold.

Once a threshold u_0 is selected, it is possible to fit the model; nonetheless, two interesting threshold search algorithms are presented in [14, 12]. The first algorithm uses the Kullback-Liebler divergence to compare the posterior predictive distribution with the assumed correct model for different thresholds [12]. The second work explores differences between the estimated parameters by use of normality tests

that search for an optimal threshold. This method is like backward step selection in linear regression [14].

Because changing the location of the Tail to a threshold larger than the original does not change the tail index, a good threshold search takes into account that estimates of the tail index α should remain nearly constant; if the sample size is not too small, above a proper threshold. Therefore, uncertainty of α carries uncertainty about u, nevertheless, above a value u_0 the Pareto Tail should approximate the Heavy-Tail [12, 14].

Proposition. *Given Heavy-Tail random variables once a threshold u_0 is surpassed the extremes are approximately a Pareto Distribution with upper bound u_0. Then for all $u \geq u_0$ the shape α does not change.*

Proof. The approximate Pareto Tail is a consequence of the Fréchet Limit for Heavy-Tail extremes. Thus,

$$X \mid \alpha, \;\; X > u_0 \sim \left(\frac{x}{u_0}\right)^{-\alpha}.$$

Recall that the Fréchet limit depends on the sequence (b_n) that scales large deviations; thus, once u_0 is large enough to stabilise extremes any value above will also stabilise extremes. Then the surplus (tail) function of $X|\ X > u \geq u_0$ is,

$$\begin{aligned} P(X > x | X > u \geq u_0) &= \frac{P(X > x)}{P(X > u)}, \\ &= \left(\frac{x}{u}\right)^{-\alpha} \left(\frac{u_0}{u_0}\right)^{-\alpha}, \\ &= \left(\frac{x}{u}\right)^{-\alpha}. \end{aligned}$$

If one is interested in the exceedance $Y = X - u > 0$ the Pareto distribution is.

$$x = y + u \Rightarrow \frac{dx}{dy} = 1.$$

Thus,

$$\begin{aligned} F_Y(y) &= \left(\frac{y+u-u}{u}\right)^{-\alpha}, \\ &= \left(1 + \frac{y}{u}\right)^{-\alpha}. \end{aligned}$$

■

This relationship motivates the threshold search this work proposes. The idea is that for valid thresholds u the Heavy-Tail can be approximated by a Pareto Distribution. Then, by implementing latent variables, it is possible to construct a Gibbs Sampler that searches for the threshold u with the most posterior probability from a collection of possible thresholds. The Gibbs Threshold Search model is the following.

To start the search propose a finite collection of thresholds of interest $u \in \{u_1, \cdots, u_K\}$ and introduce the model probability as the probability of each threshold.

$$P(u = u_k) = \beta_k, \quad k \in \{1, 2, \cdots, K\}.$$

Clearly $\sum_{k=1}^{K} \beta_k = 1$, thus, the conjugate prior of this model $\underline{\beta} = (\beta_1, \beta_2, \cdots, \beta_K)^t$ is the Dirichlet distribution. Implement the latent variables Z_i , $i \in \{1, 2, \cdots, m\}$ with the same sample size of extremes above the smallest threshold. Then, plug in each Z_i to the corresponding X_i to get m observations of (X_i, Z_i), such that the Z_i

assign a model to the X_i.

$$P(Z_i = k) = \beta_k,$$
$$Z_i = k \Rightarrow u = u_k,$$
$$f(X_i|\ \alpha, Z_i = k) = \frac{\alpha_k}{u_k}\left(\frac{x}{u_k}\right)^{-\alpha_k}.$$

Then computing probabilities with Bayes Theorem.

$$P(Z_i = k|\ X_i, \alpha) = \frac{f(X_i|\ \alpha, Z_i = k)P(Z_i = k)}{\sum_{j=1}^{K} f(X_i|\ \alpha, Z_i = j)P(Z_i = j)}.$$

This ratio computes the posterior probability of each model against the total probability of all models. It assigns a higher probability to models with more likelihood; thus, it resembles a search for a model with maximum likelihood, but in a completely different Bayesian way. This ratio is the key to the search. For a fixed threshold u the tail index conjugate model is a Gamma distribution.

$$\alpha|\underline{X_{(m)}^{(u)}} \sim Gamma(a_x, b_x),$$
$$a_x^k = m_k + a_0\ \ ;\ \ b_x^k = \sum_{i=1}^{m_k} \ln\left(\frac{x_i}{u_k}\right) + b_0,$$
$$\pi_0(\alpha) = Gamma(a_0, b_0).$$

For proper thresholds $\hat{\alpha}$ should remain around a particular value. Also, X_i can only be used in model $u = u_k$ if $X_i > u_k$; therefore, the algorithm

needs to adjust this during each iteration.

$$m_k = \sum_{i=1}^{m} \chi(Z_i = k),$$
$$\alpha_k \sim Gamma(a_x^k, b_x^k).$$

The number m_k is the sample size of model k; while, a_x^k, b_x^k update the posterior distribution of α. The likelihood function of the full model is:

$$L(\alpha, \underline{\beta}) \propto \prod_{i=1}^{m} \prod_{k=1}^{K} \big(\beta_k f(x_i|\ \alpha_k, z_i = k)\big)^{\chi(z_i=k)}.$$

Then the posterior distribution of $\underline{\beta}$ is:

$$\begin{aligned} P(\underline{\beta}|\underline{X_{(m)}}, \underline{Z_{(m)}}) &\propto L(\alpha, \underline{\beta})\pi_0(\underline{\beta}), \\ &= \beta_k^{m_k}\pi_0(\underline{\beta}) \prod_{i=1}^{m} \prod_{k=1}^{K} \big(f(x_i|\ \alpha_k, z_i = k)\big)^{\chi(z_i=k)}, \\ &\propto \prod_{k=1}^{K} \beta_k^{m_k+\tau_k}. \end{aligned}$$

The Kernel of a Dirichlet distribution where $\underline{\tau} = (\tau_1, \tau_2, \cdots, \tau_k)^t$ is the hyperparameter vector of the prior Dirichlet distribution, thus, the conjugate model is:

$$\pi_0(\underline{\beta}) \propto \prod_{k=1}^{K} \beta_k^{\tau_k},$$
$$P(\underline{\beta}|\ \underline{X_{(m)}}, \underline{Z_{(m)}}) = Dir(m_1 + \tau_1, \cdots, m_K + \tau_K).$$

The Gibbs Sampler will sample from each of the conditional distributions, after a Burn-In period, the Gibbs Sampler samples from the posterior distribution. A histogram for each of β_k will give insight

into the posterior distribution of each model's probability. The choice of threshold could be the threshold with the most posterior probability or a convex combination of the possible thresholds.

$$\hat{u}_0 = \sum_{k=1}^{K} \beta_k u_k.$$

This stochastic search variable selection finds the threshold u_k with the most posterior probability; nevertheless, it can only compare thresholds within the range of the collection. Therefore, the search should be done for different collections that seem plausible after an analysis of the Mean Life Plot, Empirical Distribution, QQ-plot, among other possible analysis [of the data] carried out before model fitting. The estimate is either the threshold with most posterior probability or a convex combination of thresholds that assigns weight according to posterior probabilities. Motivation for the model is the mixture latent variable model; that model can be found in [19].

Algorithm 2 Gibbs Threshold Search

Starting values: $\underline{\beta}^{(0)}$, $\alpha^{(t)}$, $\{u_1, \cdots, u_K\}$

for $t = 1, \cdots, T$ **do**

Generate chain:

$$Z_i^{(t)} | \alpha^{(t-1)}, X_i \sim \frac{f(X_i|\ \alpha, Z_i = k)P(Z_i = k)}{\sum_{j=1}^{K} f(X_i|\ \alpha, Z_i = j)P(Z_i = j)}$$

$$a_x^k = m_k + a_0\ ;\ b_x^k = \sum_{i=1}^{m_k} \ln\left(\frac{x_i}{u_k}\right) + b_0$$

$$m_k = \sum_{i=1}^{m} \chi(Z_i^{(t)} = k)$$

$$\alpha_k^{(t)} \sim Gamma(a_x^k, b_x^k)$$

$$\underline{\beta}^{(t)} \sim Dir(m_1 + \tau_1, \cdots, m_K + \tau_K)$$

end

Result: Chain: $\{u(0), \cdots, u(T)\}$

4.4 Posterior Distribution Simulation

This work focuses on Heavy-Tail data, thus, after performing the proper analysis and hypothesis tests, searching and selecting a threshold u; it is time to fit a model for Heavy-Tail extremes. The previous sections show that it could either be done by fitting the Fréchet distribution to the block Maxima, the Generalised Extreme Value distribution to the block Maxima, a Pareto tail for values above the threshold u, a Generalised Pareto tail to values above the threshold u or the Poisson Point Process model to values above the threshold u [1, 23].

If the hypothesis tests and empirical evidence (sample contains large deviations in almost every period observed) provide substantial evidence that an Exponential Tail does not fit these data well, then the Fréchet likelihood and Power Law intensity measure are more in accordance with the data.

On those basis, this work proposes a quasi-conjugate model to fit the Poisson Point Process model. The Poisson Point Process model accounts for the rate at which the data surpass the threshold and for the Power Law behaviour of large deviations. It also allows to fit the model to extremes of different periods [26].

This is important because the scale and location parameters should change if the number of periods changes. The ten year Maxima should at least be as extreme, if not more, as the one year Maxima. The Poisson Point Process combines the Fréchet nucleus of extremes and Pareto Tail behaviour of large deviations (i.e., values that satisfy $X > u$.).

Let $N(A_u) = m$ be a Poisson Point Process with intensity measure ν and an appropriate threshold u. The sample consists of observations

in a sample of size n larger than u observed in n_p periods.

$$\underline{X_{(m)}^{(u)}} = \{X_i \in \underline{X_{(n)}} : X_i > u\}.$$

The model depends on three parameters $\underline{\theta} = (\alpha, \theta, \gamma)^t$. The tail index $\alpha > 0$ is the parameter that shapes the tail. The smaller α is the heavier the tail is. The scale parameter $\theta > 0$ scales the extreme values to stabilise the distribution, the larger θ is the larger the percentiles of the distribution are. If the sample size is big but the number of periods is small, then θ increases by a factor, as the Maxima of a Fréchet distribution, of $n_p^{\frac{1}{\alpha}}$. The location parameter $\gamma \in \mathbb{R}$ stabilises extremes but is not affected as much by the number of periods as θ is. This number is the lower bound (minimum) value extremes may take. Once a threshold is fixed the location parameter cannot exceed the threshold (i.e., $\gamma < u$). If γ is big the percentiles will be big for any period. The likelihood function for this models is shown in chapter 3 section 2. As stated in the previous section, this likelihood mixes the properties of the Pareto tail (Peaks Over Threshold) and the Extreme Value Limit Fréchet [2]. Recall that if the lower bound of a Pareto distribution is known the conjugate model for the tail index is a Gamma distribution.

$$\alpha | \underline{X_{(m)}^{(u)}} \sim Gamma(a_x, b_x),$$

$$a_x = m + a_0 \ ; \ b_x = \sum_{i=1}^{m} \ln\left(\frac{x_i}{u}\right) + b_0,$$

$$\pi_0(\alpha) = Gamma(a_0, b_0).$$

A change of scale and/or location does not change the tail index of the Poisson Point Process, nor of the Generalised Pareto and/or the Fréchet distribution; furthermore, the decay of the tail does depend on

the possible values of X as it gives the tail its shape. However, the shape does not depend on location and/or scale. Thus, it is reasonable to think of α as being independent of γ, θ a-priori. Hence, a valid prior is:

$$\pi_0(\alpha) = Gamma(a_0, b_0) = \pi_0(\alpha|\gamma, \theta).$$

If γ, θ are fixed then,

$$\begin{aligned}
&P(\alpha|\gamma, \theta, \underline{X^{(u)}_{(m)}}) \propto \exp\left(- n_p\left(\frac{u-\gamma}{\theta}\right)^{-\alpha}\right)\cdots \\
&\alpha^m \prod_{i=1}^{m}\left(\frac{x_i-\gamma}{\theta}\right)^{-(\alpha+1)} \pi_0(\alpha|\gamma,\theta), \\
&\propto \alpha^{m+a_0-1}\exp\left(- (\alpha+1)\sum_{i=1}^{m}\ln\left(\frac{x_i-\gamma}{\theta}\right)\right)\exp\left(- n_p\left(\frac{u-\gamma}{\theta}\right)^{-\alpha}\right), \\
&\leq \alpha^{m+a_0-1}\exp\left(- \alpha\left(b_0 + \sum_{i=1}^{m}\ln\left(\frac{x_i-\gamma}{\theta}\right)\right)\right), \\
&\approx \alpha^{m+a_0-1}\exp\left(- \alpha\left(b_0 + \sum_{i=1}^{m}\ln\left(\frac{x_i}{u}\right)\right)\right).
\end{aligned}$$

To approximate the posterior distribution the model compares the conjugate gamma posterior distribution of the Pareto Tail with the posterior distribution of the Poisson Point Process likelihood. The Metropolis-Hastings acceptance probability ratio for a simulation from $\alpha_* \sim Gamma(a_x, b_x)$. Is,

$$R_\alpha = \frac{L(\alpha_*, \theta, \gamma)\pi_0(\alpha_*)\alpha^{m+a_0-1}\exp\left(- \alpha\left(b_0 + \sum_{i=1}^{m}\ln\left(\frac{x_i}{u}\right)\right)\right)}{L(\alpha, \theta, \gamma)\pi_0(\alpha)\alpha_*^{m+a_0-1}\exp\left(- \alpha_*\left(b_0 + \sum_{i=1}^{m}\ln\left(\frac{x_i}{u}\right)\right)\right)}.$$

The chain will accept the proposal if the posterior odds of the proposal

are much better than the posterior odds of the current value $\alpha^{(t)}$. This jumping ratio takes advantage of the Pareto Tail Heavy-Tail extremes have because it uses a conjugate model to approximate the posterior. This approximation is based on the Fréchet limit and its Pareto Tail. As in previous sections if the threshold is large enough, then any threshold above it will preserve the Pareto Tail. The threshold can also be a critical value, such as the maximum level of water a dam may hold.

The scale parameter does not appear to have a clear conjugate conditional posterior. The threshold does influence the scale because if the threshold is too small, the scale needs to stabilise large deviations; whereas, if the threshold is too big, the scale might be too close to zero. Nevertheless, Heavy-Tail data are characterised by large deviations, thus, the scale will adjust the period Maxima, recall that for periods with size p and n observations the number of periods is $n_p = \frac{n}{p}$.

A smaller number of periods will likely estimate a larger scale parameter. For computational simplicity the model assigns independent priors to all parameters. A possible prior for θ is a flat exponential; nonetheless, any distribution of a non-negative random variable is a valid prior, even a prior that transforms θ, but it would be necessary to adjust the distribution with the Jacobian. This algorithm accounts lack of knowledge about θ with,

$$\pi_0(\theta) = 1.$$

The conditional posterior is.

$$P(\theta|\alpha,\gamma,\underline{X_{(m)}^{(u)}}) \propto \alpha^m(\theta^{\alpha m})\exp\left(-n_p\left(\frac{u-\gamma}{\theta}\right)^{-\alpha}\right)\cdots$$
$$\pi_0(\theta|\alpha,\gamma)\prod_{i=1}^{m}(x_i-\gamma)^{-(\alpha+1)},$$
$$\propto (\theta^{\alpha m})\exp\left(-n_p\left(\frac{u-\gamma}{\theta}\right)^{-\alpha}\right)\pi_0(\theta).$$

A close look at the likelihood suggests that θ^α might be modelled as a Gamma random variable, nonetheless, most of the sample does not affect θ directly; instead, the data influence θ through α,γ. A way of approximating the posterior distribution of θ is to walk randomly around values of θ with a lot of likelihood. On that basis, this model estimates the posterior distribution of θ with a random walk and Metropolis-Hastings jumping rule for symmetric distributions.

$$\theta_* \sim N(\theta^{(t)},\sigma^2).$$

Where σ^2 is the tuneable variance of the random walk and $\theta^{(t)}$ is the previous value in the chain.

$$R_\theta = \frac{L(\alpha,\theta_*,\gamma)\pi_0(\theta_*)}{L(\alpha,\theta^{(t)},\gamma)\pi_0(\theta^{(t)})}.$$

If the prior is flat then the acceptance ratio becomes a likelihood ratio.

$$R_\theta = \frac{L(\alpha^{(t)},\theta_*,\gamma^{(t)})}{L(\alpha^{(t)},\theta^{(t)},\gamma^{(t)})}.$$

The random walk will move towards a region of high posterior density. The chain samples values that, after a burn-in period, approximate the

posterior distribution of θ informed by the other two parameters.

A location of interest is $\gamma = 0$, since the scale stabilises non-negative random variables (Fréchet limit) it is interesting to see if the posterior evidences that location moves extremes far from zero. For example, in financial applications the minimum loss is zero and extremes far from zero would require larger reserves because larger deviations lead to bigger percentiles.

Note that $\gamma < u$, thus, this has to be taken into account when approximating the posterior distribution. A prior with a lot of mass near zero is a valid starting point; furthermore, this prior distribution will help control γ to avoid sampling in regions that have no posterior likelihood (i.e., $\frac{u-\gamma}{\theta} < 0.$). Hence, this model uses a normal distribution centred at zero as the prior distribution; nevertheless, any prior over the real numbers that mixes with the posterior condition $\gamma < u$ is a valid prior.

$$\pi_0(\gamma) \propto \exp\left(-\frac{\gamma^2}{2c^2}\right).$$

The conditional posterior does not appear to have a clear conjugate model, or specific distribution.

$$
\begin{aligned}
&P(\gamma|\alpha,\theta,\underline{X^{(u)}_{(m)}}) \propto \exp\left(-n_p\left(\frac{u-\gamma}{\theta}\right)^{-\alpha}\right)\cdots \\
&\alpha^m\prod_{i=1}^{m}\left(\frac{x_i-\gamma}{\theta}\right)^{-(\alpha+1)}\pi_0(\gamma|\alpha,\theta)\chi_{(-\infty,u)}(\gamma), \\
&\propto \exp\left(-(\alpha+1)\sum_{i=1}^{m}\ln\left(\frac{x_i-\gamma}{\theta}\right)\right)\cdots \\
&\pi_0(\gamma|\alpha,\theta)\exp\left(-n_p\left(\frac{u-\gamma}{\theta}\right)^{-\alpha}\right), \\
&\propto \exp\left(-(\alpha+1)\sum_{i=1}^{m}\ln\left(\frac{x_i-\gamma}{\theta}\right)\right)\cdots \\
&\pi_0(\gamma|\alpha,\theta)\exp\left(-n_p\exp\left(-\alpha\ln\left(\frac{u-\gamma}{\theta}\right)\right)\right).
\end{aligned}
$$

This resembles - in a strange way - the kernel of the Gumbel distribution for $\rho = \ln\left(\frac{u-\gamma}{\theta}\right)$. Since the threshold u comes from a distribution whose Extreme Value Limit is a Fréchet distribution the transformation $\ln(u)$ will result in a Gumbel limit (i.e., the logarithm centres data). If one substitutes ρ, where $u-\theta e^{\rho}=\gamma$, and adjusts with $\left|\frac{d\theta}{d\rho}\right| = \theta e^{\rho}$ one gets.

$$
\begin{aligned}
&P(\rho|\alpha,\theta,\underline{X^{(u)}_{(m)}}) \propto e^{\rho}\exp\left(-n_p e^{-\alpha\rho}\right)\cdots \\
&\exp\left(-(\alpha+1)\sum_{i=1}^{m}\ln\left(\frac{x_i-u}{\theta}+e^{\rho}\right)\right).
\end{aligned}
$$

The resemblance to the Gumbel Kernel is easier to see. Therefore, a jump rule can be a random walk around γ, but instead of sampling from a normal distribution centred at γ, the proposal is a sample from

a Gumbel distribution centred at γ.

$$\gamma_* \sim Gumbel(\gamma^{(t)}, s_0),$$
$$f(\gamma) \propto \exp\big(-(e^{-\frac{\gamma-\gamma^{(t)}}{s_0}} + \frac{\gamma-\gamma^{(t)}}{s_0})\big).$$

The acceptance probability ratio is.

$$R_\gamma = \frac{L(\alpha^{(t)}, \theta^{(t)}, \gamma_*)\pi_0(\gamma_*)f(\gamma^{(t)})}{L(\alpha^{(t)}, \theta^{(t)}, \gamma^{(t)})\pi_0(\gamma^{(t)})f(\gamma_*)}.$$

The ratio approximates the posterior distribution of the lower bound of a Heavy-Tail random variable with an Exponential Tail random variable. This can also be done with other proposal distributions; nevertheless, the form of the posterior suggests that it can mix with a Gumbel distribution.

Furthermore, both $\gamma^{(t+1)}, \theta^{(t+1)}$ must fall in the support of the model.

$$\left(\frac{x_i - \gamma^{(t+1)}}{\theta^{(t+1)}}\right) > 0 \ \ \forall x_i, \quad \left(\frac{u - \gamma^{(t+1)}}{\theta^{(t+1)}}\right) > 0.$$

If that is not the case, then the chain rejects $\gamma^{(t+1)}, \theta^{(t+1)}$ and stays at the same point. If $\sum_{i=1}^{m} \ln(\left(\frac{x_i-\gamma^{(t+1)}}{\theta^{(t+1)}}\right) < 0$ frequently, then this could be evidence that the Heavy-Tail model is not appropriate. Hence, one should fit the Generalised Extreme Value distribution and/or look at the Gumbel and Inverse-Weibull models. Finally, this model approximates the posterior distribution with a Gibbs Sampler construction applied to a Metropolis Hastings algorithm. The model uses conditional posteriors to approximate the joint posterior distribution. Below is a summary of the proposal distributions as

approximations of each conditional posterior.

$$
\begin{aligned}
P(\alpha|\theta,\gamma,\underline{X_{(m)}^{(u)}}) &\approx f(\alpha|\underline{X_{(m)}^{(u)}}),\\
f(\alpha|\underline{X_{(m)}^{(u)}}) &= Gamma(a_x,b_x),\\
P(\theta|\alpha,\gamma,\underline{X_{(m)}^{(u)}}) &\approx f(\theta|\theta^{(t)}),\\
f(\theta|\theta^{(t)}) &\propto \exp\Big(-\frac{(\theta-\theta^{(t)})^2}{2c^2}\Big),\\
P(\gamma|\alpha,\theta,\underline{X_{(m)}^{(u)}}) &\approx f(\gamma|\gamma^{(t)}),\\
f(\gamma|\gamma^{(t)}) &\propto \exp\big(-(e^{-\frac{\gamma-\gamma^{(t)}}{s_0}}+\frac{\gamma-\gamma^{(t)}}{s_0})\big).
\end{aligned}
$$

The Metropolis-Hastings algorithm follows. All the variables are the ones defined above in this section.

Algorithm 3 Metropolis-Hastings Quasi-Conjugate Sampling

Starting values:= $\boldsymbol{\theta}^{(0)} = (\alpha^{(0)}, \theta^{(0)}, \gamma^{(0)})$

for $t = 1, \cdots, T$ **do**

Random Walks: Compute:

$$\alpha_* \sim Gamma(a_x, b_x) \quad R_\alpha.$$

Accept or reject:

sample $U_1 \sim Unif(0, 1)$

if $U_1 \leq R_\alpha$ **then**

$\alpha^{(t)} = \alpha_*$

else

$\alpha^{(t)} = \alpha^{(t-1)}$

end

Compute:

$$\theta_* \sim N(\theta^{(t-1)}, \sigma^2) \quad R_\theta.$$

Accept or reject:

sample $U_2 \sim Unif(0, 1)$

if $U_2 \leq R_\theta$ **then**

$\theta^{(t)} = \theta_*$

else

$\theta^{(t)} = \theta^{(t-1)}$

end

Compute:

$$\gamma_* \sim Gumbel(\gamma^{(t)}, s_0) \quad R_\gamma.$$

Accept or reject:

sample $U_3 \sim Unif(0, 1)$

if $U_3 \leq R_\gamma$ **then**

$\gamma^{(t)} = \gamma_*$

else

$\gamma^{(t)} = \gamma^{(t-1)}$

end

Check that the new points are within the support.

$$\left(\frac{x_i - \gamma^{(t)}}{\theta^{(t)}}\right) > 0 \ \forall x_i,$$

$$\left(\frac{u - \gamma^{(t)}}{\theta^{(t)}}\right) > 0.$$

end

Result: Markov Chain, $\{\underline{\theta^{(t)}}\}_{t=1}^{T}$

4.5 Posterior Based Estimates

Bayesian Inference has the advantage that the posterior distribution summarises parameter uncertainty. This is useful for Extreme Value Limit distribution because extrapolation does include parameter uncertainty. It is possible to estimate parameters and other quantities of interest via Monte Carlo Integration. The simulated Markov Chain is a sample from the joint posterior distribution [3, 19].

$$\widehat{\underline{\Theta}}_{MC} = (\underline{\theta^{(1)}}, \cdots, \underline{\theta^{(T)}}).$$

This Markov Chain is a sample of size T. By the Ergodic Theorem for MCMC, it is possible to estimate via Monte Carlo Integration. This is because samples form a MCMC, and henceforth are not independent, nonetheless, the Ergodic Theorem makes it possible to estimate these Bayes estimators for squared error loss as follows.

$$\hat{\alpha}_{BS} = \frac{1}{T}\sum_{t=1}^{T} \alpha^{(t)},$$

$$\hat{\theta}_{BS} = \frac{1}{T}\sum_{t=1}^{T} \theta^{(t)},$$

$$\hat{\gamma}_{BS} = \frac{1}{T}\sum_{t=1}^{T} \gamma^{(t)}.$$

Furthermore, histograms and empirical distributions of the simulated sample depict the posterior distribution. Therefore, the posterior distribution does not only provide a possible centre of mass, but also shows the spread of the distribution. If the posterior appears to have a known shape and concentrate in a particular region, then if one

accepts the proposed models as being valid, the posterior distribution either makes parameter uncertainty unmanageable or reduces it to workable settings [28, 29].

$$\begin{aligned} I^{\alpha}_{1-\eta} =& P(\alpha \in I|\underline{X^{(u)}_{(m)}}) = 1-\eta, \\ I^{\theta}_{1-\eta} =& P(\theta \in I|\underline{X^{(u)}_{(m)}}) = 1-\eta, \\ I^{\gamma}_{1-\eta} =& P(\gamma \in I|\underline{X^{(u)}_{(m)}}) = 1-\eta. \end{aligned}$$

Similarly for any function g of the parameters under squared error loss.

$$\widehat{g(\underline{\theta})} = \frac{1}{T}\sum_{t=1}^{T} g(\underline{\theta^{(t)}}).$$

This is particularly useful to estimate the posterior predictive distribution.

$$\begin{aligned} f(X|\underline{X^{(u)}_{(m)}}) &= \int_{\underline{\Theta}} f(X|\underline{\theta})P(\underline{\theta}|\underline{X^{(u)}_{(m)}})d\underline{\theta}, \\ &= \int_{A}\int_{\Theta}\int_{\Gamma} f(X|\alpha,\theta,\gamma)P(\alpha,\theta,\gamma|\underline{X^{(u)}_{(m)}})d\gamma d\theta d\alpha. \end{aligned}$$

Consider estimating the posterior probability that X is below a value of interest x_m, that is:

$$\widehat{P(X \leq x_m)}_{BS} = \frac{1}{T}\sum_{t=1}^{T} P(X \leq x_m|\underline{\theta^{(t)}}).$$

This calculation of percentiles accounts for parameter uncertainty by computing a weighted average of the possible distribution functions. For example, consider the task at hand is to estimate

$V@R_\upsilon(X) = \{x : P(X \leq x) = 1 - \upsilon\}$, then a Bayesian estimate accounts for parameter uncertainty rather than plugging in a point estimate of $\underline{\theta}$. Finally, simulations from the posterior predictive distribution can be computed by taking one observation as the mean of a sample from different possible distributions because the mean minimises the squared error loss and takes advantage of the Ergodic Theorem. The algorithm samples J observations from the distribution $f(x|\underline{\theta^{(t)}})$ and takes the simulation of the posterior predictive distributions as the mean of each sample of the distributions $f(x|\underline{\theta^{(t)}})$ [19, 3, 6].

$$X_j \sim f(x|\underline{\theta^{(t)}}),$$
$$X_k = \frac{1}{J}\sum_{j=1}^{J} x_j.$$

Then $X_k \sim f(x|X_{(m)}^{(u)})$, thus, one repeats the algorithm K times to get a sample of size K from the posterior predictive distribution. These samples are useful for goodness of fit tests and statistical learning techniques [18, 4].

4.6 Goodness of Fit Tests

Goodness of Fit Tests only measure how compatible the estimated model is to the data. If the assumed model is the real process that generated the data - if such a thing exists - then hypothesis tests are powerful tools. However, since knowing the real model is impossible hypothesis tests can only measure what model from a proposed collection is more compatible with the observed data. This means that one can find the model from

an assumed collection that best fits the observed data [24, 30].

It is important that to understand that hypothesis tests do not find the best model for the process that generates the data; instead, hypothesis tests find a model compatible with observed data from an assumed model collection. Therefore, if the assumed collection is not a reasonable model collection, even though hypothesis tests will find the "best" model, no model in the collection could be a reasonable choice. For example, if the data come from a process with a Heavy-Tail, but the observed sample lacks evidence of a Heavy-Tail one might fit an Exponential Tail to the data and it might even appear to be a good fit; nonetheless, future observations will include large deviations and the fitted Exponential Tail model will perform poorly in out of sample prediction. This is due to the data rather than a "bad" Goodness of Fit test.

This inherent difficulty in model selection and estimation forces statisticians to perform new inferences when new information is available. Hence, the statistical inference practice resembles the Bayesian philosophy of updating current beliefs (knowledge) with new information [30, 18].

On this basis, the Goodness of Fit tests this work proposes are based on out of sample prediction. To carry out these tests it is necessary to split the observed sampled into a training set and a validation set. The training set is used to fit the model as the previous sections in this chapter propose and the validation set is used to test the fitted model [4].

Theorem (Probability Integral Transform [6])**.** *Let X be an absolutely continuous random variable (i.e., $F(x)$ is differentiable monotone increasing.) Then $Y = F(X)$ is a standard uniform random variable.*

Proof. Since F is monotone increasing it has an inverse; furthermore, the inverse is a monotone increasing continuous function. Also, $0 \leq F(X) \leq 1$ for all $X \in \mathbb{X}$; thus, $0 \leq Y \leq 1$. Hence, for $y \in [0,1]$.

$$\begin{aligned} P(Y \leq y) &= P(F(X) \leq y), \\ &= P(X \leq F^{-1}(y)), \\ &= F(F^{-1}(y)), \\ &= y. \end{aligned}$$

Therefore the distribution function is the distribution function of a standard uniform random variable.

$$F_Y(y) = y\chi_{[0,1]}(y).$$

■

The preceding theorem enables simulation of any continuous random variable by simulating standard uniform random variables. Discrete random variables can be simulated by use of the generalised inverse,

$$\overleftarrow{F}(y) := \inf\{x : F(x) \geq y\}$$

[6]. This function looks at jumps and uses the range between jumps as the probability of that particular value. The author thinks that this theorem should be renamed as the Fundamental Theorem of Simulation because it is the basis of almost all simulation techniques. Now that it is possible to simulate samples from the posterior predictive distribution and plug-in posterior distributions, it is possible to carry out the following Goodness of Fit Tests, which are empirical rather than formal because they lack a formal Hypothesis Test

formulation.

Definition (Validation Set [4])**.** *A validation set of size m_v, usually 20% of the observed sample according to [4], are observations that are not used to fit the model.*

$$\underline{X}^{(vs)}_{(m_v)} := \{X_i : X_i \notin \underline{X}^{(u)}_{(m)}, X_i \in \underline{X}_{(n)}\}.$$

Because of the Probability Integral Transform Theorem; if the model is reasonable, then the probability integral transform of the validation set sample should resemble a standard uniform distribution.

$$\begin{aligned} P(X^{vs}|\underline{X}^{(u)}_{(m)}) &= \int_{-\infty}^{X^{vs}} f(x|\underline{X}^{(u)}_{(m)})dx, \\ &= \int_{-\infty}^{X^{vs}} \int_{\underline{\Theta}} f(x|\underline{\theta})P(\underline{\theta}|\underline{X}^{(u)}_{(m)})d\underline{\theta}dx, \\ &= F_{X|\underline{X}^{(u)}_{(m)}}(X^{vs}), \\ &= pit(X^{vs}). \end{aligned}$$

Thus, if the fitted model is a "good" fit the Probability Integral Transform should resemble a uniform distribution [18]. This test computes a large simulated sample from the posterior predictive distribution and assigns the $pit(X^{vs}_j)$ as the number of simulations smaller than X^{vs}_j divided by the total number of simulations K. Where $x^{(s)} \sim f(x|\underline{X}^{(u)}_{(m)})$.

$$pit(X^{vs}_j) = \frac{1}{K}\sum_{k=1}^{K} \chi(x^s_k)_{(-\infty, X^{vs}_j]}.$$

If the fitted model is a reasonable fit, then a histogram of the $pit(X_j^{vs})$ should look flat between $[0, 1]$. The other Goodness of Fit test is to compare test statistics from simulated samples with the same test statistics computed from the validation set sample [4, 19, 3]. Some possible test statistics are:

$$T^{(1)}(\underline{X_{(K)}}) = \sum_{k=1}^{K} \ln\left(\frac{x_k}{u}\right),$$
$$T^{(2)}(\underline{X_{(K)}}) = \min\{\underline{X_{(K)}}\},$$
$$T^{(3)}(\underline{X_{(K)}}) = V@R_{\upsilon}(X).$$

Among other possible statistics that are characteristic of Heavy-Tails. Each statistic is computed from a simulated sample of size K for n independent simulated samples. Good fits are those where the $T^{(i)}$ fall near the centre of the distribution of the simulations $T^{(i)}$. In essence, if the fit is good, then the observed $T^{(i)}$ should not be extremely unlikely events. Therefore, if the observed $T^{(i)}$ fall at the boundaries of the simulations' distribution, there is evidence of non-compatibility; nevertheless, this does not mean that model is or is not appropriate. All of these tests are sample and model dependent, as well as empirical [19, 3, 4].

Chapter 5

Simulation Experiment

All the hard work from the previous chapters is cultivated in this chapter. A big advantage of simulation studies is that it is possible to compare the fitted model with the "original" model. Inference techniques should perform at least as well with simulated data if not better than with real data [18, 19, 3]. This chapter illustrates the concepts from previous chapters with an application to a simulation study. The experiment consists of three Heavy-Tail simulations and a simulation of a random variable that is a mixture of a Heavy-Tail and an Exponential Tail. The last simulation tests the Gibbs Threshold Search. Without further ado the reader might want to jump into the action.

5.1 Heavy-Tail Data Simulation

The Probability Integral Transform Theorem (chapter 5.6) enables one to simulate independent identically distributed Fréchet random

variables. The experiment simulates a sample of size 500 with Maxima taken every 10 observations; thus, the simulation is 50 periods long.

Proposition. *The period Maxima only adjusts the scale by a factor of* $n^{\frac{1}{\alpha}}$ *where* n *is the period size.*

Proof. Let $\{X_i\}_1^n$ be independent identically distributed Fréchet random variables, then for $X^{(n)} = \max\{X_i\}_1^n$,

$$\begin{aligned} P(X^{(n)} \leq x) &= \exp\left(-\left(\frac{x-\gamma}{\theta}\right)^{-\alpha}\right)^n, \\ &= \exp\left(-n\left(\frac{x-\gamma}{\theta}\right)^{-\alpha}\right), \\ &= \exp\left(-\left(\frac{x-\gamma}{\theta n^{\frac{1}{\alpha}}}\right)^{-\alpha}\right). \end{aligned}$$

Thus, the period Maxima has the same shape (tail index), but Extremes are pushed by a factor of $n^{\frac{1}{\alpha}}$. The bigger the size of the period the bigger $\theta n^{\frac{1}{\alpha}}$ becomes to stabilise large deviations. ■

The following algorithm simulates a sample of independent identically distributed random variables.

Algorithm 4 Fréchet Simulation

Fix the parameters: $\alpha = \alpha_0$, $\theta = \theta_0$, $\gamma = \gamma_0$
for $i = 1, \cdots, m$ **do**
  Simulate single observation:

$$q \sim Unif([0,1])$$
$$x_i = \theta_0(-\ln(q)^{-\frac{1}{\alpha_0}}) + \gamma_0$$

end
Result: Independent Sample of size: m, $\underline{X_{(m)}}$

Similarly the percentile function of the Pareto distribution is:

$$q = 1 - \left(\frac{x}{u}\right)^{-\alpha} \iff x = u(1-q)^{-\frac{1}{\alpha}}.$$

Thus, it is possible to easily simulate Pareto random variables. Note that the Pareto distribution belongs to the Fréchet's domain of attraction.

$$\begin{aligned}\lim_{t\to+\infty} \frac{S(xt)}{S(t)} &= \lim_{t\to+\infty} \left(\frac{xt}{u}\right)^{-\alpha} \left(\frac{u}{t}\right)^{-\alpha}, \\ &= \lim_{t\to+\infty} \left(\frac{u}{u}\right)^{-\alpha} \left(\frac{t}{t}\right)^{-\alpha} x^{-\alpha}, \\ &= x^{-\alpha} \quad x > 0.\end{aligned}$$

This is no surprise because the Pareto Tail is a Power Law. In fact it is the tail that approximates Heavy-Tails. The experiment also simulates a sample of size 300 from a Pareto distribution with known threshold and tail index u_0, α_1. This sample is $\underline{Y_{(300)}}$, the model will fit the Poisson Point Process likelihood. The last Heavy-Tail ($Z \geq 0$) simulation comes

from a log-logistic distribution with tail index $\alpha > 0$ and scale $\beta > 0$. The distribution function is.

$$F(z) = \frac{z^\alpha}{\beta^\alpha + z^\alpha}.$$

The sample size is 200. The log-logistic distribution belongs to the domain of attraction of the Fréchet Distribution. The tail is:

$$1 - F(z) = \frac{\beta^\alpha}{\beta^\alpha + z^\alpha}.$$

The limit, rate of decay for large deviations, is:

$$\begin{aligned}
&\lim_{t\to+\infty} \frac{S(xt)}{S(t)} = \lim_{t\to+\infty} \frac{\beta^\alpha}{\beta^\alpha + (xt)^\alpha} \frac{\beta^\alpha + t^\alpha}{\beta^\alpha},\\
&= \lim_{t\to+\infty} \frac{\beta^\alpha + t^\alpha}{\beta^\alpha + t^\alpha x^\alpha},\\
&= \lim_{t\to+\infty} \frac{1 + (t/\beta)^\alpha}{1 + (xt/\beta)^\alpha}.\\
&\textit{Indeterminate type } \text{"}\frac{\infty}{\infty}\text{"}.\\
&\frac{d}{dt} 1 + (t/\beta)^\alpha = \frac{\alpha}{\beta}\left(\frac{t}{\beta}\right)^{\alpha-1},\\
&\frac{d}{dt} 1 + (xt/\beta)^\alpha = \frac{\alpha x}{\beta}\left(\frac{xt}{\beta}\right)^{\alpha-1},\\
&\lim_{t\to+\infty} \frac{\frac{\alpha}{\beta}\left(\frac{t}{\beta}\right)^{\alpha-1}}{\frac{\alpha x}{\beta}\left(\frac{xt}{\beta}\right)^{\alpha-1}} = \lim_{t\to+\infty} \frac{(t/\beta)^{\alpha-1}}{(t/\beta)^{\alpha-1}} x^{-\alpha},\\
&= x^{-\alpha}.\\
&\therefore \lim_{t\to+\infty} \frac{S(xt)}{S(t)} = x^{-\alpha}.
\end{aligned}$$

Therefore the log-logistic has a Heavy-Tail. The quantile function is:

$$F(z) = q \iff \beta\left(\frac{q}{1-q}\right)^{\frac{1}{\alpha}} = z.$$

The last sample the experiment simulates is a mixture between a Gamma distribution and a Pareto distribution. The random variable is a Gamma for values smaller than or equal to the threshold u_1 with probability p and a Pareto for values larger than u_1 with probability $1-p$. The tail is a Pareto, but not all simulations come from a Heavy-Tail. The main test for this simulation is the Gibbs Threshold Search. Simulate $W \sim F(w)$ where.

$$F(w) = \begin{cases} Gamma(a_0, b_0) & \textit{with probability} \quad p, \\ Pareto(u_1, \alpha_3) & \textit{with probability} \quad 1-p. \end{cases}$$

The sample size is 500. To simulate these data, the algorithm simulates a uniform random variable; if this number is smaller than p, then the algorithm simulates an observation from the truncated Gamma distribution. If the uniform random variable is larger than or equal to p the algorithm simulates an observation from the Pareto distribution.

5.2 Fitting the Simulated Data

This section makes Bayesian Inference and probabilistic modelling come alive. Note that all of the R code used to compute these quantities is available at https://github.com/dansal182/Thesis-Code, the code only uses the gtools library to simulate Dirichlet random variables, comments and observations are welcomed; the author apologises in advance for the

unstructured code. The four simulations are:

$$\underline{X_{(500)}} = \textit{Fréchet random sample.}$$
$$\underline{Y_{(300)}} = \textit{Pareto random sample.}$$
$$\underline{Z_{(200)}} = \textit{Log-logistic random sample.}$$
$$\underline{W_{(500)}} = \textit{Mixture random sample.}$$

Because this is a simulation experiment it is true that

$$X_i \sim Frechet(\alpha_0, \theta_0, \gamma_0),$$
$$Y_i \sim Pareto(\alpha_1, u_0),$$
$$Z_i \sim LL(\alpha_2, \beta_0),$$
$$W_i \sim F(w|\alpha_3, u_1).$$

The sample sizes above the thresholds are:

$$\underline{X_{(500)}^{(100)}} = 68,$$
$$\underline{Y_{(300)}^{(32)}} = 26,$$
$$\underline{Z_{(200)}^{(15)}} = 2,$$
$$\underline{W_{(500)}^{(11)}} = 25.$$

All parameters are known, nonetheless, their true values will be revealed after the experiment. All four random samples are fitted via EDA techniques, hypothesis tests, threshold search and Poisson Point Process likelihood. The fitted models are tested with new simulations from each distribution to carry out (Empirical) Goodness of Fit tests. The algorithms used are the ones section 5 proposes.

The EDA for the Fréchet random sample.

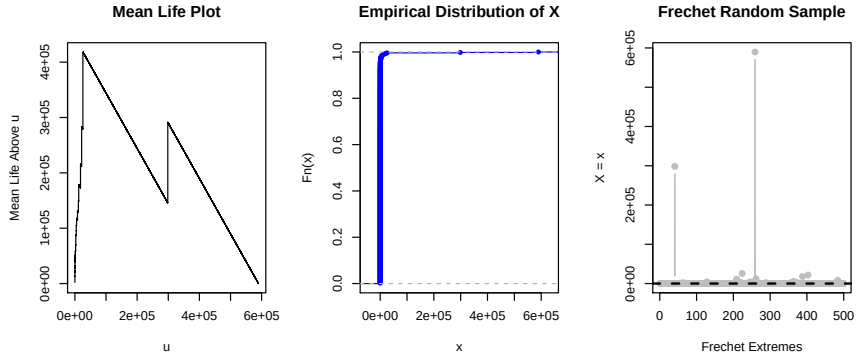

Figure 5.1. EDA Fréchet Random Sample.

The empirical distribution evidences that large deviations have significant probability; furthermore, the Mean Life Plot grows for values above 1000. Also, the observations plot shows plenty of extremes are larger than 100. This EDA suggests that the data have a Heavy-Tail. The next step is to test the data for a Heavy-Tail. The first test is the ratio test from section (5.2). It is done with a diffuse prior for $\alpha \sim Gamma(1, 1/10)$. The test R output results in.

```
Fail to Reject H0, no evidence against a Heavy-Tail.
BF = 0.
```

The second test fits the tail index of the GEV with a flat prior for the tail index ξ. This test is the interval test in section (5.2). This test is done taking the Maxima every ten observations (period size 10). If the Posterior Distribution has little or no mass (probability) in a neighbourhood of 0, then there is evidence that the data have a Heavy-Tail. The figure below summarises the results.

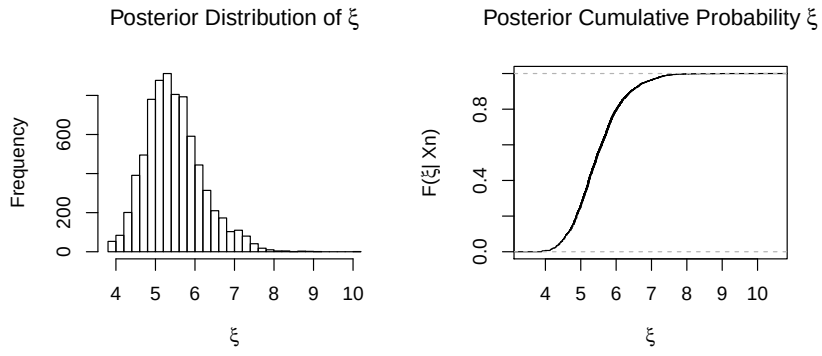

Figure 5.2. Tail Index Interval Test for the Fréchet Random Sample.

There is substantial evidence that these data have a Heavy-Tail. In fact the posterior percentiles of the tail index ξ are:

0%	25%	50%	75%	100%
3.891584	4.973710	5.392726	5.864544	10.058152

The Hypothesis Tests and EDA evidence that the data have a Heavy-Tail; furthermore, the strong tests' results evidence that this data might even not have a mean because there are plenty of large deviations and the tail index ξ posterior distribution is not near zero. On this basis, it is reasonable to fit the model subsection 5.4 proposes.

The Metropolis Hasting algorithm from subsection 5.4 fits the extremes at the selected thresholds and "observed" number of periods. The Poisson Point Process likelihood fits the model to peaks over the threshold $u = 100$ and period size ten.

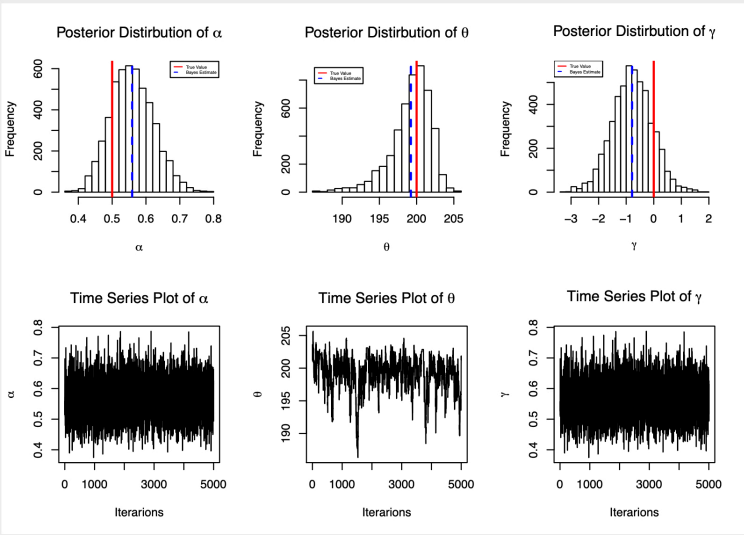

Figure 5.3. Fréchet Results.

The simulations the MH algorithm produces are enough to derive Posterior Based Estimates. These estimates are studied in the next section. It appears the the model does mix properly with the Fréchet Random Sample.

The experiment fits the other simulations with the same methodology. For data with an explicit Heavy-Tail, such as the Pareto Random Sample the threshold search does not distinguish a best threshold; instead, the threshold search suggests all thresholds are fair proposals. The EDA of the Mixture is interesting because the sample shows its mixture characteristics. It is interesting to study the mixture data for evidence of a possible Heavy-Tail.

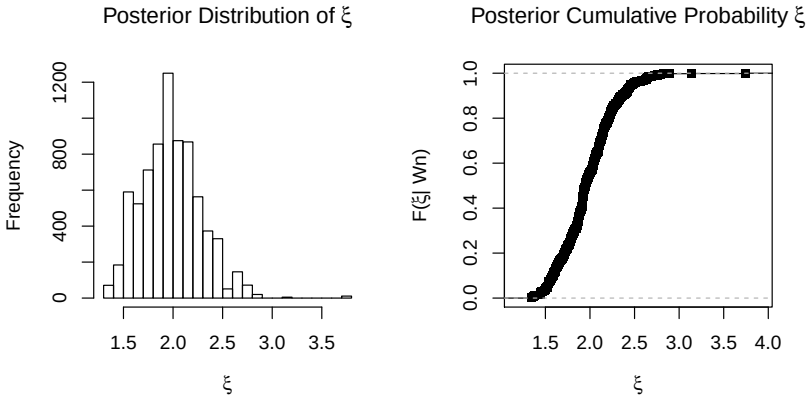

Figure 5.4. Interval Test for the Mixture Data.

The Interval tests suggest a Heavy-Tail; furthermore, the results of the ratio test are:

```
Fail to Reject H0, no evidence against a Heavy-Tail.
 BF = 5.294438e-67.
 95% Credible Test Interval
       5%          95%
   1.528103    2.459888
```

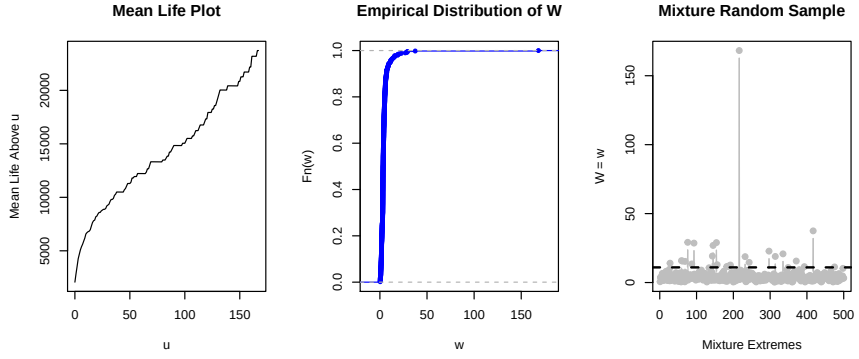

Figure 5.5. EDA of the Mixture Random Sample.

The EDA results show that the Mean Life Plot appears to be linear for values larger than 5; nevertheless, the empirical distribution function and observations plot show that the data include large deviations much bigger than 5; thus, the Gibbs Threshold Search might be able to suggest a threshold these plots do not make apparent. The search is done with collections that start with small values. After increasing the proposals (larger thresholds) the search suggests 11 as the value that starts mixing with the Pareto Tail more than smaller values. Nevertheless, the mixture random sample does show that any threshold larger than or equal to 11 is preferred to smaller values.

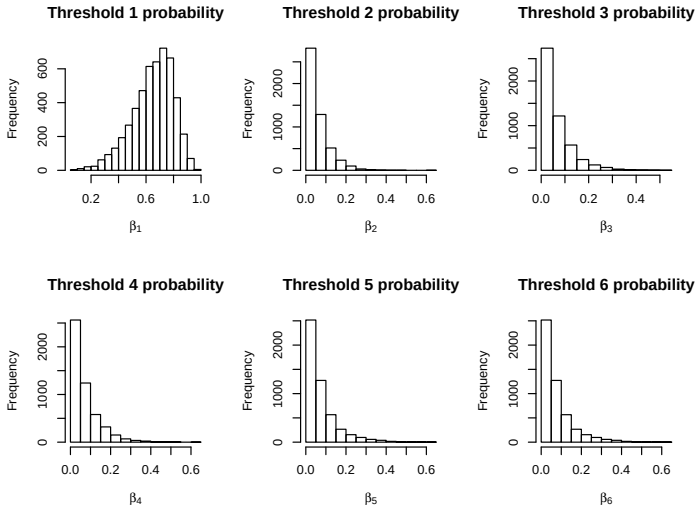

Figure 5.6. Gibbs Threshold Search for The Mixture Random Sample.

The proposed thresholds in figure 10 are $\{12, 1, 2, 4, 5, 7\}$. The interesting result is that for any value as big as 11 the search suggests that value; whereas, for values smaller than 11 the search does not concentrate posterior probability near the largest threshold that is being proposed.

The data are fitted with the threshold $u = 11$. The results of the MCMC algorithm fit of the Mixture Data are below.

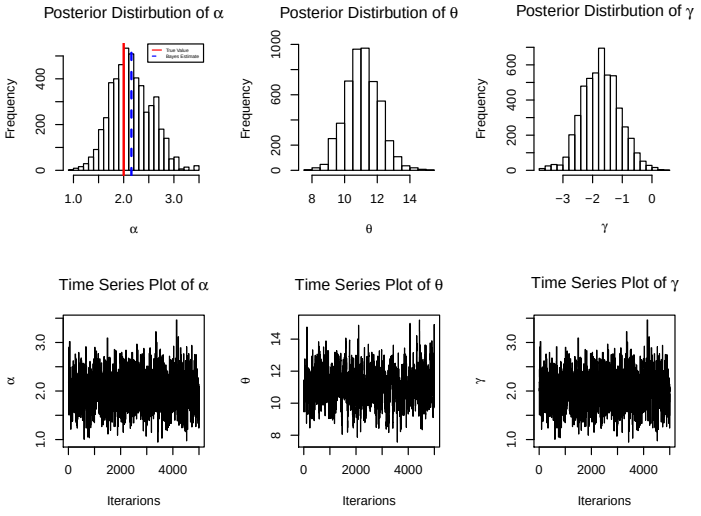

Figure 5.7. Mixture Data Posterior Distribution Simulation.

This is interesting because 11 is bigger than the sample's 95% empirical percentile; nonetheless, when fitting the model with smaller or larger thresholds the estimate does not concentrate as near to the tail index as it does for the threshold $u = 11$. Even though not all of the sample comes from a Heavy-Tail, the Gibbs Threshold Search and MH Posterior Distribution Simulation are able to identify and fit the sub-sample that does come from a Heavy-Tail distribution. The QQ-plot that tests for normality shows no compatibility with a normal distribution; also, the QQ-plot of the log data appears to be linear for values larger than 11.

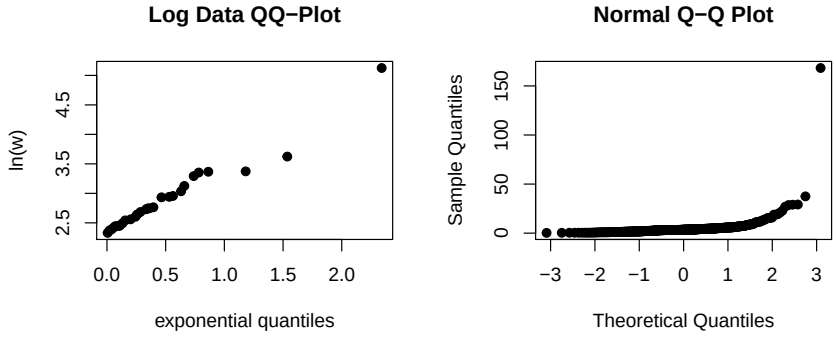

Figure 5.8. QQ-Plots Mixture Data.

Clearly there is no evidence of normality; furthermore, the log data appears to be linear. Hence, the Heavy-Tail model is a reasonable fit as the Posterior plot shows. The results of the Pareto simulation are.

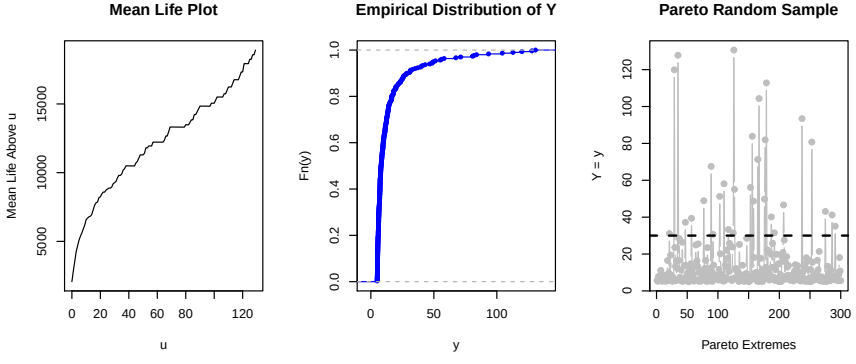

Figure 5.9. EDA Pareto Data.

The test results are:

```
Fail to Reject H0, no evidence against a Heavy-Tail.
BF = 7.706126e-51
```

```
Test Credible Intervals
      0%          25%          50%          75%         100%
1.964376     2.926547     3.273745     3.685684     7.262167
```

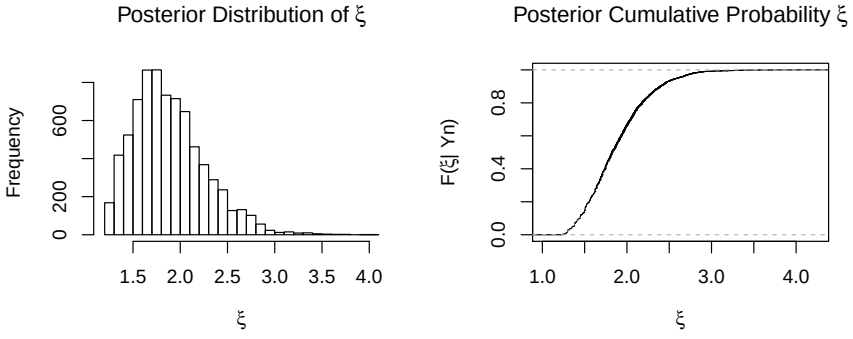

Figure 5.10. Interval Test Pareto Data.

No evidence of Heavy-tail. The EDA plots and threshold search accept a threshold value of 32 as a good mix. The MCMC fit is shown below.

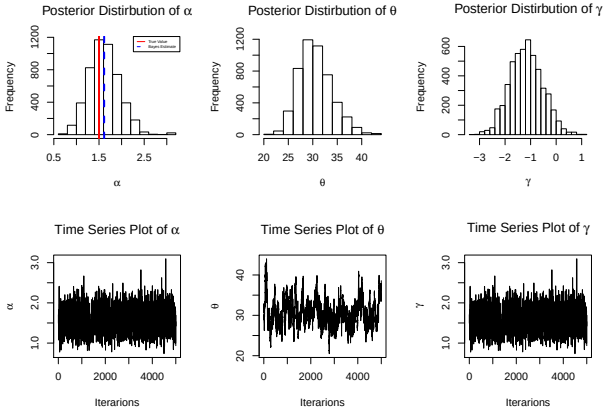

Figure 5.11. Posterior Distribution of the Pareto Data.

The experiment results for the Log-logistics sample are:

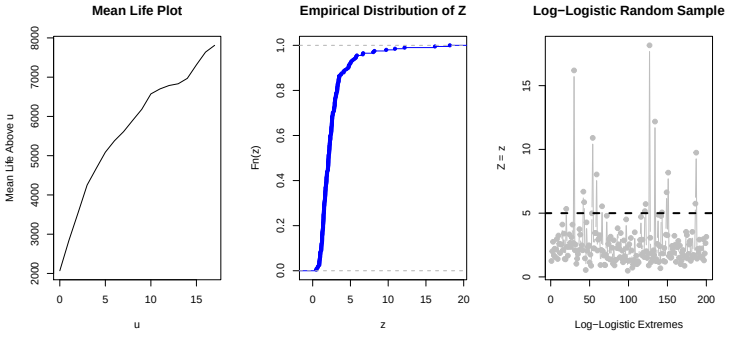

Figure 5.12. EDA Log-logis Data.

The test results are.

```
Fail to Reject H0, no evidence against a Heavy-Tail.
BF = 0.00303924
Test Credible Intervals
      0%          25%          50%          75%         100%
1.237934     1.611036     1.821019     2.095865     4.031079
```

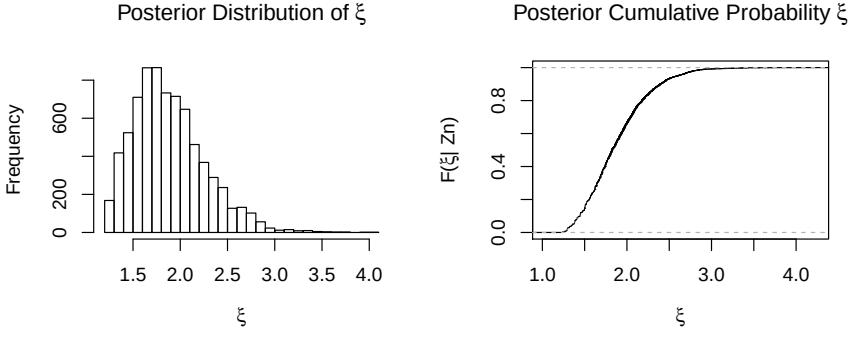

Figure 5.13. Interval Test Log-Logis Data.

No evidence against a Heavy-tail. The EDA plots and threshold search accept a threshold value of 15 as a good mix. The MCMC fit is shown below.

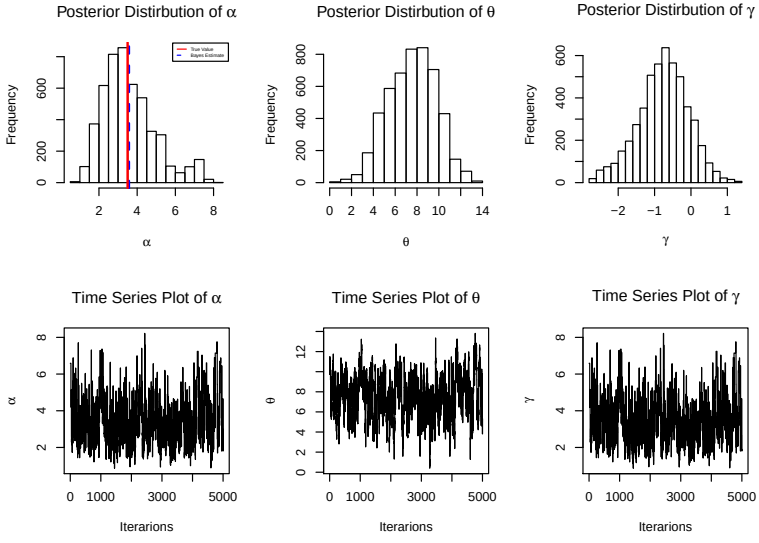

Figure 5.14. Posterior Distribution of the Log-Logis Data.

The test results and model appear to be reasonable fits to the four simulations; furthermore, the Log-logistic simulation has the lightest tail (largest tail index α). That ratio test does result in the largest Bayes Factor. It turns out that the Poisson Point Process likelihood and Peaks Over Thresholds mix well even if the sample size is not that big. This ability to extrapolate information from even small samples is a fascinating and useful characteristic of Heavy-Tail models.

5.3 Assessment of the Fitted Models

The Bayes estimate and 95% credible interval for the Fréchet Distribution with period size 10 are:

$$\begin{aligned}
\hat{\alpha}_{BS} &= 0.5601019, \\
I_{95\%}(\alpha) &= [0.4601248, 0.6659663], \\
\hat{\theta}_{BS} &= 198.5168, \\
I_{95\%}(\theta) &= [191.5671, 202.3902], \\
\hat{\gamma}_{BS} &= -0.7483885, \\
I_{95\%}(\gamma) &= [-1.8928437, 0.3790516].
\end{aligned}$$

The true value is $\alpha_0 = \frac{1}{2}$. The Bayes estimate is near this value, also, the length of the credible interval is 0.2 and its mid point is close to the Bayes estimate and real value. On this basis, the tail index is properly estimated with this methodology; furthermore, this distribution does not have a mean, hence, extrapolating to a Heavy-Tail model is necessary to study possible percentiles and functions of this random variable.

The other parameters of the 10 period Maxima are $\theta_0 = 200$ and $\gamma_0 = 0$. the confidence intervals include the true values; also, the Bayes estimates are reasonable approximations to the real values; nonetheless, percentiles of this distribution are extremely sensitive to the value of α, especially if $\alpha < 1$. The Goodness of Fit tests illustrate this. To test the probability integral transform the experiment simulates a sample from the Posterior Predictive Distribution and proceeds as in chapter (5.6). The result is:

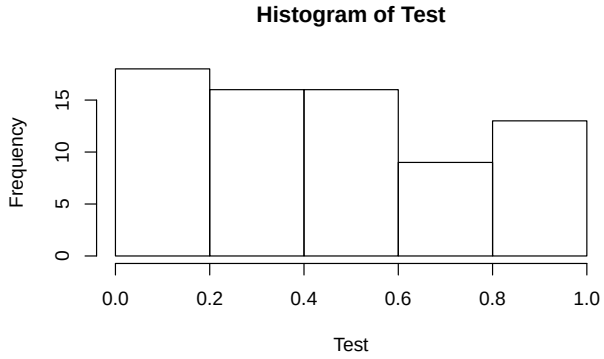

Figure 5.15. Histogram of The Probability Integral Transform Test.

The histograms intervals are:

```
  Min.    1st Qu.   Mean    3rd Qu.    Max.
0.0825  0.2047       0.4609   0.6471     0.9980 .
```

The Histogram, as expected, is not perfectly flat; nevertheless, the plot does resemble a flat rectangle. Therefore, there is some evidence that the Posterior Predictive Distribution and Test Sample result in a quasi-uniform distribution. Thus, the model appears to fit the data reasonably. The sensitivity percentiles have to the tail Index α is evident in the estimates of the 95% percentile.

$$\widehat{V@R_{95\%}}(X) = \frac{1}{T}\sum_{t=1}^{T}(\theta^{(t)}(-\ln(-0.95))^{-\frac{1}{\alpha^{(t)}}} + \gamma^{(t)}).$$

$$\widehat{V@R_{95\%}}(X) = 53421.86.$$

$$V@R_{95\%}(X) = 76016.66.$$

$$I_{95\%}(V@R_{95\%}(X)) = [16810.335, 130190.055].$$

The spread is a result of the Heavy-Tail; nonetheless, the true value is within the estimated interval. It is important to recall that Heavy-Tail Extremes are LARGE deviations. Hence, the analyst cannot expect to retrieve valuable information just from a point estimate. On that basis, the estimates should incorporate credible intervals to account for the worst case scenario. Also, these estimates should be adjusted to a new threshold if there is evidence that another threshold is a better fit, and/or the interest now is in regards to deviations larger than the new threshold. The the other samples' results are:

$$\begin{aligned}
\hat{\alpha_1}_{BS} &= 1.613611, \\
I_{95\%}(\alpha_1) &= [1.0867573, 2.1961265], \\
\alpha_1 &= 1.5. \\
\hat{\alpha_2}_{BS} &= 3.570445, \\
I_{95\%}(\alpha_2) &= [1.7351677, 6.5872371], \\
\alpha_2 &= 3.5. \\
\hat{\alpha_3}_{BS} &= 2.210936, \\
I_{95\%}(\alpha_3) &= [1.5196152, 3.0755410], \\
\alpha_3 &= 2.
\end{aligned}$$

All are reasonable estimates of the tail index α. This parameter will increase/decrease percentiles and other quantities of interest. With these estimates it is possible to estimate the Pareto Tail for different thresholds via the plug in method or accounting for uncertainty in parameter estimation via the Posterior Predictive Distribution [3]. Notice that uncertainty with respect to α increases significantly uncertainty estimating functions and percentiles of the random variable; nonetheless, the Bayesian methods this sections presents

capture the true value and provide useful information to prepare systems for eventual shocks that manifest as LARGE deviations.

The main objective of Heavy-Tail analysis is to estimate extreme value percentiles. The experiment estimates the 95% percentiles of the Pareto Tail approximation (Peaks Over Threshold) for values larger than $u = 10$.
Fréchet Simulation:

$$\begin{aligned} X^{(95\%)} &= 4000, \\ \widehat{X^{(95\%)}} &= 3022.895, \\ I_{95\%}(X^{(95\%)}) &= [933.1911, 7357.3833]. \end{aligned}$$

Pareto Simulation:

$$\begin{aligned} Y^{(95\%)} &= 73.68063, \\ \widehat{Y^{(95\%)}} &= 86.137, \\ I_{95\%}(Y^{(95\%)}) &= [42.77332, 175.23071]. \end{aligned}$$

Log-Logistic Simulation:

$$\begin{aligned} Z^{(95\%)} &= 23.53547, \\ \widehat{Z^{(95\%)}} &= 30.65674, \\ I_{95\%}(Z^{(95\%)}) &= [16.20307, 59.21696]. \end{aligned}$$

Mixture Simulation:

$$
\begin{aligned}
W^{(95\%)} &= 44.72136, \\
\widehat{W^{(95\%)}} &= 49.15939, \\
I_{95\%}(W^{(95\%)}) &= [30.09405, 81.91745].
\end{aligned}
$$

The beauty of the model is that these estimates can be computed for any threshold u. The estimated values are close to the true percentile, even though, none of the distributions was estimated with the threshold $u = 10$; furthermore, the posterior distribution of the tail index reflects the variability of each random variable. It is no accident that the random variable X with the heaviest tail reports the most uncertainty in parameter estimation; whereas, the random variable with the lightest tail (i.e., W) reports the least uncertainty in parameter estimation.

This adjustment to the inherent volatility a Heavy-Til has is a nice property of Bayesian methods. It also makes it feasible to estimate the Pareto tail for different thresholds of interest. After all, these estimates are what the analyst is looking for.

Chapter 6

Final Comments and Observations

6.1 Discussion

The 21st century presents challenges that are complex in nature. For example, climate change, tropical storms, pandemics and financial volatility. These challenges are a consequence of shocks to a complex system, such as temperature spikes, market crashes and energy shortages. These shocks impact the complex system in such an extreme way that they alter the status quo. These systems have lots of complex relationships, therefore, they usually are extremely hard to model; nevertheless, there is a need to study large deviations that affect the status quo and alter the system's functionality.

Heavy-Tail analysis is the mathematical tool that is capable of studying these shocks. These techniques have become crucial to modern financial markets among other areas of application. The

remarkable Extremal Types Theorem makes it possible to study many different processes. If the analysis of data is done with statistical models, then it does not matter what model the analyst selects; the extremes' distribution is one of the three Extreme Value distributions or a degenerate one; furthermore, if there is evidence of a Heavy-Tail, then the limit is the Fréchet distribution and the tail resembles a Power Law.

The strong theoretical background of asymptotic distributions provides a basis for statisticians, as well as data analysts to make inferences based on proven statistical methodologies such as Maximum Likelihood and Bayesian analysis. Furthermore, these strong limit results make it possible to extrapolate valuable information from small samples; moreover, these results show that the heavier the tail is the harder, if not impossible, it is to make reliable statistical inferences. Therefore, uncertainty is an inherent characteristic of Heavy-Tail analysis.

The Point Process construction is more complex than the classic construction, nevertheless, I think it is more precise and enlightening. The Point Process construction is built upon the tail's intensity measure. This construction makes it evident that sample size is not as relevant as the shape of the tail. Recall that the exponent of the Fréchet distribution is a Power Law (a consequence of polynomial decay) and that the exponent of the Gumbel distribution is an Exponential Tail (a consequence of exponential decay). The Power Law decay of a Heavy-Tail makes it a completely different model from a statistical model with an Exponential Tail.

A crucial difference is that the mean of a Heavy-Tail random variable might not exist, furthermore, if the mean does exist, then the

Power Law like tail does not allow probability to concentrate close to the mean. Thus, even though the mean might exist, observations will spread throughout the sample space rather than concentrate close to the mean. On the contrary, a random variable with an Exponential Tail does not allow probability to spread as much throughout the sample space. In fact, large deviation probabilities are exponentially bounded; thus, observations will concentrate close to the mean. As a matter of fact, an Exponential Tail makes the partial averages converge to the mean exponentially fast [5].

The careful reader will have noticed that the sample mean and sample variance have not been part of the statistical inference techniques this work proposes. This is not a coincidence, it is a consequence of Heavy-Tails. The spread of probability random variables with a Heavy-Tail have makes statistics based on moments and centres of mass, such as the sample mean much less reliable - if not useless - tools.

The Point Process construction does not only illustrate that the intensity measure is the key characteristic of a tail, it also justifies inferences based on the intensity measure. Since the intensity measure is independent from sample size it is not essential to have a large sample to justify the use of asymptotic models; instead, it is important to identify the type of decay (intensity measure). If there is not sufficient evidence to select one of the three intensity measures, then it is still possible to fit the Generalised Extreme Value and Generalised Pareto distributions. This flexibility makes Extreme Value models powerful tools because extremes, by definition, are unusual observations; thus, large sample sizes should not be expected.

The study of the unusual is challenging because Extreme Values are

rare; furthermore, it is difficult to estimate the proper type of Extreme Value Distribution. Also, Heavy-Tail models make it necessary to accept that there is more uncertainty involved in the estimation process; in fact, estimation and prediction are harder if not impossible for Heavy-Tail models. Bayesian analysis acknowledges the uncertainty present in any statistical methodology, nevertheless, a Bayesian analysis does not reduce/remove uncertainty; instead, it makes uncertainty manageable by accounting for it.

There has been an explosion of data analysis during the last few years. Most of these methods dismiss the existence of a Heavy-Tail. On that basis, most of these methods remove large deviations because the method considers large deviations from the mean to be outliers; unfortunately, removing outliers to accommodate model assumptions (Exponential Tail) makes inferences much less, if not useless, reliable - it only takes one shock (LARGE deviation) to make the model useless. For example, financial models that assume an Exponential Tail are known to perform poorly. The reality is that it is not possible to predict, without luck, accurately Heavy-Tail data. Nevertheless, Heavy-Tail techniques are capable of improving these models by providing a better fit at the tail.

To illustrate the ideas in the previous paragraph, consider the problem of Linear Regression. The objective is to see if there is a relationship between two variables. In essence, is the value of the response variable approximate to a linear function of the predictor variable.

$$Y = \beta_0 + \beta_1 x + \epsilon, \quad \epsilon \sim N(0, \sigma^2).$$

The strong assumption, more than the linear response, is that the errors have an Exponential Tail because if the errors have a Heavy-Tail the

value of Y will not concentrate near the linear response. Thus, there is no useful information in studying this relationship. A closer look at the square error $(y - \overline{y})^2$ helps understand this; if the observation y is far from the sample mean, then $(y - \overline{y})^2$ is big, thus, squared errors will affect y more than the linear response. This results in Y depending more on the random error than on the linear response. Hence, tools built on linear response assumptions would be inadequate.

Because Heavy-Tails do not amass probability in a specific region it is unwise to attempt point-wise prediction; instead, inference must focus on the intensity measure by estimating the tail index. The fact that Heavy-Tail analysis cannot provide reliable point-wise prediction does not mean that it is not useful, it only means that the focus has to be different; instead of looking at possible centres of mass and common behaviour the analyst should exploit Extreme Value models.

These models are capable of extrapolating valuable information from scarce observations, furthermore, estimations made with Extreme Value models focus on the shape of the tail. Thus, these estimates can be used to hedge against future shocks; in fact, it is possible to use these estimates to simulate future shocks to the system. These simulations can test response mechanisms, as well as motivate the development of better response mechanisms, such as stress tests to a new medicine that make evident the medicine needs to be more resistant to heat.

Additionally, it is possible to extend the Extremal Types Theorem to sequences of stationary data. If the next term in the sequence dependence on previous observations reduces significantly as the time difference increases, then it is possible to model clusters as independent. This result makes it possible to employ Extreme Value distributions without the independence assumption. It is potent

because real processes do not tend to be independent from realisation to realisation. Furthermore, it enables the study of periodic processes, such as seasonal temperature fluctuations. On that basis, there is ongoing research on how to generalise Heavy-Tail methods to a general process, the tail of a Markov process and spatial extremes among many others [1, 8].

Extreme Value Theory is a powerful tool, however, it is not flawless. It is important to remember that the real world is a mixture of complex relationships that can, at best, be approximated by mathematical models; therefore, statistical models should not be viewed as absolute truths but as a way of accounting for uncertainty and gaining deeper understanding of certain phenomena.

The more unlikely the events the model tries to approximate the harder it is to build a useful model. Therefore, inference makes subjective assumptions a part of the process. Also, models need to be updated when new information becomes available. This approach is crucial to better prepare for climate change, developments in bio-statistics and other systems where data analysis is becoming common practice.

It is true that no statistical model will fit the data perfectly, much less describe fully the phenomena under study, yet this does not mean that statistical analyses are useless; on the contrary, these models have the capacity of reducing uncertainty by providing useful information that is not trivial, such as robust estimates of fatalities during a pandemic - these estimates aid policy makers to make more informed decisions.

The reliance on statistical models will only increase in the age of data. The job of the statistician is to make sense of data not to

eliminate uncertainty; thus, statistical models cannot be left in a vacuum or seen as an absolute truth; on the contrary, it is important to constantly update and test statistical models. Therefore, assumptions and limitations of these models should be understood as well as possible before making any inference. This is sufficient reason to build better models that are capable of extrapolating important features, such as the Tail of the assumed statistical model. Statisticians must know why models work and also why they break down.

6.2 Conclusion

The hardest job in the world is making decisions; of all decisions the ones with a lot of uncertainty are the most difficult to make. On that basis, it is desirable to have a methodology that aids decision makers. In fact, Bayesian Inference is heavily influenced by decision theory; nevertheless, I think it is not fair to compare institutional decisions to the ones of individuals.

Statistical Inference permits us to better understand complex systems and, if possible, reduce uncertainty by providing valuable information that help people make better decisions based on a deeper understanding. Unfortunately, there is no method that removes subjectivity from a statistical analysis, much less, if such a thing exists, a correct model for a given problem; nonetheless, if applied properly statistical methods are capable of reducing uncertainty and pushing research, as well as solving real world problems.

The flexibility and difficulty inherent in statistical analysis are fascinating. If the model is a reasonable fit, then the information the

estimates provide is useful and may help develop new technologies, for example, machine learning done right; on the contrary, if the statistical method is not a reasonable fit, then the information it provides is no better than random searches on the Internet.

This is why I like to think of statisticians as chefs and not cooks, to do a good statistical analysis it is necessary to understand that there is no perfect model nor recipe that one applies; instead, every problem is different and requires its due diligence, whether it is incorporating experts opinions, combining different techniques or admitting that there is too much uncertainty to produce reliable estimates; it is up to the statistician to make sense of this. The ingredients used to prepare a superb new dish are out there, it is just a matter of looking into the unknown with an open attentive mind.

There has not been an extensive use of Heavy-Tail analysis in conjunction with Bayesian techniques during the current data analysis boom, especially in the machine learning context. Nevertheless, the theory is robust and the computational tools built over the last few years are capable of enhancing the scope of Heavy-Tail analysis. I believe that a lot of these (Heavy-Tail analysis) ideas can be implemented/adapted to the current wave of computational and learning techniques.

Heavy-Tail analysis motivates the statistician to focus on Extreme Values instead of ignoring them. As an analogy, I think Extreme Values are like lottery tickets. Since we do not know if we will win, we should buy as many lottery tickets as possible because it only takes one to change things forever. Who knows the key to a brighter future might be in a better understanding of large deviations.

Bibliography

[1] Stuart Coles. *An Introduction to the Statistical Modeling of Extreme Values*. Springer series in statistics, London, 2007.

[2] Resnick Sidney, I. *Heavy Tail Phenomena: Probabilistic and Statistical Modeling*. Springer Series in operations research and financial engineering, New York, 2007.

[3] Reich Brian, J. and Gosh Sujit, K. *Bayesian Statistical Methods*. CRC Press, Boca Raton, Fl, 2019.

[4] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learnig*, chapter 7, pages 219–257. Springer series in statistics, 2 edition, 2017. doi: 10.1007/b94608.

[5] Jeffrey S. Rosenthal. *A First Look At Rigorous Probability Theory*. World Scientific, Singapore, 2 edition, 2016.

[6] Ross Sheldon, M. *Introduction to Probability Models*, chapter 5, pages "281–343". Elsevier, 2 edition, 2010.

[7] Shumway Robert, H. and Stoffer David, S. *Time Series Analysis and Its Applications*, chapter 5, pages 241–284. Springer series in statistics, 4 edition, 2016.

[8] Davis Richard, A. and Thomas Mikosch. Extreme value theory for garch processes. *Advances in Applied Probability*, 2009. doi: 10.1007/978-3-540-71297-8_3.

[9] Nassim Nicholas Taleb and Pasquale Crillo. Tail risk of contagious diseases. Unpublished Nature Physics article, 2020. URL https://www.academia.edu/42307438/Tail_Risk_of_Contagious_Diseases.

[10] Sean Van der Merwe. Bayesian extreme value analysis of stock exchange data. Technical report, Department of Mathematical Statistics and Actuarial Science, University of the Free State), 2014.

[11] J. Pickands. Statistical inference using extreme order statistics. *The Annals of Statistics*, pages 119–131, 1975.

[12] D.J. De Waal A. Verster and S. van der Merwe. Selecting an optimum threshold with the kullback-leibler deviance measure, 2014.

[13] Charlie Wood. Total reinsurance capital hit record 625bn high in 2019, 1 2020. URL https://www.reinsurancene.ws/total-reinsurance-capital-hit-record-625bn-high-in-2019/.

[14] Paul Thompson, Yuzhi Cai, Dominic Reeve, and Julian Stander. Automated threshold selection methods for extreme wave analysis. *Coastal Engineering*, 56:1013–1021, 10 2009. doi: 10.1016/j.coastaleng.2009.06.003.

[15] Jem Corcoran. Modelling extremal events for insurance and finance by paul embrechts; claudia klüppelberg; thomas mikosch. *Journal of the American Statistical Association*, 97, 03 2002. doi: 10.2307/3085797.

[16] Johannes Heiny and Thomas Mikosch. Almost sure convergence of the largest and smallest eigenvalues of high-dimensional sample correlation matrices, 01 2020. ResearchGate download.

[17] Ioannis Papastathopoulos, Kirstin Strokorb, Jonathan Tawn, and Adam Butler. Extreme events of markov chains. *Advances in Applied Probability*, 49, 10 2015. doi: 10.1017/apr.2016.82.

[18] Andrew Gelman and Yulin Yao. Holes in bayesian statistics, 2 2020. URL https://arxiv.org/abs/2002.06467. Arxiv article.

[19] Andrew Gelman, John B. Carlin, Harl S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2013. doi: 10.1201/b16018.

[20] Guillermo Grabinsky. *Measure Theory*. Universidad Nacional Autónoma de México, Facultad de Ciencias, Mexico City, 2016.

[21] William Feller. *An Introduction to Probability Theory and Its Applications*, volume 2, pages 429–526. John Wiley and Sons INC, New York, 1970.

[22] Nicolas Bouleau. Splendeurs et misères des lois de valeurs extêmes. *Revue Risques - Les cahiers de l'assurance*, 4:85–92, 1991. URL https://halshs.archives-ouvertes.fr/halshs-00008305.

[23] Myriam Charras-Garrido and Pascal Lezaud. Extreme Value Analysis : an Introduction. *Journal de la Societe Française de Statistique*, 154(2):66–97, 2013. URL https://hal-enac.archives-ouvertes.fr/hal-00917995.

[24] Bruno de Finetti, Maria Carla Galavotti, and Hykel Hosni. *Philosophical Lectures on Probability*. Springer, Dordrecht, 2008. doi: 10.1007/978-1-4020-8202-3.

[25] Sidney I. Resnick and Rishin Roy. Extreme values and choice theory. In *Extreme Value Theory and Applications*, chapter 19, pages 319–336. Springer, Boston, 1994. doi: 10.1007/978-1-4613-3638-9_19.

[26] Jan Sprenger. Testing a precise null hypothesis: The case of lindley's paradox. *Philosophy of Science*, 80(5):733–744, 2013. doi: 10.1086/673730.

[27] Christian P. Robert and George Casella. The metropolis—hastings algorithm. *Monte Carlo Statistical Methods*, page 267–320, 2004.

ISSN 1431-875X. doi: 10.1007/978-1-4757-4145-2_7. URL http://dx.doi.org/10.1007/978-1-4757-4145-2_7.

[28] Alexander Ly, Josine Verhagen, and Eric-Jan Wagenmakers. An evaluation of alternative methods for testing hypotheses, from the perspective of harold jeffreys. *Journal of Mathematical Psychology*, 72:43–55, Jun 2016. ISSN 0022-2496. doi: 10.1016/j.jmp.2016.01.003. URL http://dx.doi.org/10.1016/J.JMP.2016.01.003.

[29] J. Møller, A. N. Pettitt, R. Reeves, and K. K. Berthelsen. An efficient markov chain monte carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458, Jun 2006. ISSN 0006-3444. doi: 10.1093/biomet/93.2.451. URL http://dx.doi.org/10.1093/BIOMET/93.2.451.

[30] Stephen E. Fienberg. When did bayesian inference become bayesian ? *Bayesian Anal.*, 1(1):1–40, 03 2006. doi: 10.1214/06-BA101. URL https://doi.org/10.1214/06-BA101.

[31] Stephen E. Fienberg. Introduction to r.a. fisher on inverse probability and likelihood. *Statist. Sci.*, 12(3):161, 09 1997. doi: 10.1214/ss/1030037905. URL https://doi.org/10.1214/ss/1030037905.

[32] Alexander M. Mood, Franklin A. Graybill, and Duane C. Boes. *Introduction to The Theory of Statistics*. Mcgraw Hill, 2001. ISBN 0-07-044520-6.

[33] Roger Cooke and Daan Nieboer. Heavy-tailed distributions: Data, diagnostics, and new developments. *SSRN Electronic Journal*, 03 2011. doi: 10.2139/ssrn.1811043.

[34] Werner Hürlimann. Pareto type distributions and excess-of-loss reinsurance. *Int J. Research Reviews Appl. Sciences*, 18:1–11, 01 2014.

[35] R.G. Bartle. *Elements of Real Analysis*. A Wiley Arabook. John

Wiley & Sons Incorporated, 1982. ISBN 9780471063919. URL https://books.google.com.mx/books?id=C6o1AAAACAAJ.

[36] Hadley Wickham and Garret Grolemund. *R for Data Science*. O'Reilly Media Inc., 2017. ISBN 9781491910399. doi: 10.18637/jss.v077.b01.

[37] Allen B. Downey. *Think Bayes*. O'Reilly Media Inc., 2012. URL https://greenteapress.com/wp/think-bayes/.

Appendix

Notation: iff := if and only if;
u := Threshold;
$C_K^+(f)$:= The space of continuous non-negative functions with compact support;
$\Psi_N(f)$:= Laplace Functional for the Poisson/Point Process N;
$S(x)$:= Tail of the Random Variable;
α := Tail Index;
$\pi_0(\cdot)$:= Prior distribution.

Probability Basics

Definition (Probability Space [5, 6, 21])**.** *A measure space* $(\Omega, \sigma(\Omega), \mathbb{P})$ *is a probability space if the measure* $\mathbb{P}$ *satisfies:*

$$\begin{aligned} \mathbb{P} : \sigma(\Omega) &\longrightarrow [0,1], \\ \mathbb{P}(\emptyset) &= 0, \\ \mathbb{P}(\Omega) &= 1, \\ \mathbb{P}(\bigcup_{n=1}^{\infty} A_n) &= \sum_{n=1}^{\infty} A_n. \end{aligned}$$

The basic properties of probability measures are the following.

Proposition.

$$\begin{aligned}\mathbb{P}(A \cup B) &= \mathbb{P}(A) + \mathbb{P}(B) \ \ \emptyset = A \cup B,\\ \mathbb{P}(A^c) &= 1 - \mathbb{P}(A),\\ B \subset A &\Rightarrow \mathbb{P}(B) \leq \mathbb{P}(A).\end{aligned}$$

Definition (Conditional Probability [6])**.** *The conditional probability of A given B $(A|B)$ is:*

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(AB)}{\mathbb{P}(B)}.$$

Theorem (Bayes Theorem [24])**.** *Let E be the evidence of a hypothesis and H the hypothesis regarding the process that produces the evidence, then*

$$\mathbb{P}(H|E) = \frac{\mathbb{P}(E|H)\mathbb{P}(H)}{\mathbb{P}(E)}.$$

Thus the probability of the evidence is altered by the evidence for it.

Theorem (Total Probability [6])**.** *Let $\Omega = \cup_{i=1}^{n} A_i$, such that $A_i A_j = \emptyset$ if $i \neq j$, $\cup_{i=1}^{n} A_i$ is a partition of Ω and let $B \in \sigma(\Omega)$, then*

$$\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(B|A_i)\mathbb{P}(A_i).$$

For Bayes Theorem it is possible to write the probability of evidence as the sum of the probabilities of evidence given distinct hypothesis.

$$\mathbb{P}(H|E) = \frac{\mathbb{P}(E|H)\mathbb{P}(H)}{\sum_{i=1}^{n} \mathbb{P}(E|H_i)\mathbb{P}(H_i)}.$$

Definition (Random Variable [21, 5])**.**

$$X : (\Omega, \sigma(\Omega), \mathbb{P}) \longrightarrow (\overline{\mathbb{R}}, B_{\mathbb{R}}, \mu).$$

A random variable (unfortunate name) is not a number but a measurable function that assigns numeric values to possible outcomes of random experiments defined on a probability space, this function transforms the general probability space into a probability space over the Borel σ-algebra and assigns probabilities to numeric outcomes of the random experiment. The random variable X transforms $\omega \in \Omega$ to real numbers. The set $X(\Omega) = \mathbb{X}$ is the support (domain) of the random variable X.

Definition (Cumulative Distribution Function [5, 6])**.** *The distribution function of a random variable X is the function F_X that equals the probability of $X \in (-\infty, x]$ for some number $x \in \mathbb{R}$.*

$$\begin{aligned}
F_X : \overline{\mathbb{R}} &\longrightarrow [0,1], \\
P(X \leq x) &= \mu((-\infty, x]), \\
&= \int_{(-\infty, x]} d\mu \ \textit{Measure Integral}, \\
&= \int_{-\infty}^{x} dF_X \ \textit{Riemann-Stieljes Integral}, \\
&= F_X(x) - \lim_{x \to -\infty} F_X(x), \\
&= F_X(x), \\
0 \leq F_X(x) &\leq 1 \ \ \forall x \in \mathbb{R}.
\end{aligned}$$

The next Lemma is proved rigorously in [5].

Lemma (Weak Convergence Equivalent Definitions)**.** *The convergence of non-negative bounded continuous functions $C_b^+(\mathbb{X})$ is equivalent to the convergence of bounded continuous functions $C_b(\mathbb{X})$ and of non-negative continuous functions with a compact support $C_K^+(\mathbb{X})$.*

Basic Bayesian Inference

Definition (Prior Distribution)**.** *The a-priori (before the incorporation of the sample into the model) distribution of the unknown model and/or parameter $\pi_0(\theta)$ [3].*

$$\int_{\Theta} \pi_0(\theta) d\theta = 1.$$

Definition (Posterior Distribution [3, 19])**.**

$$P(\theta|\ \underline{X_{(n)}}) = \frac{L(\underline{x_{(n)}}|\theta)\pi_0(\theta)}{\int_{\Theta} L(\underline{x_{(n)}}|\theta)\pi_0(\theta)d\theta}.$$

This distribution is the posterior distribution of θ. This is an updated belief system about the hypothesis (Model) that is linked to the data.

Definition (Loss function)**.** *The loss function quantifies the loss of an estimator [3]. Two examples are:*

$$L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2,$$
$$L(\hat{\theta}, \theta = |\hat{\theta} - \theta|.$$

Proposition. *If the loss function is square loss* $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$*, then the Bayes estimator is the expected value of the Posterior Distribution.*

$$\hat{\theta}_{BS} = E(\theta | \ \underline{X_{(n)}}).$$

Proof. Let, for notation simplicity, $\overline{\theta} = E(\theta | \ \underline{X_{(n)}})$ and $f(\theta | \ \underline{X_{(n)}}) = f(\theta)$, then the expected value can be written as:

$$\begin{aligned}
E\big(L(\hat{\theta}, \theta)\big) &= \int_{\Theta} (\hat{\theta} - \theta)^2 f(\theta) d\theta, \\
&= \int_{\Theta} (\hat{\theta} - \overline{\theta} + \overline{\theta} - \theta)^2 f(\theta) d\theta, \\
&= \int_{\Theta} ((\hat{\theta} - \overline{\theta})^2 + 2(\hat{\theta} - \overline{\theta})(\overline{\theta} - \theta) + (\overline{\theta} - \theta)^2) f(\theta) d\theta, \\
&= \int_{\Theta} ((\hat{\theta} - \overline{\theta})^2 f(\theta) d\theta + \int_{\Theta} (\theta - \overline{\theta})^2 f(\theta) d\theta \cdots \\
&+ 2(\hat{\theta} - \overline{\theta}) \int_{\Theta} (\overline{\theta} - \theta) f(\theta) d\theta.
\end{aligned}$$

The expected value of a constant is the constant; thus each integral simplifies,

$$\int_{\Theta}(\hat{\theta}-\overline{\theta})^2 f(\theta)d\theta = (\hat{\theta}-\overline{\theta})^2 \geq 0,$$

$$\int_{\Theta}(\overline{\theta}-\theta)f(\theta)d\theta = \overline{\theta} - \int_{\Theta}\theta f(\theta)d\theta,$$

$$= \overline{\theta} - \overline{\theta} = 0,$$

$$\therefore \ 2(\hat{\theta}-\overline{\theta})\int_{\Theta}(\overline{\theta}-\theta)f(\theta)d\theta = 0 \ \ \forall\hat{\theta}.$$

$$\int_{\Theta}(\theta-\overline{\theta})^2 f(\theta)d\theta = Var(\theta) \geq 0.$$

Hence, $E\big(L(\hat{\theta},\theta)\big) = (\hat{\theta}-\overline{\theta})^2 + Var(\theta) \geq Var(\theta)$. with equality iff

$$\hat{\theta} = \overline{\theta}.$$

■

Definition (Conjugate Prior)**.** *The prior distribution $\pi_0(\theta)$ is a conjugate prior if $\pi_0 \in \Upsilon(\theta)$, where $\Upsilon(\theta)$ is a family of distributions characterised by θ, then $P(\theta|\ \underline{X_{(n)}}) \in \Upsilon(\theta)$ [19].*

The proposition below will illustrate this property.

Proposition. *If the lower bound u of a Pareto distribution is known, then the Gamma distribution is a conjugate prior for the shape (Tail Index) parameter $\alpha > 0$.*

Proof. The distribution function of X is $F(x) = 1 - \left(\frac{x}{u}\right)^{-\alpha}$; thus, the density function is.

$$f(x|\alpha) = \frac{\alpha}{u}\left(\frac{x}{u}\right)^{-(\alpha+1)}.$$

Let $\pi_0(\alpha)$ be a gamma distribution. In essence,

$$\pi_0(\alpha) \propto \alpha^{a_0} e^{-\alpha b_0}.$$

Then the likelihood function is,

$$\begin{aligned} f(\underline{x_{(n)}}|\alpha) &= \prod_{i=1}^{n} \frac{\alpha}{u}\left(\frac{x_i}{u}\right)^{-(\alpha+1)}, \\ &\propto \alpha^n \exp\left(- (\alpha+1) \sum_{i=1}^{n} \ln\left(\frac{x_i}{u}\right)\right), \\ &\propto \alpha^n \exp\left(- \alpha \sum_{i=1}^{n} \ln\left(\frac{x_i}{u}\right)\right). \end{aligned}$$

The likelihood is proportional to the kernel of a Gamma distribution; hence, if it mixes with another Gamma distribution the result is another Gamma distribution.

$$\begin{aligned} f(\theta|\ \underline{X_{(n)}}) &\propto f(\underline{x_{(n)}}|\alpha)\pi_0(\alpha), \\ &\propto \alpha^n \exp\left(- \alpha \sum_{i=1}^{n} \ln\left(\frac{x_i}{u}\right)\right)\alpha^{a_0} e^{-\alpha b_0}, \\ &= \alpha^{n+a_0} \exp\left(- \alpha\left(\sum_{i=1}^{n} \ln\left(\frac{x_i}{u}\right) + b_0\right)\right). \end{aligned}$$

Therefore the Posterior distribution is a Gamma distribution with shape parameter $a_x = n + a_0$ and rate parameter $b_x = \sum_{i=1}^{n} \ln\left(\frac{x_i}{u}\right) + b_0$.

In essence:

$$\alpha| \ \underline{X_{(n)}} \sim Gamma(a_x, b_x).$$

■

Definition (Posterior Predictive Distribution)**.**

$$f(X|\underline{X_{(n)}}) = \int_{\Theta} f(x|\theta)P(\theta| \ \underline{X_{(n)}})d\theta.$$

Basic Bayesian Computational Tools

Proofs that these algorithms converge to the posterior distribution can be found in [5, 19].

Theorem (Strong Law of Large Numbers Markov Chain [19])**.** *Let* (X_n) *be a sequence of random variables that form a Markov Chain. If* $E(X) < +\infty$*, then, the sequence of partial averages* $\frac{1}{n}\sum_{i=1}^{n} X_i$ *converges pointwise to the Theoretical mean. In essence,*

$$P\left(\frac{1}{n}\sum_{i=1}^{n} X_i \to E(X)\right) = 1.$$

Both of these algorithms simulate a random walk that moves towards the region where the posterior distribution concentrates probability. They can be thought of as stochastic optimisation algorithms that search for the region with the most probability. Once, the random walk arrives to the region where probability is amassed the chain samples from the posterior distribution. With a sample from the Posterior distribution an implication of the Ergodic Theorem is that it is possible to compute expected values via Monte Carlo methods from

simulations $\theta^{(s)}$, $s \in \{1, 2, \cdots, S\}$ one gets.

$$\lim_{S \to +\infty} \frac{1}{S} \sum_{s=1}^{S} g(\theta^{(s)}) = \int_{\Theta} g(\theta) f(\theta | \ \underline{x_{(n)}}) d\theta.$$

Gibbs Sampling samples from the conditional posteriors; after a burn-in period these samples come from the posterior distribution. This algorithm is particularly useful for models with multiple parameters and known conjugate priors. As an example consider sampling from a bivariate normal distribution [18, 32]. The Gibbs sampler samples from conditional distributions as it moves to regions where the samples come from the joint distribution.

Algorithm 5 Gibbs Sampler

Starting values: $X^{(0)}$
for $t = 1, \cdots, T$ **do**
  Generate chain:

$$Y^{(t)} \sim N(\mu_{y|x^{(t-1)}}, \sigma^2_{y|x^{(t-1)}})$$
$$X^{(t)} \sim N(\mu_{x|y^{(t)}}, \sigma^2_{x|y^{(t)}}).$$

end

Metropolis-Hastings samples a random walk that uses a proposal distribution to simulate possible values and an odds ratio to determine the acceptance probability that the proposed value is a simulation from the target distribution [19, 29, 28].

The posterior odds ratio is the balanced density of the Markov-Chain; thus, after a burn-in period the random walk moves

within the posterior distribution. This algorithm is a powerful tool because it makes it possible to simulate from complicated distributions. For starters, the jumping ratio eliminates the need to compute the normalising constant, this is a huge computational advantage.

Algorithm 6 Metropolis Hastings Sampler

Starting values: $X^{(0)}$
for $t = 1, \cdots, T$ **do**
  Generate chain:

$$s \sim q(s)$$
$$R = \frac{f(s)q(x^{(t-1)})}{f(x^{(t-1)})q(s)}.$$

  if $U \sim Unif([0,1]) \leq R$ **then**
    $x^{(t)} = s$
  else
    $x^{(t)} = x^{(t-1)}$
  end
end

Convergence diagnostics can help diagnose non-convergence, but they cannot diagnose proper convergence. Good heuristics are to look at a time series plot of the Markov Chains and histograms [19, 27].

Possible test statistics are: within chain variance, correlation lags; if the chain converges to the target distribution, then the variance within chain should not be too different from the overall variance. Also, a plot of the autocorrelation function of the chain should resemble geometric decay if the chain is sampling from the stationary distribution. Recall that the stationary distribution by construction is the target (posterior) distribution [18, 18, 28, 27].

Aprendizaje Bayesiano estadístico,
para modelos de colas pesadas,
escrito por Daniel Salnikov,
se terminó de imprimir en junio de 2021
en los talleres de Diseño en Tesis.
Av Arquitectura 56-local 4,
Copilco Universidad, Coyoacán,
Ciudad de México.