Modelling operational risk using a Bayesian approach to extreme value theory

María Elena Rivera Mancía

Doctor of Philosophy

Deparment of Mathematics and Statistics

McGill University

Montréal, Québec

February 2014

A thesis submitted to McGill University in partial fullfilment of the requirements of the degree of Doctorate of Philosophy

DEDICATION

I dedicate this thesis to the strongest woman I have ever met in my life: My mom, who passed away in 2008. Her love and encouragement are still with me today.

I also dedicate this work to my family, for all the love and support they have given me throughout my life.

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to my supervisors, Professor David Stephens and Professor Johanna Nešlehová, for their invaluable support and patience for the completion of this work. This thesis would not have been possible without their continued feedback and encouragement throughout this long journey.

I thank my PhD committee members, Professor Christian Genest, Professor Jean-François Plante and Professor Russell Steele, for reading my thesis and for their insightful comments. I would also like to thank my external examiner, Professor Valérie Chavez-Demoulin, for her willingness to review my thesis and for her valuable comments and suggestions.

I am very thankful to Professor Christian Genest for kindly helping me with the French translation of the thesis abstract.

I would also like to thank the Department of Mathematics and Statistics for their financial support throughout my studies and for making my stay at McGill more enjoyable.

Special thanks to the Bank of Mexico for all that I learned during the two summer internships and for giving me the opportunity to work with real data and enrich my research.

I wish to thank my friends in Montréal for all the moments that turned my mind away from work and brought joy to my life. I must also thank those friends who are far in distance but close in heart, for their constant encouragement.

Above all I am very grateful to my family, for their trust and love during my whole life, and for teaching me what really matters in life. I cannot find the words to express all my gratitude to them.

Finally, I want to thank to all those people who have believed in me.

ABSTRACT

Extreme-value theory is concerned with the tail behaviour of probability distributions. In recent years, it has found many applications in areas as diverse as hydrology, actuarial science, and finance, where complex phenomena must often be modelled from a small number of observations.

Extreme-value theory can be used to assess the risk of rare events either through the block maxima or peaks-over-threshold method. The choice of threshold is both influential and delicate, as a balance between the bias and variance of the estimates is required. At present, this threshold is often chosen arbitrarily, either graphically or by setting it as some high quantile of the data.

Bayesian inference is an alternative to deal with this problem by treating the threshold as a parameter in the model. In addition, a Bayesian approach allows for the incorporation of internal and external observations in combination with expert opinion, thereby providing a natural probabilistic framework to evaluate risk models.

This thesis presents a Bayesian inference framework for extremes. We focus on a model proposed by Behrens et al. (2004), where an analysis of extremes is performed using a mixture model that combines a parametric form for the centre and a Generalized Pareto Distribution (GPD) for the tail of the distribution. Our approach accounts for all the information available in making inference about the unknown parameters from both distributions, the threshold included. A Bayesian analysis is then performed by using expert opinions to determine the parameters for prior distributions; posterior inference is carried out through Markov Chain Monte Carlo methods. We apply this methodology to operational risk data to analyze its performance.

The contributions of this thesis can be outlined as follows:

- Bayesian models have been barely explored in operational risk analysis. In Chapter 3, we show how these models can be adapted to operational risk analysis using fraud data collected by different banks between 2007 and 2010. By combining prior

information to the data, we can estimate the minimum capital requirement and risk measures such as the Value-at-Risk (VaR) and the Expected Shortfall (ES) for each bank.

- The use of expert opinion plays a fundamental role in operational risk modelling. However, most of time this issue is not addressed properly. In Chapter 4, we consider the context of the problem and show how to construct a prior distribution based on measures that experts are familiar with, including VaR and ES. The purpose is to facilitate prior elicitation and reproduce expert judgement faithfully.

- In Section 4.3, we describe techniques for the combination of expert opinions. While this issue has been addressed in other fields, it is relatively recent in our context. We examine how different expert opinions may influence the posterior distribution and how to build a prior distribution in this case. Results are presented on simulated and real data.

- In Chapter 5, we propose several new mixture models with Gamma and Generalized Pareto elements. Our models improve upon previous work by Behrens et al. (2004) since the loss distribution is either continuous at a fixed quantile or it has continuous first derivative at the blend point. We also consider the cases when the scaling is arbitrary and when the density is discontinuous.

- Finally, we introduce two nonparametric models. The first one is based on the fact that the GPD model can be represented as a Gamma mixture of exponential distributions, while the second uses a Dirichlet process prior on the parameters of the GPD model.

ABRÉGÉ

La théorie des valeurs extrêmes concerne l'étude du comportement caudal de lois de probabilité. Ces dernières années, elle a trouvé de nombreuses applications dans des domaines aussi variés que l'hydrologie, l'actuariat et la finance, où l'on doit parfois modéliser des phénomènes complexes à partir d'un petit nombre d'observations.

La théorie des valeurs extrêmes permet d'évaluer le risque d'événements rares par la méthode des maxima bloc par bloc ou celle des excès au-delà d'un seuil. Le choix du seuil est à la fois influent et délicat, vu la nécessité de trouver un équilibre entre le biais et la précision des estimations. À l'heure actuelle, ce seuil est souvent choisi arbitrairement, soit à partir d'un graphique ou d'un quantile élevé des données.

L'inférence bayésienne permet de contourner cette difficulté en traitant le seuil comme un paramètre du modèle. L'approche bayésienne permet en outre d'incorporer des observations internes et externes en lien avec l'opinion d'experts, fournissant ainsi un cadre probabiliste naturel pour l'évaluation des modèles de risque.

Cette thèse décrit un cadre d'inférence bayésien pour les extrêmes. Ce cadre est inspiré des travaux de Behrens et coll. (2004), dans lesquels l'étude des extrêmes est réalisée au moyen d'un modèle de mélange alliant une forme paramétrique pour le cœur de la distribution et une loi de Pareto généralisée (LPG) pour sa queue. L'approche proposée exploite toute l'information disponible pour le choix des paramètres des deux lois, y compris le seuil. Une analyse bayésienne tenant compte d'avis d'experts sur les paramètres des lois a priori est ensuite effectuée ; l'inférence a posteriori s'appuie sur une chaîne de Markov Monte-Carlo. Nous appliquons cette approche à des données relatives aux risques opérationnels afin d'analyser sa performance.

Les principales contributions de cette thèse sont les suivantes :

- On fait rarement appel aux modèles bayésiens pour l'analyse du risque opérationnel. Au chapitre 3, nous montrons comment adapter ces modèles à l'analyse du risque opérationnel au moyen de statistiques de fraudes recueillies par des banques entre

2007 et 2010. L'intégration d'information a priori aux données nous permet d'estimer le capital minimal requis pour chaque banque, ainsi que diverses mesures de risque telles que la valeur-à-risque (VaR) et le déficit prévu (DP).

- Les avis d'experts jouent un rôle clef dans la modélisation du risque opérationnel. Toutefois, cette question est souvent traitée de façon incorrecte. Au chapitre 4, nous examinons le problème dans son contexte et montrons comment choisir une loi a priori à partir de mesures que les experts connaissent bien, dont la VaR et le DP. Le but est de faciliter le choix de la loi a priori et de mieux refléter l'avis des experts.

- À la section 4.3, nous décrivons diverses techniques de synthèse d'opinions d'experts. Bien que ce problème ait déjà été abordé dans d'autres domaines, il est relativement nouveau dans notre contexte. Nous montrons comment élaborer une loi a priori à partir d'avis d'experts et mesurons leur influence sur la loi a posteriori. Des données réelles et simulées sont utilisées aux fins d'illustration.

- Au chapitre 5, nous proposons plusieurs nouveaux modèles faisant intervenir des mélanges de lois gamma et de Pareto généralisées. Ces modèles étendent les travaux de Behrens et coll. (2004) dans la mesure où la loi des pertes peut être continue à un quantile donné ou avoir une première dérivée continue au point de jonction. Nous traitons aussi les cas où l'échelle est arbitraire et la densité est discontinue.

- Enfin, nous présentons deux modèles non paramétriques. Le premier s'appuie sur le fait que le modèle LPG peut être représenté comme un mélange gamma de lois exponentielles ; dans le second, l'information a priori sur les paramètres du modèle LPG est représentée par un processus de Dirichlet.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Since its emergence, the study of operational risk has presented many issues: the combination of data sources from internal and external data and expert opinion; the elicitation of information; and since rare events occur infrequently, small data sets. Extreme Value Theory has emerged as a natural tool for quantifying operational risk. However, its application involves several challenges, mainly the characterization of tail behaviour and the inclusion of expert knowledge. To overcome these limitations, attention has turned recently to Bayesian methods which can handle all these problems in a single framework.

In this thesis, we propose a Bayesian approach that provides an appropriate framework to address the threshold[1] selection issue and allows for the inclusion of multiple expert opinion and external data, while taking into account the theoretical foundations of Extreme Value Theory.

The thesis is organized as follows. Chapter 2 reviews relevant background and concepts concerning operational risk, Extreme Value Theory and Bayesian inference. This part of the thesis has the intention to give the reader a basis for what will be discussed in later chapters. We present some historical losses that motivated the study of operational risk and gave rise to the Basel II and Basel III Accords in 2004 and 2010, respectively. We

[1]The threshold is loosely defined as the point above which losses are considered extremes and it is chosen such that the population tail can be well approximated by an extreme value model.

highlight the importance of these Accords, since they pave the way for the analysis of operational risk in a formal fashion.

We then move on to a review of Extreme Value Theory. This section introduces the foundations of Extreme Value Theory and its main results, such as the Fisher-Tippet-Gnedenko and Pickands-Balkema-de Haan theorems. In the end, the theory and methods presented in this section lead us to study the limitations of Extreme Value Theory, one of the main concerns of this thesis. We focus our attention on the Peaks Over Threshold (POT) method and the point processes approach. We find that these approaches suffer from two intrinsic limitations: Subjectivity about the threshold choice, and not accounting for threshold uncertainty in inference.

Once these problems are clearly stated, our attention turns to the range of tools available, in particular Bayesian inference. We present the Bayesian framework and the modelling strategy behind it: Formulate our beliefs in terms of the so-called prior distribution, observe the data and update our beliefs via Bayes' rule, giving place to the so-called posterior distribution. We then examine this framework in the context of extremes and emphasize its advantages compared to maximum likelihood estimation methods. In addition, we introduce some ideas about the elicitation process that will be studied in detail in Chapter 4.

In Chapter 3, we concentrate on the model introduced by Behrens et al. (2004). This model is based on a mixture distribution, which combines a parametric form for the center and a Generalized Pareto Distribution (GPD) for the tail, using all observations for inference about the unknown parameters from both distributions, the threshold included. Next, we provide various ways of choosing prior distributions for the different parameters involved, while posterior inference is carried out through Markov Chain Monte Carlo (MCMC) methods. In order to test the performance of this model, we apply this methodology to simulated data. To complete the analysis, we apply the proposed algorithm to real data in Section 3.2, consisting of bank frauds from 2007 to 2010. We start by doing an

exploratory analysis of the data to determine the presence of extremes, and then proceed to perform the Bayesian analysis using an algorithm adapted to the context of operational risk. Subsequently, we estimate the minimum capital requirement, the Value-at-Risk and the Expected Shortfall, and present some interesting findings related to these risk measures. In the second part of our analysis, we consider grouped data. Banks are classified according to their size and some other characteristics. The analysis is performed using a Reversible Markov Chain Monte Carlo (RJMCMC) algorithm and a Bayesian approach based on a mixture of Gamma distributions in which the mixing occurs over the shape parameter.

The work described in Chapters 4 and 5 attempts to fill some of the gaps in inference for extremes and to improve the Bayesian methodology presented previously. Chapter 4 is dedicated to prior analysis. We address the problem of prior sensitivity, prior elicitation and multiple expert opinion. In the first part, we study the difference of using non-informative and informative priors for extreme data. After that, we introduce an informative prior that captures expert opinion in an intuitive way, making use of quantities that experts are familiar with, and that are realistic in the context of our problem. Lastly, we explore multiple expert opinion and propose a methodology for combining multiple opinions and for updating our beliefs when new information becomes available.

Finally, in Chapter 5, we introduce a new model that we call the Blended model. This model is constructed using Gamma and Generalized Pareto elements and has the objective of improving the discontinuity problem in the model of Behrens et al. (2004). We consider different approaches according to the type of discontinuity, for instance the qth quantile or the first derivative at the blend point. In order to complete the analysis, we also consider nonparametric models and propose two different ways to construct them: (1) Represent the GPD as a mixture of Exponential distributions and use a Dirichlet process mixture formulation and (2) use a Dirichlet process prior on the parameters of the GPD model.

We conclude this thesis with a discussion of our results in Chapter 6 .

Chapter 2

Extreme Value Theory in operational risk

2.1 Operational risk

2.1.1 Introduction to operational risk

During the 1980s, several catastrophic losses ocurred in financial institutions worldwide, giving rise to a first international cooperation agreement known as Basel I Accord, issued in 1988, that sets minimum capital requirements for banks to hedge their risk exposure, particularly credit risk.

Faced with a changing environment, Basel I had limitations, so that in 2004 the Basel Committee[1] stated the New Capital Accord "Basel II", which aims to strengthen the soundness and stability of the international banking system, with more risk-sensitive capital requirements.

Basel II incorporates operational risk to the already considered credit and market risks, highlighting the importance of allocating capital for operational losses. Table 2.1 shows

[1]The Basel Committee on Banking Supervision is a committee of banking supervisory authorities which was established by the central bank Governors of the Group of Ten countries in 1975. Its recommendations have become the standard and guidelines in banking supervision and regulation in the rest of the world.

Year	Institution	Estimated Losses (Millions of USD)	Event
1995	Barings Bank	$1,300	In Singapore, a trader accumulated not reported losses for two years.
1995	Dalwa Bank	$1,100	For eleven years, a trader accumulated not reported losses.
1996	Morgan Grenfell	$640	For two years, a fund manager invested in shares of a "junk company".
1997	Natwest Bank	$145	Incorrect valuation of options.
2002	All First	$691	For three years, losses in foreign exchange trading were hidden.
2008	Société Générale	$7,000	A broker hid losses from unauthorized transactions in the futures market.

Table 2.1: *Historical operational risk losses in financial institutions worldwide*

some recent losses that motivated the study of operational risk.

A new Accord was introduced in 2010 in response to the deficiencies in financial regulation revealed by the late-2000s financial crisis. The Basel III Accord is supposed to strengthen bank capital requirements by increasing bank liquidity and bank leverage. However, the new Accord does not incorporate relevant changes regarding operational risk. Hence, we will focus on the Basel II Accord.

2.1.2 Basel II and operational risk

The Basel II Capital Accord of 2004 [2] defines operational risk as *"the risk of loss resulting from inadequate or failed internal processes, people and systems, or from external events".*

Risk events are classified into seven categories, namely:

- Internal fraud

[2]To record losses, institutions should: classify units and business lines, identify and classify the different types of loss events, and keep a historical database containing records of losses and their cost.

- External fraud
- Employment practices and workplace safety
- Clients, products, and business practice
- Damage to physical assets
- Business disruption and systems failures
- Execution, delivery and process management

The Basel II Accord also sets capital requirements to face operational losses and established 2007 as the year by which financial institutions should have implemented the calculation of capital requirements for operational risk.

The Accord provides three methods for calculating capital requirements for operational risk, which are presented in order of sophistication and risk sensitivity:

1. *The Basic Indicator Approach.* Capital is allocated using the gross income as a proxy for an institution's overall operational risk exposure, with each bank holding capital for operational risk equal to the amount of a fixed percentage α (15%), multiplied by its individual amount of positive gross income. Assume that in n out of the last 3 years the gross income was positive. Relabeling these years as $1, ..., n$, the corresponding gross income is $GI_1, ..., GI_n$. The resulting capital charge K_{BIA} under the Basic Indicator approach is given as follows.

$$K_{BIA} = \frac{\sum_{j=1}^{n} GI_j \times \alpha}{n}. \tag{2.1}$$

   This approach is simple and easy to implement, however, it does not consider specific needs and characteristics of each bank.

2. *The Standardised Approach.* Capital is determined based on predefined percentages of the average gross income of the last three years for eight business lines [3]. After determining the capital required for each business line, the aggregate is the total capital requirement. This approach differs from the Basic Indicator Approach in that bank's activities are divided into a number of standardised business units and business lines.

3. *Advanced Measurement Approach (AMA)*: The regulatory capital is determined using models developed by each institution. It considers two approaches: the internal measurement and the loss distribution approach (LDA). Under the LDA, banks use their internal data, estimate the frequency and severity distributions of operational risk events and, based on these two distributions, compute the probability distribution function of the cumulative operational loss. The capital charge is usually based on the Value-at-Risk measure (VaR), which is defined in the next section, Equation (2.3).

2.1.2.1 Implementation

In 2010, the Financial Stability Institute (FSI) carried out a survey on Basel II implementation. The survey results indicate that 112 countries have implemented or are currently planning to implement Basel II.

For operational risk the survey indicated that 80% of respondents that have adopted the Accord expected to adopt the Basic Indicator Approach.

[3]Banks' activities are divided into eight business lines by considering that gross income is a broad indicator that serves as a proxy for the scale of business operations and thus the likely scale of operational risk exposure within each of these business lines: corporate finance, trading & sales, retail banking, commercial banking, payment & settlement, agency services, asset management, and retail brokerage.

2.1.3 Loss Distribution Approach (LDA)

The Loss Distribution Approach (LDA) is a statistical approach which is very popular in actuarial sciences for computing aggregate loss distributions. Under this approach, banks adjust statistical distributions to loss data by modelling: i) the frequency of loss events and ii) their severity, then combining them to obtain the distribution of total losses.

The loss distribution is modelled as follows:

$$Z_t = \sum_{j=1}^{J} Z_t^{(j)}; \qquad Z_t^{(j)} = \sum_{i=1}^{N_t^{(j)}} X_i^{(j)}(t), \tag{2.2}$$

where

- $t = 1, 2, ...$ is discrete time in annual units.
- $Z_t^{(j)}$ is the annual loss in risk cell j, modelled as an aggregate loss over one year, with frequency $N_t^{(j)}$.
- $N_t^{(j)}$is a counting process and for each t, $X_i^{(j)}(t)$ are positive random variables representing severities, with $i = 1, ..., N_t^{(j)}$.

Under this model, the capital is defined as the Value-at-Risk at the 99.9% level quantile of the distribution for the next year annual loss Z_{t+1}

$$\mathrm{VaR}_q(Z_{t+1}) = \inf\{z \in \mathbb{R} : \ \Pr(Z_{t+1} > z) \leq 1 - q\} \tag{2.3}$$

at level $q = 0.999$.

2.1.4 Operational risk data

In the General Standards for Advanced Measurement Approaches, the Basel II Accord specifies that internal data, external data and expert opinion data should be considered into the analysis. In addition, internal control indicators and factors affecting the businesses should be used.

Operational risk data should meet specific criteria:

- *Internal data*. Internal measures must be based on a minimum five-year observation period of internal loss data, or three years when the bank first moves to the AMA. Loss data must capture all material activities and exposures from all appropriate sub-systems and geographic locations and the collected information should include the date of the risk event, any recoveries of gross loss amounts, as well as some descriptive information about its causes.

- *External data*: Banks must use relevant external data (either public data and/or pooled industry data), especially when the bank is exposed to infrequent, yet potentially severe, losses. These data should include actual loss amounts, information on the scale of business operations where the event occurred and information on the causes and circumstances of the loss events. External data are difficult to use due to different volumes and other factors.

- *Expert opinion*: A bank must use scenario analysis of expert opinion in conjunction with external data to evaluate its exposure to high-severity events. These expert assessments could be expressed as parameters of a statistical loss distribution.

2.1.5 Bank fraud

Any activity related to money entails the risk of fraud, however, in the financial sector, particularly banking, the exposure is higher. "A fraud is an action to achieve a profit or gain something illegally through deception or exploitation of a mistake made by others. Deception occurs when an individual displays a series of machinations and artifices in order to make one or more people have a false perception of reality to obtain goods or property rights of others" [4].

In some countries, most of banking fraud losses correspond to credit and debit cards. However, they also occur with other means of payment such as cheques and Internet transfers.

[4]Source: Condusef http://www.condusef.gob.mx

2.1.6 Challenges in operational risk

As it is said in McNeil (2005), "nobody doubts the importance of operational risk for the financial and insurance sector, but much less agreement exists on how to measure this risk". Undoubtedly, the study of operational risk is accompanied by several disadvantages: Possible inconsistencies in its definition, data gaps and limitations to allocate capital for it.

As a result, there have been several criticisms of Basel II on operational risk, which shows differences with respect to market and credit risks: It is more linked to process rather than product, and hence operational risk does not always arise through transactions; thus, most of the time it is not reported in the income statement. Furthermore, operational risk cannot always be reduced by diversification, or objectively assigned to a particular business line, so it is assumed inevitably as part of the company's business rather than pursuit of profit.

Additionally, there are several criticisms of the methods of measurement, mainly the questionable choice of the gross income as an indicator of risk exposure, since there may be institutions with high income and low risk, and institutions with lower income but high risk, however, the basic approach will require more capital to the first group. Some additional criticisms are the linear relationship between the indicator and risk, together with the coexistence of methodologies in which all parameters are predefined.

Because of these criticisms, in recent years the academic community has placed special emphasis on advanced measurement methods, by considering the loss distribution approach and providing a theoretical basis to implement more advanced methods.

It should be noticed that beyond the regulation for operational risk, the losses associated with such events may affect the results of the institutions and, therefore, their capitalization levels, so that the measurement of this risk should not be ignored.

2.2 Extreme Value Theory

Extreme Value Theory (EVT) is used to model and measure tail events that occur with small probability. Its main utility is that it is a quantitative method that takes into account the frequency and severity of losses.

Over the past two decades, EVT has had an important development and has been recognized as an extremely useful statistical tool for modelling "rare events". It has relevant applications in insurance, finance, risk management and meteorology, among others.

2.2.1 Basic definitions and results

Most of the results presented in this section can be found in McNeil et al. (2005, Chapter 7); proofs are given in Embrechts et al. (1997, Chapter 3).

2.2.1.1 Generalized Extreme Value Distribution

Theorem 2.1. *(Fisher-Tippett). Let X_n be a sequence of iid random variables and $M_n = \max(X_1, ..., X_n)$ be the maximum of the first n members of the sequence. If there exist norming constants $c_n > 0$, $d_n \in \mathbb{R}$ and some nondegenerate cdf H such that*

$$\frac{M_n - d_n}{c_n} \xrightarrow{d} H, \tag{2.4}$$

then H belongs to the type of one of the following cdfs:

$$\begin{cases} \textit{Type I (Gumbel):} & H(x) = \exp\left(-e^{-x}\right) \quad x \in \mathbb{R} \\ \textit{Type II (Fréchet):} & H(x) = \begin{cases} 0 & x \leq 0 \\ \exp\left(-x^{-\alpha}\right) & x > 0 \end{cases} \quad \alpha > 0 \\ \textit{Type III (Weibull):} & H(x) = \begin{cases} \exp\left(-(-x)^{\alpha}\right) & x \leq 0 \\ 1 & x > 0 \end{cases} \quad \alpha > 0 \end{cases} \tag{2.5}$$

where $\stackrel{d}{\Rightarrow}$ *denotes convergence in distribution.*

Definition 2.2. (maximum domain of attraction). If (2.4) holds for some nondegenerate cdf H and sequences of constants c_n and d_n, then F is said to be in the maximum domain of attraction of H, written $F \in \mathrm{MDA}(H)$.

The definition below provides a unified parameterization for the class of limit distributions: Gumbel, Fréchet and Weibull.

Definition 2.3. (the generalized extreme value (GEV) distribution). The distribution function of the GEV distribution is given by

$$H_{\xi,\mu,\beta}(x) = \begin{cases} \exp\left[-\left(1+\xi\frac{x-\mu}{\beta}\right)^{-\frac{1}{\xi}}\right], & \xi \neq 0 \\ \exp\left[-\exp\left(1-\frac{x-\mu}{\beta}\right)\right], & \xi = 0 \end{cases} \tag{2.6}$$

for $1+\xi(x-\mu)/\sigma > 0$, where $\mu \in \mathbb{R}$ is the location parameter, $\beta > 0$ the scale parameter and $\xi = 1/\alpha \in \mathbb{R}$ the shape parameter.

The parameter ξ in the above Definition defines the type of distribution: when $\xi > 0$ the distribution is a Fréchet distribution; when $\xi = 0$ it is a Gumbel distribution; when $\xi < 0$ it is a Weibull distribution.

2.2.1.2 Threshold Exceedances

The Generalized Pareto Distribution (GPD) is used to model the tails of distributions based on theoretical arguments. Usually, it is expressed through a distribution function that depends on two parameters.

Definition 2.4. The cdf of the GPD is given by

$$G_{\xi,\sigma}(x) = \begin{cases} 1-\left(1+\xi\frac{x}{\sigma}\right)^{-\frac{1}{\xi}}, & \xi \neq 0 \\ 1-\exp\left(-\frac{x}{\sigma}\right), & \xi = 0 \end{cases} \tag{2.7}$$

for $x \geq 0$ for $\xi \geq 0$ and for $0 \leq x \leq -\sigma/\xi$ for $\xi < 0$. In (2.7), $\sigma > 0$ and $\xi \in \mathbb{R}$ are the scale and shape parameters, respectively.

When $\xi > 0$, the Generalized Pareto distribution is a reparameterized version of the ordinary Pareto distribution with parameter $\alpha = 1/\xi$ and $\kappa = \sigma/\xi$; if $\xi < 0$, we have a Pareto type II distribution; if $\xi = 0$ we have the exponential distribution.

Definition 2.5. The right endpoint of a distribution is defined as

$$x_F = \sup \{x \in \mathbb{R} : \ F(x) < 1\}.$$

Definition 2.6. (excess distribution over threshold u). Let X be a rv with cdf F. The excess distribution over the threshold u is given by

$$F_u(x) = P(X - u \leq x \mid X > u) = \frac{F(x+u) - F(u)}{1 - F(u)}, \tag{2.8}$$

for $0 < x < x_F - u$, where $x_F \leq \infty$ is the right endpoint of F.

Definition 2.7. (mean excess function). The mean excess function of a random variable X with finite mean is given by $e(u) = E(X - u \mid X > u)$.

Theorem 2.8. *(Pickands-Balkema-de Haan). Let F be a distribution function with endpoint x_F. We can find a positive measurable function $\sigma(u)$ such that:*

$$\lim_{u \to x_F} \sup_{0 \leq x < x_F - u} \left| F_u(x) - G_{\xi,\sigma(u)}(x) \right| = 0 \tag{2.9}$$

if and only if $F \in MDA(H_\xi)$, $\xi \in \mathbb{R}$.

Hence, for high thresholds, the excess distribution function can be approximated by $G_{\xi,\sigma(u)}(x)$ for some values of ξ and σ. Thus, we may try to fit the generalized Pareto distribution to data which exceed high thresholds.

2.2.2 The Peaks Over the Threshold and the Point Process Approach

There are two inference methodologies related to Extreme Value Theory: Block Maxima (BM) and Peaks Over the Threshold (POT). The first technique consists of first dividing identically distributed observations into block of equal size and modelling the maxima of each block. On the other hand, the POT method considers models for all observations that exceed some high level. Both techniques are based on limit results: The Block Maxima method uses the Fisher–Tippett Theorem 2.1, while the POT approach uses the Pickands–Balkema–de Haan Theorem 2.8.

POT models are generally considered to be the most useful for practical applications, due to their more efficient use of the data on extreme outcomes. Moreover, these models provide insights into excess distributions over high thresholds, which are of particular interest in operational risk. Hence, we will focus on the POT method throughout the thesis.

For POT models, there are two types of analysis: Semiparametric models based on the Hill estimator, and complete parametric models based on the Generalized Pareto distribution (GPD). When used correctly, both methods are theoretically and empirically justified. However, the approach based on the Hill estimator has limited applications since the heavy-tailedness assumption is necessary, as its focus is on the case $\xi > 0$; while GPD-based models are applicable to any type of distribution, heavy tailed or not, provided it is in the MDA of H_ξ.

Under this last approach, the exceedances, i.e. observations that are larger than a threshold, are considered as a point process of exceedances, which converges weakly to a Poisson point process that allows for inference on the intensity of occurrence of such exceedances. On the other hand, the Generalized Pareto distribution provides a model for the excesses over an appropriate threshold, i.e. the magnitudes of the differences of the exceedances and the threshold.

The POT method is described in McNeil et al.(2005) as follows: Given loss data $X_1, ..., X_n$ from F, a random number N_u will exceed our threshold u; it will be convenient to relabel these data $\tilde{X}_1, ..., \tilde{X}_{N_u}$. For each of these exceedances we calculate the amount $Y_j = \tilde{X}_j - u$ of the excess loss. We wish to estimate the parameters of a GPD model by fitting this distribution to the N_u excess losses. According to Ribatet (2006), there are currently seventeen estimators available to fit a Generalized Pareto distribution. Among the most important are: the method of moments, maximum likelihood, probability weighted moments (biased and unbiased), median, Pickands, penalized maximum likelihood and moment generating function estimators. The maximum likelihood method is more commonly used and is easy to implement if the excess data can be assumed to be realizations of independent rvs, since the joint density will then be a product of marginal GPD densities.

Denoting $g_{\xi,\sigma}$ the density of the GPD, the log-likelihood may be easily calculated to be

$$\ln L\left(\xi, \sigma; Y_1, ..., Y_{N_u}\right) = \sum_{j=1}^{N_u} \ln g_{\xi,\sigma}\left(Y_j\right) = -N_u \ln \sigma - \left(1 + \frac{1}{\xi}\right) \sum_{j=1}^{N_u} \ln\left(1 + \xi \frac{Y_j}{\sigma}\right), \quad (2.10)$$

which must be maximized subject to the parameter constraints $\sigma > 0$ and $1 + \xi Y_j/\sigma > 0$ for all j. Solving the maximization problem yields a GPD model $G_{\hat{\xi},\hat{\sigma}}$ for the excess distribution F_u. If $\xi > -0.5$, the maximum likelihood estimators of ξ and σ are regular (Smith, 1985).

Next, note that $F \in \mathrm{MDA}\left(H_\xi\right)$ if and only if there exist sequences of constants $c_n > 0$, $d_n \in \mathbb{R}$ such that for all $x \in \mathbb{R}$,

$$\lim_{n\to\infty} n \cdot \left(1 - F\left(c_n x + d_n\right)\right) = \lim_{n\to\infty} n \cdot \bar{F}\left(c_n x + d_n\right) = -\ln H_\xi\left(x\right), \quad (2.11)$$

where the limit is interpreted as ∞ if $H_\xi\left(x\right) = 0$; see Embrechts et al. (1997, Proposition 3.3.2). Assuming that $F \in \mathrm{MDA}\left(H_\xi\right)$ and setting $u_n\left(x\right) = c_n x + d_n$ we thus have

$$\lim_{n\to\infty} n \cdot \bar{F}\left(u_n\left(x\right)\right) = \lambda\left(x\right), \quad (2.12)$$

where $\lambda(x) = -\ln H_\xi(x)$. Defining the point process of exceedances as

$$N_n(A) = \sum_{i=1}^{n} I\left(I_{(X_i > u_n(x))} \cdot (i/n) \in A\right) \tag{2.13}$$

for any Borel set $A \subseteq (0,1]$, it can be shown (Embrechts et al.(1997), Theorem 5.3.2) that if $\lambda(x) \in (0,\infty)$, N_n converges weakly to a homogeneous Poisson point process on $(0,1]$ with intensity $\lambda(x)$.

Now assume that $X_1, ..., X_n$ are iid observations from $F \in \text{MDA}(H_\xi)$ and that u is a high threshold of the form $u = c_n y + d_n$ for some $y \in \mathbb{R}$. We assume that the process of exceedances above u is well approximated by a Poisson point process with intensity

$$\lambda = -\ln H_\xi(y) = -\ln H_\xi\left(\frac{y - d_n}{c_n}\right). \tag{2.14}$$

Since d_n and c_n are unknown, we can replace them by the location and scaling parameters μ and β, respectively. This gives

$$\lambda = -\ln H_\xi\left(\frac{u-\mu}{\beta}\right) = -\ln H_{\xi,\mu,\beta} = \left(1 + \xi\frac{u-\mu}{\beta}\right)^{-1/\xi - 1} \tag{2.15}$$

provided $1 + \xi(u-\mu)/\beta > 0$, and by $\lambda = 0$ otherwise.

To take the sizes of the exceedances into account, this model can be extended to a (non-homogeneous) two-dimensional Poisson point process where the points (t,x) in the two-dimensional space $\mathcal{X} = (0,1] \times (u,\infty)$ are record times and magnitudes of exceedances.

For a set of the form $A = (t_1, t_2) \times (x, \infty) \subset \mathcal{X}$, the intensity measure is

$$\Lambda(A) = \int_{t_1}^{t_2}\int_{x}^{\infty} \lambda(y)\, dy\, dt = -(t_2 - t_1)\ln H_{\xi,\mu,\sigma}(x). \tag{2.16}$$

It follows from (2.16) that for any $x \geq u$ the implied one-dimensional process of exceedances of the level x is a homogeneous Poisson process with rate $\tau(x) := -\ln H_{\xi,\mu,\beta}(x)$. Now consider the excess amounts over the threshold u. The tail of the excess distribution function over the threshold u, denoted $\bar{F}_u(x)$ before, can be calculated as the ratio of the rates of exceeding the levels $u + x$ and u. We obtain

$$\bar{F}_u(x) = \frac{\tau(u+x)}{\tau(u)} = \left(1 + \frac{\xi x}{\beta + \xi(u-\mu)}\right)^{-1/\xi} = \bar{G}_{\xi,\sigma}(x) \tag{2.17}$$

for a positive scaling parameter $\sigma = \beta + \xi(u - \mu)$. This is precisely the tail of the GPD model for excesses over the threshold u used previously.

2.2.3 Threshold selection

Most of the methods involved in choosing the threshold, u, make use of the Pickands-Balkema-de Haan Theorem 2.8. In general, we face two conflicting issues. If the threshold is chosen too low it is possible to get biased estimates because the theorem does not apply. On the other hand, if the threshold is set too high then only few data points will be available and estimates will be prone to high standard errors. We therefore reduce bias by lowering the number of observations in the tail and reduce variance by increasing it. This is known as the *bias-variance tradeoff*.

The main objective is to choose a threshold so that enough events are selected to reduce the variance, without inducing bias. The theory does not propose any objective method for threshold determination; there are mainly graphical ad-hoc approaches. Some other approaches include setting the threshold as some high percentile of the data; for instance, DuMouchel (1983) suggests fitting a Generalized Pareto model to the data outside the 10th and 90th percentiles. Here, we concentrate on the most commonly used graphical methods.

2.2.3.1 Threshold choice plot

Through this graph, we analyze the stability of the model estimation based on the fit of different models, using thresholds in a given range.

Let $X - u_0 \mid X > u_0 \sim \text{GPD}_{\xi_0,\sigma_0}$. Let u_1 be any other threshold such that $u_1 > u_0$. The random variable $X - u_1 \mid X > u_1$ is also GPD with updated parameters $\sigma_1 = \sigma_0 + \xi_0 (u_1 - u_0)$ and $\xi_1 = \xi_0$ (Lemma 7.22, McNeil et al. 2005). Setting

$$\sigma_* = \sigma_1 - \xi_1 u_1, \tag{2.18}$$

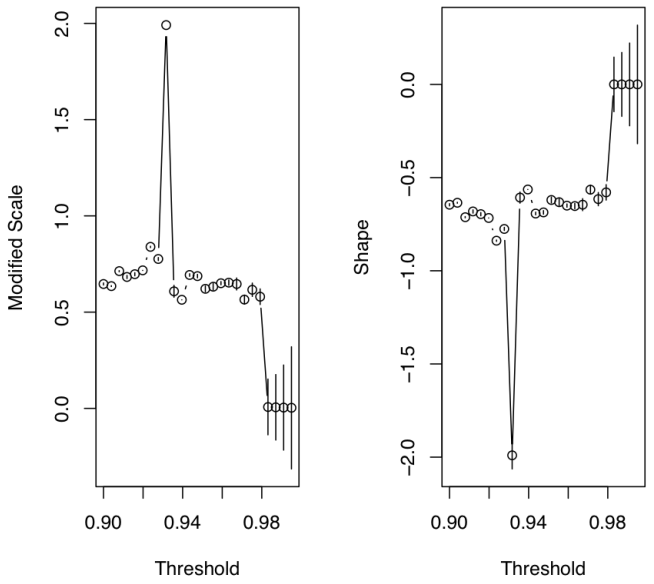

Figure 2.1: *Threshold choice plot with threshold around 0.98*

σ_* is independent of u_1. Thus, the estimates σ_* and ξ_1 are constant for every $u_1 > u_0$ if u_0 is an appropriate threshold for the asymptotic approximation.

The threshold choice plot represents the points defined by:

$$\{(u_1, \sigma_*) : u_1 \leq x_{\max}\} \quad \text{and} \quad \{(u_1, \xi_1) : u_1 \leq x_{\max}\} \tag{2.19}$$

where $x_{\max}$ is the maximum of the observations. Thus, we select the threshold at the point where estimates remain roughly constant.

Figure 2.1 shows an example of the threshold choice plot for simulated uniform data, where a threshold around 0.98 is a reasonable choice.

2.2.3.2 Mean residual life plot

This plot is also called the mean excess plot and it is based on the theoretical mean of a generalized Pareto distribution. If X is $\text{GPD}_{\xi,\sigma}$, then, provided $\xi < 1$,

$$E[X] = \frac{\sigma}{1-\xi}. \tag{2.20}$$

When $\xi \geq 1$ the theoretical mean of X is infinite. Moreover, for any $u > 0$, the excess distribution function of X is easily calculated to be

$$F_u(x) = G_{\xi,\sigma(u)}, \qquad \sigma(u) = \sigma + \xi u. \tag{2.21}$$

In particular, by equation (2.20), the mean excess function of X introduced in Definition 2.7 is then

$$e(u) = \frac{\sigma(u)}{1-\xi} = \frac{\sigma + \xi u}{1-\xi}, \tag{2.22}$$

provided $\xi < 1$, for $0 \leq u < \infty$ if $0 \leq \xi < 1$ and $0 \leq u \leq -\sigma/\xi$ if $\xi < 0$. It may be observed that the mean excess function is thus linear in u, which is a characterizing property of the GPD.

In practice, if X represents the excess over a threshold u_0 and if a GPD approximation above that threshold is good enough, the excess distribution over a higher threshold $u > u_0$ is again GPD with the same parameter ξ and scale parameter $\sigma + \xi(u - u_0)$ (Lemma 7.22, McNeil et al. 2005). Therefore, by equation (2.20) we have:

$$e(u) = \frac{\sigma + \xi(u - u_0)}{1-\xi} = \frac{\xi u}{1-\xi} + \frac{\sigma - \xi u_0}{1-\xi}, \tag{2.23}$$

where $u_0 \leq u < \infty$ if $0 \leq \xi < 1$ and $u_0 \leq u \leq u_0 - \sigma/\xi$ if $\xi < 0$.

The mean $e(u)$ is thus linear in u and can easily be estimated using the empirical sample mean. This fact is commonly used as a diagnostic for data admitting a GPD model for the excess distribution.

Given a sample $X_1, ..., X_n$, the estimator is given by

$$e_n(u) = \frac{\sum_{i=1}^{n} (X_i - u) I_{\{X_i > u\}}}{\sum_{i=1}^{n} I_{\{X_i > u\}}}, \tag{2.24}$$

where $I_{\{\cdot\}}$ is the indicator function and the sum in the denominator represents the number of observations over the threshold u.

Thus, the mean residual life plot represents points defined by:

$$\{(X_{i,n},\, e_n(X_{i,n})) : 2 \leq i \leq n\} \tag{2.25}$$

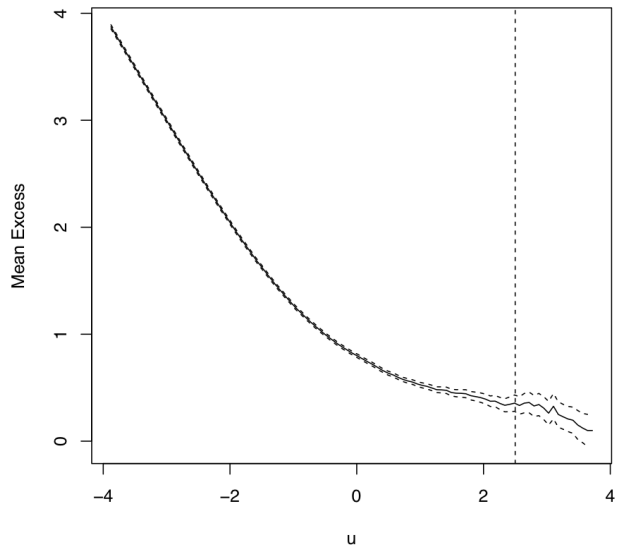

Figure 2.2: *Mean residual life plot with threshold around 2.5*

where $X_{i,n}$ denotes the $i-$th order statistic.

If the GPD model is appropriate for data exceeding a high threshold, (2.23) suggests that the mean excess plot should become increasingly "linear" for higher values of u. In general, a linear upward trend indicates a GPD model with positive shape parameter ξ; a plot tending towards the horizontal indicates a GPD with approximately zero shape parameter; a linear downward trend indicates a GPD with negative shape parameter.

An example of the mean excess plot for simulated standard normal data is displayed in Figure 2.2. The threshold is selected to be around 2.5.

2.2.3.3 Dispersion index plot

Let X be a random variable with Poisson distribution with parameter λ. Then

$$P[X = k] = \frac{e^{-\lambda}\lambda^k}{k!}, \quad k \in \mathbb{N} \tag{2.26}$$

and $E[X] = \mathrm{Var}[X]$. If the number of events follows a Poisson distribution, the ratio of the variance to the mean equals 1.

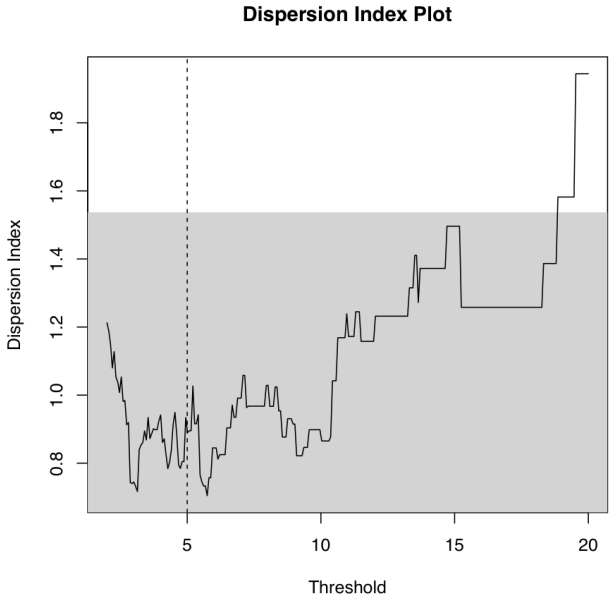

Figure 2.3: *Dispersion index plot with threshold around 5*

Under suitable conditions (i.e. $F \in \mathrm{MDA}(H)$) the exceedances over a threshold can be approximated by a GPD. Moreover, EVT also shows that the occurrences of these exceedances may be represented as a homogeneous Poisson point process. In particular, if the Poisson point process approximation above a high threshold u is valid, the number of exceedances above u in disjoint blocks of equal size constitutes an approximately i.i.d. sample from the Poisson distribution. The so-called dispersion index DI, which is the ratio of the sample variance to the sample mean, viz.

$$DI(u) = \frac{s^2}{\lambda} \tag{2.27}$$

should thus be close to 1. The dispersion index plot constitutes of points

$$\{(X_{i,n},\, \mathrm{DI}\,(X_{i,n})) :\, 2 \leq i \leq n\} \tag{2.28}$$

and we select a threshold at the point where the plot becomes close to 1. Figure 2.3 displays the dispersion index plot for the data set *ardieres* included in the R POT package, where a threshold around 5 seems to be reasonable.

2.2.4 Measuring operational risk

The POT model can be used to quantify operational risk. We will assume that the chosen threshold u satisfies a bias-variance tradeoff and that such u may be termed an unexpected loss threshold. Following the procedure described by Medova et al. (2002), we have:

- The *severity* of the losses is modelled by the Generalized Pareto distribution. The expectation of the excess loss distribution, i.e., the *expected severity* is a coherent[5] risk measure given by:
$$E\left(X-u \mid X>u\right)=\frac{\sigma+\xi u}{1-\xi}. \tag{2.29}$$

- The *number of exceedances* N_u over the threshold u and the corresponding exceedance times follow a homogeneous Poisson point process with intensity given by:
$$\lambda\left(u\right):=\left(1+\xi\frac{u-\mu}{\beta}\right)^{-\frac{1}{\xi}}. \tag{2.30}$$

- The *extra capital provision for operational risk* over the unexpected loss threshold u is estimated as the expectation of the excess loss distribution, scaled by the intensity $\lambda\left(u\right)$ of the Poisson process:
$$\lambda\left(u\right)E\left(X-u \mid X>u\right)=\lambda\left(u\right)\frac{\sigma+\xi u}{1-\xi}, \tag{2.31}$$
where u,σ,ξ and λ are the POT model parameters and time is measured in the same units as the frequency of data collection (years, months, days, etc.).

- The *total amount of capital* provided against extreme operational risks for a time period of length T will be calculated by:
$$u_T+\lambda\left(u\right)T\,E\left(X-u \mid X>u\right)=u_T+\lambda\left(u\right)T\,\frac{\sigma+\xi u}{1-\xi} \tag{2.32}$$

[5] A risk measure $R\left(\cdot\right)$ is coherent if it satisfies:

1. Monotonicity: If $X\leq Y\Rightarrow R\left(X\right)\leq R\left(Y\right)$.
2. Positive homogeneity: For all $\lambda\geq 0$ $R\left(\lambda X\right)=\lambda R\left(X\right)$.
3. Translation invariance: For all $\kappa\in\mathbb{R}$, $R\left(X+\kappa\right)$=$R\left(X\right)+\kappa$.
4. Sub-additivity: $R\left(X+Y\right)\leq R\left(X\right)+R\left(Y\right)$.

where u_T may in the first instance be considered equal to u, under the assumption of max-stability[6].

Assuming that the excess distribution above u is GPD, we have that:

$$\begin{aligned}\bar{F}(x) &= P(X > u)\, P(X > x \mid X > u) \\ &= \bar{F}(u)\, P(X - u > x - u \mid X > u) \\ &= \bar{F}(u)\, \bar{F}_u(x - u) \\ &= \bar{F}(u) \left(1 + \xi \frac{x - u}{\sigma}\right)^{-1/\xi}. \end{aligned} \tag{2.33}$$

This formula may be inverted to obtain a high quantile of the underlying distribution, which we interpret as a Value-at-Risk (VaR). For $q \geq F(u)$ we have that VaR at level q is equal to

$$\mathrm{VaR}_q = u + \frac{\sigma}{\xi}\left(\left(\frac{1-q}{\bar{F}(u)}\right)^{-\xi} - 1\right). \tag{2.34}$$

When VaR is exceeded, the actual loss can be much higher than VaR. To employ a coherent risk measure, we consider the Expected Shortfall or conditional VaR (CVaR), i.e. the expected loss once the VaR is exceeded. This measure is given by:

$$\mathrm{ES}_q = E(X \mid X > \mathrm{VaR}_q) = \frac{1}{1-q}\int_q^1 \mathrm{VaR}_x \, dx. \tag{2.35}$$

Assuming the GPD approximation above u, the Expected Shortfall can be calculated using (2.20) as

$$\mathrm{ES}_q = \mathrm{VaR}_q + E(X - \mathrm{VaR}_q \mid X > \mathrm{VaR}_q) = \frac{\mathrm{VaR}_q}{1-\xi} + \frac{\sigma - \xi u}{1-\xi}. \tag{2.36}$$

We can obtain estimates of both VaR and ES. Replacing $\bar{F}(u)$ by its empirical estimate N_u/n in (2.34), and replacing ξ and σ by their estimates, we get:

$$\widehat{\mathrm{VaR}}_q = u + \frac{\hat{\sigma}}{\hat{\xi}}\left(\left(\frac{n(1-q)}{N_u}\right)^{-\hat{\xi}} - 1\right) \tag{2.37}$$

[6]Max-stability means that if $X_1, ..., X_d$ are iid copies of X,

$$\max(X_1, ..., X_d) \overset{d}{=} \frac{X - a_d}{b_d}$$

and

$$\widehat{\mathrm{ES}}_q = \frac{\widehat{\mathrm{VaR}}_q}{1-\hat{\xi}} + \frac{\hat{\sigma} - \hat{\xi}u}{1-\hat{\xi}}. \tag{2.38}$$

2.2.5 Limitations of Extreme Value Theory

Extreme Value Theory (EVT) models are based upon an asymptotic approximation for the tail distribution, which are very flexible in terms of the allowable tail shape behaviour. The attraction of EVT based methods is that they can provide mathematically and statistically justifiable parametric models for the tails of any distribution which can give reliable extrapolations beyond the range of the observed data.

The Pickands-Balkema-de Haan theorem states that for sufficiently large values of the threshold, under certain regularity conditions, the generalized Pareto distribution (GPD) is the limit distribution of exceedances. However, one of the main issues in applying the classical GPD approach is the threshold selection (i.e. at which level of extremity into the tails of the data the GPD is a good model).

The threshold selection is a balance between reliability of the asymptotic approximation versus the sample variance of estimators. The threshold must be sufficiently high to ensure the threshold excesses are approximately GPD. However, the threshold cannot be too high as this will reduce the sample size available for inference and thus increase variability of the estimators.

As we saw in the previous sections, traditionally, data analysis with such models is performed in two steps. In the first one, the threshold u, is chosen graphically by looking at the mean excess plot (or some other graphical tools as described in Section 2.2.3) or simply setting it at some high percentile of the data. Once a suitable value has been determined, the threshold is then treated as a known fixed constant in later inference and the remaining parameters are estimated. This approach suffers from concerns over subjectivity about the threshold choice and not accounting for threshold uncertainty in

inference. Moreover, only the observations above the threshold are used in the second step.

Threshold selection is by no means an easy task, as observed by Davison and Smith (1990) and Coles and Tawn (1994), among others. Choosing the threshold through a mean excess plot or by choosing a certain percentile does not guarantee that an appropriate selection was made in order to prevent model bias which is crucial for the use of the asymptotic distribution as a model. Most of the literature has shown how threshold selection influences parameter estimation (see Smith (1987) and Coles and Tawn (1996)). Although some approaches have been developed to deal with these issues, the problem remains.

Finally, for many applications, threshold selection may be critical for the extrapolated tail behaviour, so the extra uncertainty associated with the threshold choice needs to be accounted for. In this setting, Bayesian inference offers an alternative framework which allows to overcome this uncertainty.

2.3 Bayesian inference in Extreme Value Theory

Bayesian inference is a powerful and increasingly popular statistical approach, which allows one to deal with complex problems in a conceptually simple and unified way. In Bayesian inference, parameters are random variables. Uncertainty or degree of belief with respect to the parameters is quantified by probability distributions. The basic idea of Bayesian inference is to set up a full probability model for both observed and unobserved quantities. Inference is then based on the so-called posterior density, i.e. the conditional density of the unobserved quantity conditional on the observed quantity. Additionally, it can be used to incorporate expert opinions into data analysis and to combine different data sources.

2.3.1 Bayesian framework

Bayesian techniques offer an alternative way to draw inferences from the likelihood function. As in the non-Bayesian setting, we assume data $x = (x_1, ..., x_n)$ to be realizations of a random variable whose density falls within a parametric family $\mathcal{F} = \{f(x;\theta) : \theta \in \Theta\}$ (θ is perhaps a collection of several parameters). However, parameters of a distribution are now treated as random variables, more precisely we assume that Θ is distributed according to the so-called prior density $\pi(\theta)$. The specification of this prior distribution enables us to supplement the information provided by the data—which, in extreme value analysis, is often very limited—with other sources of information.

Given $\Theta = \theta$ we model our observed data x using the probability density function $f(x;\theta)$. The likelihood function for θ is therefore $L(\theta \mid x) = f(x;\theta)$.

We can combine both the prior and the likelihood using Bayes Theorem, which states that

$$\pi(\theta \mid x) = \frac{\pi(\theta) L(\theta \mid x)}{f(x)}, \tag{2.39}$$

where

$$f(x) = \begin{cases} \int_\Theta \pi(\theta) L(\theta \mid x) d\theta & \text{if } \theta \text{ is continuous,} \\ \sum_\Theta \pi(\theta) L(\theta \mid x) & \text{if } \theta \text{ is discrete.} \end{cases} \tag{2.40}$$

Since $f(x)$ is not a function of θ, calculations (numerical and algebraic) are usually required only up to a proportionality constant. Therefore, we write Bayes theorem as

$$\pi(\theta \mid x) \propto \pi(\theta) \times L(\theta \mid x), \tag{2.41}$$

i.e.,

$$\text{posterior} \propto \text{prior} \times \text{likelihood}.$$

$\pi(\theta \mid x)$ is the posterior distribution of the parameter vector θ, $\theta \in \Theta$, i.e. the distribution of θ after the inclusion of the data. This posterior distribution is often of great interest, since the prior-posterior changes represent the changes in our beliefs after the data has

been included in the analysis, hence the posterior density can be interpreted as our updated knowledge about Θ after having observed x. Inference is typically based on reproducing all or parts of the posterior density graphically (as graphs or contour plots). Another option is to report e.g. posterior mean, mode, and quantiles.

However, many problems in Bayesian inference leave us with intractable distributions that cannot be expressed in a closed form. The posterior or joint distribution that we are interested in is often of high dimensionality, and in cases like mixture models can exhibit an exponentially increasing number of modes. Therefore, we need a way of understanding posterior densities which does not rely on being able to analytically integrate the kernel of the posterior.

To solve this problem, numerous simulation-based methods have been developed and implemented within the Bayesian paradigm, e.g. importance sampling (Ripley, 1987), Markov Chains Monte Carlo (MCMC) algorithms (Casella and Robert, 1999) and particle filtering (Doucet et al., 2001).

For this purpose, we will focus on Markov chain Monte Carlo (MCMC) methods, which can be used to generate samples from the posterior distribution. A more detailed explanation of MCMC methods is provided in Appendix A.

2.3.2 Basics of Bayesian inference for extremes

There are several reasons for preferring a Bayesian analysis of extremes over the more traditional likelihood approach. Since extreme data are (by their very nature) quite scarce, the ability to incorporate other sources of information through a prior distribution has obvious appeal. Bayes' Theorem also leads to an inference that comprises a complete distribution, meaning that the variance of the posterior distribution, for example, can be used to summarize the precision of the inference, without having to rely upon asymptotic theory. Furthermore, the concept of the predictive distribution is implicit in the Bayesian framework. This distribution describes how likely are different outcomes of a

future experiment. The predictive probability density function is given by

$$f(y \mid x) = \int_{\Theta} f(y \mid \theta)\, \pi(\theta \mid x)\, d\theta \tag{2.42}$$

when θ is continuous, and analogously when θ is discrete.

From equation (2.42), we can see that the predictive distribution is formed by weighting the possible values for θ in the future experiment $f(y \mid \theta)$ by how likely we believe they are to occur after observing the data.

For example, a suitable model for threshold excess $Y = X - u$ is $Y \sim GPD(\sigma, \xi)$. Estimation of $\theta = (\sigma, \xi)$ could be made on the basis of previous observations $x = (x_1, ..., x_n)$.

Thus, in the Bayesian framework, we would have

$$P(Y \leq y \mid x_1, ..., x_n) = \int_{\Theta} P(Y \leq y \mid \theta)\, \pi(\theta \mid x)\, d\theta. \tag{2.43}$$

Equation (2.43) gives the distribution of a future threshold excess, allowing for both parameter uncertainty and randomness in future observations.

Solving

$$P(Y \leq y \mid x_1, ..., x_n) = q \tag{2.44}$$

for y therefore gives an estimate of the Value-at-Risk at level $q \in (0, 1)$ that incorporates uncertainty due to model estimation.

Although (2.42) may seem analytically intractable, it can be approximated if the posterior distribution has been estimated using, for example, MCMC. After removal of the "burn–in" period, the MCMC procedure gives a sample $\theta_1, ..., \theta_B$ that can be regarded as realizations from the stationary distribution $\pi(\theta \mid x)$. Thus

$$P(Y \leq y \mid x_1, ..., x_n) \approx \frac{1}{B} \sum_{i=1}^{B} P(Y \leq y \mid \theta_i), \tag{2.45}$$

which we can solve for y using a numerical solver.

Another reason lending appeal to Bayesian inference for extremes is that it is less dependent on the regularity assumptions required by the theory of maximum likelihood.

For example, when $\xi < -0.5$, maximum likelihood estimation breaks down (Smith, 1985). In this case a Bayesian approach provides a feasible alternative.

2.3.3 Combining different data sources

As mentioned in Section 2.1.4, Basel II requires that operational risk models include the use of several different sources of information: internal data, relevant external data, scenario analysis and factors reflecting the business environment and internal control systems.

Additionally, it is widely recognized that estimation of operational risk distributions cannot be done exclusively using historical data given the limited number of data available and the difficulty of predicting future losses in a banking environment which is constantly changing.

Thus, combining different sources of information is critical for estimation of operational risk, especially for low-frequency/high-severity risks.

Shevchenko (2011) considers three ways that have been proposed to process different data sources of information:

1. Ad-hoc procedures.
2. Bayesian methods.
3. General non-probabilistic methods.

We focus on the first two procedures only. General non-probabilistic methods are mainly based on the Dempster-Shafer theory and probabilistic boxes. More on these subjects can be found in Ferson et al. (2003).

2.3.3.1 Ad-hoc procedures

One suggested method for the ad-hoc combining is based on mixing of distributions, where the loss distribution is expressed as:

$$w_1 F_{SA}(x) + w_2 F_{int}(x) + (1 - w_1 - w_2) F_{ext}(x), \tag{2.46}$$

where $F_{SA}(x)$, $F_{int}(x)$ and $F_{ext}(x)$ are the distributions identified by scenario analysis, internal data and external data, respectively, using expert specified weights w_1 and w_2.

After that, we can apply the *minimum variance principle*, where the combined estimator is a linear combination of the individual estimators obtained from internal data, external data and expert opinion separately, with the weights chosen to minimize the variance of the combined estimator.

Consider two unbiased independent estimates $\hat{\theta}_1$ and $\hat{\theta}_2$ for a parameter θ, i.e. $E\left[\hat{\theta}_m\right] = \theta$; $\text{Var}\left[\hat{\theta}_m\right] = \sigma_m^2$, $m = 1, 2$.

The combined unbiased estimator is:

$$\hat{\theta}_{tot} = w_1\hat{\theta}_1 + w_2\hat{\theta}_2, \;\; w_1 + w_2 = 1 \tag{2.47}$$

and

$$\text{Var}\left(\hat{\theta}_{tot}\right) = w_1^2\sigma_1^2 + (1 - w_1)^2\sigma_2^2 \tag{2.48}$$

Then choose weights to minimize $\text{Var}\left(\hat{\theta}_{tot}\right)$, namely

$$\hat{w}_1 = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}, \quad \hat{w}_2 = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}. \tag{2.49}$$

This technique can be easily extended to combine three or more estimators.

$$\hat{\theta}_{tot} = w_1\hat{\theta}_1 + \cdots + w_K\hat{\theta}_K, \quad w_1 + \cdots + w_K = 1 \tag{2.50}$$

and

$$w_i = \frac{1/\sigma_i^2}{\sum_{k=1}^{K}(1/\sigma_k^2)}. \tag{2.51}$$

Heuristically, this can be applied to almost any quantity, including a distribution parameter or distribution characteristic such as mean, variance or quantile. The assumption that the estimators are unbiased estimators for θ is probably reasonable when combining

estimators from different experts (or from expert and internal data). However, it is certainly questionable if applied to combine estimators from the external and internal data (Shevchenko, 2011).

2.3.3.2 Bayesian methods

Different methodologies have been proposed in the literature in order to combine different sources of information. Here, we only show part of the methodology suggested by Lambrigger et al. (2007). They consider the following assumptions:

- Loss frequency and severity are modelled by parametric distributions (e.g. Poisson for the frequency or Pareto, lognormal, etc. for the severity). In any case, the parameter vector θ has to be estimated.

- Before we have any internal data, only external data are available and the best prediction for the parameter θ is given by the belief in the available external knowledge. The parameter of interest is modelled by a prior distribution corresponding to a random vector Θ.

- The true specific parameter θ_0 is treated as a realization of a random vector Θ, where Θ corresponds to the whole data (including external and internal data) and θ stands for the unknown parameter of the specific entity being considered.

- As time passes, internal data $X = (X_1, ..., X_K)'$ and expert opinions $\Delta = (\Delta_1, ..., \Delta_M)$ about θ become available.

- X and Δ are assumed to be conditionally independent given Θ, with joint density

$$h(\mathbf{x}, \delta \mid \theta) = h_1(\mathbf{x} \mid \theta)\, h_2(\delta \mid \theta), \tag{2.52}$$

  where h_1 and h_2 are the conditional densities of X and Δ given Θ.

- Assuming that observations and expert opinions are conditionally independent and identically distributed, given $\Theta = \theta$, we have that

$$h_1(\mathbf{x} \mid \theta) = \prod_{k=1}^{K} f_1(x_k \mid \theta), \tag{2.53}$$

$$h_2(\delta \mid \theta) = \prod_{m=1}^{M} f_2(\delta_m \mid \theta), \tag{2.54}$$

  where f_1 and f_2 are the marginal densities of a single internal observation and a single expert opinion, respectively.

Now, by considering the unconditional joint density of X and Δ, denoted by $h(\mathbf{x}, \delta)$ and using Bayes' theorem,

$$h(\mathbf{x}, \delta \mid \theta)\, \pi(\theta) = \pi(\theta \mid \mathbf{x}, \delta)\, h(\mathbf{x}, \delta) \tag{2.55}$$

and hence the posterior density satisfies

$$\pi(\theta \mid \mathbf{x}, \delta) \propto \pi(\theta)\, h(\mathbf{x}, \delta \mid \theta). \tag{2.56}$$

The posterior density is then

$$\pi(\theta \mid \mathbf{x}, \delta) \propto \pi(\theta) \prod_{k=1}^{K} f_1(x_k \mid \theta) \prod_{m=1}^{M} f_2(\delta_m \mid \theta). \tag{2.57}$$

2.3.4 Prior elicitation

Elicitation is another part of the process of statistical modelling. Garthwaite et al. (2004) define elicitation as *"the process of formulating a person's knowledge and beliefs about one or more uncertain quantities into a (joint) probability distribution for those quantities*". In a Bayesian context, it arises most usually as a method for specifying the prior distribution for one or more unknown parameters of a statistical model. However, there are some other contexts in which elicitation is important.

2.3.4.1 The elicitation process

The elicitation of information is not an easy task, even if it only aims to extract the expert's beliefs about an event or particular hypothesis. Even when the expert is familiar with probabilities and their meaning, it is not easy to assign a probability value to an event accurately.

It is convenient to think of the elicitation process as a task that involves a facilitator who assists the expert to formulate his knowledge in a probabilistic way. Sometimes, if the expert is familiar with statistical concepts, then there may be no formal need for a facilitator, although this is rare in practice.

To carry out this process, four stages have been identified:

1. Setup. It consists of selecting and training the expert(s), identifying what aspects of the problem to elicit, etc.

2. Elicitation. It is the extraction of information from the expert and it is the core of the process.

3. Fit. A probability distribution is fitted to the information obtained in the second stage.

4. Evaluation. This stage involves assessing the adequacy of the elicitation, with the option of returning to the second stage and eliciting more summaries from the expert(s).

Despite the difficulties to carry out the elicitation process, it is a valuable tool since it facilitates decision-making and allows to make inference. Furthermore, it brings the analysis closer to the application by demanding attention to what is being modelled, and what is reasonable to believe about it. Similarly, it helps to summarize the posterior distribution by meaningful quantities.

To decide whether the process has been performed successfully, it is important to distinguish between the quality of an expert's knowledge and the accuracy with which this knowledge has been translated into a probabilistic form.

A successful elicitation process faithfully represents the opinion of the person being elicited, which does not necessarily mean that it is the correct view.

2.3.4.2 Elicitation and extreme events

It is reasonable to hope that experts should provide relevant prior information about extremal behaviour, since they have specific knowledge of the characteristics of the data under study.

Unfortunately, when we consider unlikely events, these are hard to evaluate. In this case, expert opinion plays an important role due to the scarcity of data. It has been observed that in general people are capable of estimating proportions, modes and medians of samples. However, when dealing with highly skewed distributions several errors may occur.

To avoid these problems, the questions should be formulated in an intuitive way. Steinhoff and Baule (2006) suggest the following questions:

1. Which events of severity between x and y do you remember?
2. How many of those events have happened in a year on average?
3. How many of those events happened in a good year (minimum) and in a bad year (maximum)?

They also point out that these questions are quite easy if, on average, there is more than one event a year. However, the crucial losses are usually the low-frequency, high-impact events that occur very rarely. In such case, they suggest some changes to the questions as follows:

2a. How many extreme events happened in the past 10 or xx years?

3a. How many extreme events can happen in the chosen time range in worst- and best-case scenarios?

To obtain more accurate results, the last two questions might be changed into:

- How many years will we have to wait, all things being equal, to observe an event of severity x (or above)?

2.3.4.3 Distribution fitting

As we saw in the previous paragraphs, expressing prior beliefs directly in terms of the distribution parameters is not a simple job. Depending on the questions we ask, expert opinions on potential losses and corresponding probabilities are often expressed as:

- Opinion on the distribution parameter;
- Opinions on the number of losses with the amount to be within some ranges;
- Separate opinions on the frequency of the losses and quantiles of the severity;
- Opinion on how often the loss exceeding some level may occur.

Usually, if the expected values for the quantiles (or mean) and their uncertainties are estimated by the expert then it is possible to fit the priors.

It is very unusual that experts express their beliefs in terms of the distribution parameters. Instead, they provide some quantities such as quantiles or other risk characteristics. If we consider this situation, it might be better to assume some priors for these quantities that will imply a prior for the parameters.

Let $\theta = (\theta_1, ..., \theta_n)$ be the model parameters and $d_i = g_i(\theta)$, $i = 1, ..., n$ be risk characteristics that can be elicited from experts, such as specific quantiles, expected values, return level, etc. Assuming that experts specify the joint prior $\pi(d_1, ..., d_n)$, we can obtain the prior for $\theta_1, ..., \theta_n$ by using the transformation method as follows:

$$\pi(\theta) = \pi(g_1(\theta), ..., g_n(\theta)) \left| \frac{\partial(g_1(\theta), ..., g_n(\theta))}{\partial(\theta_1, ..., \theta_n)} \right|, \tag{2.58}$$

where the second factor on the right-hand side is the Jacobian determinant of the transformation.

Since there is possible dependence between parameters, it is helpful to choose characteristics such that independence can be assumed.

Coles and Tawn (1996) and Coles and Powell (1996) elicit prior information in terms of extreme quantiles, arguing that this is a scale on which an expert is most likely to be able to accurately quantify their prior beliefs about extremal behaviour.

If the prior for quantiles $q_1 < \cdots < q_n$ (for specific probability values $p_1 < \cdots < p_n$) is to be specified, in order to respect the ordering we can work instead with the differences

$$d_1 = q_1 - e_1, \ d_2 = q_2 - q_1, ..., d_n = q_n - q_{n-1},$$

where e_1 is a physical lower end point for the process variable.

Under this setting it is reasonable to assume independence between these differences and impose constraints $d_i > 0$, $i = 2, ..., n$.

If the experts assign marginal priors $\pi(d_1), ..., \pi(d_n)$ then the full joint prior is

$$\pi(d_1, ..., d_n) = \pi(d_1) \times \cdots \times \pi(d_n). \tag{2.59}$$

Hence, the prior for $\theta_1, ..., \theta_n$ can be calculated using (2.58).

2.3.5 Advantages and challenges of Bayesian inference in EVT

Recently, Bayesian inference has been focused on the development of mixture type models, which typically treat the threshold as a model parameter to be estimated, and so also automatically accounts for the uncertainty associated with the threshold selection. For example, Behrens et al. (2004) use a truncated Gamma distribution for observations below the threshold and the GPD approach for observations exceeding the threshold. Tancredi et

al. (2006) propose to model extreme and non-extreme data with a distribution composed of a piecewise constant density from a low threshold up to an unknown end point and a GPD with threshold for the remaining tail part.

In both examples, Bayesian inference is used for fitting the mixture model as it can take advantage of any expert prior information, which can be important in tail estimation due to the inherent sparsity of extremal data. There are distinct benefits and potential drawbacks to the mixture modelling approach when compared to the classical fixed threshold method. The principal advantages are that the threshold is estimated avoiding the often subjective choice in the classical approach and the uncertainty associated with the estimation is accounted for in inference, which is rather challenging for the fixed threshold method. The automated threshold estimation is a major benefit when trying to automate fitting the GPD to multiple datasets.

The principal drawbacks are the added complexity of estimating the additional parameters and the fit in the bulk of the distribution (or the alternate tail) may have an influence on the tail fit. It is clear that different parameters values could give similar model fits.

However, in most of the cases, Bayesian inference provides reliable parameter estimates, the threshold included. This will be studied more closely in the next chapters.

2.4 Conclusions of the Chapter

In this chapter we have studied the importance of measuring operational risk, as well as the integrated risk framework established by Basel II, which defines the guidelines to measure the capital adequacy of financial institutions. Furthermore, we have presented the fundamentals of Extreme Value Theory through various definitions and theorems, along with practical aspects for estimating and assessing statistical models for heavy-tailed events.

One of the most important aspects that has been addressed in this chapter is the

subjectivity in the threshold selection by using graphical methods. This is one of the most controversial aspects of Extreme Value Theory and, possibly, its main weakness.

Moreover, we have shown how Bayesian inference is a powerful alternative to handle the threshold issue and to include expert opinion, as stated in Basel II. Similarly, we have introduced the concept of elicitation and showed how this process may be carried out, playing a fundamental role in the estimation due to the scarcity of data.

All the material presented in this chapter will help us to understand the work presented in subsequent chapters.

Chapter 3

A Bayesian model for operational risk

The use of Bayesian methods in extreme value modelling has recently become more common. Several models have been proposed in the literature. For example, Pickands (1994) discusses Bayesian estimation of extreme quantiles assuming independent non-informative priors among parameters. Bermudez et al. (2001) propose a Bayesian predictive approach for the choice of the threshold. Behrens et al. (2004) develop an extreme value mixture model by combining a parametric bulk model below the threshold with a GPD above the threshold. Several models have been derived from this last mixture model. For instance, Mendes and Lopes (2004) and Zhao et al. (2009) propose a mixture with a normal distribution for the bulk, with both tails represented by separate threshold models. Carreau and Bengio (2009) propose a hybrid Pareto by splicing a normal distribution with a GPD and setting continuity constraints on the density and on its first derivative at the threshold. Cabras and Castellanos (2010) consider a semiparametric bulk model spliced with a GPD upper tail. Do Nascimento et al. (2011) extended the model of Behrens et al. (2004) by defining the bulk distribution as a weighted mixture of Gamma densities.

Due to its notable influence, for the purposes of this thesis, we focus our attention

on the model developed by Behrens et al. (2004). We refer the reader to Scarrott and MacDonald (2012) for a detailed review of Bayesian methods in extreme value modelling.

3.1 General model

The model proposed in this chapter is based on the work of Behrens et al. (2004), where the uncertainty in the threshold selection is incorporated by choosing a prior, possibly flat, for it to compose the model.

Consider $X_1, ..., X_n$ independent and identically distributed observations and u the threshold. We assume that observations below the threshold come from a certain distribution with parameters η, denoted by $F_H(\cdot \mid \eta)$, while those above the threshold come from a GPD, denoted by $F_G(x \mid u, \sigma, \xi)$. Therefore, the distribution function F_B, of any observation X, can be written as:

$$F_B(x \mid \eta, u, \sigma, \xi) = \begin{cases} F_H(x \mid \eta), & x < u, \\ F_H(u \mid \eta) + [1 - F_H(u \mid \eta)] F_G(x \mid u, \sigma, \xi), & x \geq u. \end{cases} \tag{3.1}$$

The likelihood is:

$$L(\theta; x) = \prod_A f_H(x \mid \eta) \prod_B [1 - F_H(u \mid \eta)] \left[\frac{1}{\sigma} \left(1 + \frac{\xi(x_i - u)}{\sigma} \right)_+^{-\frac{1+\xi}{\xi}} \right] \tag{3.2}$$

for $\xi \neq 0$ and

$$\prod_A f_H(x \mid \eta) \prod_B [1 - F_H(u \mid \eta)] \left[\frac{1}{\sigma} \exp\left\{ \left(\frac{x_i - u}{\sigma} \right) \right\} \right] \tag{3.3}$$

for $\xi = 0$, where $\theta = (\eta, u, \sigma, \xi)$, $x = (x_1, ..., x_n)$, $A = \{i : x_i < u\}$ and $B = \{i : x_i \geq u\}$.

Figure 3.1 is a representation of this model. We can observe that the density has a discontinuity at the threshold u. Depending on the parameters, the density jump can be larger or smaller, and in each case the choice of which observations will be considered as exceedances can be more obvious or less evident.

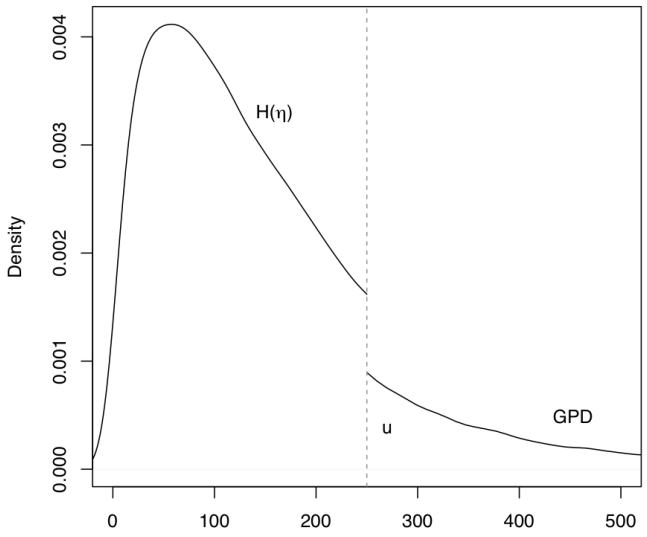

Figure 3.1: *Representation of the mixture model*

3.1.1 Priors

Prior for parameters below the threshold

To model observations below the threshold through a parametric form, it is always better to try to obtain a conjugate prior to simplify the problem analytically.

Given that operational losses are non-negative, a convenient choice is the Gamma distribution. We have $\eta = (\alpha, \beta)$ and reparameterizing as $\mu = \alpha/\beta$, we can set $\alpha \sim \text{Gamma}(a, b)$ and $\mu \sim \text{Gamma}(c, d)$ where a, b, c and d are known hyperparameters. Therefore the joint prior of $\eta = (\alpha, \beta)$ is:

$$\pi(\eta) = \frac{b^a}{\Gamma(a)} \alpha^{a-1} e^{-b\alpha} \frac{d^c}{\Gamma(c)} \left(\frac{\alpha}{\beta}\right)^{c-1} e^{-d\alpha/\beta} \left(\frac{\alpha}{\beta^2}\right). \tag{3.4}$$

Prior for the threshold

To set up a prior distribution for the threshold, we can assume that u follows a truncated normal distribution with parameters (μ_u, σ_u), truncated from below at e_1 (the minimum of the data). Setting μ_u at some high data percentile and σ_u large enough to represent a

fairly non-informative prior, the density becomes, for $u > e_1$,

$$\pi\left(u \mid \mu_u, \sigma_u^2, e_1\right) = \frac{1}{\sqrt{2\pi\sigma_u^2}} \frac{\exp\left\{-\frac{1}{2}\left(\frac{u-\mu_u}{\sigma_u}\right)^2\right\}}{\Phi\left[-\left(\frac{e_1-\mu_u}{\sigma_u}\right)\right]}. \tag{3.5}$$

A continuous uniform prior is an alternative. A discrete distribution can also be assumed. In this case, u could take any value between certain high data percentiles which is convenient for applications as it facilitates posterior computation.

Prior for the GPD parameters

Coles and Tawn (1996) point out that increasing σ or ξ leads to a longer-tailed distribution, so a priori negative dependence between these parameters is expected. To avoid the assumption of independent priors on the GPD parameters, we use an elicitation method as outlined in Section 2.3.4.3.

In their paper, Behrens et al. (2004) call $q = u + \frac{\sigma}{\xi}\left(p^{-\xi} - 1\right)$ the return level[1] and elicitation of the prior information is done in terms of (q_1, q_2, q_3), in the case of location–scale parameterization of the GPD, for specific values of $p_1 > p_2 > p_3$. We consider more appropriate to treat q as the Value-at-Risk, since it is derived in terms of the GPD parameters. We should also notice that in this case p is a fixed value and the probability of exceeding the threshold is not considered. We will return to this issue later in Chapter 4, when we discuss prior sensitivity. For now, our main concern is to provide some insight into how the priors are chosen.

In Behrens et al. (2004), the prior for ξ and σ is computed by choosing two levels $q_1 < q_2$ and setting $d_1 = q_1$ and $d_2 = q_2 - q_1$ to be Gamma. Concretely, we have the following Gamma distributions with known hyperparameters: $d_1 = q_1 \sim \text{Gamma}(a_1, b_1)$

[1]The return level is the level that is expected to be exceeded, on average, once every $1/p$ years (called the return period), where p is the probability of the extreme event occurring. It is obtained by computing the inverse of the GEV, although in their paper Behrens et al. (2004) define the return level in terms of the GPD parameters.

and $d_2 = q_2 - q_1 \sim \text{Gamma}(a_2, b_2)$. The hyper parameters a_1, b_1, a_2 and b_2 are typically obtained from the experts' information; this will be treated in detail in Chapter 4.

In order to derive the marginal prior distribution for σ and ξ, we have:

$$\pi(\sigma,\xi) \propto \pi(d_1)\,\pi(d_2)\left|\frac{\partial(d_1,d_2)}{\partial(\sigma,\xi)}\right| = \left|\frac{\partial d_1}{d\sigma}\times\frac{\partial d_2}{d\xi}\right| - \left|\frac{\partial d_1}{d\xi}\times\frac{\partial d_2}{d\sigma}\right|. \tag{3.6}$$

To find the Jacobian $\left|\frac{\partial(d_1,d_2)}{\partial(\sigma,\xi)}\right|$:

$$\begin{aligned}
\frac{\partial d_1}{d\sigma} &= \frac{p_1^{-\xi}}{\xi} - \frac{1}{\xi}, \\
\frac{\partial d_2}{d\sigma} &= \frac{p_2^{-\xi}}{\xi} - \frac{p_1^{-\xi}}{\xi}, \\
\frac{\partial d_1}{d\xi} &= -\frac{\sigma p_1^{-\xi}}{\xi^2} + \frac{\sigma p_1^{-\xi}\ln p_1}{\xi} + \frac{\sigma}{\xi^2}, \\
\frac{\partial d_2}{d\xi} &= -\frac{\sigma p_2^{-\xi}}{\xi^2} + \frac{\sigma p_2^{-\xi}\ln p_2}{\xi} + \frac{\sigma p_1^{-\xi}}{\xi^2} - \frac{\sigma p_1^{-\xi}\ln p_1}{\xi}.
\end{aligned} \tag{3.7}$$

Hence

$$\frac{\partial d_1}{d\sigma}\times\frac{\partial d_2}{d\xi} - \frac{\partial d_1}{d\xi}\times\frac{\partial d_2}{d\sigma} = \frac{\sigma\,(p_1p_2)^{-\xi}\,\ln p_2}{\xi^2} - \frac{\sigma p_2^{-\xi}\ln p_2}{\xi^2} + \frac{\sigma p_1^{-\xi}\ln p_1}{\xi^2} - \frac{\sigma\,(p_1p_2)^{-\xi}\,\ln p_1}{\xi^2}, \tag{3.8}$$

and therefore

$$\begin{aligned}
\pi(\sigma,\xi) \propto\ & \left[u + \tfrac{\sigma}{\xi}\left(p_1^{-\xi} - 1\right)\right]^{a_1-1} \exp\left[-b_1\left\{u + \tfrac{\sigma}{\xi}\left(p_1^{-\xi} - 1\right)\right\}\right] \\
& \times \left[u + \tfrac{\sigma}{\xi}\left(p_2^{-\xi} - p_1^{-\xi}\right)\right]^{a_2-1} \exp\left[-b_2\left\{\tfrac{\sigma}{\xi}\left(p_2^{-\xi} - p_1^{-\xi}\right)\right\}\right] \\
& \times \left|-\tfrac{\sigma}{\xi^2}\left[(p_1p_2)^{-\xi}\left(\log p_2 - \log p_1\right) - p_2^{-\xi}\log p_2 + p_1^{-\xi}\log p_1\right]\right|
\end{aligned}. \tag{3.9}$$

If we consider the case $\xi = 0$, it is possible to assign a positive probability to this point and then the prior distribution would be a mixture between the prior distribution corresponding to the case $\xi = 0$ and the prior distribution corresponding to the case $\xi \neq 0$.

In finance, there are some risk measures used in practice and it would be convenient to specify priors in terms of these measures. This will be discussed in the next chapter.

3.1.2 Posterior inference

From the likelihood (3.2) and the prior distributions specified before, using the Bayes theorem we can obtain the posterior distribution, which is given as follows (Section 3.4,

Behrens et al. 2004):

$$\begin{aligned}\log p(\theta \mid x) = \; & K + \sum_{i=1}^{n} I\left(x_i < u\right)\left[\alpha \log\beta - \log\Gamma\left(\alpha\right) + \left(\alpha - 1\right)\log x_i - \beta x_i\right] \\ & \sum_{i=1}^{n} I\left(x_i \geq u\right)\log\left[1 - \int_0^u \tfrac{\beta^\alpha}{\Gamma(\alpha)} t^{\alpha-1} e^{-\beta t} dt\right] - \sum_{i=1}^{n} I\left(x_i \geq u\right)\log\sigma \\ & -\tfrac{1+\xi}{\xi}\sum_{i=1}^{n} I\left(x_i \geq u\right)\log\left[1 + \tfrac{\xi(x_i - u)}{\sigma}\right] \\ & +(a-1)\log\alpha - b\alpha + (c-1)\log\left(\tfrac{\alpha}{\beta}\right) - d\left(\tfrac{\alpha}{\beta}\right) + \log\left(\tfrac{\alpha}{\beta^2}\right) \\ & +\tfrac{1}{2\sigma_u}\left(u - \mu_u\right)^2 + (a_1 - 1)\log\left[u + \tfrac{\sigma}{\xi}\left(p_1^{-\xi} - 1\right)\right] \\ & -b_1\tfrac{\sigma}{\xi}\left(p_1^{-\xi} - 1\right) + (a_2 - 1)\log u + \tfrac{\sigma}{\xi}\left(p_2^{-\xi} - p_1^{-\xi}\right) - b_2\tfrac{\sigma}{\xi}\left(p_2^{-\xi} - p_1^{-\xi}\right) \\ & +\log\left|-\tfrac{\sigma}{\xi^2}\left[\left(p_1 p_2\right)^{-\xi}\left(\log p_2 - \log p_1\right) - p_2^{-\xi}\log p_2 + p_1^{-\xi}\log p_1\right]\right| \end{aligned} \tag{3.10}$$

In (3.10) K is the normalizing constant and $\theta = (\eta, u, \sigma, \xi)$. We only show the posterior with the likelihood for the case $\xi \neq 0$ and with a normal prior for the threshold. However, the case $\xi = 0$ can be considered in the model as well.

As expected, the posterior has no closed form and the implementation of Markov Chain Monte Carlo methods is required. Computations will be done via the Metropolis–Hastings steps within a blockwise algorithm. The algorithm is shown in Appendix A.

3.1.3 Performance in simulations

The algorithm was tested first with simulated data, generated for fixed values of α, β, u, σ and ξ, based on different characteristics such as skewness, tail behaviour and number of observations available for estimation of these parameters.

For each scenario we considered a sample size of $n = 1000$ and different combinations of parameters, by varying the value of the shape parameter (ξ); data below the threshold were simulated using a Gamma distribution.

100,000 Iterations					
Parameter	Posterior mean	Posterior median	Posterior std	Credible interval	True value
α	1.342	1.342	0.002	(1.268, 1.438)	1.350
β	0.292	0.290	0.003	(0.257, 0.330)	0.290
u	1.492	1.389	0.068	(1.211, 2.221)	1.400
σ	3.684	3.677	0.072	(3.170, 4.209)	3.500
ξ	-0.080	-0.083	0.003	(-0.182, 0.035)	-0.100

Table 3.1: *Posterior MCMC estimates for n=1000 simulated data from a Gamma-GPD mixture, with* $\xi = -0.1$

We display here the results for the following combination of parameters.

$$\begin{aligned} p &= 0.1, \\ \alpha &= 1.35, 0.5, \\ \beta &= 0.29, 0.2, \\ \sigma &= 3.5, \\ \xi &= -0.1, 0.2. \end{aligned}$$

The value of u is defined automatically by the sample size and the quantile p. We started by drawing a number n of observations from a Gamma(α, β) and u was defined as the $1-p$ quantile of this sample; n_1 observations below this value were retained, and the sample size was completed by drawing $n_2 = n - n_1$ observations from $u + \text{GPD}_{\sigma,\xi}$.

Results for $\xi = -0.1$ are shown in Table 3.1 and Figures 3.2 and 3.3. In all cases, the estimates are very close to the true values and convergence seems to be achieved. We also display the estimates for $\xi = 0.2$ in Table 3.2. Again, the estimates are close to the true values.

	100,000 Iterations				
Parameter	Posterior mean	Posterior median	Posterior std	Credible interval	True value
α	0.510	0.508	0.020	(0.493, 0.512)	0.5
β	0.203	0.202	0.014	(0.197, 0.212)	0.2
u	5.944	6.312	2.319	(5.998, 6.451)	6.0
σ	3.625	3.629	0.665	(3.484, 3.810)	3.5
ξ	0.187	0.178	0.121	(0.177, 0.191)	0.2

Table 3.2: *Posterior MCMC estimates for n=1000 simulated data from a Gamma-GPD mixture, with* $\xi = 0.2$

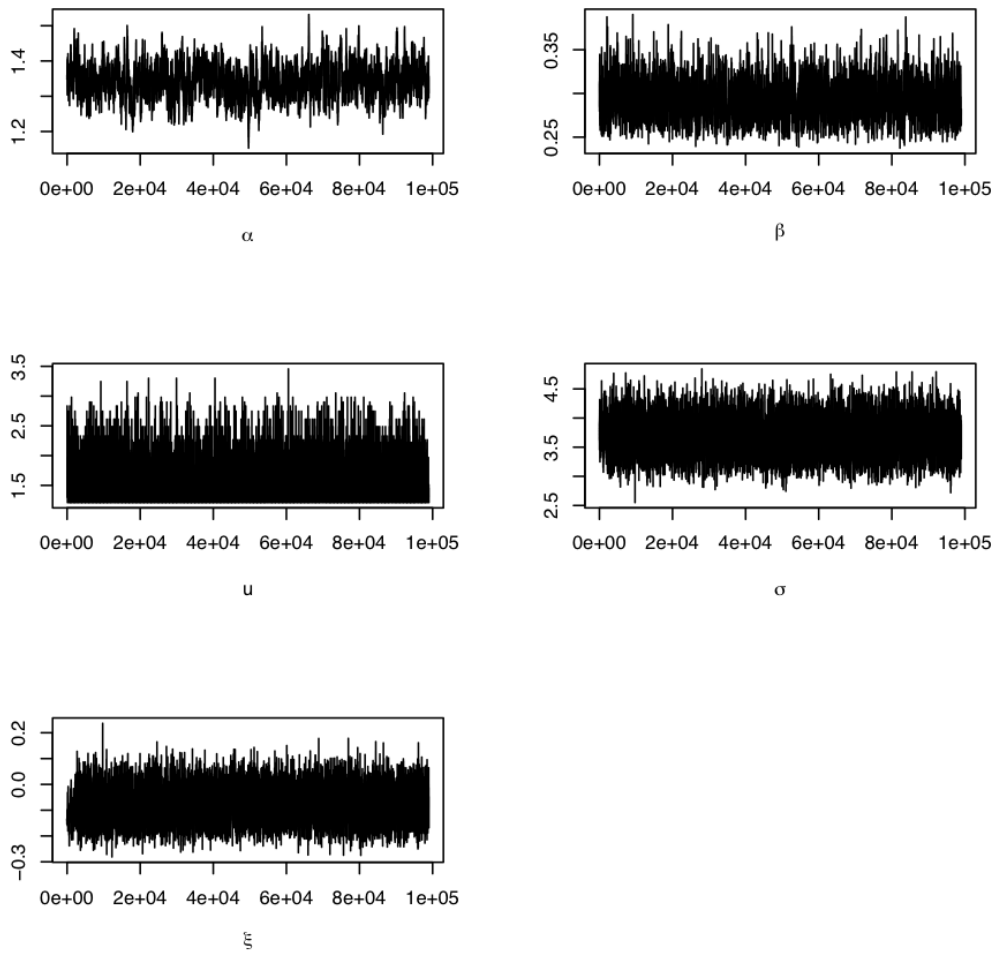

Figure 3.2: *Trace plots of the MCMC samples from the posterior density for n=1000 simulated data from a Gamma-GPD mixture, with* $\xi = -0.1$. *100,000 iterations after burn-in*

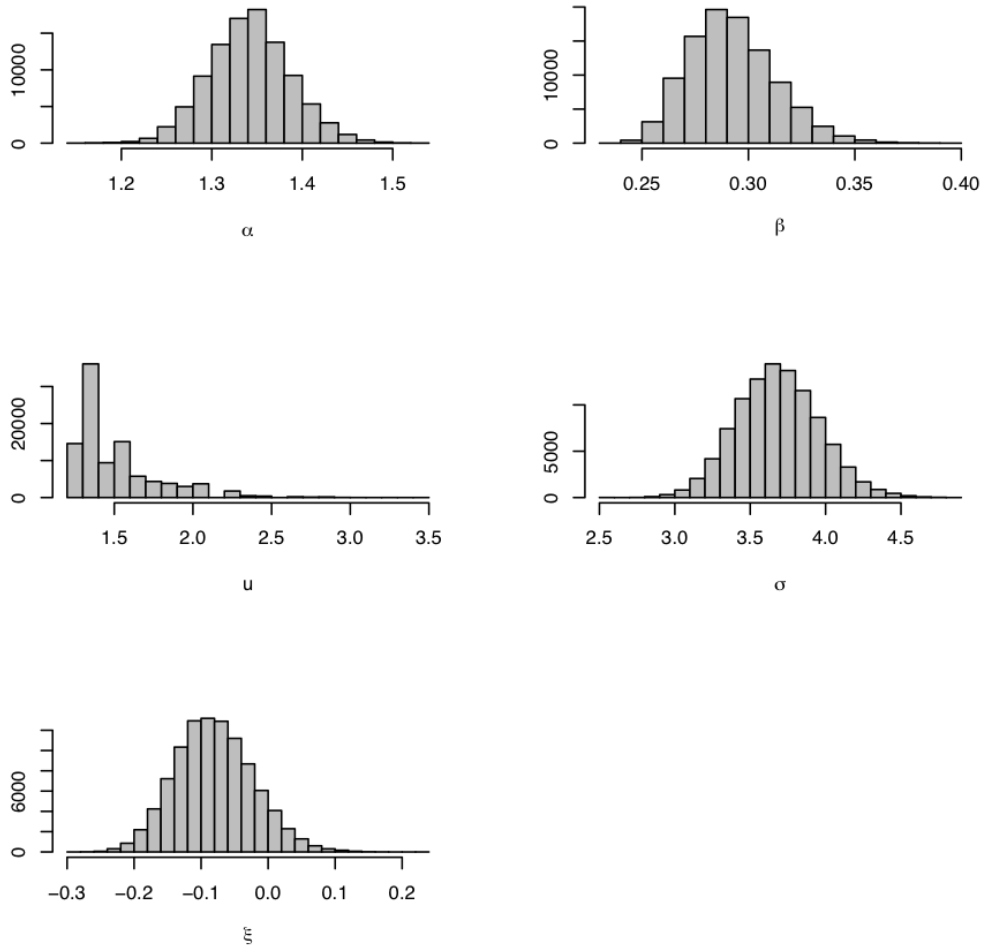

Figure 3.3: *Histograms of the MCMC samples from the posterior density for n=1000 simulated data from a Gamma-GPD mixture, with* $\xi = -0.1$. *100,000 iterations after burn-in*

3.2 An application to operational risk data

In this section, the objective is to make use of the Bayesian model presented in the previous section to analyze operational risk data.

3.2.1 Data description

Unfortunately, in many countries the record of operational risk losses has not yet been formalized. The number of institutions that report their operational losses and the avail-

ability of information are still limited, so obtaining data to carry out this work was not an easy task.

In an ideal scenario, we should have a database containing all risk events and their corresponding Basel II categories. However, the available data are still insufficient and most of them have been collected for a short period of time.

In spite of this fact, it was possible to collect fraud data, one of the most frequent operational risk events that has been a constant concern in the financial sector. The data consist of 626 observations for fraud losses in 41 banks in Mexico, recorded between January 2007 and April 2010. These data were obtained during a summer internship at the Bank of Mexico and were used for research purposes only. The names of the banks were changed into capital letters to preserve their anonymity.

Due to the large differences in banks sizes and their respective losses, these were scaled by the asset size of each bank at the time of the event considered and multiplied by one million[2].

3.2.2 Exploratory analysis

As a first approach, we determined the main descriptive statistics (Table 3.3). Similarly, the behaviour of the data over time was observed by using graphical tools (Figure 3.4). In all cases, one may identify the presence of possible extreme events. Also, we observe a concave shape in the quantile-quantile plot when it is compared to the exponential

[2]To perform the scaling of data it is important to consider the work of Shih et al. (2000); they found a relationship between the size of the institution and the magnitude of its losses. This paper highlights some aspects:

- The size of an institution is related to the magnitude of its loss, but the relationship is not necessarily linear.
- The size represents a very small proportion (about 5%) of the variability in the loss severity.

The scaling performed is a simple approach to the relationship between the size of the institution and the magnitude of its losses, as this is a subject of study itself, due to the heterogeneity between the banks considered. Particularly, for the fraud risk, the types of frauds and their effects may differ widely from one institution to another.

The use of scaled data allowed that not only losses corresponding to large banks exceeded the threshold, but also smaller banks were incorporated into the analysis.

distribution, indicating heavy-tailedness. This fact is supported by the value of the kurtosis and appearance of the histogram of losses.

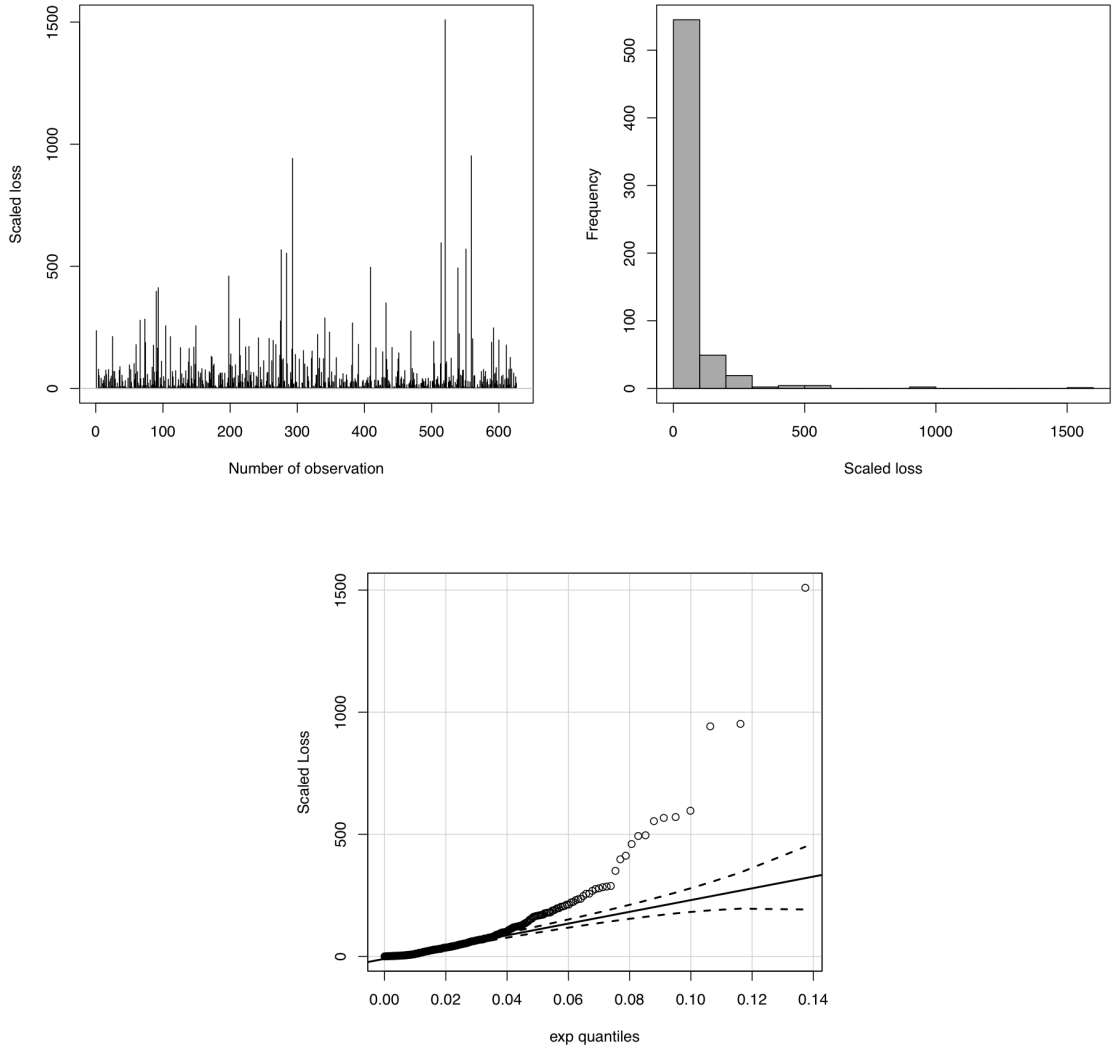

Figure 3.4: *Top left: Monthly fraud losses (scaled by the asset size) in 41 banks from 01/2007 to 04/2010. Top right: Histogram of scaled fraud losses from 01/2007 to 04/2010. Bottom: Exponential Q-Q plot for n=626 scaled fraud losses. We can infer the heavy-tailedness of the data in all cases.*

Number of observations: 626	Standard deviation: 110.793
Mean: 51.930	Median: 19.890
Minimum: 0.0002	Skewness: 6.604
Maximum: 1509.410	Kurtosis: 66.678
Percentile (q)	**Percentile value** (X_q)
$q = 0.25$	$X_{0.25} = 3.241$
$q = 0.50$	$X_{0.50} = 19.892$
$q = 0.75$	$X_{0.75} = 54.134$
$q = 1.00$	$X_{1.00} = 1509.416$

Table 3.3: *Descriptive statistics for n=626 fraud losses in 41 banks, recorded from 01/2007 to 04/2010*

Parameter	Runs		
	10,000	25,000	100,000
	Mean, Median, Std	Mean, Median, Std	Mean, Median, Std
α	0.553, 0.552, 0.029	0.553, 0.553, 0.029	0.552, 0.552, 0.029
β	0.014, 0.014, 0.001	0.014, 0.014, 0.001	0.014, 0.014,0.001
u	70.176, 72.980, 17.703	70.011, 72.980, 19.231	71.310, 72.980, 19.072
σ	93.033, 93.922, 19.847	93.392, 94.141, 20.673	94.173, 94.593,21.109
ξ	0.097, 0.092, 0.059	0.097, 0.091, 0.058	0.095, 0.089, 0.060
λ	38.147, 33.934, 13.255	38.411, 33.934, 13.538	37.752, 33.933, 13.523

Table 3.4: *Posterior MCMC estimates for n=626 fraud losses in 41 banks, recorded from 01/2007 to 04/2010*

3.2.3 Results

Two chains were run in R with 10,000, 25,000 and 100,000 iterations. The initial values for the first chain were chosen by using classical estimators: α and β were set as the maximum likelihood estimators, u = data 70th percentile, and σ and ξ the maximum likelihood estimates (MLE's). For the second chain, starting values were chosen far from the maximum likelihood estimates used for the first chain. Run time varied from half an hour to five hours in R.

The posterior mean, median and standard deviation of α, β, σ, u, ξ are shown in Table 3.4. We also estimate the rate of exceedances, i.e., the number of observations exceeding the threshold u per year. This quantity is denoted by λ and it will not be used in further analysis, although it is displayed in Table 3.4 as another model parameter.

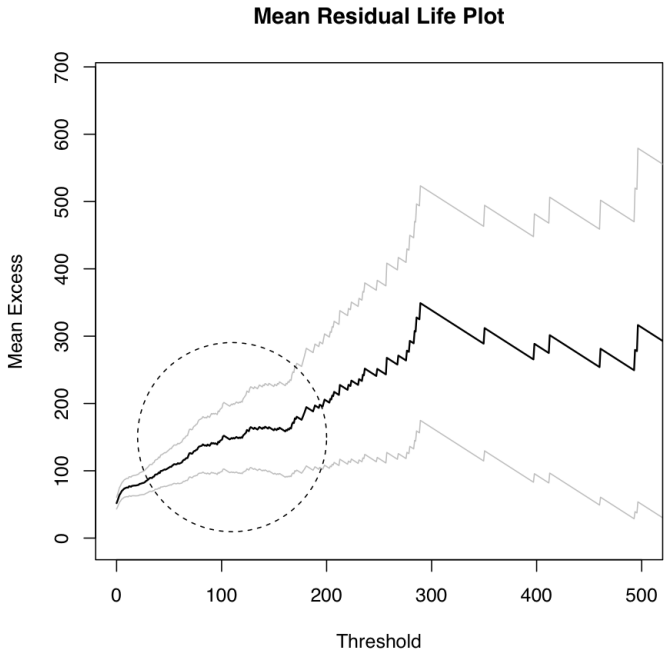

Figure 3.5: *Mean excess plot (fraud data) Circled area represents the posterior range of u*

Figure 3.5 shows the mean excess plot used in classical Extreme Value Theory for the threshold selection. From the plot, one can see that the posterior range (circled area) contains the value of u where the function becomes linear, about $u \approx 110$.

In order to compare both chains, trace plots are displayed in Figure 3.6. Notice that regardless of the starting values, the chains show convergence to the same value. A visual examination of the ergodic mean behaviour seems to indicate that the chain actually converges. Also, the histogram shows the distribution of the parameters for the first chain (Figure 3.7).

Although the estimates seem to be stable, in order to study convergence in a more formal way, we used the Gelman–Rubin statistic to compare the first and second chains (Table 3.5 and Figure 3.8). Also, the effective sample size and the Heidelberg and Welch diagnostics were used for the first chain (the one corresponding to the MLE's as initial values, Tables 3.6 and 3.7).

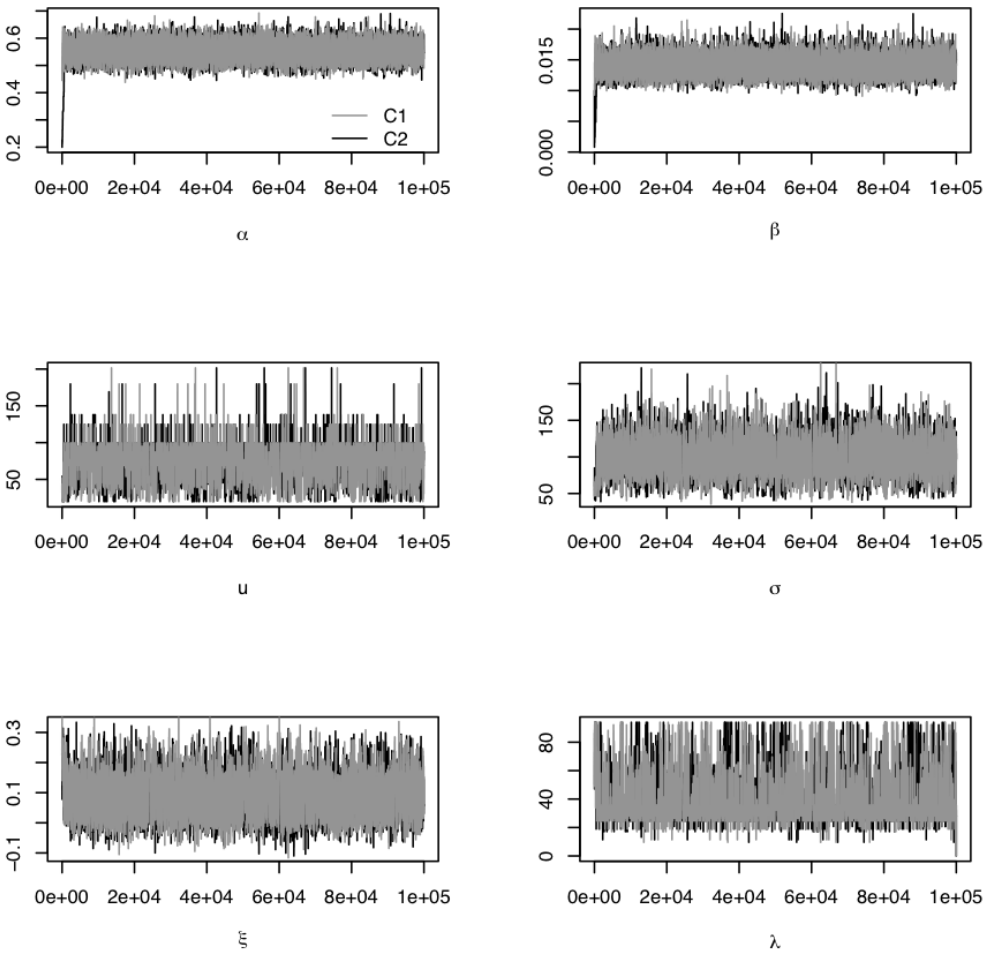

Figure 3.6: *Trace plots of the MCMC samples from the posterior density for n=626 fraud losses in 41 banks, recorded from 01/2007 to 04/2010. Two chains for 100,000 iterations were run in* R*: Gray colour is the first chain and black colour is the second chain*

	Parameter					
Runs	α	β	u	σ	ξ	λ
10,000	1.000	1.010	1.070	1.050	1.050	1.140
	(1.000)	(1.020)	(1.150)	(1.090)	(1.100)	(1.400)
25,000	1.000	1.000	1.000	1.000	1.000	1.000
	(1.000)	(1.000)	(1.010)	(1.000)	(1.000)	(1.000)
100,000	1.000	1.000	1.000	1.000	1.000	1.000
	(1.000)	(1.000)	(1.010)	(1.010)	(1.010)	(1.010)

Table 3.5: *Gelman-Rubin statistic for the MCMC posterior estimates for n=626 fraud losses in 41 banks, recorded from 01/2007 to 04/2010. (Scale reduction factor and its 97.5% quantile)*

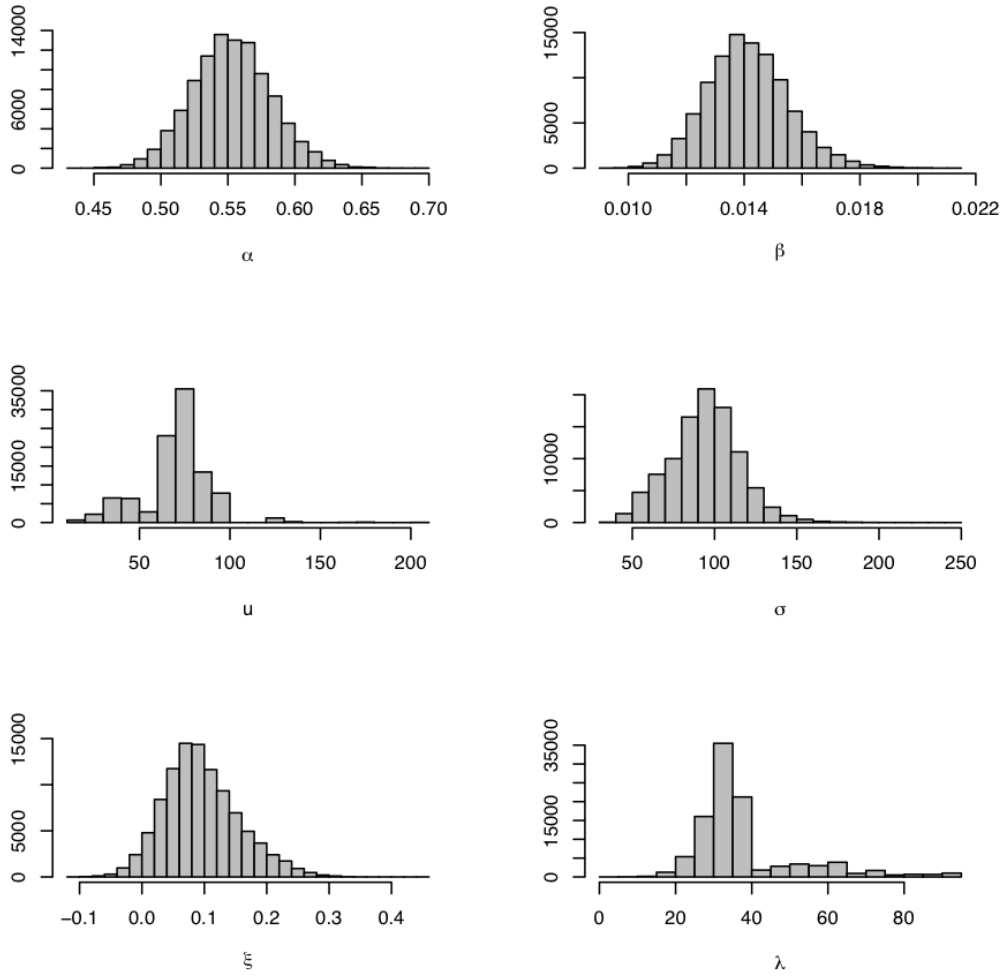

Figure 3.7: *Histograms of the MCMC samples from the posterior density for n=626 fraud losses in 41 banks, recorded from 01/2007 to 04/2010. 100,000 iterations*

	Parameter					
Runs	α	β	u	σ	ξ	λ
10,000	639.790	573.720	122.370	128.560	158.040	82.210
25,000	1535.950	1417.400	309.280	327.280	342.040	232.110
100,000	5196.060	4262.690	1239.950	1409.310	1450.460	926.520

Table 3.6: *Effective sample size for the MCMC posterior estimates for n=626 fraud losses in 41 banks, recorded from 01/2007 to 04/2010*

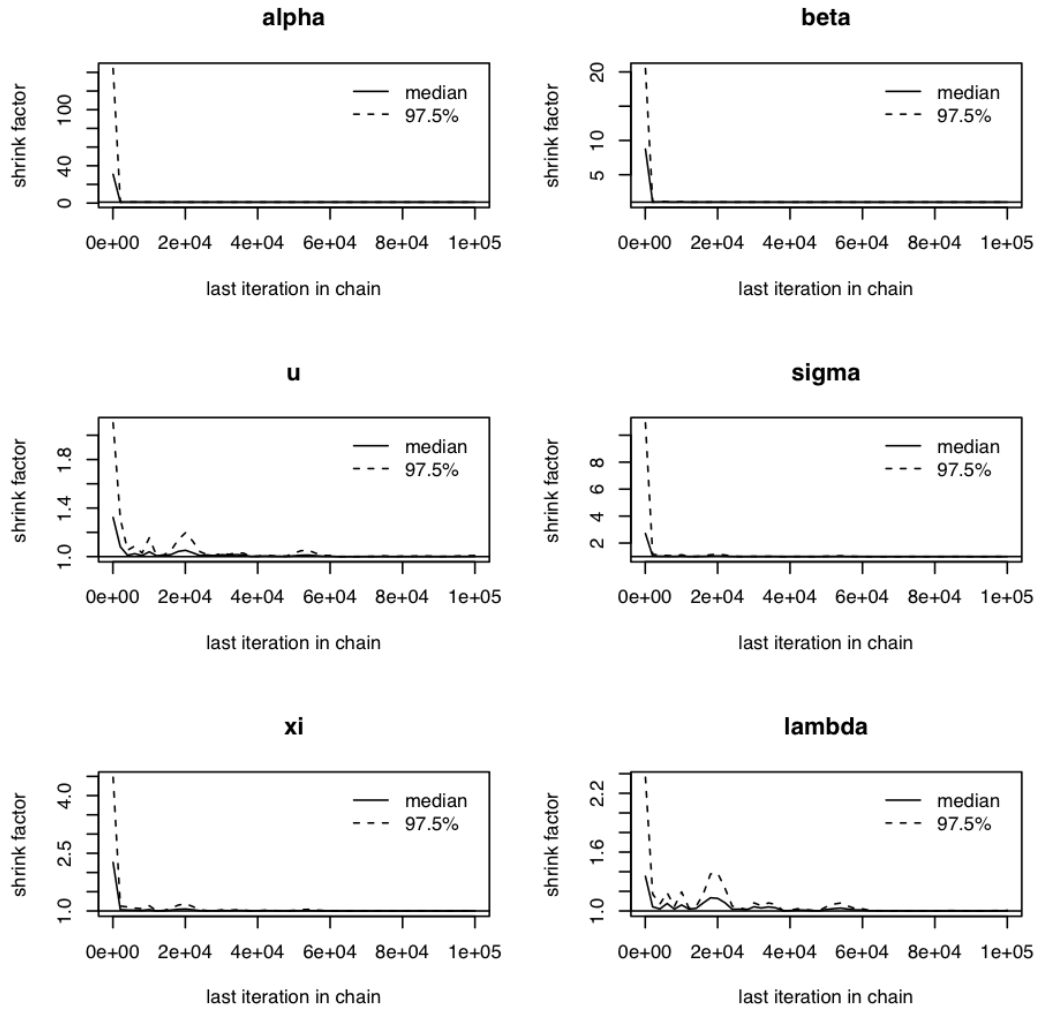

Figure 3.8: *Gelman plots for the MCMC posterior estimates for n=626 fraud losses in 41 banks, recorded from 01/2007 to 04/2010. 100,000 iterations*

Results reported in Table 3.5 clearly indicate that the behaviour of both chains is basically the same. The value of the Gelman–Rubin statistic is very close to 1 in all cases, which occurs when the pooled within-chain variance dominates the between-chain variance, meaning that at that point, all chains have escaped the influence of their starting points and have traversed all of the target distribution.

The Heidelberg and Welch diagnostic, reported in Table 3.7, also indicates that the first chain, which will be used in the analysis, achieves stationarity.

These results are essential to achieve an adequate estimation of the parameters.

Runs	Test	Result (p-value/ Halfwidth)					
		Parameter					
		α	β	u	σ	ξ	λ
10,000	Stat	passed	passed	passed	passed	passed	passed
		(0.321)	(0.151)	(0.911)	(0.838)	(0.883)	(0.953)
	Halfw	passed	passed	passed	passed	passed	passed
		(0.002)	(0.001)	(2.687)	(2.967)	(0.008)	(1.957)
25,000	Stat	passed	passed	passed	passed	passed	passed
		(0.475)	(0.611)	(0.383)	(0.276)	(0.225)	(0.217)
	Halfw	passed	passed	passed	passed	passed	passed
		(1.320e-03)	(7.590e-05)	(2.780)	(3.220)	(9.440e-03)	(2.130)
100,000	Stat	passed	passed	passed	passed	passed	passed
		(0.852)	(0.939)	(0.884)	(0.854)	(0.897)	(0.859)
	Halfw	passed	passed	passed	passed	passed	passed
		(8.200e-04)	(4.470e-05)	(1.240)	(1.320)	(3.680e-03)	(1.000)

Table 3.7: *Heidelberg and Welch diagnostic for the MCMC posterior estimates for n=626 fraud losses in 41 banks, recorded from 01/2007 to 04/2010*

3.2.4 Operational risk measurement

Once the estimates are sufficiently accurate, we can compute the minimum capital requirement for different banks. The operational risk VaR and ES at levels 0.95, 0.99 and 0.999 can be computed as well from the MCMC samples, using formulas (2.37) and (2.38). Since data were scaled, to determine the capital requirement for each bank, data were returned to their original scale by multiplying them by the asset size and dividing by a million.

The values obtained with the different methods were compared to the current capital requirement, which in most banks is calculated following the Basic Indicator Approach (BIA). Tables 3.8–3.10 display the results for all banks.

Notice that, in general, the estimate of the minimum capital requirement obtained from both the classical Peaks Over the Threshold and Bayesian inference is similar. However, in the Bayesian approach we have incorporated the threshold uncertainty and prior information, leading to smaller estimates. Also notice that the capital requirement using the BIA seems to overestimate (underestimate) the requirement. Figure 3.9 compares the capital requirement using the BIA and the Bayesian estimate.

Bank	Capital requirement					Equity	Proportion of equity capital				
	BIA	POT	Bayesian (runs)				BIA	POT	Bayesian(runs)		
			10000	25000	100000	capital			10000	25000	100000
A	107	169	154.44	154.46	154.72	2422.19	4.42%	6.98%	6.38%	6.38%	6.39%
B	112	67	60.91	60.92	61.02	2554.29	4.38%	2.62%	2.38%	2.38%	2.39%
C	2	3	2.69	2.69	2.69	412. 97	0.48%	0.73%	0.65%	0.65%	0.65%
D	5	7	6.19	6.19	6.20	420.63	1.19%	1.66%	1.47%	1.47%	1.47%
E	282	266	242.43	242.48	242.88	4520.69	6.24%	5.88%	5.36%	5.36%	5.37%
F	322	314	285.97	286.02	286.51	9388.67	3.43%	3.34%	3.05%	3.05%	3.05%
G	4805	4904	4471.30	4472.09	4479.60	153078.65	3.14%	3.20%	2.92%	2.92%	2.93%
H(*)	2	5	4.57	4.57	4.58	439.00	0.46%	1.14%	1.04%	1.04%	1.04%
I	6239	4470	4076.17	4076.90	4083.75	103496.22	6.03%	4.32%	3.94%	3.94%	3.95%
J	14	29	26.86	26.87	26.91	762.68	1.84%	3.80%	3.52%	3.52%	3.53%
K(*)	71	126	114.63	114.65	114.84	3497.53	2.03%	3.60%	3.28%	3.28%	3.28%
L(*)	1	3	2.77	2.77	2.78	690.29	0.14%	0.43%	0.40%	0.40%	0.40%
M	2272	2232	2035.11	2035.47	2038.89	40489.16	5.61%	5.51%	5.03%	5.03%	5.04%
N	147	157	143.22	143.25	143.49	2200.09	6.68%	7.14%	6.51%	6.51%	6.52%
O	36	71	65.02	65.03	65.14	1052.59	3.42%	6.75%	6.18%	6.18%	6.19%
P(*)	39	64	58.71	58.73	58.82	1523.23	2.56%	4.20%	3.85%	3.86%	3.86%
Q	53	39	35.84	35.84	35.90	4298.29	1.23%	0.91%	0.83%	0.83%	0.84%
R	9	28	25.82	25.82	25.86	689.20	1.31%	4.06%	3.75%	3.75%	3.75%
S(*)	22	52	47.75	47.76	47.84	1182.59	1.86%	4.40%	4.04%	4.04%	4.05%
T(*)	37	147	134.33	134.35	134.58	2101.42	1.76%	7.00%	6.39%	6.39%	6.40%
U	4	1	1.06	1.06	1.06	148.81	2.69%	0.67%	0.71%	0.71%	0.71%
V	38	43	39.00	39.00	39.07	1508.55	2.52%	2.85%	2.59%	2.59%	2.59%
W	2049	1644	1498.64	1498.91	1501.43	40099.08	5.11%	4.10%	3.74%	3.74%	3.74%
X	843	868	791.35	791.49	792.82	43254.84	1.95%	2.01%	1.83%	1.83%	1.83%
Y(*)	237	322	294.06	294.11	294.60	7371.61	3.22%	4.37%	3.99%	3.99%	4.00%
Z	175	300	273.30	273.35	273.81	3365.10	5.20%	8.92%	8.12%	8.12%	8.14%
AA	88	102	92.60	92.62	92.78	1993.36	4.41%	5.12%	4.65%	4.65%	4.65%
BB	196	256	233.37	233.41	233.80	4196.14	4.67%	6.10%	5.56%	5.56%	5.57%
CC(*)	111	130	118.62	118.64	118.84	4302.47	2.58%	3.02%	2.76%	2.76%	2.76%
DD	60	130	118.38	118.41	118.60	1058.72	5.67%	12.28%	11.18%	11.18%	11.20%
EE(*)	17	82	74.58	74.59	74.72	1680.43	1.01%	4.88%	4.44%	4.44%	4.45%
FF	27	59	53.99	54.00	54.09	1601.32	1.69%	3.68%	3.37%	3.37%	3.38%
GG(*)	8	15	13.58	13.58	13.60	513.85	1.56%	2.92%	2.64%	2.64%	2.65%
HH	3527	2496	2276.19	2276.59	2280.42	80529.98	4.38%	3.10%	2.83%	2.83%	2.83%
II	1086	678	618.04	618.15	619.19	25590.18	4.24%	2.65%	2.42%	2.42%	2.42%
JJ	23	44	39.93	39.94	40.01	798.56	2.88%	5.51%	5.00%	5.00%	5.01%
KK(*)	19	26	23.32	23.32	23.36	724.57	2.62%	3.59%	3.22%	3.22%	3.22%
LL(*)	1	8	7.16	7.17	7.18	385.56	0.26%	2.07%	1.86%	1.86%	1.86%
MM	43	53	48.27	48.28	48.36	925.07	4.65%	5.73%	5.22%	5.22%	5.23%
NN	7	7	6.65	6.65	6.67	495.64	1.41%	1.41%	1.34%	1.34%	1.34%
OO	2	7	6.69	6.69	6.70	992.76	0.20%	0.71%	0.67%	0.67%	0.67%

Table 3.8: *Minimum capital requirement (Millions of pesos) for fraud losses in 41 banks*

Bank	$\mathbf{VaR}_{0.95}$				$\mathbf{VaR}_{0.99}$				$\mathbf{VaR}_{0.999}$			
	POT	**Bayesian (Runs)**			**POT**	**Bayesian (Runs)**			**POT**	**Bayesian (Runs)**		
		10000	**25000**	**100000**		**10000**	**25000**	**100000**		**10000**	**25000**	**100000**
A	8.26	8.25	8.26	8.28	18.55	15.78	15.79	15.82	46.13	28.69	28.67	28.68
B	3.26	3.25	3.26	3.27	7.32	6.22	6.23	6.24	18.19	11.32	11.31	11.31
C	0.14	0.14	0.14	0.14	0.32	0.27	0.27	0.28	0.80	0.50	0.50	0.50
D	0.33	0.33	0.33	0.33	0.74	0.63	0.63	0.63	1.85	1.15	1.15	1.15
E	12.97	12.95	12.96	13.00	29.13	24.77	24.78	24.83	72.41	45.04	45.01	45.03
F	15.30	15.28	15.29	15.33	34.36	29.22	29.24	29.29	85.42	53.13	53.09	53.12
G	239.28	238.87	239.12	239.69	537.16	456.86	457.11	458.03	1335.51	830.67	830.13	830.49
H(*)	0.24	0.24	0.24	0.25	0.55	0.47	0.47	0.47	1.37	0.85	0.85	0.85
I	218.13	217.76	217.99	218.51	489.70	416.49	416.72	417.55	1217.50	757.27	756.77	757.10
J	1.44	1.44	1.44	1.44	3.23	2.74	2.75	2.75	8.02	4.99	4.99	4.99
K(*)	6.13	6.12	6.13	6.14	13.77	11.71	11.72	11.74	34.24	21.30	21.28	21.29
L(*)	0.15	0.15	0.15	0.15	0.33	0.28	0.28	0.28	0.83	0.51	0.51	0.51
M	108.91	108.72	108.83	109.10	244.49	207.94	208.05	208.47	607.86	378.08	377.83	378.00
N	7.66	7.65	7.66	7.68	17.21	14.63	14.64	14.67	42.78	26.61	26.59	26.60
O	3.48	3.47	3.48	3.49	7.81	6.64	6.65	6.66	19.42	12.08	12.07	12.08
P(*)	3.14	3.14	3.14	3.15	7.05	6.00	6.00	6.01	17.54	10.91	10.90	10.91
Q	1.92	1.91	1.92	1.92	4.31	3.66	3.66	3.67	10.70	6.66	6.65	6.66
R	1.38	1.38	1.38	1.38	3.10	2.64	2.64	2.64	7.71	4.80	4.79	4.80
S(*)	2.56	2.55	2.55	2.56	5.74	4.88	4.88	4.89	14.26	8.87	8.87	8.87
T(*)	7.19	7.18	7.18	7.20	16.14	13.72	13.73	13.76	40.12	24.96	24.94	24.95
U	0.06	0.06	0.06	0.06	0.13	0.11	0.11	0.11	0.32	0.20	0.20	0.20
V	2.09	2.08	2.09	2.09	4.68	3.98	3.99	3.99	11.65	7.24	7.24	7.24
W	80.20	80.06	80.14	80.34	180.04	153.13	153.21	153.52	447.62	278.42	278.23	278.35
X	42.35	42.28	42.32	42.42	95.07	80.86	80.90	81.06	236.37	147.02	146.92	146.98
Y(*)	15.74	15.71	15.73	15.76	35.33	30.05	30.06	30.12	87.83	54.63	54.59	54.62
Z	14.63	14.60	14.62	14.65	32.83	27.93	27.94	28.00	81.63	50.77	50.74	50.76
AA	4.96	4.95	4.95	4.96	11.13	9.46	9.47	9.49	27.66	17.20	17.19	17.20
BB	12.49	12.47	12.48	12.51	28.04	23.84	23.86	23.91	69.70	43.36	43.33	43.35
CC(*)	6.35	6.34	6.34	6.36	14.25	12.12	12.13	12.15	35.43	22.04	22.02	22.03
DD	6.34	6.32	6.33	6.35	14.22	12.10	12.10	12.13	35.36	21.99	21.98	21.99
EE(*)	3.99	3.98	3.99	4.00	8.96	7.62	7.62	7.64	22.28	13.86	13.85	13.85
FF	2.89	2.88	2.89	2.89	6.49	5.52	5.52	5.53	16.13	10.03	10.02	10.03
GG(*)	0.73	0.73	0.73	0.73	1.63	1.39	1.39	1.39	4.05	2.52	2.52	2.52
HH	121.81	121.60	121.73	122.02	273.45	232.57	232.70	233.17	679.87	422.87	422.59	422.77
II	33.07	33.02	33.05	33.13	74.25	63.15	63.18	63.31	184.60	114.82	114.74	114.79
JJ	2.14	2.13	2.14	2.14	4.80	4.08	4.08	4.09	11.93	7.42	7.41	7.42
KK(*)	1.25	1.25	1.25	1.25	2.80	2.38	2.38	2.39	6.96	4.33	4.33	4.33
LL(*)	0.38	0.38	0.38	0.38	0.86	0.73	0.73	0.73	2.14	1.33	1.33	1.33
MM	2.58	2.58	2.58	2.59	5.80	4.93	4.93	4.94	14.42	8.97	8.96	8.97
NN	0.36	0.36	0.36	0.36	0.80	0.68	0.68	0.68	1.99	1.24	1.24	1.24
OO	0.36	0.36	0.36	0.36	0.80	0.68	0.68	0.68	2.00	1.24	1.24	1.24

Table 3.9: *VaR at different levels (Millions of pesos) for fraud losses in 41 banks*

Bank	$ES_{0.95}$				$ES_{0.99}$				$ES_{0.999}$			
	POT	Bayesian (Runs)			POT	Bayesian (Runs)			POT	Bayesian (Runs)		
		10000	25000	100000		10000	25000	100000		10000	25000	100000
A	11.60	12.16	12.16	12.18	21.89	19.69	19.69	19.73	49.46	32.60	32.58	32.59
B	4.57	4.79	4.80	4.81	8.63	7.76	7.77	7.78	19.51	12.86	12.85	12.85
C	0.20	0.21	0.21	0.21	0.38	0.34	0.34	0.34	0.86	0.57	0.57	0.57
D	0.46	0.49	0.49	0.49	0.88	0.79	0.79	0.79	1.98	1.31	1.31	1.31
E	18.21	19.08	19.10	19.13	34.36	30.9	30.92	30.97	77.65	51.17	51.14	51.16
F	21.48	22.51	22.53	22.56	40.53	36.45	36.47	36.53	91.59	60.36	60.33	60.35
G	335.82	351.95	352.20	352.77	633.71	569.94	570.19	571.11	1432.06	943.75	943.21	943.57
H(*)	0.34	0.36	0.36	0.36	0.65	0.58	0.58	0.58	1.46	0.97	0.96	0.96
I	306.15	320.85	321.07	321.60	577.71	519.57	519.80	520.64	1305.51	860.36	859.86	860.19
J	2.02	2.11	2.12	2.12	3.81	3.42	3.43	3.43	8.60	5.67	5.67	5.67
K(*)	8.61	9.02	9.03	9.04	16.25	14.61	14.62	14.64	36.71	24.19	24.18	24.19
L(*)	0.21	0.22	0.22	0.22	0.39	0.35	0.35	0.35	0.89	0.58	0.58	0.58
M	152.85	160.19	160.30	160.57	288.43	259.41	259.52	259.94	651.80	429.55	429.30	429.46
N	10.76	11.27	11.28	11.30	20.30	18.26	18.26	18.29	45.87	30.23	30.21	30.22
O	4.88	5.12	5.12	5.13	9.21	8.29	8.29	8.30	20.82	13.72	13.72	13.72
P(*)	4.41	4.62	4.62	4.63	8.32	7.48	7.49	7.50	18.81	12.39	12.39	12.39
Q	2.69	2.82	2.82	2.83	5.08	4.57	4.57	4.58	11.48	7.56	7.56	7.56
R	1.94	2.03	2.03	2.04	3.66	3.29	3.29	3.30	8.27	5.45	5.45	5.45
S(*)	3.59	3.76	3.76	3.77	6.77	6.09	6.09	6.10	15.29	10.08	10.07	10.08
T(*)	10.09	10.57	10.58	10.6	19.04	17.12	17.13	17.16	43.02	28.35	28.34	28.35
U	0.08	0.08	0.08	0.08	0.15	0.14	0.14	0.14	0.34	0.22	0.22	0.22
V	2.93	3.07	3.07	3.08	5.53	4.97	4.97	4.98	12.49	8.23	8.23	8.23
W	112.56	117.96	118.05	118.24	212.40	191.03	191.11	191.42	479.98	316.32	316.13	316.26
X	59.44	62.29	62.33	62.44	112.16	100.87	100.92	101.08	253.45	167.03	166.93	167.00
Y(*)	22.09	23.15	23.16	23.2	41.68	37.48	37.5	37.56	94.18	62.07	62.03	62.05
Z	20.53	21.51	21.53	21.56	38.73	34.84	34.85	34.91	87.53	57.69	57.65	57.67
AA	6.96	7.29	7.29	7.31	13.12	11.80	11.81	11.83	29.66	19.55	19.53	19.54
BB	17.53	18.37	18.38	18.41	33.08	29.75	29.76	29.81	74.74	49.26	49.23	49.25
CC(*)	8.91	9.34	9.34	9.36	16.81	15.12	15.13	15.15	37.99	25.04	25.02	25.03
DD	8.89	9.32	9.32	9.34	16.78	15.09	15.10	15.12	37.92	24.99	24.97	24.98
EE(*)	5.60	5.87	5.87	5.88	10.57	9.51	9.51	9.53	23.89	15.74	15.73	15.74
FF	4.06	4.25	4.25	4.26	7.65	6.88	6.89	6.90	17.29	11.40	11.39	11.39
GG(*)	1.02	1.07	1.07	1.07	1.92	1.73	1.73	1.73	4.35	2.87	2.86	2.86
HH	170.96	179.17	179.29	179.59	322.60	290.14	290.26	290.73	729.01	480.43	480.16	480.34
II	46.42	48.65	48.68	48.76	87.59	78.78	78.81	78.94	197.95	130.45	130.37	130.42
JJ	3.00	3.14	3.15	3.15	5.66	5.09	5.09	5.10	12.79	8.43	8.42	8.43
KK(*)	1.75	1.84	1.84	1.84	3.30	2.97	2.97	2.98	7.47	4.92	4.92	4.92
LL(*)	0.54	0.56	0.56	0.57	1.02	0.91	0.91	0.92	2.29	1.51	1.51	1.51
MM	3.63	3.80	3.80	3.81	6.84	6.15	6.16	6.17	15.46	10.19	10.18	10.19
NN	0.50	0.52	0.52	0.52	0.94	0.85	0.85	0.85	2.13	1.40	1.40	1.40
OO	0.50	0.53	0.53	0.53	0.95	0.85	0.85	0.85	2.14	1.41	1.41	1.41

Table 3.10: *Expected Shortfall at different levels (Millions of pesos) for fraud losses in 41 banks*

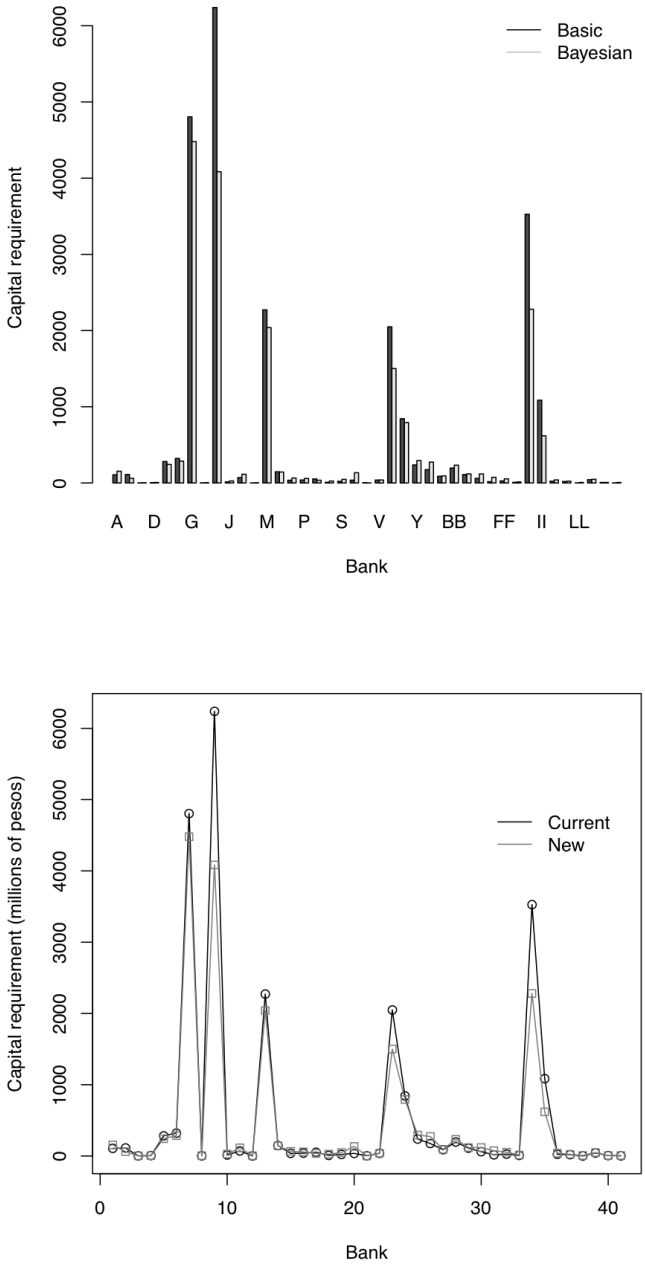

Figure 3.9: *Top: Bar plots of the Minimum capital requirement for fraud losses in each bank, using the Basic Indicator Approach (Dark gray) and the Bayesian approach (Gray). Bottom: Minimum capital requirement trend for fraud losses in each bank, using the Basic Indicator Approach (Dark gray) and the Bayesian approach (Gray)*

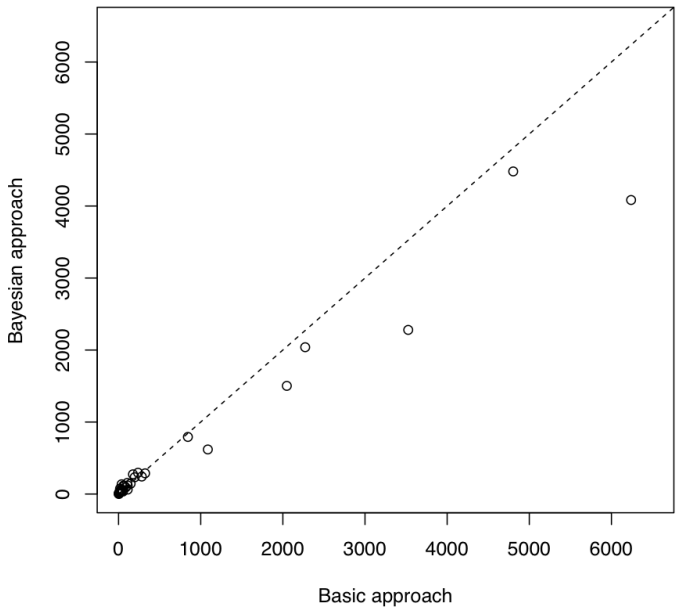

Figure 3.10: *Minimum capital requirement for fraud losses in each bank using the Basic Indicator Approach vs. Bayesian approach*

For the estimates of Value-at-Risk for operational risk, we obtained similar values with both methods for VaR 95%. However, for VaR 99% and 99.9%, the POT method provides much higher estimates.

Finally, in Figure 3.9, we confirm that in some cases the current capital estimate is much higher with respect to the Bayesian one. Figure 3.10 shows the current capital estimate (Bayesian Indicator Approach) against the Bayesian estimate.

3.2.5 Grouped data

The forty-one banks considered were classifed into four groups according to their size and origin: BAC, FIL, G-7 and MED. Different models were fitted to each group.

The first group includes those banks linked to department stores while the second one corresponds to subsidiary banks. G-7 groups the seven largest banks and MED consists of medium size banks.

Parameter	100,000 Runs	
	BAC	All
	Mean, Median, Std	Mean, Median, Std
α	0.682, 0.678, 0.097	0.552, 0.552, 0.029
β	0.012, 0.011, 0.002	0.014, 0.014,0.001
u	178.683, 182.513, 32.320	71.31, 72.980, 19.072
σ	322.281, 307.589, 127.466	94.173, 94.593, 21.109
ξ	-0.673, -0.645, 0.315	0.095, 0.089, 0.060

Table 3.11: *Posterior MCMC estimates for BAC data (n=75)*

If we observe the scaled losses in the different groups (Figure 3.11), we may notice that there are much fewer threshold exceedances in the first two groups than in the last two. Also, we should highlight the fact that we have only 75 observations for BAC and 50 for FIL, while the number of observations is larger for G-7 and MED (270 and 231, respectively). This makes the distribution of BAC and FIL not really heavy-tailed when all data are pooled.

We might expect that the fitted distributions differ from group to group, as it is shown in Tables 3.11–3.14, where the shape parameter of BAC clearly indicates that its distribution is not heavy-tailed as it is for the other groups. Figure 3.12 shows the fitted densities using the estimated parameters of the Gamma and GPD. In the first two cases, the use of the GPD does not appear appropriate, however for the last two groups the estimated densities seem to fit the data well. Something similar happens when we estimate the parameters using all data together, in this case the estimated density shows a good fit with the data.

Among other things, one might think that a bad fit is due to the fact that $\xi < 0$ imposes an upper bound on the density. This does not seem to be appropriate for the BAC data. For FIL data, the threshold seems a bit too low. Additionally, in both cases, we have a small number of observations.

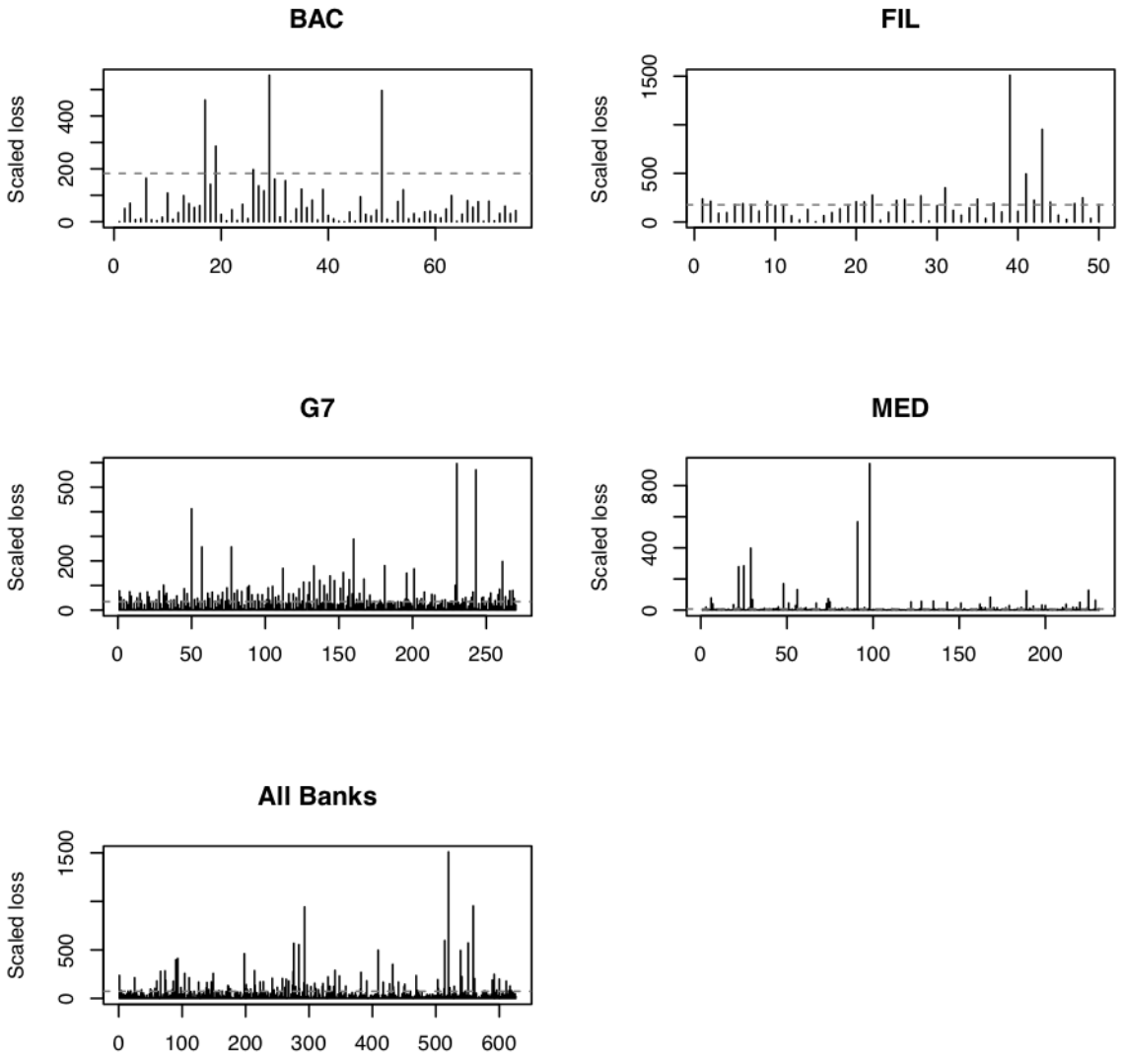

Figure 3.11: *Scaled losses for fraud data subgroups: BAC (75 observations), FIL (50 observations), G-7 (270 observations) and MED (231 observations). (Dashed gray line is the threshold value)*

Parameter	100,000 Runs	
	FIL	All
	Mean, Median, Std	Mean, Median, Std
α	0.971, 0.956, 0.203	0.552, 0.552, 0.029
β	0.005, 0.005, 0.001	0.014, 0.014, 0.001
u	177.035, 176.351, 11.984	71.310, 72.980, 19.072
σ	56.023, 53.075, 19.651	94.173, 94.593, 21.109
ξ	0.147, 0.145, 0.065	0.095, 0.089, 0.060

Table 3.12: *Posterior MCMC estimates for FIL data (n=50)*

Parameter	100,000 Runs	
	G7	All
	Mean, Median, Std	Mean, Median, Std
α	0.839, 0.834, 0.080	0.552, 0.552, 0.029
β	0.019, 0.019, 0.003	0.014, 0.014,0.001
u	38.024, 34.264, 11.710	71.31, 72.980, 19.072
σ	34.918, 34.482, 4.679	94.173, 94.593, 21.109
ξ	0.180, 0.179, 0.049	0.095, 0.089, 0.060

Table 3.13: *Posterior MCMC estimates for G7 data (n=270)*

Parameter	100,000 Runs	
	MED	All
	Mean, Median, Std	Mean, Median, Std
α	0.864, 0.860, 0.111	0.552, 0.552, 0.029
β	0.149, 0.147, 0.036	0.014, 0.014, 0.001
u	6.930, 6.742, 2.502	71.310, 72.980, 19.072
σ	18.102, 16.024, 8.695	94.173, 94.593, 21.109
ξ	0.351, 0.354, 0.093	0.095, 0.089, 0.060

Table 3.14: *Posterior MCMC estimates for MED data (n=231)*

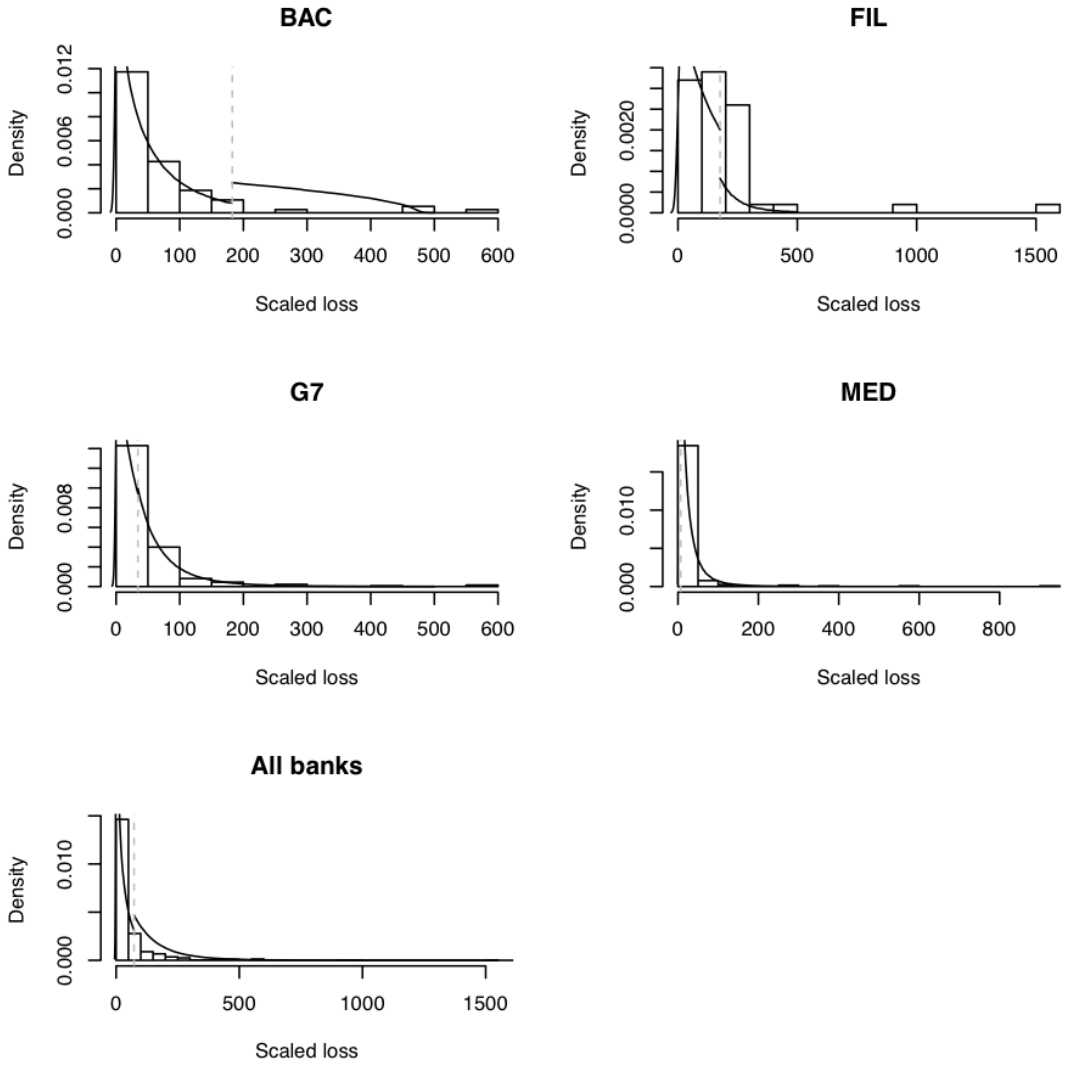

Figure 3.12: *Original data and fitted densities using the MCMC posterior estimates of the Gamma and GPD parameters (dashed gray line is the threshold value)*

3.2.5.1 Introducing a Reversible Jump Markov Chain Monte Carlo algorithm

Due to the poor fit of the GPD for BAC and FIL data, we consider two models :

1. A mixture of a Gamma distribution and the GPD (Original model, M).
2. A Gamma distribution (M').

To determine which of these models is more appropriate for the different data sets, we introduce a reversible jump step. The algorithm to do so is as follows:

1. Update the parameters θ and θ' conditional on the model M or M', respectively, using the Metropolis–Hastings algorithm.
2. Update the model conditional on the current parameter values using the following steps:
   a) Propose to move from model M to M'.
   b) Accept this proposed move with probability A.

where:

$$M = \text{Model 1},$$
$$M' = \text{Model 2},$$
$$\theta = \{\alpha, \beta, u, \sigma, \xi\},$$
$$\theta' = \{\alpha', \beta'\}.$$

Auxiliary variables:

$$u = \{u_1, u_2, u_3\}.$$

A function such that:

$$\alpha' = \alpha,$$

$$\beta' = \beta,$$

$$u_1 = u,$$

$$u_2 = \sigma,$$

$$u_3 = \xi.$$

This is the identity function and therefore $|J| = 1$.

In order to preserve the support of u, σ and ξ we can set:

$$u_i \sim \mathrm{TN}\left(a_i, b_i, \mu_i, \sigma_i\right), \quad i = 1, 2 \qquad \text{and} \qquad u_3 \sim \mathrm{N}\left(0, \sigma_3\right),$$

where TN denotes the truncated normal distribution with mean μ_i, standard deviation σ_i, and lower and upper truncation points a_i and b_i, respectively.

The acceptance probability is

$$\min\left\{1, A\right\}$$

with

$$A = \frac{\pi\left(M', \theta' \mid x\right) P\left(M \mid M'\right) p\left(u_1, u_2, u_3\right)}{\pi\left(M, \theta \mid x\right) P\left(M' \mid M\right)}, \tag{3.11}$$

where

- π denotes the posterior distribution over parameter and model space.
- P is the probability of proposing to move to model M given the current state of the chain is M', and viceversa.
- p is a proposal distribution.

See section A.5 of the Appendix for more details about RJMCMC methods.

For performing the RJ step:

$$\alpha = \alpha',$$

$$\beta = \beta',$$

$$u = u_1,$$

$$\sigma = u_2,$$

$$\xi = u_3.$$

with acceptance probability:

$$\min\left\{1, A^{-1}\right\}.$$

3.2.5.2 Results

Results for 20,000 iterations are shown in Tables 3.15–3.18. Figure 3.13 shows the behaviour of the chain when introducing the reversible jump step. We may notice that in most cases the chain explores Model 2 (M') and immediately goes to Model 1 (M) and stays there, except for the FIL data.

For G7 and MED data we expect Model 1 to be more suitable than Model 2. For FIL and BAC data, Model 2 would seem to be a more suitable choice. Nonetheless, for BAC data, the algorithm indicates that Model 1 is more appropriate. That might be a consequence of the chosen distribution for the auxiliary variables u_1, u_2 and u_3.

Parameter	20,000 Runs
	BAC
	Mean, Median, Std
α	0.582, 0.580, 0.057
β	0.009, 0.009, 8.364e-05
u	180.947, 182.513, 30.576
σ	309.573, 330.354, 79.067
ξ	-0.651, -0.648, 0.218

Table 3.15: *Posterior RJMCMC estimates for BAC data (n=75)*

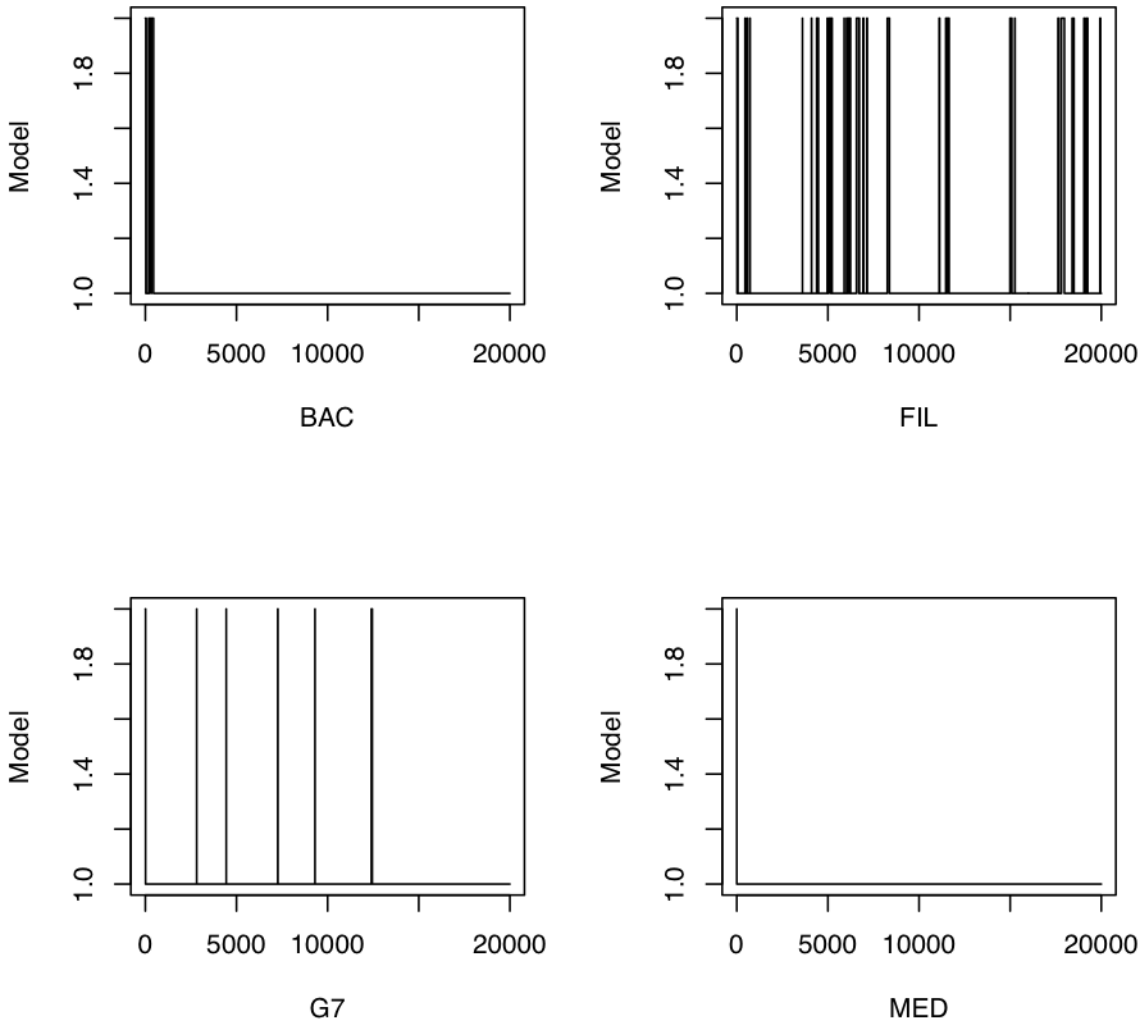

Figure 3.13: *Jumps between Models 1 and 2 for the different fraud data subgroups*

Parameter	20,000 Runs FIL Mean, Median, Std
α	0.897, 0.880, 0.171
β	0.054, 0.054, 0.001
u	194.388, 196.087, 5.543
σ	58.737, 58.442, 1.806
ξ	0.167, 0.169, 0.197

Table 3.16: *Posterior RJMCMC estimates for FIL data (n=50)*

Parameter	20,000 Runs G7 Mean, Median, Std
α	0.856, 0.854, 0.071
β	0.019, 0.019, 0.002
u	54.914, 54.408, 8.646
σ	29.465, 29.478, 1.254
ξ	0.193, 0.198, 0.007

Table 3.17: *Posterior RJMCMC estimates for G7 data (n=270)*

Parameter	20,000 Runs MED Mean, Median, Std
α	0.874, 0.890, 0.105
β	0.158, 0.146, 0.025
u	7.142, 6.946, 5.223
σ	21.997, 26.316, 8.555
ξ	0.360, 0.350, 0.099

Table 3.18: *Posterior RJMCMC estimates for MED data (n=231)*

Tables 3.19–3.22 display the estimates of VaR and ES at different levels, for each group, before and after introducing the reversible jump (RJ) step. We can observe that the RJ step introduces some variation in the estimates and leads to higher estimates of VaR at different levels for all the groups, except for G7.

	Before RJMCMC			After RJMCM		
Bank	$\mathbf{VaR}_{0.95}$	$\mathbf{VaR}_{0.99}$	$\mathbf{VaR}_{0.999}$	$\mathbf{VaR}_{0.95}$	$\mathbf{VaR}_{0.99}$	$\mathbf{VaR}_{0.999}$
E	16.97	33.44	40.42	17.35	34.95	42.34
J	1.88	3.71	4.48	1.92	3.87	4.69
U	0.07	0.15	0.18	0.08	0.15	0.19
V	2.73	5.38	6.50	2.79	5.62	6.81
OO	0.47	0.92	1.11	0.47	0.96	1.17

Table 3.19: *VaR at different levels before and after RJMCMC-BAC data*

	Before RJMCMC			After RJMCMC		
Bank	$\mathbf{VaR}_{0.95}$	$\mathbf{VaR}_{0.99}$	$\mathbf{VaR}_{0.999}$	$\mathbf{VaR}_{0.95}$	$\mathbf{VaR}_{0.99}$	$\mathbf{VaR}_{0.999}$
B	5.10	7.25	11.34	5.39	7.84	12.72
K(*)	9.61	13.65	21.34	10.15	14.75	23.93
L(*)	0.23	0.33	0.52	0.25	0.36	0.58
P(*)	4.92	6.99	10.93	5.20	7.55	12.26
S(*)	4.00	5.69	8.89	4.23	6.14	9.97
T(*)	11.26	15.99	25.01	11.89	17.28	28.04
Y(*)	24.64	35.01	54.75	26.03	37.83	61.39
CC(*)	9.94	14.12	22.09	10.50	15.26	24.76
JJ	3.35	4.75	7.44	3.54	5.14	8.34
KK(*)	1.95	2.78	4.34	2.06	3.00	4.87
LL(*)	0.60	0.85	1.33	0.63	0.92	1.50
NN	0.56	0.79	1.24	0.59	0.86	1.39

Table 3.20: *VaR at different levels before and after RJMCMC-FIL data*

	Before RJMCMC			After RJMCMC		
Bank	$\mathbf{VaR}_{0.95}$	$\mathbf{VaR}_{0.99}$	$\mathbf{VaR}_{0.999}$	$\mathbf{VaR}_{0.95}$	$\mathbf{VaR}_{0.99}$	$\mathbf{VaR}_{0.999}$
G	149.87	262.95	493.66	131.43	222.90	416.50
I	136.63	239.71	450.03	119.82	203.20	379.70
M	68.22	119.68	224.69	59.82	101.45	189.57
W	50.23	88.13	165.46	44.05	74.71	139.60
HH	76.30	133.86	251.31	66.91	113.47	212.03
X	26.53	46.54	87.37	23.26	39.45	73.71
II	20.72	36.35	68.24	18.17	30.81	57.57

Table 3.21: *VaR at different levels before and after RJMCMC-G7 data*

	Before RJMCMC			After RJMCMC		
Bank	$\mathbf{VaR}_{0.95}$	$\mathbf{VaR}_{0.99}$	$\mathbf{VaR}_{0.999}$	$\mathbf{VaR}_{0.95}$	$\mathbf{VaR}_{0.99}$	$\mathbf{VaR}_{0.999}$
A	2.09	4.90	13.07	3.11	7.58	20.45
C	0.04	0.09	0.23	0.05	0.13	0.36
D	0.08	0.20	0.52	0.12	0.30	0.82
F	3.86	9.08	24.20	5.76	14.04	37.87
H(*)	0.06	0.15	0.39	0.09	0.22	0.61
N	1.94	4.55	12.12	2.89	7.03	18.97
O	0.88	2.06	5.50	1.31	3.19	8.61
Q	0.48	1.14	3.03	0.72	1.76	4.75
R	0.35	0.82	2.18	0.52	1.27	3.42
Z	3.69	8.68	23.13	5.51	13.42	36.19
AA	1.25	2.94	7.84	1.87	4.55	12.26
BB	3.15	7.41	19.75	4.70	11.46	30.90
DD	1.60	3.76	10.02	2.39	5.81	15.68
EE(*)	1.01	2.37	6.31	1.50	3.66	9.88
FF	0.73	1.71	4.57	1.09	2.65	7.15
GG(*)	0.18	0.43	1.15	0.27	0.67	1.80
MM	0.65	1.53	4.09	0.97	2.37	6.39

Table 3.22: *VaR at different levels before and after RJMCMC-MED data*

3.2.5.3 An alternative for FIL and BAC data

This section is based on the paper by Venturini et al. (2008), where a Bayesian approach for the estimation of tail probabilities of heavy-tailed distributions is proposed, based on a mixture of Gamma distributions in which the mixing occurs over the shape parameter. The procedure is as follows.

Let Y be a positive random variable. The Gamma Shape Mixture (GSM) model is defined as:

$$f\left(y \mid w_1, ..., w_J, \theta\right) = \sum_{j=1}^{J} w_j \, f_j\left(y \mid \theta\right), \tag{3.12}$$

where $f_j\left(y \mid \theta\right) = \frac{\theta^j}{\Gamma(j)} y^{j-1} e^{-\theta y}$, is the density function of a Gamma(j, θ) random variable, with parameters $j > 0$ (shape) and $\theta > 0$ (scale). We assume that the number of components J is known and fixed, whereas $w = (w_1, ..., w_J)$ is an unknown vector of mixture weights. The GSM model has two useful properties:

1. $1/\theta$ is a scale parameter for the whole model, since $f\left(y \mid w_1, ..., w_J, \theta\right) = \theta \cdot f(\theta \cdot$

$y|w_1, ..., w_J, 1)$.

2. Its moments are convex combinations of the moments of Gamma(j, θ) variables, so that the mth moment is given by

$$E\left[Y^m \mid w_1, ..., w_J, \theta\right] = \sum_{j=1}^{J} w_j E\left[Y_j^m \mid \theta\right] = \sum_{j=1}^{J} w_j \frac{\prod_{l=1}^{m}(j+l-1)}{\theta^m}.$$

We assume that θ and w are independent a priori and we specify the following conjugate prior distributions:

$$\theta \sim \text{Gamma}\left(\alpha, \beta\right),$$
$$w = (w_1, ..., w_J) \sim D_J\left(\frac{1}{J}, ..., \frac{1}{J}\right),$$

where $D_J\left(a_1, ..., a_J\right)$ denotes the Dirichlet distribution with parameters $a_1, ..., a_J > 0$.

In practice, setting the prior hyperparameters equal to $1/J$ tends to produce posterior distributions where only a small subset of the J mixture weights will have high prior probability to be selected at each iteration of the MCMC.

Given a sample $y = (y_1, ..., y_n)$ of i.i.d. observations from (3.12), the likelihood is

$$L\left(w, \theta \mid y\right) = \prod_{i=1}^{n} \sum_{j=1}^{J} w_j \, f_j\left(y_i \mid \theta\right). \tag{3.13}$$

This expression is however intractable because it includes J^n different terms.

Given $y = (y_1, ..., y_n)$ from (3.12), we can associate to each y_i an integer x_i between 1 and J that identifies the component of the mixture generating observation y_i . Thus, the variable x_i takes value j with prior probability w_j, $1 \leq j \leq J$. The vector $x = (x_1, ..., x_n)$ of component labels is the missing data part of the sample since it is not observed.

Suppose the missing data $x_1, ..., x_n$ were available. Then the model could be written as

$$p\left(y_1, ..., y_n \mid x_1, ..., x_n, \theta\right) = \frac{\theta^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} \Gamma\left(x_i\right)} \left(\prod_{i=1}^{n} y_i^{x_i - 1}\right) e^{-\theta \sum_{i=1}^{n} y_i}. \tag{3.14}$$

Thus, using (3.14) and the priors, the posterior distribution is

$$p\left(w_1, ..., w_J, \theta \mid y_1, ..., y_n, x_1, ..., x_n\right) \propto \left(\prod_{j=1}^{J} w_j^{(1/J)+n_j-1}\right) \theta^{\alpha + \sum_{i=1}^{n} x_i} e^{-\left(\beta + \sum_{i=1}^{n} y_i\right)\theta}, \tag{3.15}$$

where $n_j = \sum_{i=1}^{n} \mathbb{I}(x_i = j)$, $j = 1, ..., J$. The main consequence of this conditional decomposition is that, for a given missing data vector $(x_1, ..., x_n)$, the conjugacy is preserved and, therefore, the simulation can be easily performed, conditional on the missing data $x_1, ..., x_n$.

To compute the posterior distribution, after having integrated out θ, the full conditional distribution of the mixture weights is given by

$$p(w_1, ..., w_J \mid y_1, ..., y_n, x_1, ..., x_n) \propto \prod_{j=1}^{J} w_j^{(1/J)+n_j-1}, \tag{3.16}$$

that is, the Dirichlet distribution $D_J\left(\frac{1}{J} + n_1, ..., \frac{1}{J} + n_J\right)$.

The full conditional probability of the ith missing label is then given by

$$p\left(x_i \mid y, x_{(-i)}, w\right) = \frac{\sum_{j=1}^{J} w_j \frac{y_i^{j-1}\left(\alpha + \sum_{(-i)} x_r\right)_j}{\Gamma(j)\left(\beta + \sum_{r=1}^{n} y_r\right)^j} \mathbb{I}(x_i = j)}{\sum_{k=1}^{J} w_k \frac{y_i^{k-1}\left(\alpha + \sum_{(-i)} x_r\right)_k}{\Gamma(k)\left(\beta + \sum_{r=1}^{n} y_r\right)^k}}, \tag{3.17}$$

where $x_{(-i)}$ is the $x = (x_1, ..., x_n)$ vector with the ith element deleted, $\sum_{(-i)} x_r$ denotes the sum of all the component labels except for the ith one, and $(n)_k = n(n+1)\cdots(n+k-1)$ is the Pochhammer symbol. We assume that α is an integer, for computation speed, and to avoid overflow errors. The integration of θ implies that the missing data are no longer independent.

For a given value of J, a strategy for choosing α and β is as follows:

1. Compute $\widetilde{\theta} = J/\max(y_1, ..., y_n)$ and check that $1/\widetilde{\theta} \leq \min(y_1, ..., y_n)$; the idea is that, on average, θ should take values that allow the set of Gamma distributions in (3.12) to completely span the range of observed values (the last Gamma distribution should have a mean not smaller than the maximum observation and the first Gamma distribution a mean not greater than the minimum observation). Hence $\widetilde{\theta}$ is a candidate for the prior mean α/β.

2. Choose a value ω for the weight of the prior information. Values between 0.2 and 0.5 are usually reasonable choices. Set β to

$$\omega \cdot \sum_{i=1}^{n} y_i / 1 - \omega. \tag{3.18}$$

3. Set α to be the closest integer to the quantity $\widetilde{\theta} \cdot \beta$.

Regarding the choice of J, a small value of J can create a severe limitation to the model, as the set of densities available in the class being mixed may not be sufficiently rich with elements that have a large mean. On the other hand, too large a value does not cause serious difficulties as the fit is often robust when there are several Gamma distributions in the class that can serve as building blocks for a particular mixture component. However, too large a J may cause numerical problems.

θ	Mean, median
BAC	0.685, 0.642
FIL	0.462, 0.462

Table 3.23: *Estimates of θ from the GSM procedure*

	RJMCMC			GSM		
Bank	$\mathbf{VaR}_{0.95}$	$\mathbf{VaR}_{0.99}$	$\mathbf{VaR}_{0.999}$	$\mathbf{VaR}_{0.95}$	$\mathbf{VaR}_{0.99}$	$\mathbf{VaR}_{0.999}$
E	16.97	33.44	40.42	20.41	34.09	37.28
J	1.88	3.71	4.48	2.26	3.78	4.13
U	0.07	0.15	0.18	0.09	0.15	0.16
V	2.73	5.38	6.50	3.28	5.48	6.00
OO	0.47	0.92	1.11	0.56	0.94	1.03

Table 3.24: *VaR at different levels using the RJMCMC and GSM procedure-BAC data*

For the BAC and FIL data, the above model was fitted using a Gibbs sampler. J was chosen by trial and error, the best value was about 380 and 700, respectively. This value was used to compute $\widetilde{\theta}$, α and β, following the strategy above. Results are shown in Table 3.23. VaR estimates at different levels for the RJ step and the GSM procedure are displayed in tables 3.24–3.25, while fitted densities are shown in Figures 3.14 and 3.15.

	RJMCMC			GSM		
Bank	$\mathbf{VaR}_{0.95}$	$\mathbf{VaR}_{0.99}$	$\mathbf{VaR}_{0.999}$	$\mathbf{VaR}_{0.95}$	$\mathbf{VaR}_{0.99}$	$\mathbf{VaR}_{0.999}$
B	5.10	7.25	11.34	9.73	17.73	21.05
K(*)	9.61	13.65	21.34	18.31	33.37	39.61
L(*)	0.23	0.33	0.52	0.44	0.81	0.96
P(*)	4.92	6.99	10.93	9.38	17.09	20.29
S(*)	4.00	5.69	8.89	7.63	13.90	16.50
T(*)	11.26	15.99	25.01	21.46	39.11	46.42
Y(*)	24.64	35.01	54.75	46.98	85.61	101.61
CC(*)	9.94	14.12	22.09	18.95	34.53	40.99
JJ	3.35	4.75	7.44	6.38	11.63	13.80
KK(*)	1.95	2.78	4.34	3.72	6.79	8.06
LL(*)	0.60	0.85	1.33	1.14	2.09	2.48
NN	0.56	0.79	1.24	1.06	1.94	2.30

Table 3.25: *VaR at different levels using the RJMCMC and GSM procedure-FIL data*

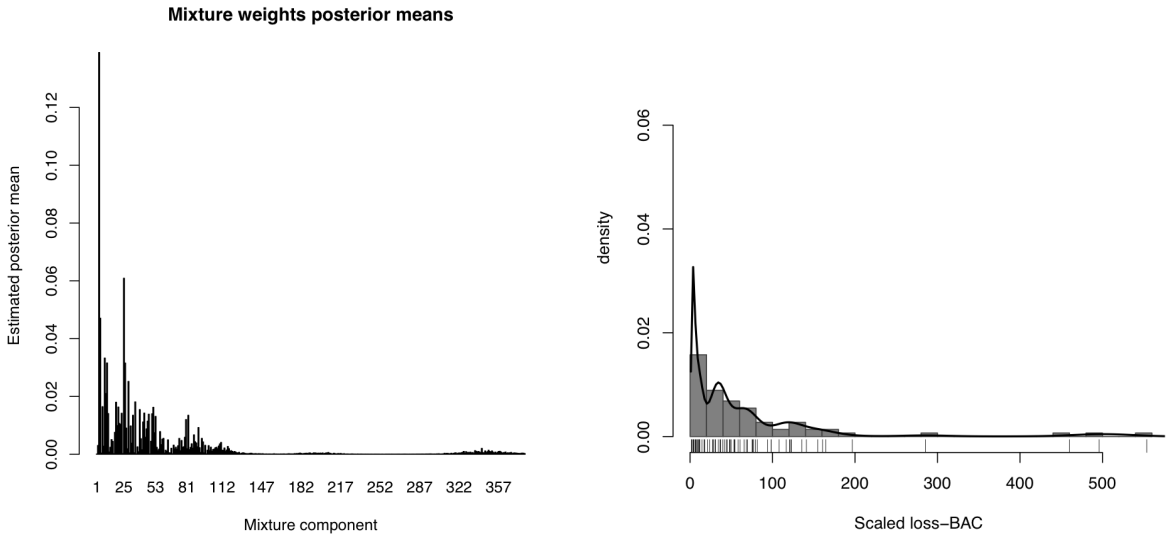

Figure 3.14: *GSM for BAC data*

The GSM procedure yields higher estimates of VaR for both groups and the density is well approximated by the mixture. It appears that, in this case, the GSM model represents a good alternative for fitting the data.

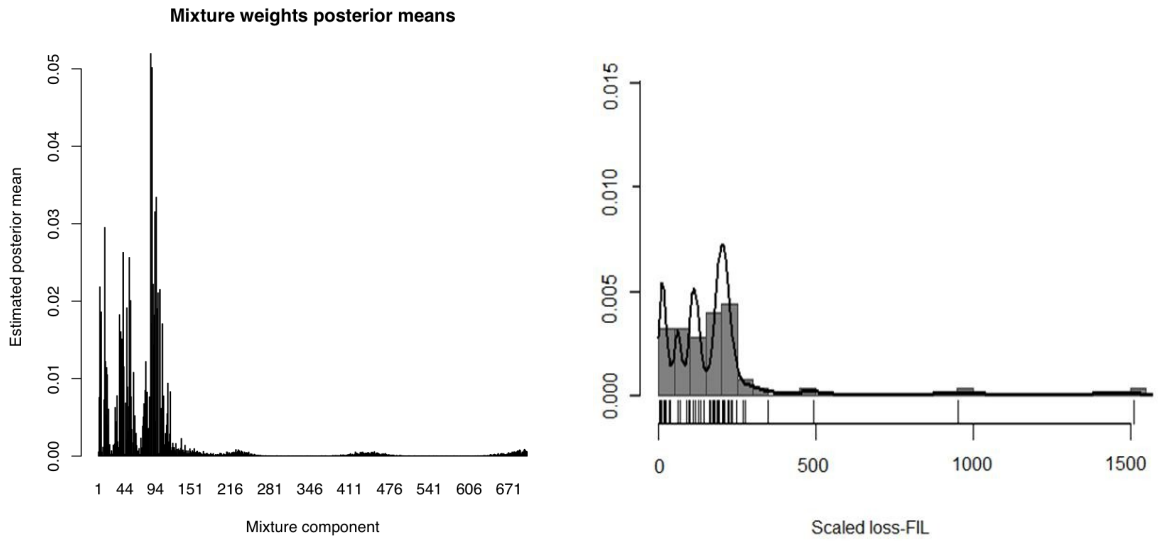

Figure 3.15: *GSM for FIL data*

3.3 Conclusions of the Chapter

In this chapter we have introduced a mixture model developed by Behrens et al. (2004) that has significantly influenced the development of new Bayesian models. This model uses a Gamma distribution for observations below the threshold, and a GPD for those above it. One of its main features is that the threshold is considered as another parameter in the model.

By using real and simulated data, we have identified the potential of this model for the estimation of operational risk measures, particularly the Value-at-Risk and the Expected Shortfall. At the same time, we have confirmed the poor performance of the Basic Indicator approach.

Furthermore, we have explored model selection using RJMCMC methods. As we have seen, these methods allow us to select the most suitable model while performing parameter estimation. It is worth pointing out that the selection of the "best" model does not mean that it is generally the most appropriate to model our data. Thus, in some cases, other alternatives should be considered, as it was the case for FIL and BAC data, where the GSM model presented a better overall performance than previous models.

This leads us to consider that Bayesian models can be used in many different ways in the study of extremes, with the advantage of considering the uncertainty implicit in the parameters and the prior information available.

Chapter 4

Prior elicitation and analysis

An important question that arises in our Bayesian analysis is: How sensitive are the posterior results to variations in the prior?

The issue of prior selection is one of the most controversial aspects of Bayesian theory. An interesting discussion about these issues can be found in Irony and Singpurwalla (1997).

It is well-known that the prior can have an impact on posterior results when observations are scarce, as it is the case for operational risk data. Different priors may lead to different posterior estimates and hence it is essential to find an appropriate prior by considering the context of the problem.

In the Bayesian analysis of extremes, authors such as Cabras et al. (2010) or Castellanos and Cabras (2007) have used non-informative priors for the GPD parameters, specifically the Jeffreys prior (see Appendix B), because of their simplicity. Jeffreys priors are useful in situations in which we would like the observations to "speak for themselves"; however, this assumption is not always realistic, especially when a considerable part of the analysis is based on expert opinion.

In this chapter we consider two important risk measures which are based on the shape and location parameters of the GPD: The Value-at-Risk (VaR) and the Expected Shortfall (ES). Both measures were defined in Chapter 2. Recall that if the excess distribution of

a loss variable X above a high threshold u is approximated by a GPD with parameters σ and ξ, and if $1-q < P(X > u)$,

$$\begin{aligned} \text{VaR}_q &= u + \frac{\sigma}{\xi}\left(\left(\frac{1-q}{\bar{F}(u)}\right)^{-\xi} - 1\right), \quad (4.1) \\ \text{ES}_q &= \frac{\text{VaR}_q}{1-\xi} + \frac{\sigma - \xi u}{1-\xi}. \quad (4.2) \end{aligned}$$

These measures are commonly used in finance and risk management. Therefore, experts are familiar with them.

In Section 4.1, we study the prior and posterior distributions of these two risk measures when using the Jeffreys prior. In Section 4.2, we study the prior elicitation from expert opinion in terms of VaR.

4.1 The non-informative prior

In this section, we study the performance of the Jeffreys prior for the GPD parameters. A complete derivation of this prior is provided in Appendix B and leads to

$$\pi(\sigma, \xi) \propto \sigma^{-1}(1+\xi)^{-1}(1+2\xi)^{-1/2}, \quad -0.5 < \xi, \quad \sigma > 0. \quad (4.3)$$

For $\xi = 0$, $\pi(\sigma, \xi)$ corresponds to the Jeffreys's prior for the scale parameter of the exponential distribution.

Equation (4.3) implies that σ and ξ are a priori independent. Notice that only the marginal prior for σ is improper.

We obtained samples from the prior distribution of $\text{VaR}_{0.99}$ and $\text{ES}_{0.99}$ by considering different scenarios. In all cases we set a truncated normal for the threshold, using different data sets for the lower bound e_1 in (3.5): The fraud data described in Section 3.2.1, and samples from Exponential(0.1), Log-normal(0,1) and Gamma(2,0.5) distributions. The improper prior of σ was approximated by a Log-normal$(0, 1.25)$. The prior for ξ was derived from Equation (4.3) by introducing an appropriate normalizing constant.

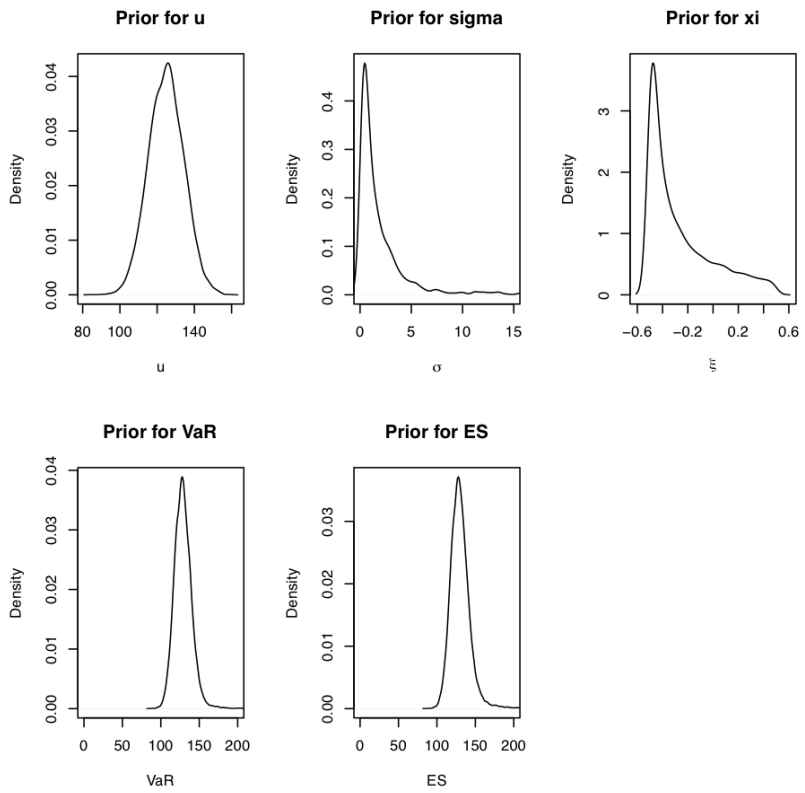

Figure 4.1: *Prior distributions:* $\sigma \sim LN(0, 1.25)$ *;* $u \sim TN(min\,(fraud\ data))$*;* $\xi \sim$ *Jeffreys prior*

To study the effect of these parameters on the prior distribution of VaR and ES, we fixed one of them and kept the priors specified above unchanged for the rest.

We fixed values of 1, 10 and 100 for σ; quantiles 0.5, 0.7 and 0.9 for the threshold u; and -0.3, 0.05 and 0.45 for ξ. Then, we combined all these scenarios and compared the prior distributions obtained. Results for $\text{VaR}_{0.99}$ and $\text{ES}_{0.99}$ can be observed in Figure 4.1 and Figures D.1–D.12 shown in Appendix D.

From the graphs, one can note the same pattern in all data sets when we fix a specific parameter. The choice of the prior for the scale σ and shape ξ parameters strongly influences the distribution of VaR and ES.

The Jeffreys prior for ξ (Figure 4.1) yields values of ξ very close to -0.5 and most of them are concentrated around this number, which is completely unrealistic in practice. On the other hand, although the prior for σ is reasonable, we should notice that for the

prior in equation (4.3), σ and ξ are independent. As it was pointed out in Section 3.1.1, a priori negative dependence between these parameters is expected (Coles and Tawn, 1996). This shows us that, at least in the context of extreme data, it is not appropriate to adopt a non-informative prior.

In the next section, we propose a subjective prior, based on the VaR measure.

4.2 Elicitation

As it has been mentioned in the previous chapters, one of the requirements of the Advanced Measurement Approach (AMA) is the inclusion of expert opinion. Additionally, the scarcity of data has prompted us to turn to experts through the elicitation of information, which is not an easy task.

In this section, the objective is to facilitate prior elicitation and use a prior distribution that reflects expert opinion faithfully, providing realistic values for the parameters involved.

For that purpose, we have chosen the Value-at-Risk. Based on the prior for σ and ξ proposed in the previous chapter, we introduce the following prior.

Consider two high quantiles $q_1 < q_2$ (usually $q_1 = 0.99$ and $q_2 = 0.999$) and the Value-at-Risk at those levels. Recall that if the GPD approximation is used above a high threshold u and $1 - q_1 < P(X > u)$,

$$\mathrm{VaR}_{q_1} = u + \frac{\sigma}{\xi}\left(k_1^{-\xi} - 1\right), \quad \mathrm{VaR}_{q_2} = u + \frac{\sigma}{\xi}\left(k_2^{-\xi} - 1\right), \tag{4.4}$$

where $k_i = (1 - q_i)/P(X > u)$, $i = 1, 2$.

Since we need to guarantee that $u < \mathrm{VaR}_{q_1} < \mathrm{VaR}_{q_2}$, we can follow a similar approach to the one discussed in Section 3.1.1. More specifically, we work with the differences

$$d_1 = (\mathrm{VaR}_{q_1} - u) = \frac{\sigma}{\xi}\left(k_1^{-\xi} - 1\right), \quad d_2 = (\mathrm{VaR}_{q_2} - u) - (\mathrm{VaR}_{q_1} - u) = \frac{\sigma}{\xi}\left(k_2^{-\xi} - k_1^{-\xi}\right)$$

and assume that $\pi(d_i) \sim \mathrm{Gamma}(a_i, b_i)$, $i = 1, 2$ and that d_1 is independent of d_2. In comparison to the prior for ξ and σ from Section 3.1.1, this approach guarantees that

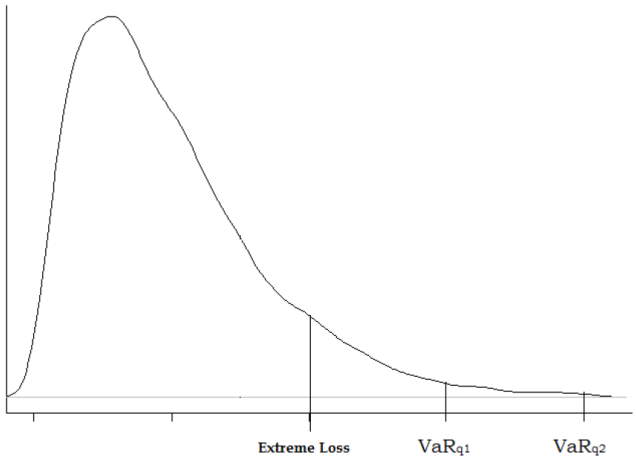

Figure 4.2: *Loss distribution used in the elicitation process*

indeed $\text{VaR}_{q_1} > u$. The derivation of $\pi(\sigma, \xi)$ is analogous to the transformation method from Section 3.1.1. We get:

$$\begin{aligned}\pi(\sigma,\xi) \propto\ & \left[\tfrac{\sigma}{\xi}\left(k_1^{-\xi}-1\right)\right]^{a_1-1} \exp\left[-b_1\left\{\tfrac{\sigma}{\xi}\left(k_1^{-\xi}-1\right)\right\}\right] \\ & \times \left[\tfrac{\sigma}{\xi}\left(k_2^{-\xi}-k_1^{-\xi}\right)\right]^{a_2-1} \exp\left[-b_2\left\{\tfrac{\sigma}{\xi}\left(k_2^{-\xi}-k_1^{-\xi}\right)\right\}\right] \\ & \times \left|-\tfrac{\sigma}{\xi^2}\left[(k_1k_2)^{-\xi}(\log k_2-\log k_1)-k_2^{-\xi}\log k_2+k_1^{-\xi}\log k_1\right]\right|.\end{aligned} \tag{4.5}$$

The hyperparameters a_i and b_i are determined by measures of location and variability in prior belief. Experts are asked for estimates of the median and 90% quantiles of each of the d_i's. Parameter estimates may be obtained by solving numerically for a_i and b_i.

We illustrate the elicitation process with a simple example. Suppose we have an expert who can provide us with information about quantiles. We could start by showing him/her the plot in Figure 4.2. This plot displays the loss distribution and three different points:

1. The point above which a loss is considered extreme (EL).

2. The VaR at level q_1=0.99.

3. The VaR at level q_2=0.999.

Elicited Value		a_i	b_i
$q_{0.5,1} = 10,$	$q_{0.9,1} = 100$	0.339	0.001
$q_{0.5,2} = 5,$	$q_{0.9,2} = 45$	0.362	0.023

Table 4.1: *Gamma parameters obtained from elicited quantiles (fictitious expert)*

Next, we ask him/her to think of how different these quantities are from one another. That is, how large are the intervals $[\text{EL}, \text{VaR}_{q_1}]$ and $[\text{VaR}_{q_1}, \text{VaR}_{q_2}]$.

Then, we ask him/her for the median and 90% quantiles of these differences. We denote these values by: $q_{0.5,1}$(median of the first difference), $q_{0.5,2}$ (median of the second difference), $q_{0.9,1}$ (0.9 quantile of the first difference) and $q_{0.9,2}$ (0.9 quantile of the second difference). Once we know these quantiles, we can obtain the values of a_i and b_i by solving the following equations:

$$F_x\left(q_{0.5,1}, a_1, b_1\right) = 0.5 \quad \text{and} \quad F_x\left(q_{0.9,1}, a_1, b_1\right) = 0.9, \tag{4.6}$$

$$F_x\left(q_{0.5,2}, a_2, b_2\right) = 0.5 \quad \text{and} \quad F_x\left(q_{0.9,2}, a_2, b_2\right) = 0.9, \tag{4.7}$$

where F_x is a Gamma(a_i, b_i). This can be solved numerically. For instance, we may use the function `get.gamma.par` from the `rriskDistributions` R package. Table 4.1 shows the results for some specific quantiles.

Now, keeping the priors previously proposed for the rest of the parameters, we can derive the posterior distribution:

$$\begin{aligned}
\log p(\theta \mid x) = \; & K + \sum_{i=1}^{n} I\left(x_i < u\right)\left[\alpha \log\beta - \log\Gamma\left(\alpha\right) + \left(\alpha - 1\right)\log x_i - \beta x_i\right] \\
& \sum_{i=1}^{n} I\left(x_i \geq u\right)\log\left[1 - \int_0^u \tfrac{\beta^\alpha}{\Gamma(\alpha)} t^{\alpha-1} e^{-\beta t} dt\right] - \sum_{i=1}^{n} I\left(x_i \geq u\right)\log\sigma \\
& -\tfrac{1+\xi}{\xi}\sum_{i=1}^{n} I\left(x_i \geq u\right)\log\left[1 + \tfrac{\xi(x_i - u)}{\sigma}\right] \\
& +(a-1)\log\alpha - b\alpha\alpha + (c-1)\log\left(\tfrac{\alpha}{\beta}\right) - d\left(\tfrac{\alpha}{\beta}\right) + \log\left(\tfrac{\alpha}{\beta^2}\right) \\
& +\tfrac{1}{2\sigma_u}\left(u\mu_u\right)^2 + (a_1 - 1)\log\left[\tfrac{\sigma}{\xi}\left(k_1^{-\xi} - 1\right)\right] \\
& -b_1\tfrac{\sigma}{\xi}\left(k_1^{-\xi} - 1\right) + (a_2 - 1)\log\tfrac{\sigma}{\xi}\left(k_2^{-\xi} - k_1^{-\xi}\right) - b_2\tfrac{\sigma}{\xi}\left(k_2^{-\xi} - k_1^{-\xi}\right) \\
& +\log\left|\tfrac{\sigma}{\xi^2}\left[\left(k_1 k_2\right)^{-\xi}\left(\log k_2 - \log k_1\right) - k_2^{-\xi}\log k_2 + p_1^{-\xi}\log k_1\right]\right|,
\end{aligned} \tag{4.8}$$

Hyperparameter	a_1	b_1	a_2	b_2
Expert 1	0.1	0.005	0.9	0.0302

Table 4.2: *Hyperparameters for a fictitious expert*

	10,000 runs		
Parameter	Mean	Median	Std
σ	66.494	65.533	15.056
ξ	0.074	0.048	0.077
$\text{VaR}_{0.99}$	171.740	141.086	97.729
$\text{ES}_{0.99}$	254.513	223.150	114.117

Table 4.3: *Fraud data: Prior MCMC estimates, using the opinion of a fictitious expert and the prior in Equation* (4.5)

	10,000 runs		
Parameter	Mean	Median	Std
α	0.559	0.558	0.031
β	0.015	0.014	0.002
u	49.404	43.777	21.685
σ	69.061	64.355	19.288
ξ	0.202	0.207	0.068
$\text{VaR}_{0.99}$	363.684	360.392	34.825
$\text{ES}_{0.99}$	532.386	527.596	54.286

Table 4.4: *Posterior MCMC estimates, using the opinion of a fictitious expert for n=626 fraud losses in 41 banks, recorded from 01/2007 to 04/2010*

where α, β, a_1, a_2, b_1 and b_2 are hyperparameters from the Gamma distribution and K is a normalizing constant.

To study the performance of the prior, we obtained samples of $\text{VaR}_{0.99}$ and $\text{ES}_{0.99}$ from the prior and the corresponding posterior. Again, we used the real data described in Section 3.2.1. This time the main purpose is to study the behaviour of the posterior when the prior is elicited from an expert. For simplicity, we pooled all the data together rather than look at each bank separately, as in Chapter 3. It is important to mention that since real experts were not available, we chose the hyperparameters so that the MCMC algorithm was reasonably efficient. Table 4.2 shows the hyperparameters of a fictitious expert while Tables 4.3 and 4.4 and Figures 4.3–4.7 display the results of this exercise.

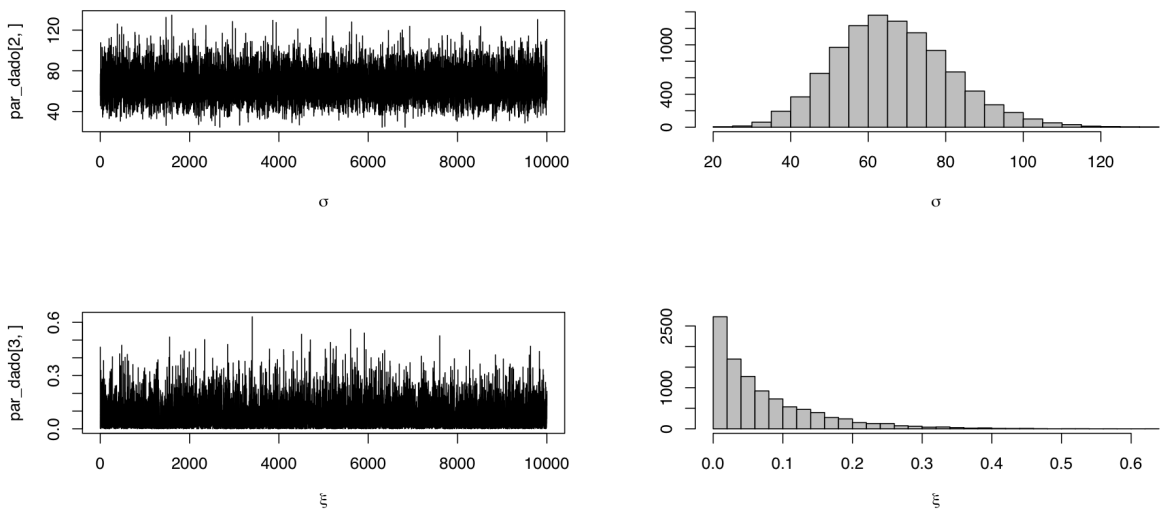

Figure 4.3: *Trace plots and histograms of prior samples, using the opinion of a fictitious expert and the prior in Equation* (4.5)

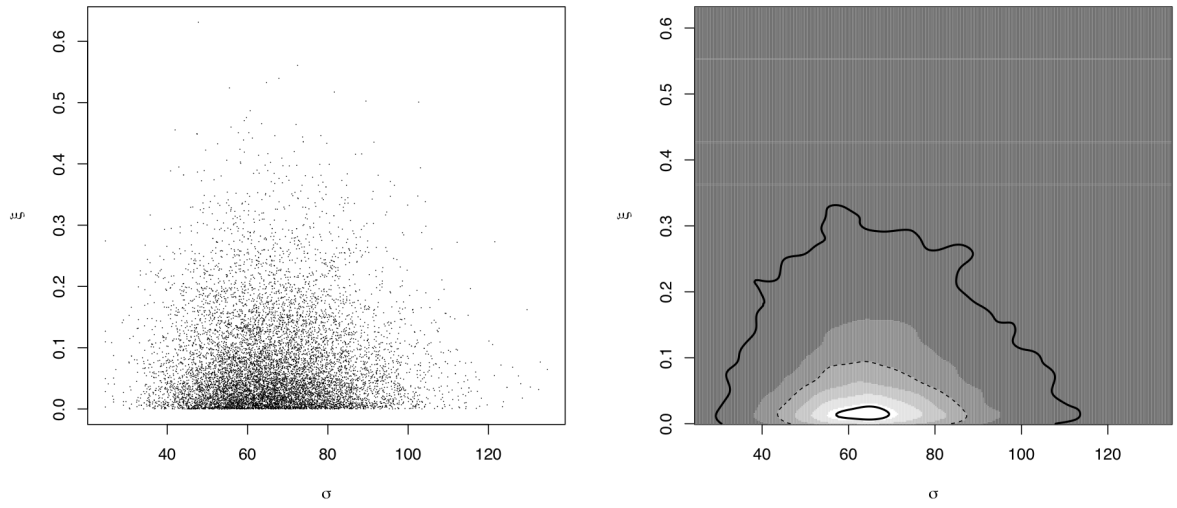

Figure 4.4: *Prior samples (left) and contour plot (right) of the joint distribution of σ and ξ (truncated at 0), using the opinion of a fictitious expert and the prior in Equation 4.5*

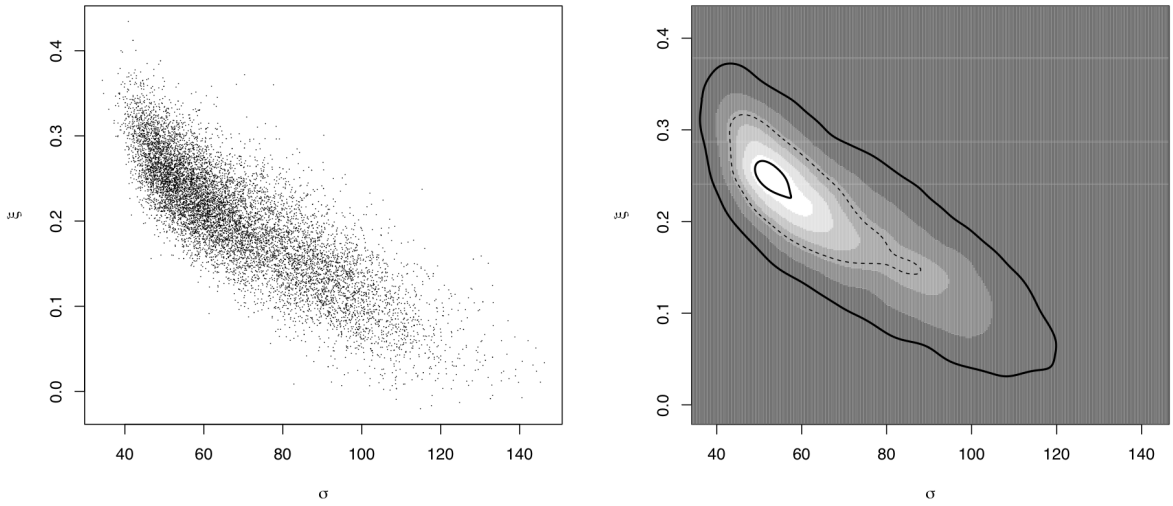

Figure 4.5: *Posterior samples (left) and contour plots (right) of the joint posterior distribution of σ and ξ, using the opinion of a fictitious expert for n=626 fraud losses in 41 banks, recorded from 01/2007 to 04/2010*

Parameter	Effective Sample Size
σ	9357.655
ξ	9735.053
$\text{VaR}_{0.99}$	8878.431
$\text{ES}_{0.99}$	9122.453

Table 4.5: *Effective Sample Size of prior samples, using the opinion of a fictitious expert and the prior in Equation* (4.5)

From the figures, the contour plot of the prior is similar to that of a non-informative prior; however, when look at the posterior, we can observe how the prior influences the result. We can notice that there has been a slight reduction in uncertainty from prior to posterior. The location of the density has also changed and the posterior distribution captures the negative correlation between parameters σ and ξ.

Some convergence diagnostics are also shown in Tables 4.5–4.8. In all cases convergence seems to be achieved and the effective sample size is large enough.

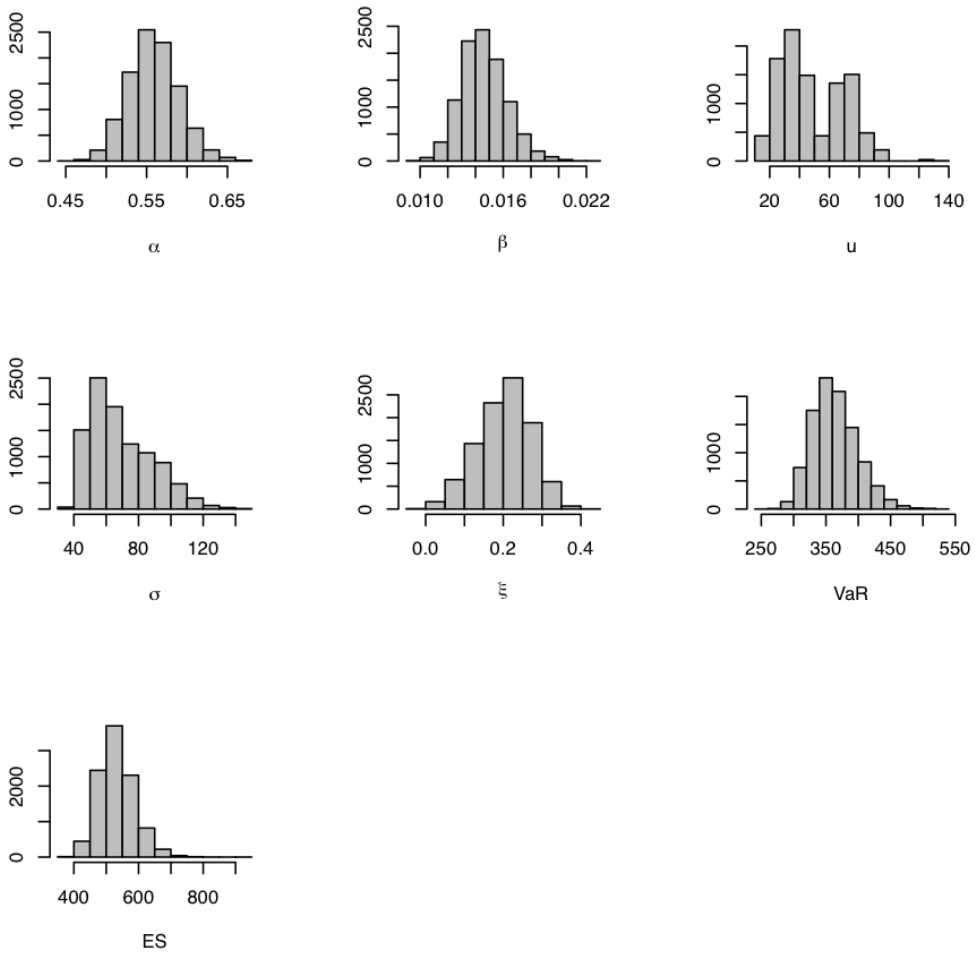

Figure 4.6: *Histograms of posterior samples, using the opinion of a fictitious expert for n=626 fraud losses in 41 banks, recorded from 01/2007 to 04/2010*

	Stationarity		Halfwidth Mean	
Parameter	Result	p-value	Result	Halfwidth
u	passed	0.889	passed	1.029
σ	passed	0.916	passed	0.280
ξ	passed	0.240	passed	0.001
$\text{VaR}_{0.99}$	passed	0.909	passed	2.103
$\text{ES}_{0.99}$	passed	0.913	passed	2.348

Table 4.6: *Heidelberg and Welch diagnostics for prior samples, using the opinion of a fictitious expert and the prior in Equation 4.5*

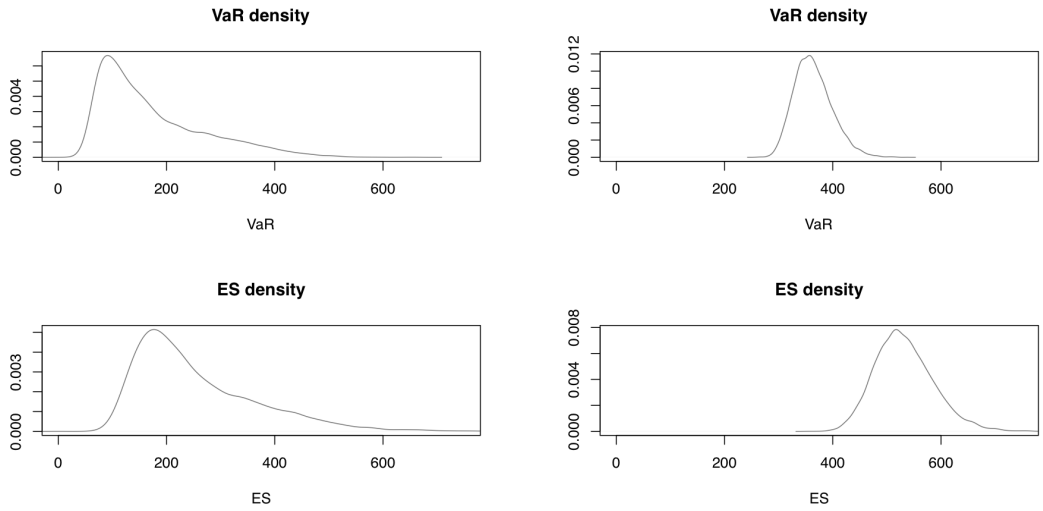

Figure 4.7: $\mathrm{VaR}_{0.99}$ *and* $\mathrm{ES}_{0.99}$ *prior (left) and posterior (right) distribution, using the opinion of a fictitious expert for n=626 fraud losses in 41 banks, recorded from 01/2007 to 04/2010*

Parameter	Effective Sample Size
α	4636.1564
β	3881.205
u	949.799
σ	874.672
ξ	1188.032
$\mathrm{VaR}_{0.99}$	1941.437
$\mathrm{ES}_{0.99}$	6504.822

Table 4.7: *Effective Sample Size of posterior samples, using the opinion of a fictitious expert for n=626 fraud losses in 41 banks, recorded from 01/2007 to 04/2010*

	Stationarity		Halfwidth Mean	
Parameter	Result	p-value	Result	Halfwidth
α	passed	0.788	passed	8.240e-04
β	passed	0.844	passed	5.110e-05
u	passed	0.427	passed	1.410
σ	passed	0.319	passed	1.290
ξ	passed	0.393	passed	4.060e-03
$\mathrm{VaR}_{0.99}$	passed	0.206	passed	1.550
$\mathrm{ES}_{0.99}$	passed	0.255	passed	1.320

Table 4.8: *Heidelberg and Welch diagnostics for posterior samples, using the opinion of a fictitious expert for n=626 fraud losses in 41 banks, recorded from 01/2007 to 04/2010*

4.3 Elicitation from multiple experts

As pointed out in the paper by Jenkinson (2005), whilst understanding the opinions of one expert is useful, there is an underlying statistical principle that the more information we have got, the better the results will be. Hence, it can be preferable to elicit the opinions of several experts.

According to Genest (1984b), Genest and Zidek (1986) and McConway (1981), among others, there are two possible ways of combining multiple opinions: Elicit the distributions from each expert and combine them mathematically (mathematical approach), or elicit a consensus distribution (behavioural approach).

Mathematical methods are divided into two types: axiomatic approaches and Bayesian approaches. The two main axiomatic approaches are:

1. The linear opinion pool.

2. The logarithmic opinion pool.

The linear opinion pool. Let $\pi_j(\theta)$, $j = 1, 2, ..., k$ be the j-th expert's probability density function and w_j be the weight attached to the j-th expert's opinion, with $w_j > 0$ and $\sum_{j=1}^{k} w_j = 1$. The linear opinion pool is the weighted arithmetic mean of the densities

$$\pi(\theta) = \sum_{j=1}^{k} w_j \pi_j(\theta). \tag{4.9}$$

This method satisfies the marginalization property (Genest, 1984b and McConway, 1981), which states that for a multivariate θ, the marginal probability from the combined density for any of the components is the same as what would be achieved if the elicited marginal distributions for that component were combined.

The logarithmic opinion pool. This method consist of the weighted geometric mean of the densities:

$$\pi(\theta) = n(w) \prod_{j=1}^{k} \pi_j(\theta)^{w_j}, \tag{4.10}$$

where $n(w)$ is a normalizing constant, i.e.:

$$n^{-1}(w) = \int_{\Theta} \prod_{j=1}^{k} \pi_j(\theta)^{w_j} \, d\theta$$

and the weigthts w_j are nonnegative and sum up to one.

This approach satisfies the external Bayesian (EB) principle (Genest, 1984a), which refers to how a decision maker updates the combined distribution when new information becomes available.

For simplicity, in this work we will focus on the linear opinion pool approach.

4.3.1 Prior on σ and ξ for multiple experts

We had previously defined the prior distribution for the GPD parameters (σ and ξ) in terms of the Value-at-Risk (Equation 4.5).

We can now use the two axiomatic approaches to obtain a prior for multiple experts. Under this assumption, we can elicit the VaR and work with the differences as follows:

$$d_{1j} = \left(\text{VaR}_{q_{1j}} - u\right), \;\; d_{2j} = \left(\text{VaR}_{q_{2j}} - u\right) - \left(\text{VaR}_{q_{1j}} - u\right)$$

for each expert j.

We can keep the assumption $\pi(d_{ij}) \sim \text{Gamma}(a_{ij}, b_{ij}), \;\; i = 1, 2; \;\; j = 1, ..., k.$

Let $\pi_j(\sigma, \xi)$ be the prior for expert j as defined in Equation (4.5). Then, under the linear opinion pool, the prior for k experts can be written as:

$$\pi(\sigma, \xi) = \sum_{j=1}^{k} w_j \pi_j(\sigma, \xi), \tag{4.11}$$

where w_j is the weight assigned to expert j. Under the logarithmic pool approach, the density is given by:

$$\pi(\sigma, \xi) = n \prod_{j=1}^{k} \pi_j(\sigma, \xi)^{w_j}. \tag{4.12}$$

Opinion	Similar		Different		Very different	
Hyp.	Expert 1	Expert 2	Expert 1	Expert 2	Expert 1	Expert 2
a_{1i}	0.1	0.15	0.1	5	0.1	10.5
b_{1i}	0.005	0.007	0.005	0.252	0.005	0.53
a_{2i}	0.9	1	0.9	9	0.9	20
b_{2i}	0.03	0.033	0.03	0.302	0.03	0.671

Table 4.9: *Sets of hyperparameters for two fictitious experts with similar, different and very different opinions*

4.3.2 Multiple experts and real data

In order to compare the behaviour of the posterior density of $\text{VaR}_{0.99}$ and $\text{ES}_{0.99}$ under the linear opinion pool, we analyze the real data from Section 3.2.1 for different scenarios. Recall that, for simplicity, we pooled all the data together and not looking at each bank separately. We assume that we have two experts and three different cases:

1. Two experts with similar opinions and weights $w_1 = w_2 = 0.5$
2. Two experts with different opinions and weights $w_1 = w_2 = 0.5$
3. Two experts with very different opinions and weights $w_1 = w_2 = 0.5$

Similarities and differences in opinions are expressed through the hyperparameters. Table 4.9 shows three different sets of hyperparameters, according to the case considered.

Tables 4.10–4.15 and Figures 4.8–4.10 show the results for different scenarios. When both experts have similar opinions, prior and posterior parameter estimates do not show great variation with respect to the original estimates (using only one expert opinion); if experts have different opinions, the estimates vary slightly; however, when experts have very different opinion, the estimates vary considerably. In this last case, one may notice that the combined posterior resembles closely one of the prior densities (Figure 4.10).

	10,000 runs		
Parameter	Mean	Median	Std
σ	66.409	65.399	14.824
ξ	0.073	0.049	0.075
$\text{VaR}_{0.99}$	173.455	144.803	95.923
$\text{ES}_{0.99}$	255.781	227.924	110.288

Table 4.10: *Fraud data: Prior MCMC estimates for the linear opinion pool for two fictitious experts with similar opinions and weights* $w_1 = w_2 = 0.5$ *(10,000 runs-1,000 after thinning)*

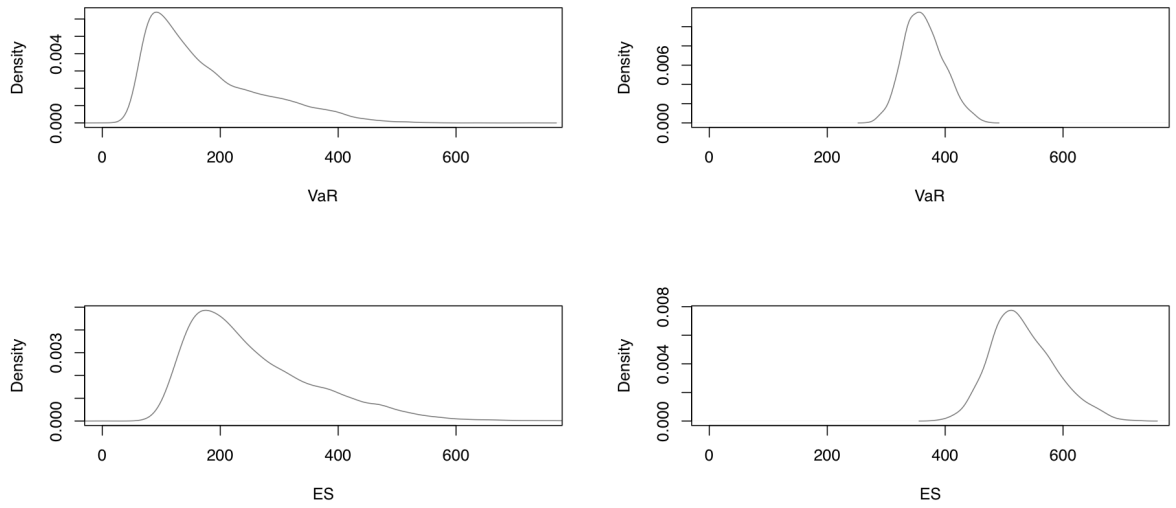

Figure 4.8: *Fraud data:* $\text{VaR}_{0.99}$ *and* $\text{ES}_{0.99}$ *prior (left) and posterior (right) distribution for the linear opinion pool for two fictitious experts with similar opinions and weights* $w_1 = w_2 = 0.5$ *(10,000 runs-1,000 after thinning)*

	10,000 runs		
Parameter	Mean	Median	Std
α	0.559	0.558	0.031
β	0.015	0.014	0.002
u	50.147	43.777	20.217
σ	69.477	65.102	17.740
ξ	0.201	0.208	0.0643
$\text{VaR}_{0.99}$	364.224	361.466	33.656
$\text{ES}_{0.99}$	532.409	525.920	53.799

Table 4.11: *Fraud data: Posterior MCMC estimates for the linear opinion pool for two fictitious experts with similar opinions and weights* $w_1 = w_2 = 0.5$ *(10,000 runs-1,000 after thinning)*

	10,000 runs		
Parameter	Mean	Median	Std
σ	27.805	26.499	3.328
ξ	0.004e-01	9.3e-07	0.009
$\text{VaR}_{0.99}$	107.827	87.858	63.369
$\text{ES}_{0.99}$	135.712	114.484	63.778

Table 4.12: *Fraud data: Prior MCMC estimates for the linear opinion pool for two fictitious experts with different opinions and weights* $w_1 = w_2 = 0.5$ *(10,000 runs-1,000 after thinning)*

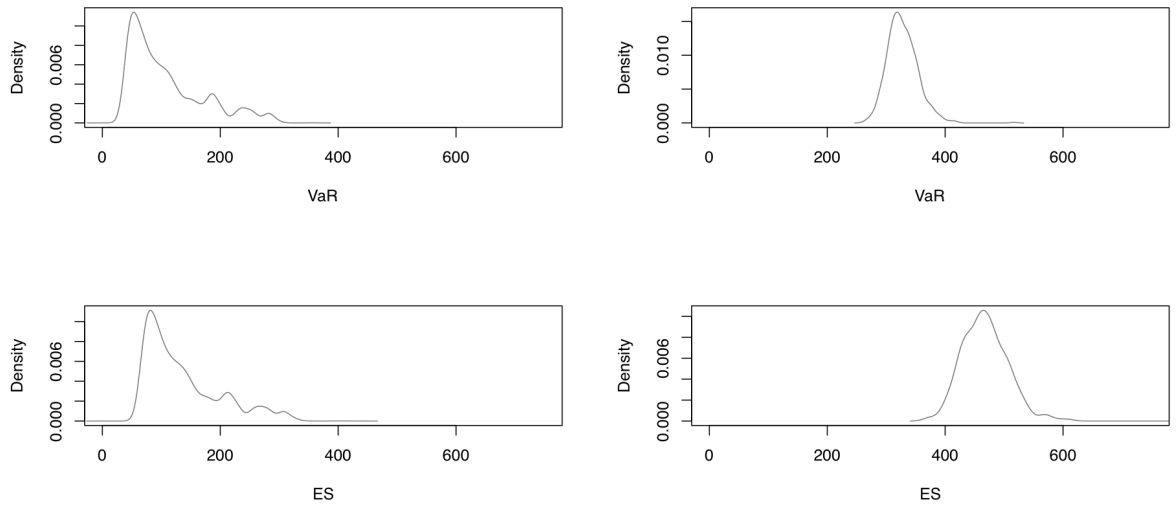

Figure 4.9: *Fraud data:* $\text{VaR}_{0.99}$ *and* $\text{ES}_{0.99}$ *prior (left) and posterior (right) distribution for the linear opinion pool for two fictitious experts with different opinions and weights* $w_1 = w_2 = 0.5$ *(10,000 runs-1,000 after thinning)*

	10,000 runs		
Parameter	Mean	Median	Std
α	0.561	0.561	0.033
β	0.015	0.015	0.002
u	38.081	36.228	16.606
σ	59.627	57.432	12.075
ξ	0.185	0.187	0.048
$\text{VaR}_{0.99}$	327.787	325.256	25.771
$\text{ES}_{0.99}$	467.983	465.486	40.546

Table 4.13: *Fraud data: Posterior MCMC estimates for the linear opinion pool for two fictitious experts with different opinions and weights* $w_1 = w_2 = 0.5$ *(10,000 runs-1,000 after thinning))*

	10,000 runs		
Parameter	Mean	Median	Std
σ	65.747	64.972	15.602
ξ	0.072	0.047	0.075
$\text{VaR}_{0.99}$	170.194	140.679	95.520
$\text{ES}_{0.99}$	251.542	223.487	111.087

Table 4.14: *Fraud data: Prior MCMC estimates for the linear opinion pool for two fictitious experts with very different opinions and weights* $w_1 = w_2 = 0.5$ *(10,000 runs-1,000 after thinning)*

	10,000 runs		
Parameter	Expert 1	Expert 2	Both experts
α	0.567	0.561	0.555
β	0.0152	0.015	0.014
u	25.012	39.786	45.399
σ	49.898	56.734	62.127
ξ	0.152	0.387	0.369
$\text{VaR}_{0.99}$	279.074	445.949	448.077
$\text{ES}_{0.99}$	384.766	795.887	773.989

Table 4.15: *Fraud data: Posterior MCMC estimates for the linear opinion pool for two fictitious experts with very different opinions and weights* $w_1 = w_2 = 0.5$ *(10,000 runs-1,000 after thinning)*

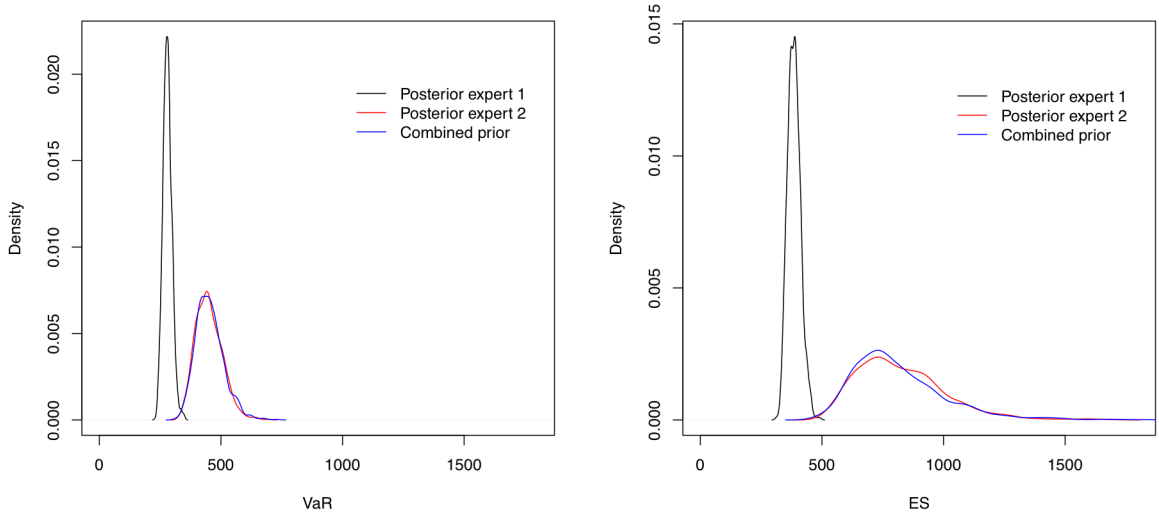

Figure 4.10: *Fraud data: Posterior distribution of* $\text{VaR}_{0.99}$ *and* $\text{ES}_{0.99}$ *for the linear opinion pool for two fictitious experts with very different opinions and weights* $w_1 = w_2 = 0.5$ *(10,000 runs-1,000 after thinning)*

Opinion	Very different	
Hyperparameters	Expert 1	Expert 2
a_{1i}	24	0.005
b_{1i}	1.21	2e-04
a_{2i}	26	0.008
b_{2i}	0.87	3e-04

Table 4.16: *Set of hyperparameters for two experts with very different opinions for n=1000 simulated data from a Gamma(100,0.09)*

To determine if this problem is related to the data set, we simulated 1000 observations from a Gamma(100,0.09). Again, the differences in opinions are expressed through the hyperparameters. Table 4.16 shows the hyperparameters used in this analysis for fictitious experts.

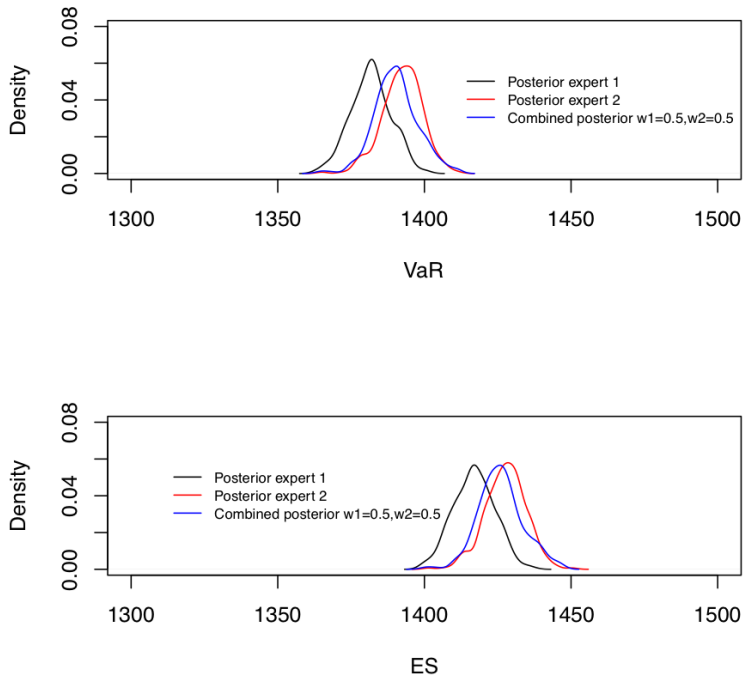

Figure 4.11: $\text{VaR}_{0.99}$ *and* $\text{ES}_{0.99}$ *posterior distribution (linear opinion pool for two fictitious experts with very different opinions) for n=1000 simulated data from a Gamma(100,0.09)*

In this case the posterior does not exhibit the same behaviour; however, it is still close to one of the prior densities. (Figure 4.11 and Table 4.17).

Gamma(100,0.09)-10,000 runs			
$\text{VaR}_{0.99}$	Mean	Median	Std
Posterior Expert 1	1381.758	1381.749	6.891
Posterior Expert 2	1392.442	1392.782	6.820
Posterior $w_1 = 0.5$, $w_2 = 0.5$	1390.320	1390.025	7.468
True value	1385.806		

Table 4.17: $\text{VaR}_{0.99}$ *estimates and true value (linear opinion pool for two fictitious experts with very different opinions) for n=1000 simulated data from a Gamma(100,0.09)*

4.3.3 A new prior for multiple experts

In order to improve the prior proposed in Section 4.2, we consider a different prior, based on the same elicited quantities. Consider the ratio

$$r_1 = \frac{(\text{VaR}_{q_1} - u)}{(\text{VaR}_{q_2} - u)} = \frac{\frac{\sigma}{\xi}\left(k_1^{-\xi} - 1\right)}{\frac{\sigma}{\xi}\left(k_2^{-\xi} - 1\right)} = \frac{k_1^{-\xi} - 1}{k_2^{-\xi} - 1}.$$

Then

$$\begin{aligned}
\frac{dr_1}{d\xi} &= \frac{-\left(k_1^{-\xi}\ln k_1\right)\left(k_2^{-\xi} - 1\right) + \left(k_1^{-\xi} - 1\right)\left(k_2^{-\xi}\ln k_2\right)}{\left(k_2^{-\xi} - 1\right)^2} \\
&= \frac{-\left(k_1k_2\right)^{-\xi}\ln k_1 + k_1^{-\xi}\ln k_1 + \left(k_1k_2\right)^{-\xi}\ln k_2 - k_2^{-\xi}\ln k_2}{\left(k_2^{-\xi} - 1\right)^2} \\
&= \frac{-\left(k_1k_2\right)^{-\xi}\left(\ln k_1 - \ln k_2\right) + k_1^{-\xi}\ln k_1 - k_2^{-\xi}\ln k_2}{\left(k_2^{-\xi} - 1\right)^2}.
\end{aligned}$$

If we choose a Beta distribution for the ratio r_1, we have:

$$\pi\left(\xi \mid u, \sigma\right) = \pi\left(r_1\right)\left|\frac{\partial r_1}{\partial \xi}\right| \propto r_1^{a_1 - 1}\left(1 - r_1^{b_1 - 1}\right)\left|\frac{\partial r_1}{\partial \xi}\right|. \tag{4.13}$$

We also choose an Inverse Gaussian for σ and a truncated normal for u.

For the simulated data (Gamma(100,0.09)), using the prior specified above and the hyperparameters for fictitious experts in Table 4.16, we obtain the densities shown in Figures 4.12 and 4.13.

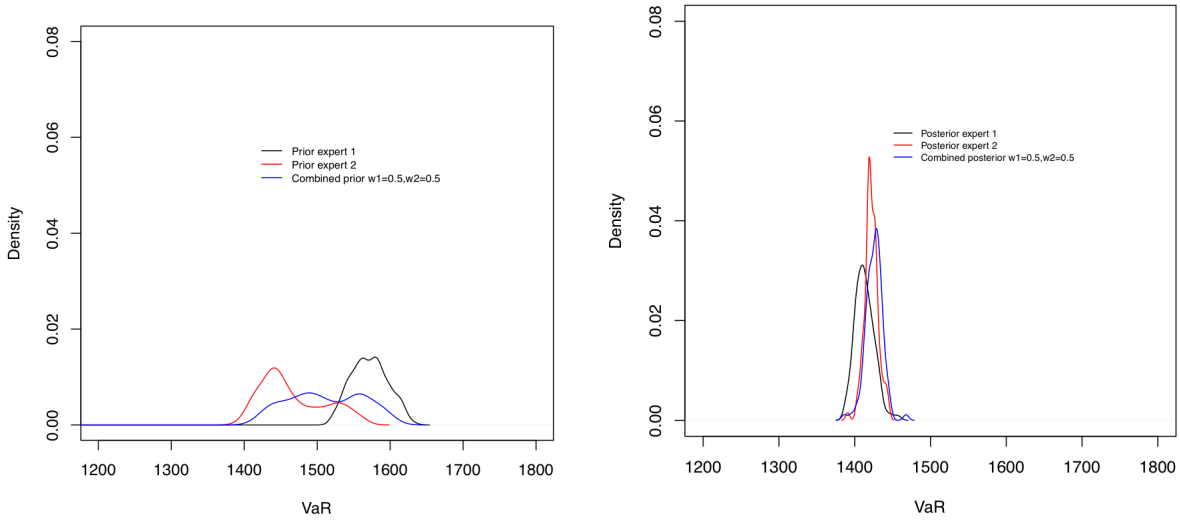

Figure 4.12: *Prior (left) and posterior (right) distribution of* $\text{VaR}_{0.99}$*, using the prior from Equation* (4.13) *and the linear opinion pool for two fictitious experts with very different opinions. Expert 1 (black), expert 2 (red) and combined distribution (blue)*

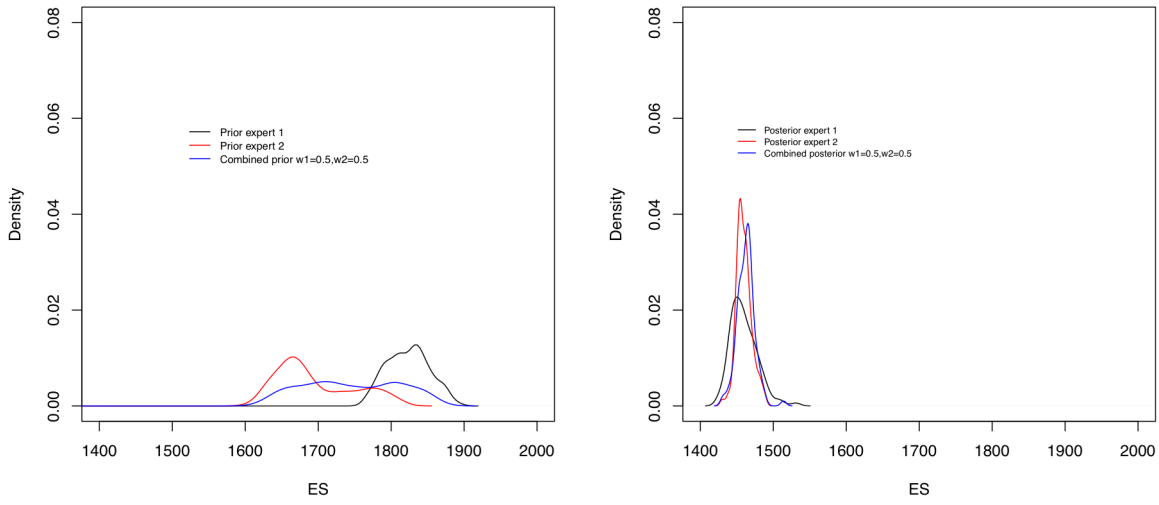

Figure 4.13: *Prior (left) and posterior (right) distribution of* $\text{ES}_{0.99}$*, using the prior from Equation* (4.13) *and the linear opinion pool for two fictitious experts with very different opinions. Expert 1 (black), expert 2 (red) and combined distribution (blue)*

4.3.4 Posterior analysis for more than two experts

So far, we have considered only the case when the prior is based on two experts' opinion. We have observed how the differences between these opinions may influence the prior and posterior behaviour. To conclude this chapter, we study the posterior behaviour when several experts express their opinions. We consider five experts with different opinions. Again, we resort to the Gamma simulated data, using three sets of hyperparameters (Table 4.18). The results obtained are shown in Figure 4.14.

As can be seen from the figures, even when the posteriors for each expert are different, all lead to similar posterior distributions.

Table 4.19 shows the estimates of the Value-at-Risk at 99.5% level. It can be seen from this table that experts whose opinion is far from the true value influence the combined posterior in all cases. However, the combined estimate is still acceptable.

To complete this study, we perform the analysis in two different subsets of the Gamma$(100, 0.09)$ data, using the same hyperparameters. Results are shown in Tables 4.20 and 4.21. This time, the results depend on the subset we are working with; however, we see again how experts whose opinion is far from the true value influence the combined posterior.

The results of this analysis indicate that when dealing with several opinions, one should pay special attention to prior specification.

	Set 1					Set 2					Set 3				
Hyp.	E1	E2	E3	E4	E5	E1	E2	E3	E4	E5	E1	E2	E3	E4	E5
a_{1i}	0.001	25	12	8	1	3	28	13	11	3	0.001	0.28	0.15	0.11	0.3
b_{1i}	0.3	13	14	7	2	3.3	16	17	10	4	0.3	0.16	0.17	0.1	0.4
a_{2i}	10	6	8	8	0.9	12	9	10	11	2.9	0.12	0.9	0.1	0.11	0.29
b_{2i}	0.3	1	2	5	0.7	2.3	4	4	8	2.7	0.23	0.4	0.4	0.8	0.27

Table 4.18: *Sets of hyperparameters for five different experts (E1,E2,E3,E4,E5)*

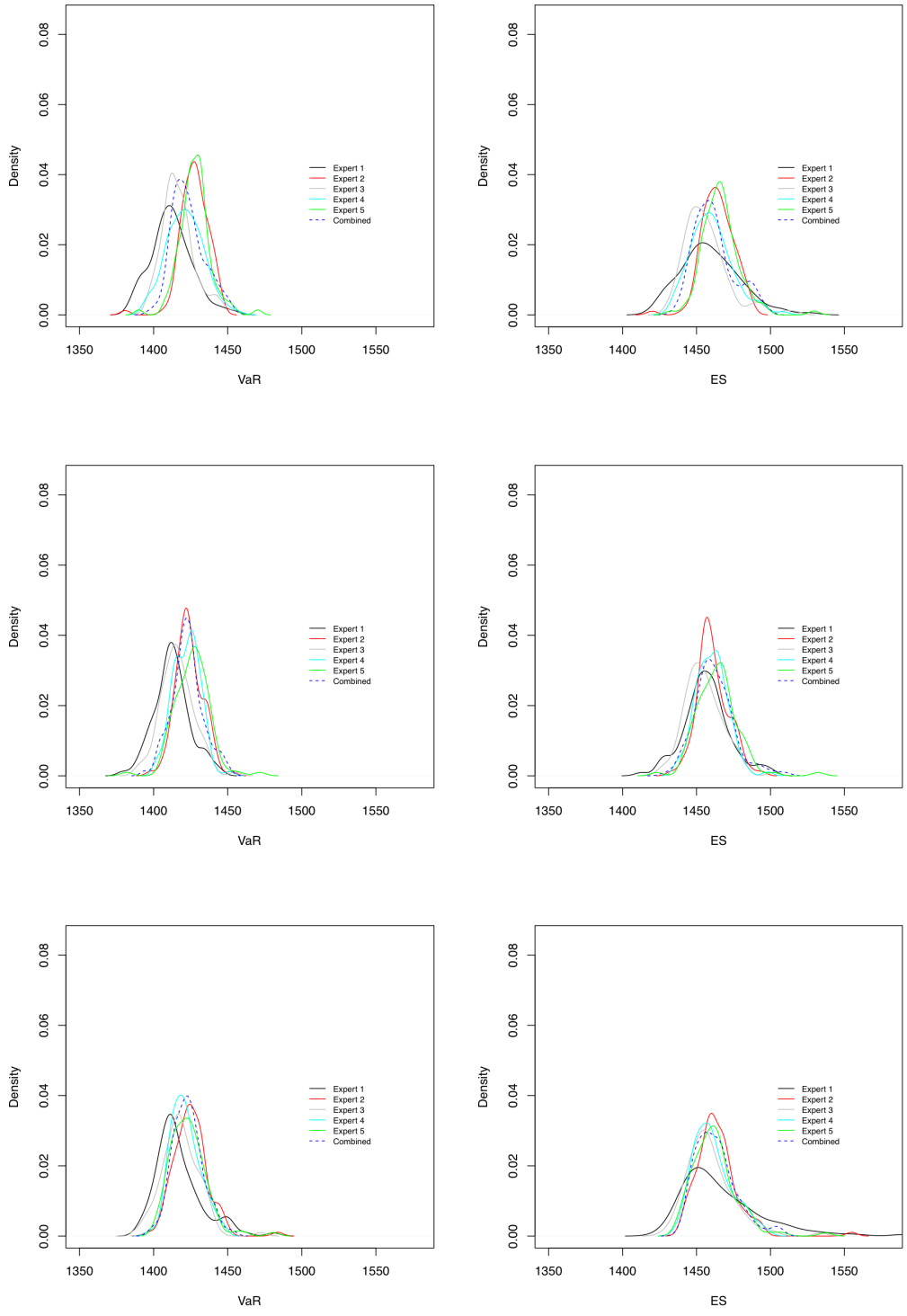

Figure 4.14: *Posterior distribution of* $\text{VaR}_{0.995}$ *and* $\text{ES}_{0.995}$ *for the first set (top), second set (middle) and third set (bottom) of hyperparameters for n=1000 simulated data from a Gamma(100,0.09), using the opinion of five experts (E1,E2,E3,E4,E5) with equal weights*

	Set 1	Set 2	Set 3
Expert	Mean, Median, Std	Mean, Median, Std	Mean, Median, Std
E1	1411.990, 1411.600, 13.680	1412.450, 1412.050, 12.230	1416.880, 1413.290, 15.860
E2	1427.560, 1427.710, 9.410	1424.060, 1422.710, 9.190	1425.120, 1424.370, 11.850
E3	1416.900, 1415.200, 11.370	1415.940, 1415.400, 11.210	1418.020, 1417.360, 12.680
E4	1421.230, 1421.320, 12.290	1422.240, 1423.200, 9.010	1421.380, 1420.520, 9.840
E5	1427.770, 1427.410, 10.610	1426.070, 1426.540, 10.710	1423.240, 1422.370, 12.390
Combined	1423.330, 1421.310, 11.030	1422.690, 1421.970, 11.740	1422.930, 1422.540, 10.080
True value		1418.134	

Table 4.19: *Parameter estimates of $VaR_{0.995}$ for n=1000 simulated data from a Gamma(100,0.09), using the opinion of five experts (E1,E2,E3,E4,E5) with equal weights (10,000 runs)*

	Set 1	Set 2	Set 3
Expert	Mean, Median, Std	Mean, Median, Std	Mean, Median, Std
E1	1371.820, 1370.186, 6.989	1443.776, 1441.065, 55.335	1405.031, 1397.473, 31.126
E2	1391.487, 1389.019, 11.620	1395.270, 1393.940, 10.823	1399.606, 1396.492, 17.876
E3	1386.927, 1383.362, 8.986	1388.259, 1385.408, 16.212	1400.394, 1397.394, 24.305
E4	1385.866, 1385.433, 6.877	1392.034, 1389.880, 13.309	1396.940, 1394.319, 17.206
E5	1396.547, 1397.347, 13.535	1395.127, 1393.501, 11.729	1399.752, 1397.355, 17.270
Combined	1389.755, 1387.520, 7.271	1398.320, 1393.878, 16.025	1400.570, 1393.845, 18.116

Table 4.20: *Parameter estimates of $VaR_{0.995}$ for the first subset of n=1000 simulated data from a Gamma(100,0.09), using the opinion of five experts (E1,E2,E3,E4,E5) with equal weights (10,000 runs)*

	Set 1	Set 2	Set 3
Expert	Mean, Median, Std	Mean, Median, Std	Mean, Median, Std
E1	1431.522, 1423.112, 48.875	1452.904, 1445.230, 46.153	1463.631, 1457.973, 52.967
E2	1429.530, 1427.390, 19.159	1432.792, 1431.263, 21.051	1439.916, 1434.600, 30.502
E3	1414.734, 1413.327, 19.829	1413.011, 1410.485, 17.976	1426.850, 1420.401, 27.519
E4	1430.454, 1429.945, 25.231	1424.375, 1425.571, 19.345	1431.188, 1427.123, 24.997
E5	1439.861, 1438.333, 21.899	1434.694, 1435.101, 19.474	1442.770, 1441.365x, 23.989
Combined	1438.283, 1433.558, 23.646	1432.958, 1427.611, 23.489	1436.812, 1434.325, 27.353

Table 4.21: *Parameter estimates of $VaR_{0.995}$ for the second subset of n=1000 simulated data from a Gamma(100,0.09), using the opinion of five experts (E1,E2,E3,E4,E5) with equal weights (10,000 runs)*

4.3.5 Updating the weighting of the experts' opinions

The Bayesian model we introduced before allows the incorporation of expert opinion into the analysis (via the prior distributions) for model parameters. These parameters can be updated as new data become available. Experts may reassess their opinion to incorporate new information (for example, new policies or controls) and this can be done as follows.

Using equation (2.39), let $\pi_j(\theta)$ be the prior density for θ and $L(\theta \mid x)$ be the likelihood. Then, the posterior density is:

$$p_j(\theta \mid x) = \frac{\pi_j(\theta) L(\theta \mid x)}{K_j}, \tag{4.14}$$

where $K_j = \int_{-\infty}^{\infty} \pi_j(\theta) L(\theta \mid x) d\theta$.

Now, let $\pi(\theta) = \sum_{j=1}^{J} w_j \pi_j(\theta)$ be a mixture prior given. Hence, the posterior density is

$$p(\theta \mid x) = \frac{\sum_{j=1}^{J} w_j \pi_j(\theta) L(\theta \mid x)}{K} = \frac{\sum_{j=1}^{J} \frac{w_j K_j \pi_j(\theta) L(\theta \mid x)}{K_j}}{K} = \frac{\sum_{j=1}^{J} w_j K_j p_j(\theta \mid x)}{K}. \tag{4.15}$$

From (4.15), we require $\sum_{j=1}^{J} w_j K_j / K = 1$, which implies $K = \sum_{j=1}^{J} w_j K_j$. Therefore, the posterior distribution is given by

$$p(\theta \mid x) = \sum_{j=1}^{J} w_j' p_j(\theta \mid x), \tag{4.16}$$

where $w_j' = \{w_j K_j\} / \sum_{h=1}^{J} w_h K_h$.

4.3.6 Prior weights updating

Once we have performed the analysis and observed the results, we might want to evaluate how accurate experts are in their prediction, and adjust the weight assigned to each expert in future exercises, based on their past performance. To do so, one can use a measure of divergence, such as the Kullback-Leibler divergence[1].

[1]Let $g(x)$ and $h(x)$ be two probability density functions defined over the same support S. The Kullback and Leibler divergence between the two distributions is defined as:

$$D(g \,||\, h) = \int_S g(x) \ln \frac{g(x)}{h(x)} dx$$

Suppose that, for each risk assessment, we base our prior information on the opinion of n experts. In each exercise we might do the following:

- Follow the opinion of the majority of experts.
- After a pre-determined number of assessments, see which of the experts get it right most of the time and then follow their advice.

Although these strategies work well in some cases, the first one fails when only a few experts make good predictions. The second one fails when there is an expert that performs well for the first evaluations but is wrong after that.

In this section, we will consider the Multiplicative Weights approach (Arora et al., 2012), which allows us to consider the opinion of all experts, but weighting each expert's opinion according to his/her past performance.

Multiplicative update algorithms were proposed in game theory in the early fifties. One of the first ideas was that at each step each player observes actions taken by his/her opponent in previous stages, updates his beliefs about his opponents' strategies, and chooses myopic pure best responses against these beliefs. However, the multiplicative update rule was rediscovered in Computational Geometry in the late 1980s, while the weighted majority algorithm has been independently discovered in operations research and statistical decision making.

Initially, the algorithm assigns equal weights to all experts. As time goes on, some experts are seen as making better predictions than others, and the algorithm increases their weight proportionally. The algorithm is as follows.

Weighted majority algorithm. At every step t, we have a weight w_i^t assigned to expert i. Initially equal weights for all i. At step $t+1$, for each i such that expert i was found to have predicted the quantity of interest incorrectly, we set

$$w_i^{t+1} = (1-\epsilon)\, w_i^t \quad \text{(update rule)}$$

Our prediction for step $t+1$ is the opinion of a weighted majority of the experts.

The following result (Theorem 1.1, Arora et al. 2012) shows that the number of mistakes the algorithm makes is bounded above.

Theorem 4.1. *After t steps, let m_i^t be the number of mistakes of expert i and m^t be the number of mistakes our algorithm has made. Then we have the following bound for every i:*

$$m^t \leq \frac{2\,ln\,n}{\epsilon} + 2\,(1+\epsilon)\,m_i^t.$$

We have adapted this algorithm to our problem by considering the KL divergence. We propose the following procedure to perform the updating:

1. Set $w_j^0 = 1/n$ for $i = 1, ..., n$.

2. After assessment t, update the weights as $w_i^t = \frac{(1-\epsilon_i)w_i^{t-1}}{\sum_j (1-\epsilon_j)w_j^{t-1}}$. Here, $\epsilon_j = {}^{KL}/_{10^N}$, where KL corresponds to the Kullback-Leibler divergence between the posterior distribution of VaR of expert j and the combined posterior distribution (including the opinion of all experts), and N is the number of digits left of the decimal point of $\max_j(\text{KL}_j)$.

This procedure assigns more weight to experts whose opinion is closer to the mixture of opinions and penalizes those who are far from the majority, based on the the KL divergence.

We can apply this algorithm to the simulated data Gamma(100,0.09), and to the the posterior distribution of $\text{VaR}_{0.99}$ for different sets of hyperparameters (Table 4.18). In this case, $n = 5$ and $w_j^0 = 1/5 = 0.2$ for $j = 1, ..., 5$.

Table 4.22 summarizes the results for the different sets of hyperparameters. From it, we can notice that as KL increases, w_j^1 decreases. That is, those experts whose KL divergence is large are penalized in the next risk assessment by assigning them smaller weights. On the other hand, experts with smaller KL divergence gain more credibility and get larger weights in the new assessment.

	Set 1				Set 2			Set 3		
Expert	w_j^0	KL_j	ϵ_j	w_j^1	KL_j	ϵ_j	w_j^1	KL_j	ϵ_j	w_j^1
E1	0.2	0.418	0.042	**0.205**	0.176	0.176	**0.220**	0.450	0.450	**0.159**
E2	0.2	1.638	0.164	**0.179**	0.155	0.155	**0.225**	0.330	0.330	**0.194**
E3	0.2	0.663	0.066	**0.199**	0.355	0.355	**0.172**	0.147	0.147	**0.246**
E4	0.2	0.293	0.029	**0.208**	0.335	0.335	**0.177**	0.029	0.029	**0.281**
E5	0.2	0.237	0.024	**0.209**	0.227	0.227	**0.206**	0.584	0.584	**0.120**

Table 4.22: *Prior weights updating for different sets of hyperparameters for n=1000 simulated data from a Gamma(100,0.09), using the opinion of five experts (E1,E2,E3,E4,E5) with equal weights*

	Set 1				Set 2			Set 3		
Expert	w_j^0	KL_j	ϵ_j	w_j^1	KL_j	ϵ_j	w_j^1	KL_j	ϵ_j	w_j^1
E1	0.2	0.178	0.178	**0.235**	0.062	0.062	**0.205**	0.061	0.061	**0.204**
E2	0.2	0.357	0.357	**0.183**	0.041	0.041	**0.208**	0.054	0.054	**0.205**
E3	0.2	0.199	0.199	**0.227**	0.093	0.093	**0.197**	0.030	0.030	**0.211**
E4	0.2	0.609	0.609	**0.111**	0.106	0.106	**0.194**	0.202	0.202	**0.173**
E5	0.2	0.139	0.139	**0.244**	0.097	0.097	**0.196**	0.044	0.044	**0.207**

Table 4.23: *Prior weights updating for different sets of hyperparameters for the first subset of n=1000 simulated data from a Gamma(100,0.09), using the opinion of five experts (E1,E2,E3,E4,E5) with equal weights*

In order to study the variation of the weights for different data, we obtain new weights for the two subsets of the Gamma(100,0.09) that were used in Section 4.3.4. Tables 4.23 and 4.24 display the results. Once again, one may notice that w_j^1 is inversely proportional to KL, and therefore experts whose KL divergence is large get smaller weights, while those with small KL divergence are assigned larger weights.

According to Basel II, operational risk has to be measured on a yearly basis. The method presented in this section allows us not only to update prior information as new information becomes available, but also the weights assigned to experts are based on the accuracy of their past opinions.

	Set 1				Set 2			Set 3		
Expert	w_j^0	KL_j	ϵ_j	w_j^1	KL_j	ϵ_j	w_j^1	KL_j	ϵ_j	w_j^1
E1	0.2	0.330	0.330	**0.167**	0.182	0.182	**0.214**	0.109	0.109	**0.218**
E2	0.2	0.022	0.022	**0.243**	0.144	0.144	**0.224**	0.401	0.401	**0.146**
E3	0.2	0.039	0.039	**0.239**	0.319	0.319	**0.178**	0.292	0.292	**0.174**
E4	0.2	0.190	0.190	**0.202**	0.190	0.190	**0.213**	0.065	0.065	**0.228**
E5	0.2	0.401	0.401	**0.149**	0.348	0.348	**0.171**	0.042	0.042	**0.234**

Table 4.24: *Prior weights updating for different sets of hyperparameters for the second subset of n=1000 simulated data from a Gamma(100,0.09), using the opinion of five experts (E1,E2,E3,E4,E5) with equal weights*

	Prior distribution				
Expert	Year 0	Year 1	Year 2	...	Year n
E1	$\pi_1(\theta)$	$p_1(\theta \mid x_1)$	$p_1(\theta \mid x_1, x_2)$	...	$p_1(\theta \mid x_1, x_2, ..., x_n)$
E2	$\pi_2(\theta)$	$p_2(\theta \mid x_1)$	$p_2(\theta \mid x_1, x_2)$	...	$p_2(\theta \mid x_1, x_2, ..., x_n)$
E3	$\pi_3(\theta)$	$p_3(\theta \mid x_1)$	$p_3(\theta \mid x_1, x_2)$	...	$p_3(\theta \mid x_1, x_2, ..., x_n)$
E4	$\pi_4(\theta)$	$p_1(\theta \mid x_1)$	$p_4(\theta \mid x_1, x_2)$	...	$p_4(\theta \mid x_1, x_2, ..., x_n)$
E5	$\pi_5(\theta)$	$p_1(\theta \mid x_1)$	$p_5(\theta \mid x_1, x_2)$	...	$p_5(\theta \mid x_1, x_2, ..., x_n)$

Table 4.25: *Prior distributions for different years*

4.3.7 Posterior distribution updating

We will assume now that the risk assessment is performed every year and, as time passes, more data become available. This affects the resulting posterior distribution and some adjustments have to be made. The more information we have, the better we are able to predict.

Under this situation, the prior distribution of each expert for the next year should be based on their posterior distribution from the previous year. That is, we start with opinions expressed in terms of the prior distribution $\pi_i(\theta)$ for year 0. For the next year, the prior of expert i is replaced by $p_i(\theta \mid x_1)$, which is the posterior distribution after including the data collected during the first year. We can continue this procedure for the upcoming years, as it is ilustrated in Table 4.25. We expect that after many years, as the number of observations increases, all of the experts will have similar posterior distributions.

To explore this, we use the simulated data Gamma(100,0.09) and the real data described in Section 3.2.1. For the simulated data, we assume that new information becomes

	Year 0			Year 1			Year 2			Year 3
Expert	w_j^0	KL_j	ϵ_j	w_j^1	KL_j	ϵ_j	w_j^2	KL_j	ϵ_j	w_j^3
E1	0.2	0.418	0.042	**0.205**	0.319	0.319	**0.169**	0.067	0.007	**0.206**
E2	0.2	1.638	0.164	**0.179**	0.043	0.043	**0.237**	1.201	0.120	**0.182**
E3	0.2	0.663	0.066	**0.199**	0.314	0.314	**0.170**	0.111	0.011	**0.205**
E4	0.2	0.293	0.029	**0.208**	0.228	0.228	**0.191**	0.278	0.028	**0.201**
E5	0.2	0.237	0.024	**0.209**	0.056	0.056	**0.233**	0.077	0.008	**0.206**

Table 4.26: *Prior weights for different experts in a 3-year period. n=1000 simulated data from a Gamma(100,0.09), using the opinion of five experts (E1,E2,E3,E4,E5) with equal weights*

available every year during 3 years. We start with 1,000 observations for year 0. For the next year, we add 400 observations to the data set. In the second year, we also add 400 observations. Finally, in the third year, 500 observations are added to the data set. Following the notation in Table 4.25, we have:

- x_1: Observations for year 0 (1000);
- x_2: Observations for year 1 (400);
- x_3: Observations for year 2 (400);
- x_4: Observations for year 3 (500).

Tables 4.26 and 4.27 show the results for the period considered, while Figure 4.15 displays the posterior distributions. We can observe that every year, the estimated VaR is closer to the true value. Additionally, the posterior distributions present less variability and look more similar to each other.

The same procedure is applied to the real data set from Section 3.2.1 to analyze the posterior behaviour when new data are added. We have 626 observations available from year 2007 to 2010. We divide our data into four different data sets:

- x_1: Observations for 2007 (154);
- x_2: Observations for 2008 (193);
- x_3: Observations for 2009 (212);

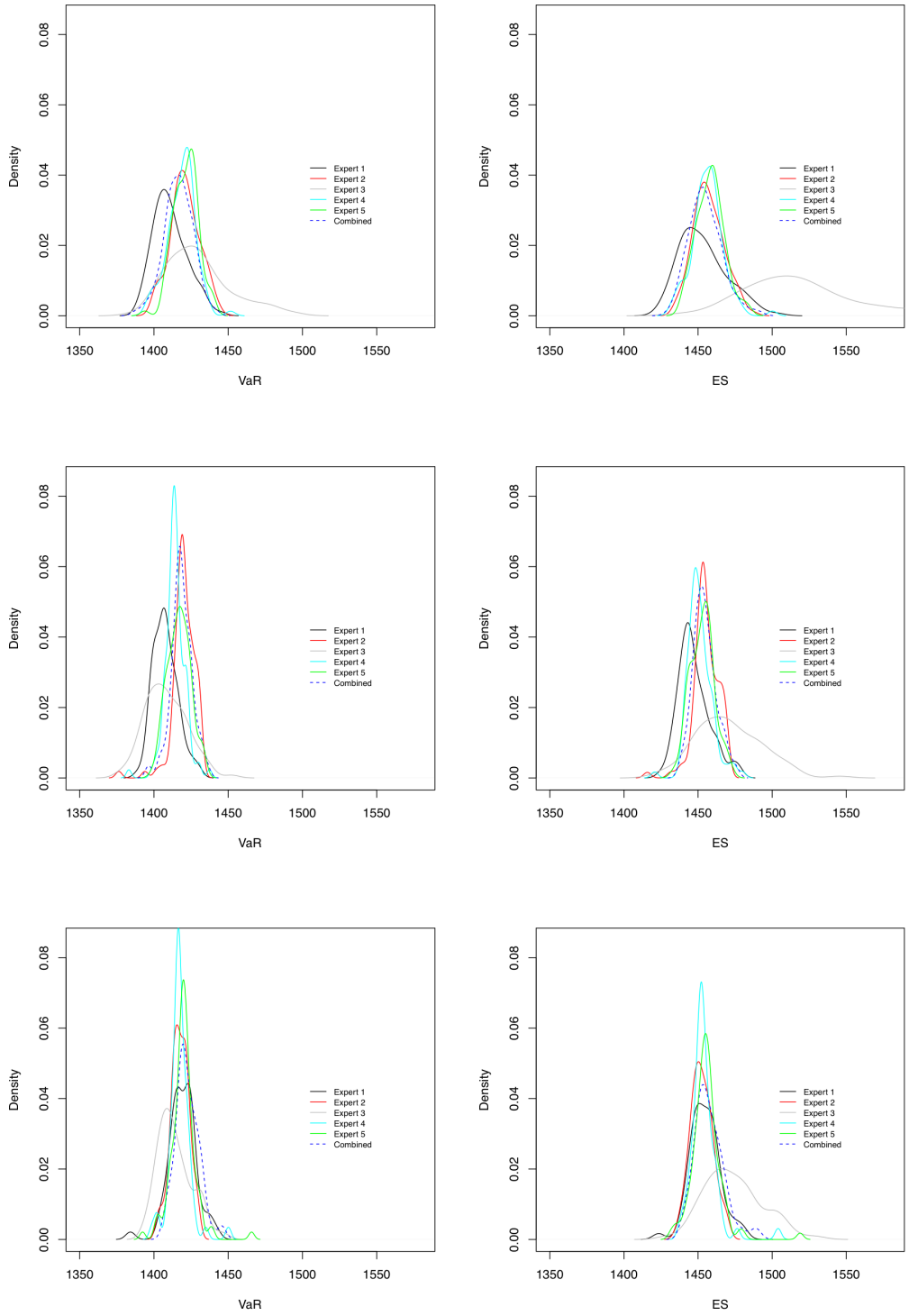

Figure 4.15: *Posterior distribution of* $\text{VaR}_{0.995}$ *and* $\text{ES}_{0.995}$ *for different experts in the first year (top), second year (middle) and third year (bottom). n=1000 simulated data from a Gamma(100,0.09), using the opinion of five experts (E1,E2,E3,E4,E5) with equal weights*

	Year 0	Year 1	Year 2	Year 3
Expert	Mean, Median, Std	Mean, Median, Std	Mean, Median, Std	Mean, Median, Std
E1	1411.990, 1411.600, 13.680	1410.843, 1409.158, 11.276	1407.612, 1406.914, 8.211	1419.249, 1419.307, 9.353
E2	1427.560, 1427.710, 9.410	1420.800, 1420.384, 9.77	1420.044, 1419.911, 7.914	1417.47, 1417.476, 6.057
E3	1416.900, 1415.200, 11.370	1429.660, 1426.730, 22.121	1408.145, 1406.73, 14.041	1414.015, 1411.566, 11.477
E4	1421.230, 1421.320, 12.290	1418.614, 1419.563, 9.414	1414.167, 1413.947, 6.911	1417.266, 1416.81, 7.299
E5	1427.770, 1427.410, 10.610	1421.409, 1421.879, 8.259	1416.877, 1416.652, 7.586	1422.44, 1420.701, 7.881
Comb.	1423.330, 1421.310, 11.030	1416.476, 1416.917, 9.658	1418.539, 1417.987, 7.229	1418.561, 1418.467, 8.679
True value	1418.134			

Table 4.27: *Posterior distribution for different experts in a 3- year period. n=1000 simulated data from a Gamma(100,0.09), using the opinion of five experts (E1,E2,E3,E4,E5) with equal weights*

	2007	2007-2008			2007-2009			2007-2010		
Expert	w_j^0	KL_j	ϵ_j	w_j^1	KL_j	ϵ_j	w_j^2	KL_j	ϵ_j	w_j^3
E1	0.2	2.141	0.214	0.167	0.029	0.029	0.201	0.031	0.031	0.214
E2	0.2	0.037	0.004	0.211	0.030	0.030	0.200	0.117	0.117	0.195
E3	0.2	0.493	0.049	0.202	0.009	0.009	0.205	0.117	0.117	0.195
E4	0.2	0.052	0.005	0.211	0.078	0.078	0.190	0.182	0.182	0.180
E5	0.2	0.127	0-013	0.209	0.0135	0.0135	0.204	0.023	0.023	0.216

Table 4.28: *Prior weights for different experts in a 3-year period. n=626 fraud losses in 41 banks, recorded from 01/2007 to 04/2010*

- x_4: Observations for 2010 (67).

We can observe the results for different years in Tables 4.28–4.29. Posterior distributions are shown in Figures 4.16–4.19. From the plots one may notice that, at the beginning, expert 4 is far from the rest. However, as more data become available during the next years, his distribution gradually resembles the combined distribution. Something similar happens to the other expert distributions. As expected, in the analysis corresponding to the year 2010, we can observe how the distribution of VaR and ES for all of the experts are more uniform than in previous years.

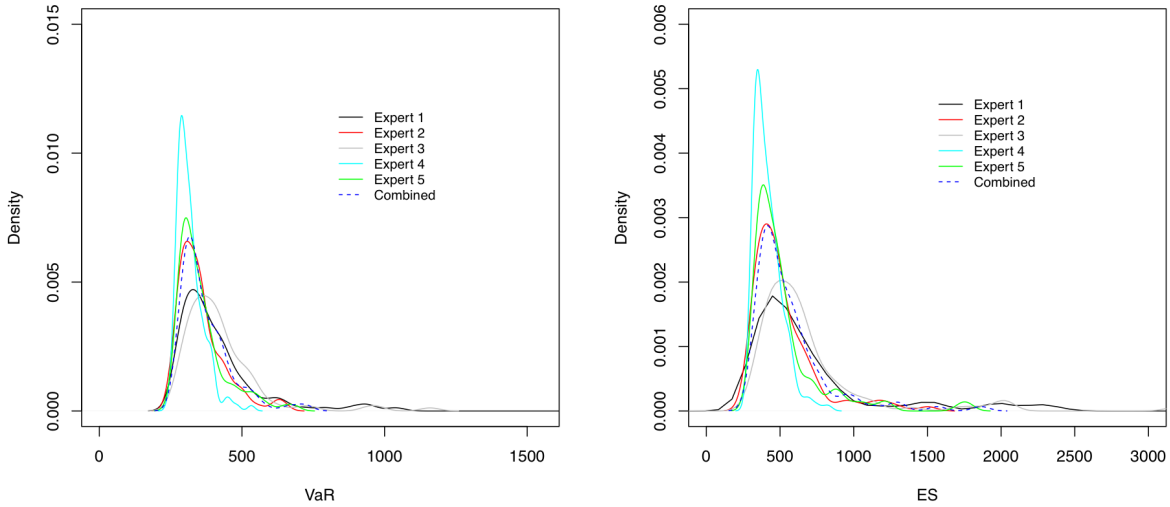

Figure 4.16: *Fraud data: Posterior distribution of* $\text{VaR}_{0.99}$ *and* $\text{ES}_{0.99}$ *for different experts in 2007*

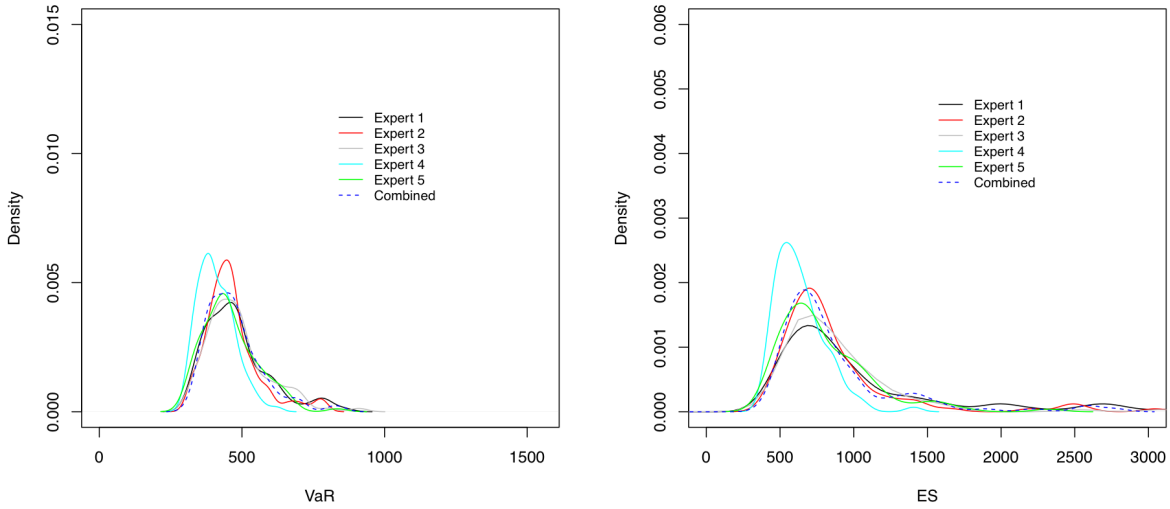

Figure 4.17: *Fraud data: Posterior distribution of* $\text{VaR}_{0.99}$ *and* $\text{ES}_{0.99}$ *for different experts in 2008*

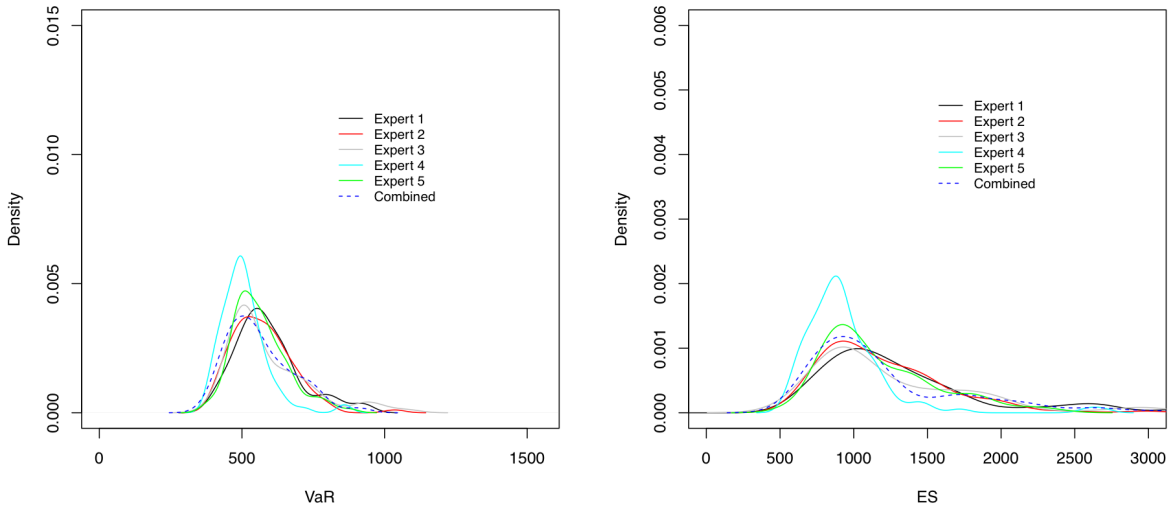

Figure 4.18: *Fraud data: Posterior distribution of* $\text{VaR}_{0.99}$ *and* $\text{ES}_{0.99}$ *for different experts in 2009*

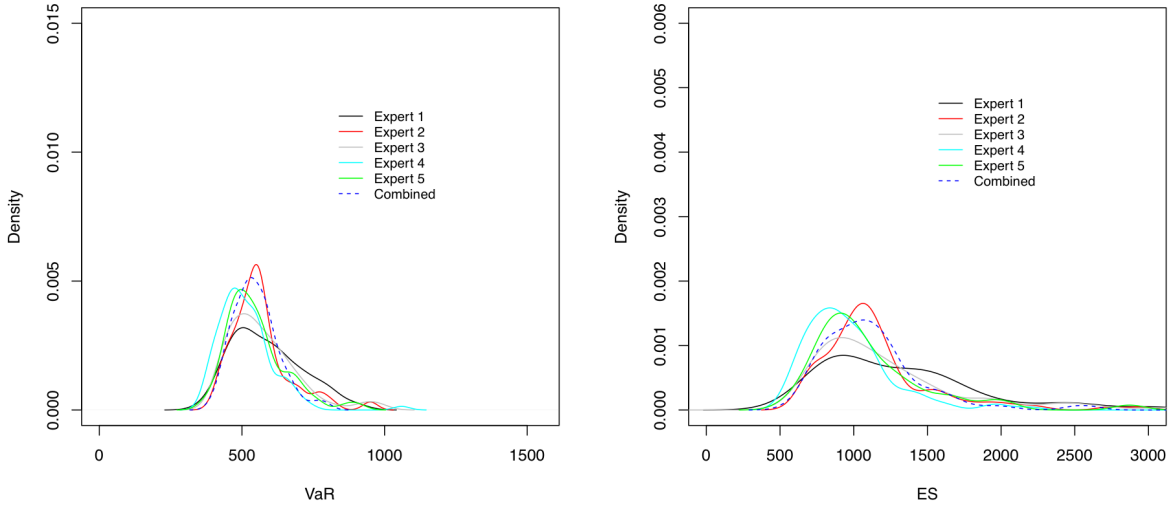

Figure 4.19: *Fraud data: Posterior distribution of* $\text{VaR}_{0.99}$ *and* $\text{ES}_{0.99}$ *for different experts in 2010*

	2007	2007-2008	2007-2009	2007-2010
Expert	Mean, Median, Std	Mean, Median, Std	Mean, Median, Std	Mean, Median, Std
E1	428.748, 364.665, 214.425	480.516, 462.743, 114.047	590.205, 567.618, 116.473	587.406, 569.175, 122.972
E2	355.307, 336.308, 81.129	464.560, 447.699, 95.808	570.429, 559.874, 104.638	564.206, 549.007, 104.701
E3	422.933, 395.480, 140.720	482.356, 469.746, 109.211	589.436, 537.179, 144.232	570.985, 541.690, 121.582
E4	318.049, 306.768, 46.8490	409.576, 396.502, 63.460	507.235, 496.151, 82.829	513.787, 497.401, 98.330
E5	356.672, 334.556, 84.574	459.366, 439.861, 96.358	562.389, 550.084, 94.883	548.625, 527.648, 104.754
Combined	371.900, 336.464, 89.656	471.559, 455.680, 98.280	562.042, 542.345, 114.736	547.475, 535.806, 78.192

Table 4.29: *Posterior distribution for different experts in a 3- year period. n=626 fraud losses in 41 banks, recorded from 01/2007 to 04/2010*

4.4 Conclusions of the Chapter

The conclusions drawn by the discussion of this chapter reveal the importance of considering expert opinion in the analysis of extremes. The first important finding in this chapter was that non-informative priors are not appropriate for modelling extreme data when few observations are available. To handle this issue, we have proposed two different ways of constructing a subjective prior, based on risk measures that experts are familiar with.

In all cases considered, we could observe how expert opinion influences the posterior behaviour and how the differences in these opinions may lead to different conclusions. Hence, the relevance of an appropriate elicitation process.

Additionally, in Section 4.3.4, we have provided the tools for prior analysis when opinions from multiple experts are available, particularly some methods to combine different distributions and assign weights to experts according to their past performance. This becomes particularly important in situations where data sets are limited, as it is typically the case of operational risk data. However, as we saw in the previous analysis, as more data become available, we have less uncertainty about the unknown parameters and our estimates seem to be more accurate. Moreover, the analysis is carried out using all the information available, either expert opinion or collected data.

Chapter 5

Extending the GPD mixture model

In Chapter 3 we presented a Bayesian model that takes into account observations above and below the threshold u. However, in that model the density has a discontinuity at the threshold. The jump can be larger or smaller depending on the parameters.

In this chapter, we introduce a new model that allows to handle the discontinuity issue in different ways. We consider different approaches according to the sort of discontinuity we are dealing with. We start by introducing a model where the threshold u plays the role of the qth quantile and that has a continuous density at that point. We also consider a second model where the threshold plays the role of the quantile of the overall distribution at which the Gamma tail is replaced by the GPD tail, not a fixed quantile for a pre-specified q. This time we require the derivative of the density to be continuous at the blend point. Next, we present a model with arbitrary scaling for the GPD density and, again, we require this model to have continuous derivative at the blend point. Lastly, we introduce a more general model where the scaling is arbitrary but that can be implemented even if the density and its derivative are discontinuous.

Finally, we explore a Bayesian nonparametric framework. We start by considering a model where the GPD is represented as a mixture of Exponentials and use a Dirichlet process mixture formulation that allows for a flexible density specification. We provide

the details for simulation from the DPM model using the Pólya urn scheme and MCMC sampling. After that, we introduce a second Bayesian nonparametric model that uses a Dirichlet process prior on the parameters of the GPD model. We also provide the sampling scheme and a possible extension of this model. We finally introduce a nonparametric version of the model with arbitrary scaling that can be implemented even if the density and its first derivative are discontinuous. As for the other models, we provide the details of its implementation.

It is worth pointing out that we have implemented most of these models for different data sets; however, in order to be brief, we only present the most interesting findings. Nonetheless, we provide the details of all the models as extra information. Another extension for mixture models is provided in the next chapter.

5.1 Blended Gamma-GPD model

Recall the blended model with Gamma and Generalized Pareto elements blended to be continuous at an upper quantile, q (where $q = 0.9, 0.99$ say), which is a special case of the general model introduced in Section 3.1. Suppose that u is defined such that $F_H(u;\alpha,\beta) = q$, with $F_H(x;\alpha,\beta)$ denoting the Gamma cdf, where for $x >0$,

$$F_H(x;\alpha,\beta) = \int_0^x f_H(t;\alpha,\beta)dt, \tag{5.1}$$

with

$$f_H(x;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}, \quad x > 0.$$

The blended cdf, $F_B(x;u,\sigma,\xi)$, takes the form

$$F_B(x;u,\sigma,\xi) = \begin{cases} F_H(x;\alpha,\beta), & x \le u, \\ F_H(u;\alpha,\beta) + (1 - F_H(u;\alpha,\beta))\, F_G(x;u,\sigma,\xi), & x > u, \end{cases} \tag{5.2}$$

where $F_G(x;u,\sigma,\xi)$ is the shifted generalized Pareto cdf viz.

$$F_G(x;u,\sigma,\xi) = 1 - \left(1 + \xi\frac{(x-u)}{\sigma}\right)^{-1/\xi}, \quad x > u$$

with density

$$f_G\left(x;u,\sigma,\xi\right)=\frac{1}{\sigma}\left(1+\frac{\xi\left(x-u\right)}{\sigma}\right)^{-\frac{1}{\xi}-1},$$

where the support is

$$\begin{cases} x>u & \text{if} \quad \xi\geq 0, \\ u\leq x\leq u-\sigma/\xi & \text{if} \quad \xi<0, \end{cases}$$

with parameters $u,\sigma>0$, and ξ. We may assume that $\xi\geq 0$ if we require the support of the blended distribution to be unbounded from above.

5.1.1 A model with continuous density at the qth quantile

Observe that the blended model satisfies

$$F_H(u;\alpha,\beta)=q. \tag{5.3}$$

In this model, we require the density to be continuous at $x=u$. The density is

$$f_B(x;u,\sigma,\xi)=\begin{cases} f_H(x;\alpha,\beta), & x\leq u, \\ \left(1-q\right)f_G\left(x;u,\sigma,\xi\right), & x>u. \end{cases} \tag{5.4}$$

and therefore we require

$$f_H(u;\alpha,\beta)=\left(1-q\right)f_G\left(u;u,\sigma,\xi\right)=\frac{1-q}{\sigma}. \tag{5.5}$$

This defines a second equality constraint to reduce the number of free parameters. We regard the GPD parameters (u,σ,ξ) as the working parameters, and thus solve equations (5.3) and (5.5) simultaneously for (α,β). This must be done numerically, and leaves (α,β) determined by q, u and σ.

5.1.2 A model with continuous first derivative at the blend point

In this version of the model, u plays the role of the quantile of the overall distribution at which the Gamma tail is replaced by the GPD tail; it is not a fixed qth quantile for a

pre-specified q. Again, we require the blended model to have continuous density at $x = u$. Up to proportionality, the density is

$$f_B(x; u, \sigma, \xi) \propto \begin{cases} f_H(x; \alpha, \beta), & x \leq u, \\ (1 - F_H(u; \alpha, \beta))\, f_G(x; u, \sigma, \xi), & x > u. \end{cases} \tag{5.6}$$

where α and β are such that

$$f_H(u; \alpha, \beta) = (1 - F_H(u; \alpha, \beta))\, f_G(u; u, \sigma, \xi) = \frac{1 - F_H(u; \alpha, \beta)}{\sigma}. \tag{5.7}$$

This defines one equality constraint to reduce the number of free parameters. For a further constraint, we also require the derivative of the density to be continuous at $x = u$. Up to proportionality, the derivative is

$$\dot{f}_B(x; u, \sigma, \xi) \propto \begin{cases} \dot{f}_H(x; \alpha, \beta), & x \leq u, \\ (1 - F_H(u; \alpha, \beta))\, \dot{f}_G(x; u, \sigma, \xi), & x > u. \end{cases} \tag{5.8}$$

Hence we require

$$\dot{f}_H(u; \alpha, \beta) = (1 - F_H(u; \alpha, \beta))\, \dot{f}_G(u; u, \sigma, \xi) = -\frac{(1 - F_H(u; \alpha, \beta))\,(1 + \xi)}{\sigma^2}. \tag{5.9}$$

Note that

$$\dot{f}_H(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \left[(\alpha - 1)\, x^{\alpha-2} e^{-\beta x} - \beta x^{\alpha-1} e^{-\beta x}\right] = \left[\frac{\alpha - 1}{x} - \beta\right] f_H(x; \alpha, \beta). \tag{5.10}$$

Because of (5.7), (5.9) thus reduces to

$$\left[\frac{\alpha - 1}{u} - \beta\right] = -\frac{(1 + \xi)}{\sigma},$$

yielding that

$$\beta = \frac{\alpha - 1}{u} + \frac{1 + \xi}{\sigma} = \beta(\alpha, u, \sigma, \xi),$$

say. Equation (5.7) then becomes

$$f_H(u; \alpha, \beta(\alpha, u, \sigma, \xi)) = \frac{1 - F_H(u; \alpha, \beta(\alpha, u, \sigma, \xi))}{\sigma}, \tag{5.11}$$

which defines the second equality constraint. We regard the parameters (u, σ, ξ) as the working parameters, and thus solve equation (5.11) for α. The proportionality constant is defined by requiring $f_B(x)$, defined by (5.6), to integrate to 1. This gives the normalizing constant

$$\frac{1}{F_H(u) + (1 - F_H(u))} = 1$$

so that

$$f_B(x; u, \sigma, \xi) = \begin{cases} f_H(x; \alpha, \beta), & x \leq u, \\ (1 - F_H(u; \alpha, \beta))\, f_G(x; u, \sigma, \xi), & x > u. \end{cases} \tag{5.12}$$

We note the necessary constraint $\sigma > u(1 + \xi)$ which guarantees that $\beta(\alpha, u, \sigma, \xi) > 0$ for any value of $\alpha > 0$.

5.1.3 A model with continuous first derivative with arbitrary scaling

In this version of the model, the density is

$$f_B(x; u, \sigma, \xi) \propto \begin{cases} f_H(x; \alpha, \beta), & x \leq u, \\ \omega\,(1 - F_H(u; \alpha, \beta))\, f_G(x; u, \sigma, \xi), & x > u, \end{cases} \tag{5.13}$$

for some $\omega > 0$ which is a further parameter of the model. To assure the continuity of f_B,

$$f_H(u; \alpha, \beta) = \omega\,(1 - F_H(u; \alpha, \beta))\, f_G(u; u, \sigma, \xi) = \frac{\omega\,(1 - F_H(u; \alpha, \beta))}{\sigma}. \tag{5.14}$$

This defines one equality constraint to reduce the number of free parameters. For a further constraint, we also require the derivative of the density to be continuous at $x = u$. Up to proportionality, the derivative is

$$\dot{f}_B(x; u, \sigma, \xi) \propto \begin{cases} \dot{f}_H(x; \alpha, \beta), & x \leq u, \\ \omega\,(1 - F_H(u; \alpha, \beta))\, \dot{f}_G(x; u, \sigma, \xi), & x > u. \end{cases} \tag{5.15}$$

As before, we have that

$$\beta = \frac{\alpha - 1}{u} + \frac{1 + \xi}{\sigma} = \beta\left(\alpha, u, \sigma, \xi\right),$$

say. Equation (5.14) then becomes

$$f_H\left(u; \alpha, \beta\left(\alpha, u, \xi, \sigma\right)\right) = \frac{\omega\left(1 - F_H(u; \alpha, \beta\left(\alpha, u, \sigma, \xi\right))\right)}{\sigma}, \tag{5.16}$$

which defines the second equality constraint. We regard the GPD parameters (u, σ, ξ) as the working parameters, and thus solve equation (5.16) for α. The proportionality constant is defined by requiring $f_B(x)$ to integrate to 1: with the density defined by (5.13), we have that the normalizing constant is

$$\frac{1}{F_H\left(u; \alpha, \beta\right) + \omega\left(1 - F_H(u; \alpha, \beta)\right)}$$

so that

$$f_B(x; u, \sigma, \xi) \propto \begin{cases} \dfrac{f_H(x; \alpha, \beta)}{F_H\left(u\right) + \omega\left(1 - F_H(u; \alpha, \beta)\right)}, & x \leq u, \\ \dfrac{\omega\left(1 - F_H(u; \alpha, \beta)\right) f_G\left(x; u, \sigma, \xi\right)}{F_H\left(u\right) + \omega\left(1 - F_H(u; \alpha, \beta)\right)}, & x > u, \end{cases} \tag{5.17}$$

again with the constraint $\sigma > u(1 + \xi)$ which guarantees that $\beta(\alpha, u, \sigma, \xi) > 0$ for any $\alpha > 0$. At $x = u$, we have that

$$F_B\left(u; u, \sigma, \xi\right) = \frac{F_H\left(u\right)}{F_H\left(u\right) + \omega\left(1 - F_H(u; \alpha, \beta)\right)}.$$

5.1.4 A model with discontinuous density with arbitrary scaling

The scaling from the model from Section 5.1.3 can be implemented even if the density and its derivative are discontinuous, as in the original version of the model. Suppose that f_B takes the form

$$f_B(x; u, \sigma, \xi) \propto \begin{cases} f_H(x; \theta), & x \leq u, \\ \omega\left(1 - F_H(u; \theta)\right) f_G\left(x; u, \sigma, \xi\right), & x > u, \end{cases} \tag{5.18}$$

where f_H is the density for the left-hand component (not necessarily Gamma), active when $x < u$, and for some $\omega > 0$. The proportionality constant is defined by requiring $f_B(x)$ defined in (5.18) to integrate to 1. The normalizing constant can be computed to be

$$\frac{1}{F_H(u;\theta) + \omega\left(1 - F_H(u;\theta)\right)}$$

so that

$$f_B(x;u,\theta,\sigma,\xi) \propto \begin{cases} \dfrac{f_H(x;\theta)}{F_H(u;\theta) + \omega\left(1 - F_H(u;\theta)\right)}, & x \leq u, \\ \dfrac{\omega\left(1 - F_H(u;\theta)\right) f_G(x;u,\sigma,\xi)}{F_H(u;\theta) + \omega\left(1 - F_H(u;\theta)\right)}, & x > u. \end{cases} \tag{5.19}$$

At $x = u$, we have that

$$F_B(u;u,\theta,\sigma,\xi) = \frac{F_H(u;\theta)}{F_H(u;\theta) + \omega\left(1 - F_H(u;\theta)\right)}.$$

From (5.19), it is apparent that this model can be thought of as resulting from a mixture representation

$$f_B(x) = \pi \tilde{f_H}(x) + (1-\pi) f_G(x),$$

where f_H has support $(0,u)$ and f_G has support (u,∞) with

$$\tilde{f_H}(x) = \frac{f_H(x;\theta)}{F_H(u;\theta)} \quad x \leq u$$

and

$$\pi = \frac{F_H(u;\theta)}{F_H(u;\theta) + \omega\left(1 - F_H(u;\theta)\right)}.$$

In the conventional version of this model, we fix ω =1.

5.1.5 Real data example

The four models proposed above were tested on different data sets. In this section, we illustrate the results for the fraud data described in Section 3.2.1 (Figure 5.1). We concentrate our attention on the model with continuous first derivative at the blend point (Section 5.1.2) and the model with discontinuous density with arbitrary scaling (Section 5.1.4), due to their performance. MCMC was used to sample the parameters from these

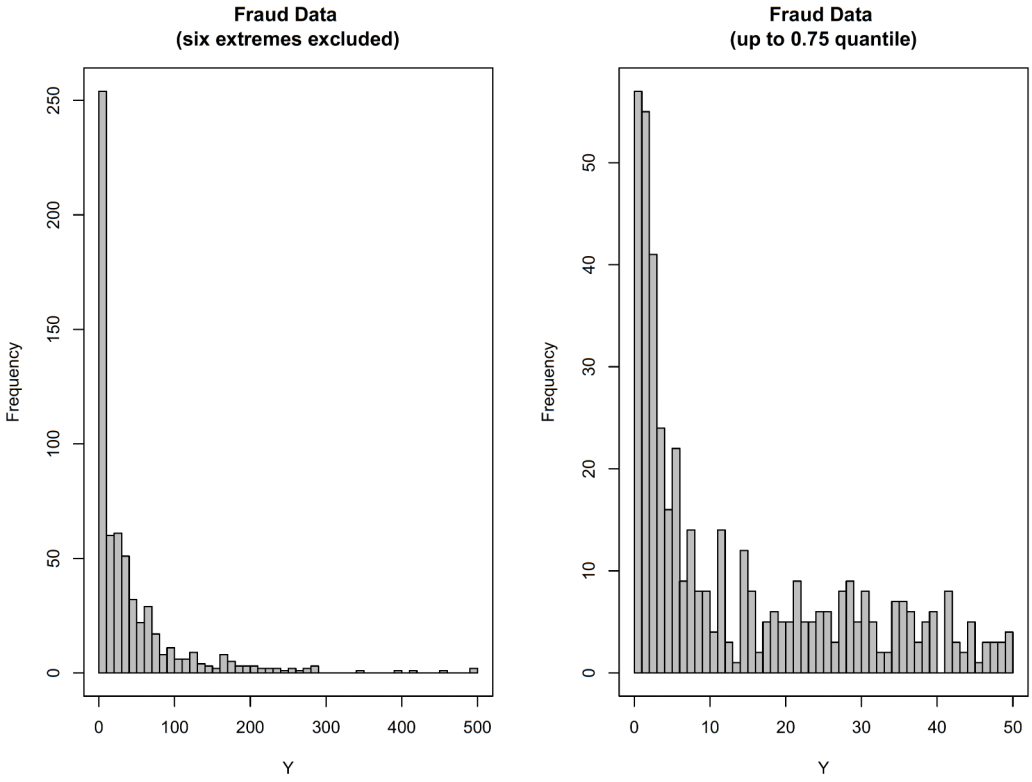

Figure 5.1: *Fraud data: Histograms of data with cut-offs at 500 (left) and 50 (right)*

models. Using uniform priors on all parameters, 2000 samples from the posterior for the parameters were obtained.

Figure 5.2 displays the fitted densities for the model from Section 5.1.2, while Figure 5.3 shows the corresponding posterior distributions. As we can see from the figures, the model has a good performance overall.

Now, for the model from Section 5.1.4, we can assume a Gamma model for f_H, and that

(a) ω is fixed to be 1. This is the usual model.

(b) ω is allowed to vary as a free parameter.

For model (a), with $\omega = 1$, and model (b), where ω is allowed to vary, we obtain the posteriors shown in Figures 5.4–5.6.

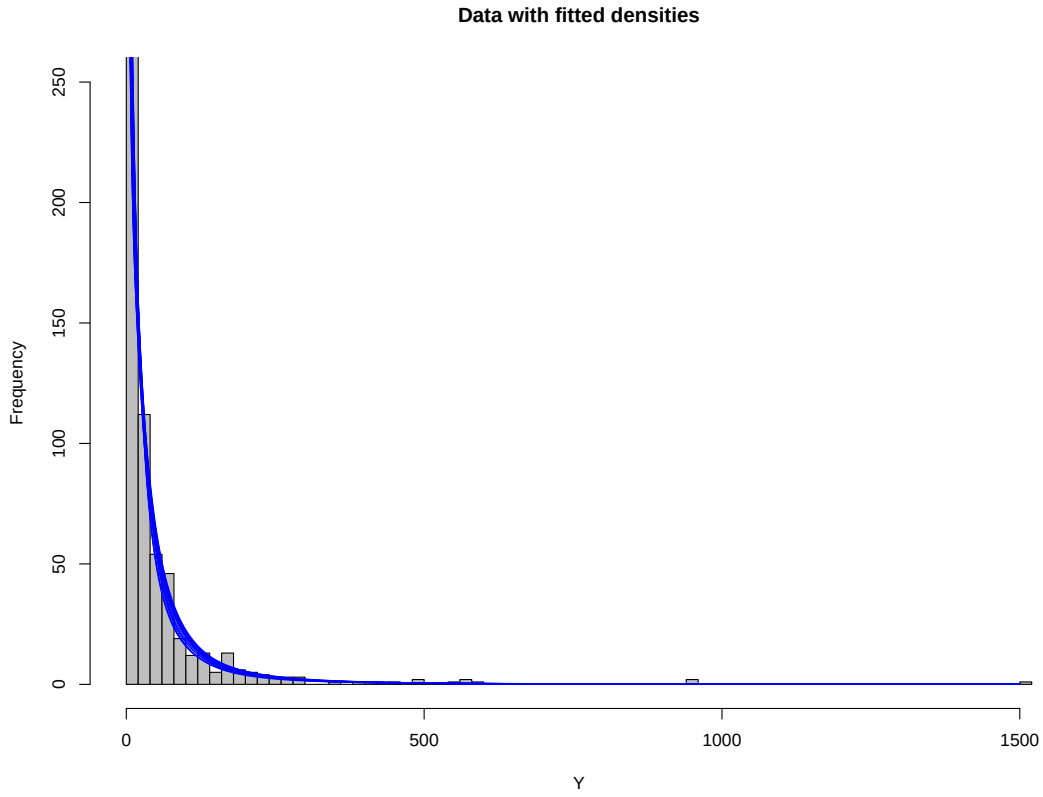

Figure 5.2: *Fraud data: Fitted densities for the model with continuous first derivative at the blend point*

For these parameters, the results are broadly similar; the most different are the posterior for β in the Gamma model, and the posterior for u. In model (b), the posterior for ω is as follows: It is clear from Figure 5.6 that $\omega > 1$ is heavily supported in the analysis. Figure 5.7 uses an estimated distribution function formed from the Bayesian analysis to carry out a goodness of fit assessment using a P-P plot; it seems that the model where ω varies is preferred.

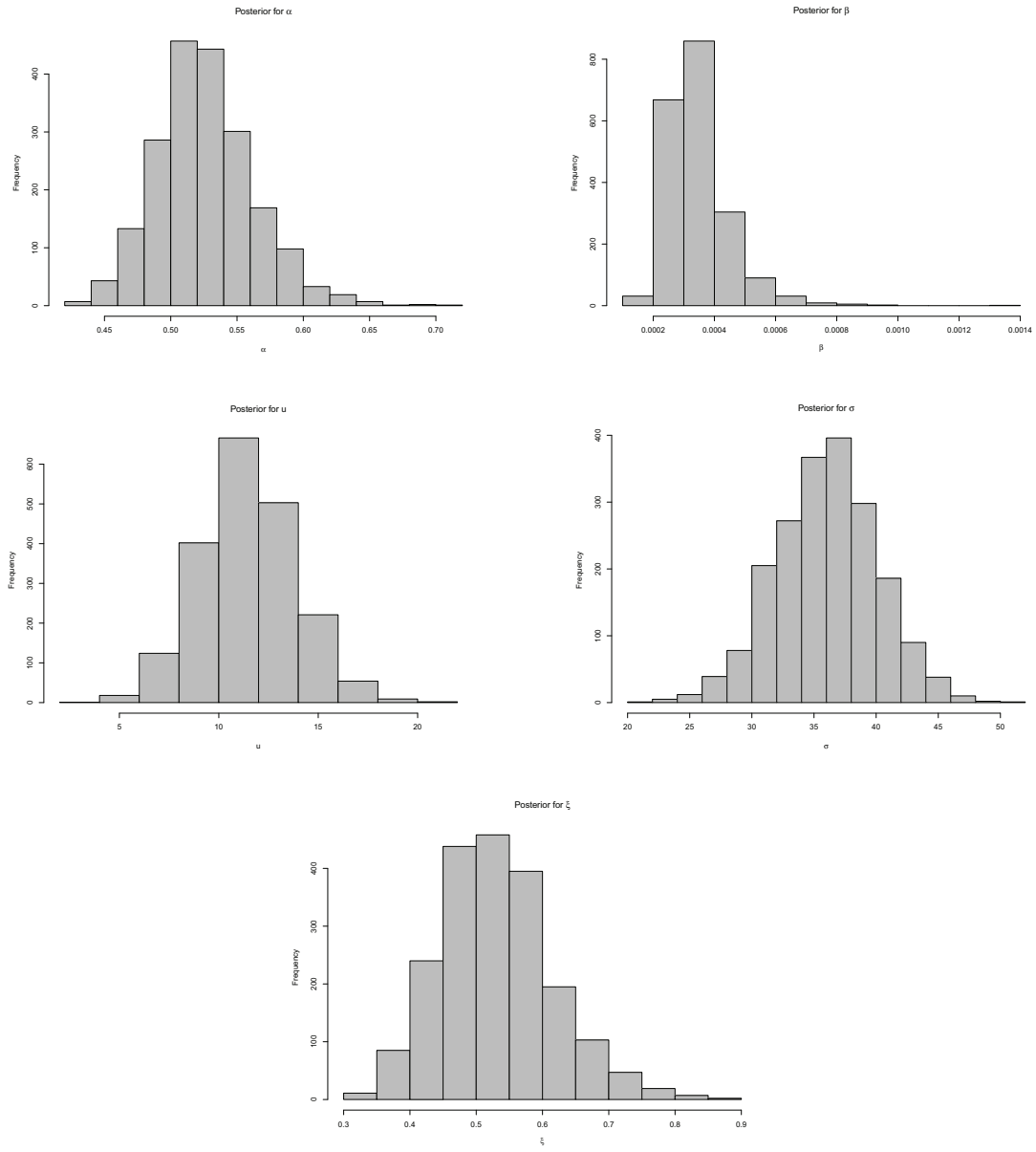

Figure 5.3: *Fraud data: Posterior histograms for the model with continuous first derivative at the blend point. Top: α and β. Middle: u and σ Bottom: ξ*

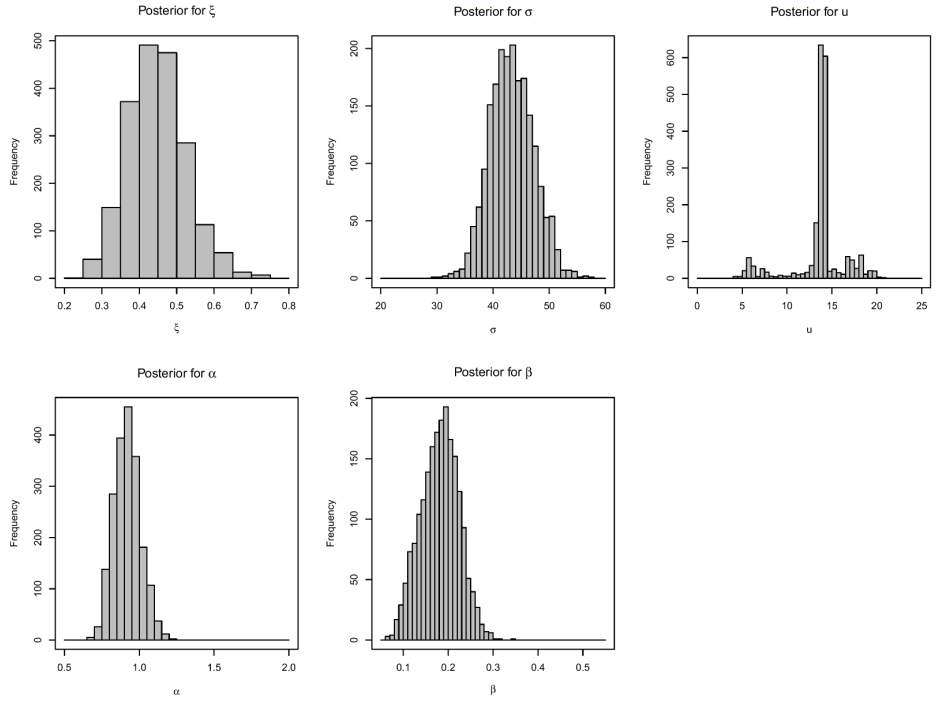

Figure 5.4: *Fraud data: Posterior histograms for* $\omega = 1$

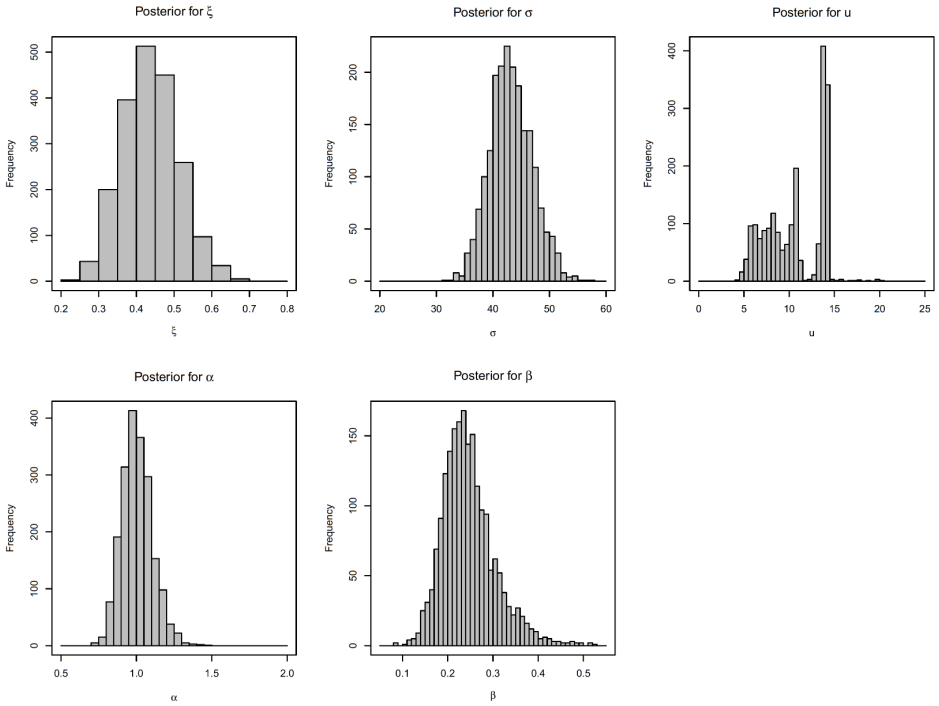

Figure 5.5: *Fraud data: Posterior histograms for* ω *varying*

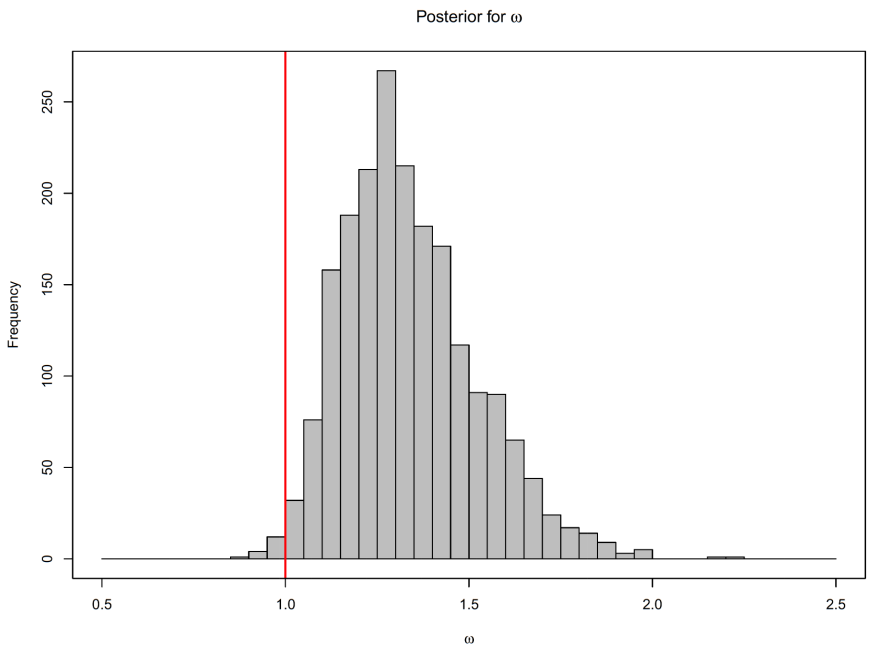

Figure 5.6: *Fraud data: Posterior histogram for ω in model (b) based on f_B*

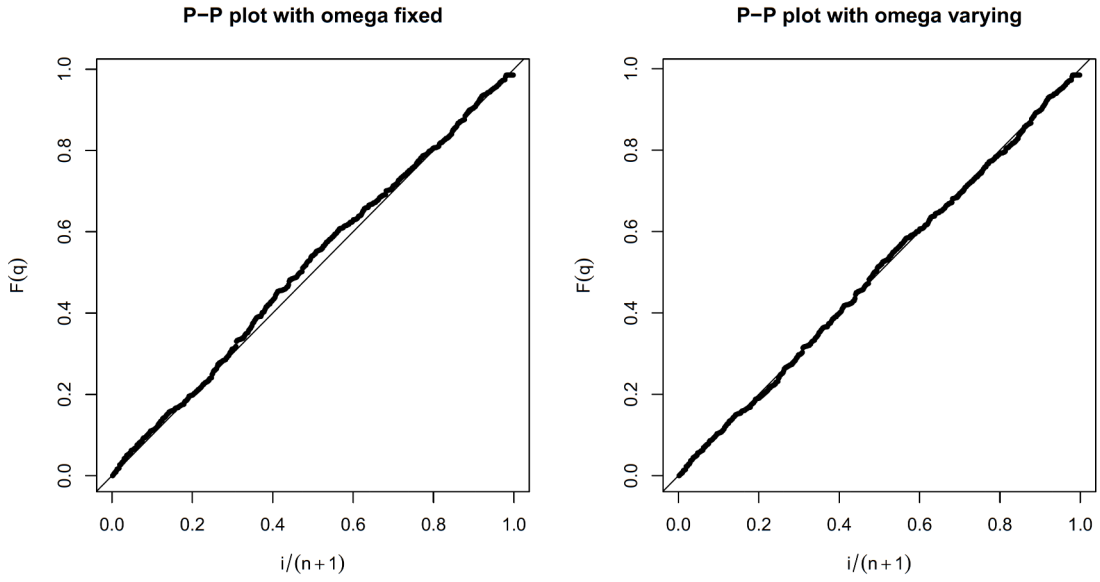

Figure 5.7: *P-P plots for data using Bayesian estimates from fitted models. Left panel is model (a) ($\omega = 1$), right panel is model (b) (ω varying)*

5.2 A Bayesian nonparametric model

It is well-known that the GPD model can be represented as a Gamma mixture of Exponential distributions. Recall that if

$$f_{X|\theta}(x \mid \theta) \equiv \text{Exponential}(\theta) \quad x > 0$$
$$f_{\theta}(\theta \mid \gamma, \beta) \equiv \text{Gamma}(\gamma, \beta)$$

then for $x > 0$

$$\begin{aligned} f_X(x|\gamma,\beta) &= \int_0^\infty \theta e^{-\theta x} \frac{\beta^\gamma}{\Gamma(\gamma)} \theta^{\gamma-1} e^{-\beta\theta}\, d\theta \\ &= \frac{\beta^\gamma}{\Gamma(\gamma)} \int_0^\infty \theta^{(\gamma+1)-1} e^{-(\beta+x)\theta}\, d\theta \\ &= \frac{\beta^\gamma}{\Gamma(\gamma)} \frac{\Gamma(\gamma+1)}{(\beta+x)^{\gamma+1}} \\ &= \frac{\gamma}{\beta} \frac{1}{(1+x/\beta)^{\gamma+1}} \end{aligned}$$

so setting $\gamma = 1/\xi$ and $\beta = \sigma/\xi$ yields the GPD(σ, ξ). Adding the fixed location shift parameter u is straightforward, this will be discussed later.

This suggests a natural Bayesian nonparametric model based on a Dirichlet process mixture formulation. Consider a random sample $X_1, ..., X_n$ where $X_i \sim$ Exponential(θ_i), $i = 1, ..., n$ and

$$\theta_1, ..., \theta_n \overset{i.i.d}{\sim} F_\theta,$$

where F_θ is constructed as follows. Let DP(ν, G_0) be a Dirichlet process with parameter ν and base distribution G_0, taken to be the Gamma distribution with parameters$(1/\xi, \sigma/\xi)$. The parameter ν is a precision hyperparameter. The CDF F_θ is then a random (almost surely discrete) CDF with prior specifying that for any set B,

$$E[F_\theta(B)] = G_0(B), \qquad V[F_\theta(B)] = \frac{G_0(B)(1 - G_0(B))}{\nu + 1}.$$

As $\nu \to \infty$, the unconditional distribution of X becomes the GPD(σ, ξ) distribution. For finite ν however, the Bayesian nonparametric specification imparts additional flexibility.

This model is termed a Dirichlet process mixture (DPM) model; it allows for a flexible density specification.

The parameter ν determines the degree of "clustering" in the Dirichlet Process model, that is, the number of distinct values in the collection $\theta_1, ..., \theta_n$. If K is the (random) number of distinct values, recalling the result of Antoniak (1974), we have that

$$p(K \mid n, \nu) = \frac{S_{n,K} n! \nu^K \Gamma(\nu)}{\Gamma(\nu + n)}, \quad K = 1, ..., n$$

where $S_{n,k}$ denotes the Stirling number of the first kind. It holds that (Teh, 2010)

$$E[K \mid n, \nu] = \nu (\psi(\nu + n) - \psi(\nu)),$$

$$V[K \mid n, \nu] = \nu (\psi(\nu + n) - \psi(\nu)) + \nu^2 (\psi'(\nu + n) - \psi'(\nu)),$$

where $\psi(\cdot)$ and $\psi'(\cdot)$ are the digamma and trigamma functions, respectively.

5.2.1 Simulation from the DPM model

To simulate from the DPM model (Mahmoud, 2008), we first must specify how to simulate the sample $\theta_1, ..., \theta_n$; this is achieved via the Pólya urn scheme operating on the implicit clusters that form part of the Dirichlet process specification. The algorithm proceeds as follows:

Step 1. Set $i = 1$; set $c_1 = 1$ as the first cluster label.

Step 2. For $i = 2, 3, ..., n$: let K_{i-1} denote the number of distinct values in $c_1, ..., c_{i-1}$, and let $n_1(i-1), ..., n_{K_{i-1}}(i-1)$ denote the counts of each of the distinct cluster labels. Then generate

$$c_i \mid c_1, ..., c_{i-1} \sim \frac{\nu}{\nu + i - 1} \delta_{K_{i-1}+1} + \frac{1}{\nu + i - 1} \sum_{k=1}^{K_{i-1}} n_k (i-1) \delta_k,$$

that is, with probability $\nu / (\nu + i - 1)$ generate a new cluster label $K_{i-1} + 1$, but with probability

$$\frac{n_k (i-1)}{\nu + i - 1},$$

set $c_i = k$. When $i = n$, the values $c_1, ..., c_n$ form K_n clusters of labels $1, 2, ..., K_n$, with $n_1(n), ..., n_{K_n}(n)$ in each cluster respectively.

Step 3. Let $K \equiv K_n$ denote the number of distinct values in $c_1, ..., c_n$. For $k = 1, ..., K$, generate $\vartheta_k \sim G_0$ independently, and for $i = 1, ..., n$ set

$$\theta_i = \sum_{k=1}^{K} \vartheta_k 1_k (c_i) .$$

The $\vartheta_1, ..., \vartheta_K$ values represent the cluster "locations" for the K clusters to which the n data belong. The final step involves generating $X_i \mid \theta_i$.

Step 4. Simulate $X_i \mid \theta_i \sim$ Exponential(θ_i).

5.2.2 Posterior inference under the DPM model

Given data $x_1, \ldots, x_n$ that are assumed to arise from a data generating process to be modelled using a DPM, we seek a MCMC sampling scheme to perform posterior inference. This can be achieved using a standard strategy that samples in turn

(i) the parameters $\theta_1, \ldots, \theta_n$;

(ii) the GPD parameters σ, ξ;

(iii) the Dirichlet process precision parameter ν;

from their full conditional distributions.

5.2.2.1 Sampling the θ parameters

In this step, we define the sampling weights as follows.

$$f_X (x \mid \sigma, \xi) = \int_0^\infty f_{X|\theta} (x \mid \theta) f_\theta (\theta \mid \sigma, \xi) \, d\theta \tag{5.20}$$

is the marginal density for X after integrating out over the mixing distribution, and $p_i (\theta \mid x_i, \sigma, \xi)$ is the posterior density derived from datum x_i, formed as

$$p_i (\theta \mid x_i, \sigma, \xi) \propto f_{X|\theta} (x_i \mid \theta) \pi_\theta (\theta \mid \sigma, \xi) . \tag{5.21}$$

In the conjugate formulation we have utilized, both equations (5.20) and (5.21) can be computed analytically: we have that

$$f_X\left(x \mid \sigma, \xi\right) \equiv \mathrm{GPD}\left(x \mid \sigma, \xi\right), \tag{5.22}$$

as in the initial calculation, and

$$p_i(\theta|x_i, \xi, \sigma) \propto \theta e^{-\theta x_i} \theta^{1/\xi - 1} e^{-\sigma\theta/\xi} = \theta^{1/\xi + 1 - 1} e^{-(\sigma/\xi + x_i)\theta}, \tag{5.23}$$

so that

$$p_i(\theta|x_i, \xi, \sigma) \equiv \mathrm{Gamma}(1/\xi + 1, \sigma/\xi + x_i). \tag{5.24}$$

Next, using a Pólya urn scheme that samples the parameter for datum i from its full conditional posterior distribution we have that

$$\theta_i \mid x_i, \theta_{(-i)}, \sigma, \xi, \nu \sim w_0 p_i\left(\theta \mid x_i, \sigma, \xi\right) + \sum_{j \neq i} w_j \delta_{\theta_j}, \tag{5.25}$$

where

$$w_0 = \frac{\nu f_X\left(x_i \mid \sigma, \xi\right)}{\nu f_X\left(x_i \mid \sigma, \xi\right) + \sum_{j \neq i} f_{X|\theta}\left(x_i \mid \theta_j\right)}, \qquad w_j = \frac{f_{X|\theta}\left(x_i \mid \sigma, \xi\right)}{\nu f_X\left(x_i \mid \sigma, \xi\right) + \sum_{j \neq i} f_{X|\theta}\left(x_i \mid \theta_j\right)}, \quad j \neq i.$$

In summary, for θ_i, we first sample from the set $\{0, 1, 2, ..., i-1, i+1, ..., n\}$ with probabilities $\{w_0, w_1, w_2, \ldots, w_{i-1}, w_{i+1}, \ldots, w_n\}$; if the value 0 is obtained, we sample from θ_i from (5.24), whereas if index j is sampled, we set θ_i equal to θ_j.

In this scheme, the clustering of the θ values is not utilized in the sampling, although it is of course present; obtaining index 0 and sampling from (5.24) generates a completely new θ and a new cluster, but otherwise an existing value of θ is used. As before, we let K denote the number of distinct values of θ, $\vartheta_1, ..., \vartheta_K$ denote these distinct values, and $c_1, ..., c_n$ denote the cluster labels denoting which datum belongs to which cluster.

5.2.2.2 Sampling the GPD parameters

The full conditional for (σ, ξ) given all other parameters is proportional to

$$\left\{\prod_{k=1}^{K} f_\theta\left(\vartheta_k \mid \sigma, \xi\right)\right\} \pi\left(\sigma, \xi\right). \tag{5.26}$$

where $f_\theta(\cdot \mid \sigma, \xi) \equiv \text{Gamma}(1/\xi, \sigma/\xi)$, that is, proportional to the likelihood of the **distinct** θ values multiplied by the prior $\pi(\sigma, \xi)$. The distribution cannot be sampled directly, so Metropolis-Hastings updates must be used.

5.2.2.3 Sampling the DP precision parameter

The full conditional for ν given $\vartheta_1, ..., \vartheta_K$ can be updated using an auxiliary variable method designed by Escobar and West (1995). Suppose that the prior distribution for ν is Gamma(a, b). To update ν, we introduce η with conditional distribution

$$\eta \mid \nu \sim \text{Beta}(\nu + 1, n), \tag{5.27}$$

and then update ν as

$$\nu \mid \eta, K \sim \begin{cases} \text{Gamma}(a + K, b - \log(\eta)) & \text{with prob. } \dfrac{a + K - 1}{a + K - 1 + n(b - \log(\eta))}, \\ \text{Gamma}(a + K - 1, b - \log(\eta)) & \text{with prob. } \dfrac{n(b - \log(\eta))}{a + K - 1 + n(b - \log(\eta))}. \end{cases} \tag{5.28}$$

5.2.3 Introducing the offset u

If the GPD model pertains beyond a threshold u, then the model can be readily extended for the right tail model. We have $X_1, ..., X_n$ independent where for $X_i > u$ we have

$$X_i \mid u, \theta_i \sim \text{Shifted exponential}(\theta_i, u)$$

with density $\theta_i e^{-\theta_i(x-u)}$, $x > u$. As before, we assume a Gamma$(1/\xi, \sigma/\xi)$ model for θ. For $X_i \leq u$, we might assume that $X_i \sim \text{Gamma}(\alpha, \beta)$, independent of θ_i. That is, we are appealing to the continuous mixture representation

$$f_X(x \mid u) = \int_0^\infty f_{X|\theta}(x \mid u, \theta)\, \pi(\theta)\, d\theta \equiv \begin{cases} \text{Gamma}(\alpha, \beta), & x \leq u, \\ \text{GPD}(u, \sigma, \xi), & x > u. \end{cases} \tag{5.29}$$

For a Bayesian nonparametric specification, we again assume

$$\theta_1, ..., \theta_n \overset{i.i.d.}{\sim} F_\theta, \quad F_\theta \sim DP\left(\nu, G_0\left(\sigma, \xi\right)\right). \tag{5.30}$$

where $G_0\left(\sigma, \xi\right)$ is a GPD base distribution. The generative model is therefore represented by the following simulation scheme:

1. Simulate $\theta_1, ..., \theta_n$ from the DP model using the Pólya urn scheme.

2. For $i = 1, ..., n$, generate $z_i \sim \text{Uniform}(0, 1)$.

   a) If $z_i \leq F_G(u; \alpha, \beta)$, then simulate X_i from the Gamma(α, β) distribution truncated at u; this may be done by cdf inversion or by rejection sampling.

   b) If $z_i > F_G(u; \alpha, \beta)$, then simulate $V_i \sim \text{Exponential}(\theta_i)$, and set $X_i = V_i + u$.

For inference, the previous algorithm can be implemented with amendments. The most notable changes occur in the sampling of the θ_is from $\theta_i \mid x_i, \theta_{(-i)}$ in the Pólya urn scheme. The two cases need separate consideration:

- $x_i \leq u$: we have that the sampling weights are defined as proportional to

  $$\nu f_G\left(x_i \mid \alpha, \beta, u\right)$$

  for w_0, where $f_G(\cdot \mid \alpha, \beta, u)$ is the Gamma(α, β) density truncated at u, and w_j is proportional to $f_G(x_i \mid \alpha, \beta, u)$ for each $j \neq i$. Therefore the sum of the sampling weights is $\left(\nu + n - 1\right) f_G(x_i \mid \alpha, \beta, u)$. The posterior measure $p_i\left(\theta \mid x_i, u, \sigma, \xi\right)$ is given by

  $$p_i\left(\theta \mid x_i\right) \propto f_X\left(x_i \mid u\right) \pi_\theta\left(\theta \mid \sigma, \xi\right) \propto \pi_\theta\left(\theta \mid \sigma, \xi\right). \tag{5.31}$$

  That is, we may sample θ_i in this case using a prior Pólya urn scheme

  – with probability $\nu / \left(\nu + n - 1\right)$, sample $\theta_i \sim \text{Gamma}\left(1/\xi, \sigma/\xi\right)$;

  – with probability $w_j = 1 / \left(\nu + n - 1\right)$, $j \in \{1, 2, ..., i-1, i+1, ..., n\}$ set $\theta_i = \theta_j$;

- $x_i > u$: in this case, the scheme reverts to the one from Section 5.2.2.1.

Conditional on $\theta_1, ..., \theta_n$, the sampling of the Generalized Pareto parameters (σ, ξ) and precision parameter ν proceeds as before. To sample the Gamma parameters (α, β) from their full conditional distribution given the data and u, the likelihood

$$\left\{ \prod_{i\,:\,x_i \leq u} f_G(x_i; \alpha, \beta) \right\} \{1 - F_G(u; \alpha, \beta)\}^{n-n_1} \tag{5.32}$$

is used, in conjunction with a suitable prior. Finally, to sample the parameter u from its full conditional density given the other parameters, we utilize the likelihood

$$\left\{ \prod_{i\,:\,x_i \leq u} f_G(x_i; \alpha, \beta) \right\} \left\{ \prod_{i\,:\,x_i > u} \theta_i \exp\{-\theta_i (x_i - u)\} \right\}. \tag{5.33}$$

5.3 A second Bayesian nonparametric model

A second Bayesian nonparametric approach uses a Dirichlet process prior on the parameters of the GPD model. We assume that for a fixed threshold u,

$$f_X(x|\varphi, u) = \iint_{\Omega} f_X\left(x \mid u, \sigma, \xi\right) F(d\xi, d\sigma), \tag{5.34}$$

where $\Omega = (-1/2, \infty) \times \mathbb{R}^+$, and

$$F \sim \mathrm{DP}\left(\nu, \mathrm{G}_0\left(\varphi\right)\right)$$

and $G_0\left(\varphi\right)$ is some base measure with hyperparameters φ and ν is the precision parameter. For example, we might choose that G_0 is a product of two Gamma distributions with parameters (α_ξ, β_ξ) and $(\alpha_\sigma, \beta_\sigma)$ respectively, with the prior for ξ relocated to $(-1/2, \infty)$. The parametric analysis is recovered when $\nu \to \infty$. An equivalent analysis is recovered when the scale mixture of Exponentials is used, that is,

$$f_X(x|\varphi, u) = \iint_{\Omega} \left\{ \int_0^\infty f_{X|\theta}(x|\theta, u) \pi(\theta|\sigma, \xi) d\theta \right\} F(d\xi, d\sigma), \tag{5.35}$$

where $f_{X|\theta}\left(x \mid \theta, u\right)$ is Exponential (θ) shifted by u, and $\pi\left(\theta \mid \sigma, \xi\right)$ is Gamma $(1/\xi, \sigma/\xi)$. In the initial formulation, we consider the case $u = 0$.

5.3.1 Priors

Reasonable choices for the hyperparameters seem to be

$$(\alpha_\xi, \beta_\xi) = (6, 5)\,, \quad (\alpha_\sigma, \beta_\sigma) = (3, 0.1)\,.$$

Note that the Jeffreys' prior for ξ (see Section 4.1), given by

$$\frac{1}{\pi}\frac{1}{(1+\xi)\sqrt{1+2\xi}}, \quad -\frac{1}{2} < \xi < \infty \tag{5.36}$$

is reasonably well approximated by Gamma$(1/4, 1/4)$ relocated to the range $(-1/2, \infty)$.

5.3.2 Inference

Given data $x_1, ..., x_n$ assumed to arise from a data generating process to be modelled using a DPM, we seek a MCMC sampling scheme to perform posterior inference. This can be achieved using the strategy from Section 5.2.2. We sample

(i) the parameters $(\sigma_1, \xi_1), ..., (\sigma_n, \xi_n)$,

(ii) the hyperparameters $(\alpha_\sigma, \beta_\sigma)$ and (α_ξ, β_ξ),

(iii) the Dirichlet process precision parameter ν,

from their full conditional distributions. We use the notation $\zeta_i = (\xi_i, \sigma_i)$.

5.3.2.1 Sampling the ζ parameters

This step is achieved using a Pólya urn scheme that samples the pair of parameters for datum i from its full conditional posterior distribution

$$\zeta_i \mid x_i, \zeta_{(-i)} \sim w_0 p_i\left(\zeta \mid x_i\right) + \sum_{j \neq i} w_j \delta_{\zeta_j}, \tag{5.37}$$

suppressing the dependence on the hyperparameters, where

$$w_0 = \frac{\nu f_X\left(x_i\right)}{\nu f_X\left(x_i\right) + \sum_{j \neq i} f_{X|\zeta}\left(x_i \mid \zeta_j\right)} \quad \text{and} \quad w_j = \frac{f_{X|\zeta}\left(x_i \mid \zeta_j\right)}{\nu f_X\left(x_i\right) + \sum_{j \neq i} f_{X|\zeta}\left(x_i \mid \zeta_j\right)}, \quad j \neq i \tag{5.38}$$

define the sampling weights, where

$$f_X(x) = \int_{-1/2}^{\infty} \int_0^{\infty} f_{X|\zeta}(x \mid \zeta)\, \pi_\zeta(\zeta)\, d\zeta \tag{5.39}$$

is the marginal density for X after integrating out over the mixing distribution, and $pi_i(\zeta \mid x_i)$ is the posterior density derived from datum x_i, formed as

$$p_i(\zeta \mid x_i) \propto f_{X|\zeta}(x_i \mid \zeta)\, \pi_\zeta(\zeta). \tag{5.40}$$

Here, equations (5.39) and (5.40) cannot be computed analytically. For (5.39) we can use numerical integration effectively, as this is merely a bivariate integral. This can be achieved using quadrature, or using Monte Carlo, using the following 'Rao-Blackwellized' estimation strategy:

- Sample $\zeta_l = (\sigma_l, \xi_l),\ l = 1, \ldots, L$ independently from $\pi_\zeta(\cdot)$.
- Compute the density estimate

  $$\widehat{f}_X(x) = \frac{1}{L} \sum_{l=1}^{L} f_{X|\zeta}(x|\zeta_l)$$

  on a fixed, fine grid of x values, to form a look-up table.
- For any desired x_i, use interpolation from the look-up table.

For (5.40), we must use MCMC sampling.

As before, for ζ_i, we first sample from the set $\{0, 1, 2, ..., i-1, i+1, ..., n\}$ with probabilities given by $w_0, w_1, w_2, ..., w_{i-1}, w_{i+1}, ..., w_n$; if the value 0 is obtained, we sample ζ_i from (5.40) using MCMC, whereas if index $j > 0$ is sampled, we set ζ_i, equal to ζ_j .

As before, the clustering of the ζ values is not utilized in the sampling, although it is of course present; obtaining index 0 and sampling from (5.40) generates a completely new ζ and a new cluster, but otherwise an existing value of ζ is used. As before, we let K denote the number of distinct values of ζ. Denote these distinct values by $\vartheta_1, ..., \vartheta_K$, and denote the cluster labels by $c_1, ..., c_n$, indicating which datum belongs to each cluster. Let $\vartheta_k \subset (\sigma_k^*, \xi_k^*)$.

5.3.2.2 Sampling the hyperparameters

If these parameters are to be updated (which is not really necessary or advisable) we have from our formulation that $(\alpha_\sigma, \beta_\sigma)$ and (α_ξ, β_ξ) are conditionally independent given all other parameters. The full conditional for (α_ξ, β_ξ) is proportional to

$$\left\{\prod_{k=1}^{K} f_\xi(\xi_k^*|\alpha_\xi, \beta_\xi)\right\} \pi(\alpha_\xi, \beta_\xi), \tag{5.41}$$

where $f_\xi(\cdot|\alpha_\xi, \beta_\xi)$ is Gamma(α_ξ, β_ξ) that is, proportional to the likelihood of the **distinct** values ξ^* multiplied by the prior for these parameters, $\pi(\alpha_\xi, \beta_\xi)$. The distribution cannot be sampled directly, so Metropolis-Hastings updates must be used.

5.3.3 Extending the second Bayesian nonparametric model

For a fixed $u > 0$ the formulation proceeds as before. We have

$$f_X(x|\varphi, u) = \iint_\Omega f_X(x|\xi, \sigma, u) F(d\xi, d\sigma),$$

where $\Omega = (-1/2, \infty) \times \mathbb{R}^+$, and $F \sim \mathrm{DP}(\alpha, \mathrm{G}_0(\varphi))$ and $G_0(\varphi)$ is some base measure with hyperparameters φ. The support of this density is clearly (u, ∞).

Most of the details go through as before for implementing the DP components, but we additionally condition on u. To sample the ζs, we again consider a Pólya urn scheme that samples the pair of parameters for datum i from its full conditional posterior distribution

$$\zeta_i|x_i, \zeta_{(-i)}, u \sim w_0 p_i(\zeta|x_i, u) + \sum_{j \neq i} w_j \delta_{\zeta_j},$$

suppressing the dependence on the hyperparameters, where

$$w_0 = \frac{\nu f_X(x_i|u)}{\nu f_X(x_i|u) + \sum_{l \neq i} f_{X|\zeta}(x_i|\zeta_l, u)} \qquad \text{and} \qquad w_j = \frac{f_{X|\zeta}(x_i|\zeta_j, u)}{\nu f_X(x_i|u) + \sum_{l \neq i} f_{X|\zeta}(x_i|\zeta_l, u)}, \ j \neq i$$

define the sampling weights, where

$$f_X(x|u) = \int_{-1/2}^{\infty} \int_0^\infty f_{X|\zeta}(x|\zeta, u) \pi_\zeta(\zeta) \, d\zeta \tag{5.42}$$

is the marginal density for X after integrating out over the mixing distribution, and $p_i(\zeta|x_i)$ is the posterior density derived from datum x_i, formed as

$$p_i(\zeta|x_i, u) \propto f_{X|\zeta}(x_i|\zeta, u)\pi_\zeta(\zeta). \tag{5.43}$$

For (5.42) we can again use numerical integration effectively, as this is merely a bivariate integral; to compute for each u, note that

$$f_X(x|u) = f_X(x - u),$$

where the right hand side is the density computed for $u = 0$ from (5.39). For (5.43), we again use MCMC sampling. Sampling the hyperparameters and the DP precision parameter proceed as before.

5.4 A nonparametric version of the model from Section 5.1.4

Suppose again, as in Section 5.1.4, that the data density f_B takes the form

$$f_B(x; u, \phi, \xi, \sigma) = \begin{cases} \dfrac{f_H(x;\phi)}{F_H(u;\phi) + \omega(1 - F_H(u;\phi))}, & x \leq u, \\ \dfrac{\omega(1 - F_H(u;\phi))f_G(x; u, \xi, \sigma)}{F_H(u;\phi) + \omega(1 - F_H(u;\phi))}, & x > u. \end{cases}$$

A nonparametric specification for the component f_G of this model can be considered, specifically the DPM from Section 5.3.3.

5.4.1 Implementing the DPM in the model from Section 5.1.4

The only complicated step involves sampling the $\zeta = (\sigma, \xi)$ from their full conditional density as indicated by Equations (5.37) and (5.38). Recall that we need to sample ζ_i from its full conditional density

$$\zeta_i|x_i, \zeta_{(-i)}, u \sim w_0 p_i(\zeta|x_i, u) + \sum_{j \neq i} w_j \delta_{\zeta_j}$$

for each $i = 1, \ldots, n$, where

$$w_0 = \frac{\nu f_X(x_i|u)}{\nu f_X(x_i|u) + \sum_{l \neq i} f_{X|\zeta}(x_i|\zeta_l, u)} \qquad \text{and} \qquad w_j = \frac{f_{X|\zeta}(x_i|\zeta_j, u)}{\nu f_X(x_i|u) + \sum_{l \neq i} f_{X|\zeta}(x_i|\zeta_l, u)}, \ j \neq i$$

with

$$f_X(x|u) = \int_{-1/2}^{\infty} \int_0^{\infty} f_{X|\zeta}(x|\zeta, u)\pi_\zeta(\zeta)\, d\zeta$$

and

$$p_i(\zeta|x_i, u) \propto f_{X|\zeta}(x_i|\zeta, u)\pi_\zeta(\zeta),$$

the full conditional posterior for ζ given x_i.

The two cases $x_i \leq u$ and $x_i > u$ need separate consideration:

- $x_i \leq u$: in this case, the density for x does not depend on ζ, so using the same logic as in Section 5.2.3, we have that the sampling weights are proportional to

  $$\nu f_H(x_i|\phi, u)$$

  for w_0, where $f_H(.|\phi, u)$ is the density f_H truncated to $(0, u)$, and w_j is $f_H(x_i|\phi, u)$ for each $j \neq i$. Therefore the sum of the sampling weights is $(\nu + n - 1)f_H(x_i|\phi, u)$. The posterior measure $p_i(\zeta|x_i, u,)$ is given by

  $$p_i(\zeta|x_i, u) \propto f_H(x_i|\phi, u)\pi_\zeta(\zeta) \propto \pi_\zeta(\zeta).$$

  That is, we may sample ζ_i in this case using a prior Pólya urn scheme

  - with probability $\nu/(\nu + n - 1)$, sample $\zeta_i \sim \pi_\zeta(\cdot)$
  - with probability $w_j = 1/(\nu + n - 1)$, $j \in \{1, ..., i - 1, i + 1, ..., n\}$, set $\zeta_i = \zeta_j$.

- $x_i > u$: in this case, the standard scheme applies. We have by assumption that the data density is

  $$\frac{\omega(1 - F_H(u; \phi))f_G(x; u, \zeta)}{F_H(u; \phi) + \omega(1 - F_H(u; \phi))} = M f_G(x; \zeta, u)$$

say, so therefore

$$w_0 \propto \nu M f_G(x_i|u) = \nu M \int_{-1/2}^{\infty} \int_0^{\infty} f_{X|\zeta}(x_i|\zeta, u)\pi_\zeta(\zeta)\, d\zeta$$

and

$$w_j \propto M f_G(x_i|\zeta_j, u), \qquad j \neq i.$$

That is, the constant M cancels, and we may sample ζ_i as follows. Let

$$S_i = \sum_{j \neq i} f_{X|\zeta}(x_i|\zeta_j, u)$$

then

- with probability $\nu f_G(x_i|u)/(\nu f_G(x_i|u)+S_i)$, sample $\zeta_i \sim p_i(\zeta|x_i, u)$ using Metropolis-Hastings;
- with probability $w_j = 1/(\nu f_G(x_i|u) + S_i)$, $j \in \{1, 2, ..., i-1, i+1, ..., n\}$ set $\zeta_i = \zeta_j$.

5.4.2 Updating the remaining parameters

5.4.2.1 Parameters of the density f_H

The parameters ϕ of the component f_H can be updated in the usual way for a parametric analysis. Conditional on the parameters $\zeta_1, \ldots, \zeta_n$, ω and u, we have that the full conditional likelihood takes the form

$$\left\{ \prod_{i\,:\,x_i \leq u} \frac{f_H(x_i;\phi)}{F_H(u;\phi) + \omega(1 - F_H(u;\phi))} \right\} \left\{ \prod_{i\,:\,x_i > u} \frac{\omega(1 - F_H(u;\phi)) f_G(x_i;\zeta_i, u)}{F_H(u;\phi) + \omega(1 - F_H(u;\phi))} \right\} \tag{5.44}$$

which, if n_G denotes the number of $x_i > u$, is proportional to

$$\left\{ \prod_{i\,:\,x_i \leq u} f_H(x_i;\phi) \right\} \frac{(1 - F_H(u;\phi))^{n_G}}{[F_H(u;\phi) + \omega(1 - F_H(u;\phi))]^n}$$

A prior on ϕ completes the specification.

5.4.2.2 Threshold parameter u and scaling parameter ω

The parameter u is updated using the conditional likelihood in equation (5.44), in conjunction with a suitable prior. The same is true for scaling parameter ω. Both updates are achieved using Metropolis-Hastings.

5.4.2.3 Dirichlet process parameters

The hyperparameters of the Dirichlet process base measure, and the Dirichlet process precision parameter, are updated as in previous sections.

5.4.3 Real data example

As before, we tested the Bayesian nonparametric models on different data sets. We only show the results for the fraud data from Section 3.2.1. This time we focus on the first Bayesian nonparametric model and on the nonparametric version of the model from Section 5.1.4.

Following the procedure from Sections 5.2.1 and 5.4.1, we obtained 2000 samples from the posterior distributions using MCMC. Again, we assume a Gamma model for f_H.

Results for the first Bayesian nonparametric model are shown in Figures 5.8–5.10, while Figures 5.11–5.13 display the results for the nonparametric version of the model from Section 5.1.4. In Figure 5.12, we can observe samples of u, ω and f_H model parameters. We can also assess the model performance using the P-P plots. From the figures one may notice that the model fit is good in general, although it can be improved by introducing the ω parameter.

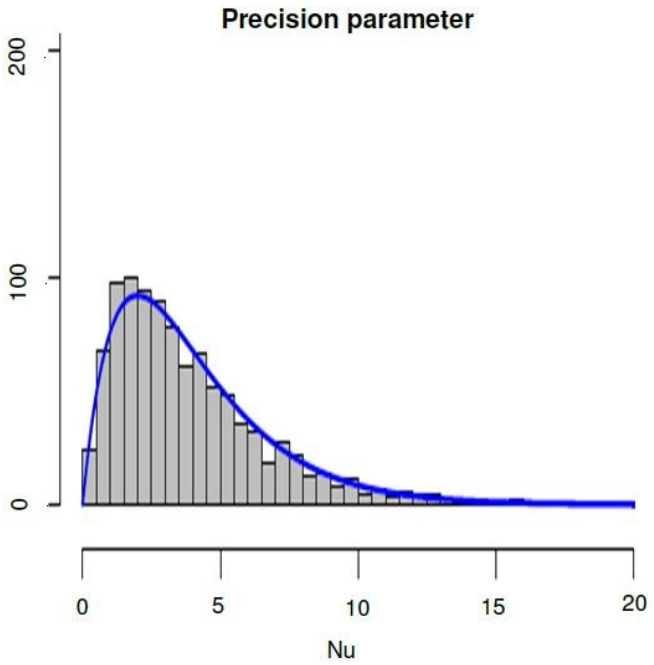

Figure 5.8: *Fraud data: Posterior samples of the precision parameter for the first Bayesian nonparametric model. The blue line is a Gamma prior for ν in terms of frequencies*

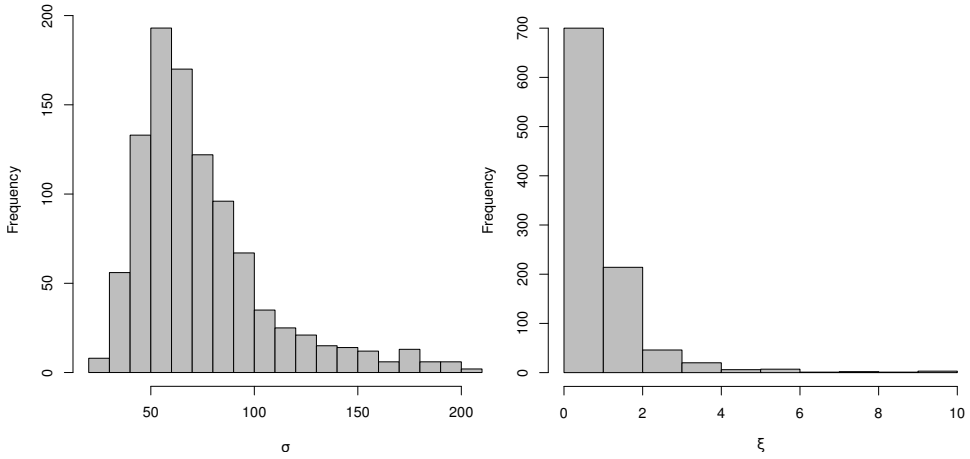

Figure 5.9: *Fraud data: Histograms of the posterior MCMC estimates for the first Bayesian nonparametric model*

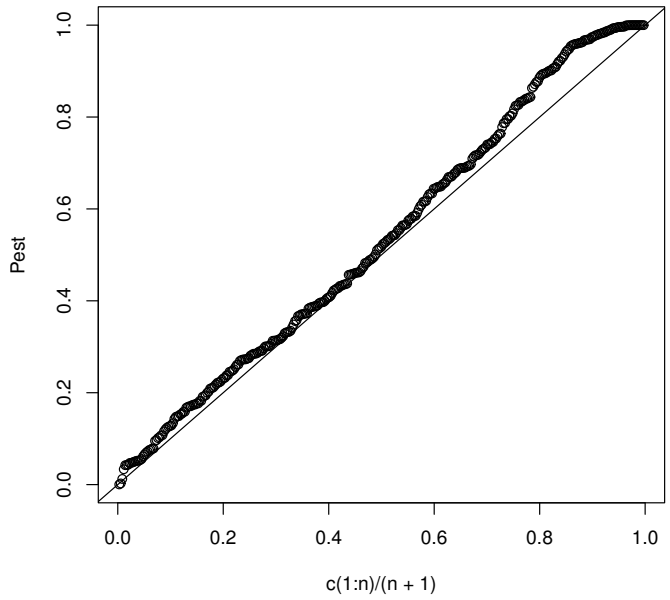

Figure 5.10: *Fraud data: P-P plot using the posterior MCMC estimates from the first Bayesian nonparametric model*

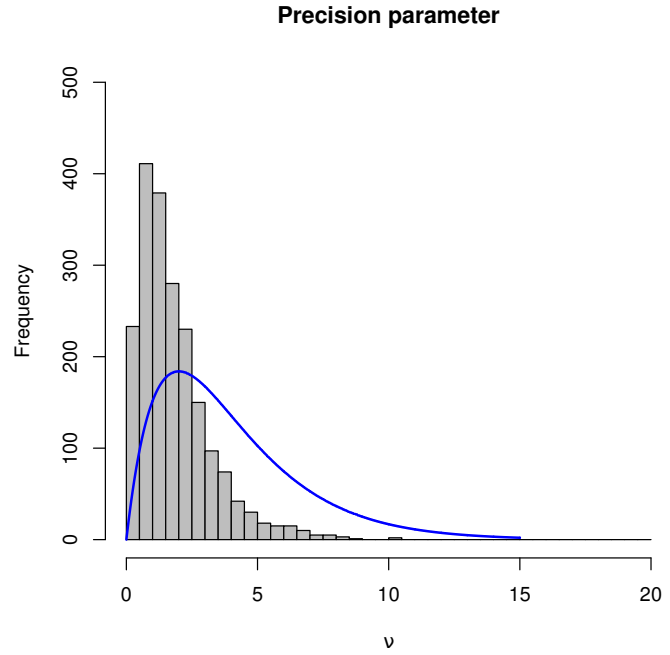

Figure 5.11: *Fraud data: Posterior samples of the precision parameter for the nonparametric version of the model with discontinuous density with arbitrary scaling. The blue line is a Gamma prior for ν in terms of frequencies*

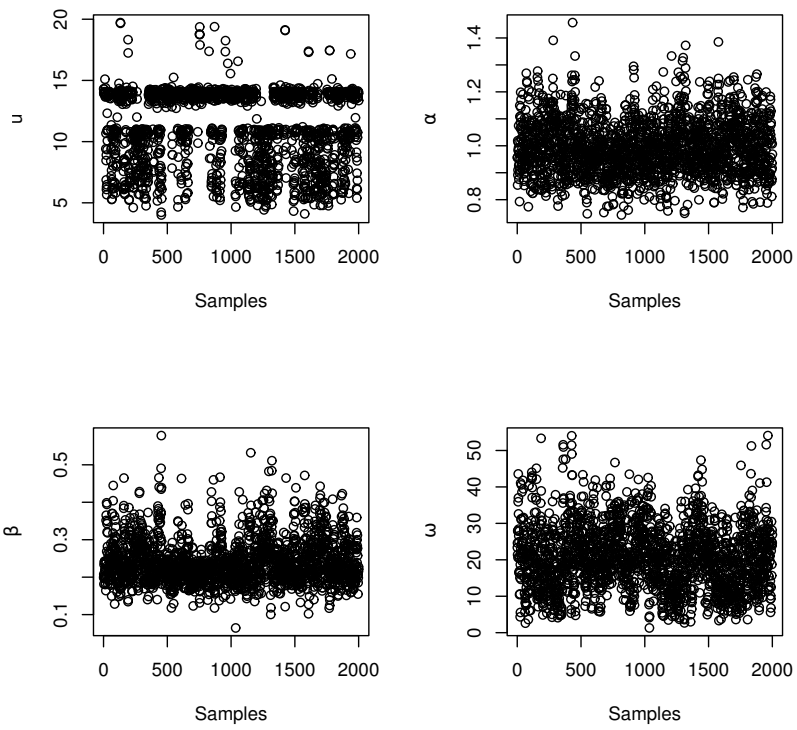

Figure 5.12: *Fraud data: Posterior parameter samples from the nonparametric version of the model with discontinuous density with arbitrary scaling*

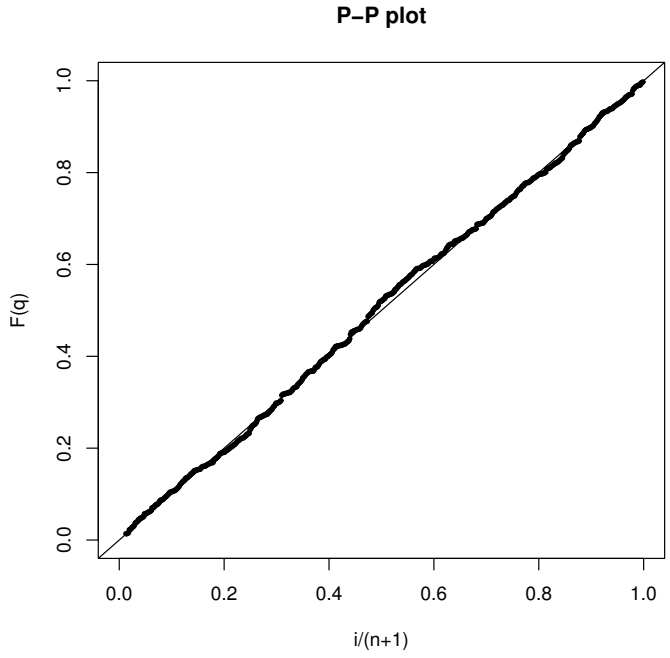

Figure 5.13: *Fraud data: P-P plot using the posterior MCMC estimates from the nonparametric version of the model with discontinuous density with arbitrary scaling*

5.5 Conclusions of the Chapter

In this chapter, we have proposed several new models. The first model has a density continuous at the qth quantile, fixing the discontinuity issue in the model by Behrens et al. (2004) in a simple way. Nonetheless, q is a fixed quantity. Thus, in the second model we require the first derivative to be continuous at the blend point, so that q is not fixed anymore. The third model has a continuous first derivative, but this time the scaling is arbitrary, i.e., the density is scaled by some $\omega > 0$, which gives greater flexibility. The last Bayesian model is based on a discontinuous density with arbitrary scaling. One of the main features of this model is that it can be implemented even if the density and its derivative are discontinuous.

As for the Bayesian nonparametric models, we have recalled some properties of the GPD, in particular that it can be represented as a Gamma mixture of Exponentials, suggesting a model based on a Dirichlet Process Mixture representation. We have also introduced a second Bayesian nonparametric model that differs from the previous one in that it uses a Dirichlet process prior on the parameters of the GPD. Similarly, we have applied this last model to the case when the density is discontinuous and the scaling is arbitrary.

As we have seen in our analysis, all these models can be easily implemented using MCMC methods. For the fraud data, we have observed that the proposed models have a good overall performance, providing a powerful alternative to deal with the discontinuity issue in the model introduced by Behrens et al. (2004).

Chapter 6

Conclusions and future research

6.1 Contributions of the thesis

Extreme Value Theory based models have been widely used in many fields as they supply a statistically justifiable and flexible method for extrapolating tail distributions. In the context of operational risk, EVT offers an interesting view on the quantitative measurement of risk. As Embrechts (2002) pointed out: "The main virtue of EVT is that it gives the user a critical view on and a methodological toolkit for issues like skewness, fat tails, rare events, stress scenarios,..."; however, as we have discussed through this thesis, there are still serious limitations that have to be overcome to make it viable; for instance, this methodology requires a good choice of a threshold to have a good performance, which most of the times is problematic. Furthermore, the Basel II requirements for implementing the Advanced Measurement Approach are not completely met.

This dissertation has investigated how a Bayesian analysis of extremes offers and alternative to the standard classical one, in the sense that it is able to account for the uncertainty associated with the threshold choice by considering it as another parameter of the model to be estimated. It also offers the ability to supplement information provided by the data with other sources of information, through the prior distribution. In addition, the

output of a Bayesian analysis provides posterior information which can also be exploited for extrapolation, via the posterior predictive distribution.

The first contribution of this work was presented in Chapter 3, where both the model proposed and its application to operational risk data showed how Bayesian inference allows for parameter estimation to determine the loss distribution, by using all data and prior information through a Bayesian model. Our study has shown the importance of considering the threshold uncertainty in the estimation, as the GPD parameters are very sensitive with respect to this value. We also observed how the inclusion of expert opinion is an important input, particularly when we have small data sets, as shown in Chapter 3.

Another interesting discovery was that the Basic Indicator Approach –the most commonly used method for estimating the minimum capital requirement–, although is a good start, rarely provides a good estimate, since it does not take into account either the data or expert opinion.

One of the most significant findings to emerge from this study was that expert opinion represents a valuable but underestimated source of information. In many cases, to simplify the problem, it is assumed that the information experts can provide is not significant and statisticians tend to use non-informative priors instead of priors based on expert judgment. Nevertheless, as shown in Chapter 4, many of the non-informative prior distributions are unrealistic. It was further noted that expert opinion can dramatically influence the estimates in the context of extreme events and hence, one should pay special attention to the elicitation process.

Furthermore, we addressed a common problem in practice: the combination of multiple expert opinions. Although this topic has been widely studied, little research has focused on operational risk. Our work thus makes a noteworthy contribution in the sense that it provides the tools for combining and updating multiple opinions on extreme data. Moreover, we could notice how the inclusion of different opinions may influence the results and only after some years, when more data become available, the influence of experts is

limited.

The last contribution of this thesis was the introduction of several new models that allow to deal with the discontinuity problem observed in previous models. Our models are a modification on a previously introduced model (Behrens et al., 2004). In this last Chapter, we provided the theory and the sampling scheme for all the different versions of our model. This can be used in order to improve the performance of Bayesian models and parameter estimates, providing a more realistic approach to the loss distribution. We also explored Bayesian nonparametric models, giving more flexibility to the analysis of extreme events and laying the foundation for future work in such models.

6.2 Future research

The findings of this study have a number of important implications for future practice, and we strongly advocate that Bayesian approaches to operational risk modelling should be considered as a serious alternative for practitioners in banks and financial institutions.

It is worth pointing out that these models may be improved and extended in a number of directions. Currently, the use of Bayesian models in extremes is an active research area. For instance, some authors have extended these models to include covariates in a GPD model formulation (Cabras et al., 2011) while others are treating this problem in a nonparametric setting (Wang et al., 2011).

As part of our future research, we have considered the problem of dealing with losses caused by different effects. In the next section we develop some ideas and present some of our findings for the study of mixture distributions in EVT.

6.2.1 Finite mixture distributions and EVT

Extreme value theory is usually applied to single loss distributions; however, there is the case where losses are caused by two or more independent effects, each happening with a

certain probability. In this situation, we should consider mixture models, i.e. distributions of the form $F(x) = \sum_{i \in I} w_i F_i(x)$ for $x \in \mathbb{R}$, where $w_i \geq 0$, $i \in I$ and $\sum_{i \in I} w_i = 1$; the index set I is assumed to be finite or countable and $\{F_i : i \in I\}$ is a family of univariate distribution functions. But this leads to a different problem. Recently, some authors have investigated the performance of the POT method under these circumstances. For instance, Nešlehová et al. (2006) point out the potential shortcomings of the classical POT model in the presence of *"data contamination"*, that is, observations within the sample, which do not follow the same distribution as the rest of the data. They provide some examples based on mixtures of Pareto distributions, showing that very high losses are driven for one of the mixture components, which leads to incorrect estimates of high quantiles.

In this section we explore whether extreme value theory can be applied to mixture distributions. We start by reviewing some results from the literature, particularlly a theorem in the paper by Kang and Serfozo (1999), which shows that for mixture distributions, the limiting distribution is determined by one or more distributions among the mixtures whose tails dominate the other tails. They also derive bounds and rates of convergence.

Definition 6.1. Suppose F is a mixture of the form

$$F(x) = \sum_{i \in I} w_i F_i(x), \quad x \in \mathbb{R}, \tag{6.1}$$

where $\{F_i : i \in I\}$ is a finite or countable collection of distributions and the w_i are non-negative constants such that $\sum_{i \in I} w_i = 1$.

We say that the tail of the distribution F^* with right endpoint x^* *dominates* those of $\{F_i : i \in I\}$ if, for each $i \in I$, the right endpoint of F_i is x^* and

$$\lim_{x \to x^*} \frac{\bar{F}_i(x)}{\bar{F}^*(x)} = r_i,$$

for some finite $r_i \geq 0$, and this limit is uniform in i in case I is an infinite set. The coefficients r_i are called the tail ratios. F^* and F_i are called *tail equivalent* if $r_i > 0$, and the tail of F^* *strictly dominates* that of F_i when $r_i = 0$.

Theorem 6.2. *Suppose F is a mixture of the form (6.1) and there is a distribution $F^* \in \{F_i : i \in I\}$ whose tail dominates those of $\{F_i : i \in I\}$ with tail ratios such that $\gamma \equiv \sum_i w_i r_i$ is positive. Then the following statements are equivalent.*

(i) *$F \in \mathrm{MDA}(H)$ with normalizing constants c_n, d_n.*

(ii) *$F^* \in \mathrm{MDA}(H)$ with normalizing constants c_n^*, d_n^*.*

When these statements hold, the normalizing constants are related as follows.

$$\left.\begin{array}{lll} c_n = c_n^*, & d_n = d_n^* + c_n^* log\,\gamma & \text{if } H \text{ is Gumbel} \\ c_n = \gamma^{1/\alpha} c_n^*, & d_n = d_n^* = 0 & \text{if } H \text{ is Fréchet} \\ c_n = \gamma^{-1/\alpha} c_n^*, & d_n = d_n^* & \text{if } H \text{ is Weibull} \end{array}\right\} \tag{6.2}$$

Proof of Theorem 6.2. For simplicity, let $x_n = c_n x + d_n$. We first show that statement (ii) is equivalent to

$$n\bar{F}^*(x_n) \to -\gamma^{-1} \log H(x), \quad x \in \mathbb{R}, \tag{6.3}$$

where c_n, d_n are given by (6.2). Suppose that (ii) holds and note that it is equivalent to

$$n\bar{F}^*(c_n^* x + d_n^*) \to -\log H(x). \tag{6.4}$$

by (2.11). If H is Gumbel, i.e. $H(x) = e^{-e^{-x}}$, then $x_n = c_n^*(x + \log\gamma) + d_n^*$ and so by (6.4),

$$n\bar{F}^*(x_n) \to -\log H(x + \log\gamma) = -\gamma^{-1} \log H(x).$$

If H is Fréchet or Weibull, it follows by a similar argument that

$$x_n = \begin{cases} c_n^*\left(x\gamma^{1/\alpha}\right) + d_n^* & \text{if } H \text{ is Fréchet,} \\ c_n^*\left(x\gamma^{-1/\alpha}\right) + d_n^* & \text{if } H \text{ is Weibull,} \end{cases}$$

respectively, and hence

$$n\bar{F}^*(x_n) \to \left\{ \begin{array}{c} -\log H\left(x\gamma^{1/\alpha}\right) \\ -\log H\left(x\gamma^{-1/\alpha}\right) \end{array} \right\} = -\gamma^{-1} \log H(x).$$

The last equality is due to the two forms of H. This proves that (ii) implies (6.3). Conversely, suppose that (6.3) holds and define c_n^*, d_n^* by (6.2). Then arguing as above, it follows that, for the three forms of H,

$$n\bar{F}^*\left(c_n^* x + d_n^*\right) \to \left\{ \begin{array}{c} -\gamma^{-1}\log H\left(x - \log\gamma\right) \\ -\gamma^{-1}\log H\left(x\gamma^{-1/\alpha}\right) \\ -\gamma^{-1}\log H\left(x\gamma^{1/\alpha}\right) \end{array} \right\} = -\log H\left(x\right).$$

Thus $F^* \in \mathrm{MDA}\left(H\right)$, which proves that (ii) is equivalent to (6.3).

We now consider the equivalence of (i) and (ii). Because of (2.11), this is equivalent to showing that (6.3) is equivalent to which is equivalent to

$$n\bar{F}\left(x_n\right) \to -\log H\left(x\right). \tag{6.5}$$

By the definition of F,

$$n\bar{F}\left(x_n\right) = n\bar{F}^*(x_n)\sum_{i\in I} w_i \frac{\bar{F}_i(x_n)}{\bar{F}^*(x_n)}. \tag{6.6}$$

Since the tail of F^* dominates those of the F_i, we have that $\bar{F}_i(x_n)/\bar{F}^*(x_n) \to r_i$ as $n \to \infty$ for $i \in I$. Furthermore, this convergence is uniform in i when I is an infinite set. Thus, by the dominated convergence theorem, we have

$$\sum_i w_i \frac{\bar{F}_i(x_n)}{\bar{F}^*(x_n)} \to \sum_i w_i r_i = \gamma. \tag{6.7}$$

This implies that (6.3) is equivalent to (6.5) by (6.6). □

6.2.1.1 Simulation study

In the previous section we studied the theoretical foundations of extreme value theory for mixture distributions. In order to compare the differences between theory and practice, we carry out a simulation study. We divide our study in two parts: 1. Behaviour of the GPD scale and shape parameters when dealing with mixture distributions and 2. Estimation of high quantiles for a particular mixture distribution.

6.2.1.2 GPD scale and shape parameters for mixture distributions

In this part of our study, we simulate data from three different (two-component) mixtures with $I = \{1, 2\}$:

1. F_1 is $N(100, 20^2)$ and F_2 is $N(150, 25^2)$.

2. F_1 is $N(700, 50^2)$ and F_2 is Gamma$(100, 0.1)$.

3. F_1 is $t_{(2,10)}$ and F_2 is $t_{(5,15)}$.

We start by exploring the tail dominance of these distributions. For a mixture of Gaussians we have:

$$\begin{aligned}\lim_{x\to\infty}\frac{\bar{F}_i(x)}{\bar{F}_j(x)} &= \lim_{x\to\infty}\frac{f_i(x)}{f_j(x)} = \lim_{x\to\infty}\frac{\frac{1}{\sigma_i\sqrt{2\pi}}\exp\left[-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right]}{\frac{1}{\sigma_j\sqrt{2\pi}}\exp\left[-\frac{(x-\mu_j)^2}{2\sigma_j^2}\right]} \\ &= \frac{\sigma_j}{\sigma_i}\exp\left[-\frac{\mu_i^2}{2\sigma_i^2}+\frac{\mu_j^2}{2\sigma_j^2}\right]\lim_{x\to\infty}\exp\left[\left(\frac{1}{\sigma_j^2}-\frac{1}{\sigma_i^2}\right)\frac{x^2}{2}+\left(\frac{\mu_i}{\sigma_i^2}-\frac{\mu_j}{\sigma_j^2}\right)x\right].\end{aligned}$$

Thus $\bar{F}_i(x)/\bar{F}_j(x)$ converges to 0 as $x\to\infty$ if

$$\left(\frac{1}{\sigma_j^2}-\frac{1}{\sigma_i^2}\right)<0 \quad \text{or} \quad \mu_i<\mu_j \quad \text{for} \quad \sigma_i=\sigma_j.$$

Therefore, the tail of F_j strictly dominates that of F_i when $\sigma_i<\sigma_j$ or $\mu_i<\mu_j$ for $\sigma_i=\sigma_j$.

Hence, for the mixture $N(100, 20^2)$ -$N(150, 25^2)$, the tail is dominated by $N(150, 25^2)$. Similar results can be derived for the other mixtures.

In order to observe the behaviour of these distributions in practice, we simulate 1,000 observations from each mixture and assign equal weights $w_i = 0.5 \;\; i = 1, 2$ to each component. The corresponding densities are shown in Figure 6.1.

We approximate the tail of each mixture using the POT method and maximum likelihood for the estimation of the scale (σ) and shape (ξ) parameters. The threshold is set as the 0.7, 0.8 and 0.9 empirical quantile. Results are displayed in Tables 6.1-6.6.

From Table 6.1, we can observe that, under the classic GPD estimation, the estimate of the ξ parameter—denoted by $\hat{\xi}$—is always larger for the mixture than for the single normals. The Bayesian estimates present a similar behaviour and, again, the $\hat{\xi}$ parameter is larger for the mixture (Table 6.2). In most cases, $\hat{\xi}$ is far from the theoretical value of 0, since the normal distribution belongs to the Gumbel domain of attraction.

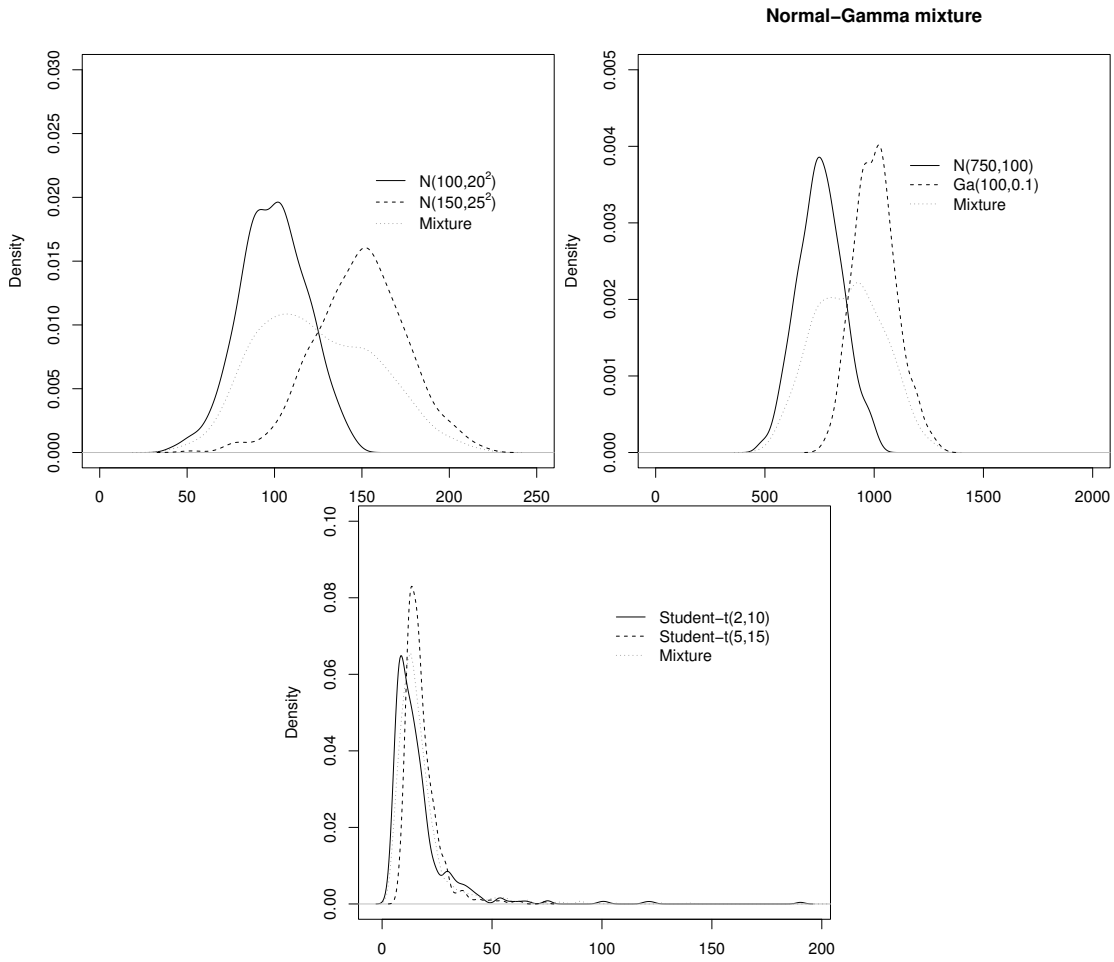

Figure 6.1: *Simulated mixture densities*

For the Normal-Gamma mixture we can see in Table 6.3 that the shape parameter of the mixture is larger in almost all cases under the classical GPD estimation, except for $u = 0.7$ quantile. However, for the Bayesian estimates, the $\hat{\xi}$ parameter of the mixture is smaller compared to the shape parameter of the single distributions.

On the other hand, for the mixture of noncentral-t distributions, from Tables 6.5 and 6.6 we can see that $\hat{\xi}$ behaves as previously, being larger for the mixture in all cases, except for $u = 0.9$ quantile. The theoretical values of $\xi = 0.5, 0.2$, respectively, are better approximated under the Bayesian approach.

Now, in order to observe the effect of increasing the sample size, we simulate 2,000 observations from the previous noncentral t-distributions (Figure 6.2). Results for the

Distribution	$u = q_{0.7}$	σ	ξ	$u = q_{0.8}$	σ	ξ	$u = q_{0.9}$	σ	ξ
Mixture	146.685	24.522	-0.226	157.513	19.737	-0.179	172.163	13.005	-0.043
$N(100, 20^2)$	109.585	15.897	-0.307	115.648	14.258	-0.318	124.307	11.859	-0.336
$N(150, 25^2)$	164.468	18.652	-0.260	172.338	15.619	-0.226	181.785	14.622	-0.276

Table 6.1: *Classic GPD estimation for the mixture of normals.*

	Mixture	$N(100, 20^2)$	$N(150, 25^2)$
Parameter	Mean, median, std	Mean, median, std	Mean, median, std
u	152.839, 152.777, 1.042	100.610, 100.444, 1.498	170.189, 172.163, 4.969
σ	12.242, 11.541, 2.615	18.330, 18.461, 1.507	22.127, 22.127, 1.789
ξ	0.0168, 0.012, 0.109	-0.286, -0.291, 0.051	-0.259, -0.264, 0.056

Table 6.2: *Bayesian parameter estimates for the mixture of normals.*

Distribution	$u = q_{0.7}$	σ	ξ	$u = q_{0.8}$	σ	ξ	$u = q_{0.9}$	σ	ξ
Mixture	974.868	130.720	-0.365	1030.631	98.825	-0.299	1092.69	79.450	-0.293
$N(700, \sqrt{50})$	806.842	94.180	-0.385	841.525	82.837	-0.402	889.185	74.334	-0.519
Gamma(100,0.1)	806.842	86.709	-0.260	1083.393	84.352	-0.307	1131.591	80.534	-0.402

Table 6.3: *Classic GPD estimation for the Normal-Gamma mixture.*

	Mixture	N(700,50)	Gamma(100,0.1)
Parameter	Mean, median, std	Mean, median, std	Mean, median, std
u	910.686, 910.141, 3.024	779.329, 778.683, 2.625	1008.302, 1008.372, 0.638
σ	119.698, 118.807, 6.255	60.753, 60.330, 6.067	64.387, 64.130, 4.883
ξ	-0.212, -0.209, 0.040	-0.007, -0.011, 0.091	0.012, 0.008, 0.070

Table 6.4: *Bayesian parameter estimates for the Normal-Gamma mixture.*

classic GPD estimation are displayed in Table 6.7. This time we can notice that the shape parameter value of the mixture is between the other two parameters' values.

Bayesian estimates are presented in Table 6.8. Again, the shape parameter value of the mixture is between the other two parameters' values. Hence, one may infer that increasing the sample size decreases the effect of the dominant tail and leads to better approximations. Nonetheless, in practice, we usually have a small number of observations.

Distribution	$u = q_{0.7}$	σ	ξ	$u = q_{0.8}$	σ	ξ	$u = q_{0.9}$	σ	ξ
Mixture	18.149	6.728	0.511	21.099	8.203	0.524	27.940	14.188	0.372
$t_{(2,10)}$	17.131	8.664	0.414	20.169	12.919	0.266	30.490	11.656	0.460
$t_{(5,15)}$	19.273	5.658	0.223	21.790	5.879	0.254	26.042	7.460	0.220

Table 6.5: *Classic GPD estimation for the mixture of Student-t distributions (1,000 observations).*

	Mixture	$t_{(2,10)}$	$t_{(5,15)}$
Parameter	Mean, median, std	Mean, median, std	Mean, median, std
u	17.571, 17.270, 1.520	17.63, 17.471, 1.430	19.084, 17.751, 0.632
σ	6.278, 6.183, 0.713	6.378, 6.319, 0.716	5.645, 5.568, 0.910
ξ	0.521, 0.518, 0.079	0.519, 0.528, 0.094	0.232, 0.221, 0.116

Table 6.6: *Bayesian parameter estimates for the mixture of Student-t distributions (1,000 observations).*

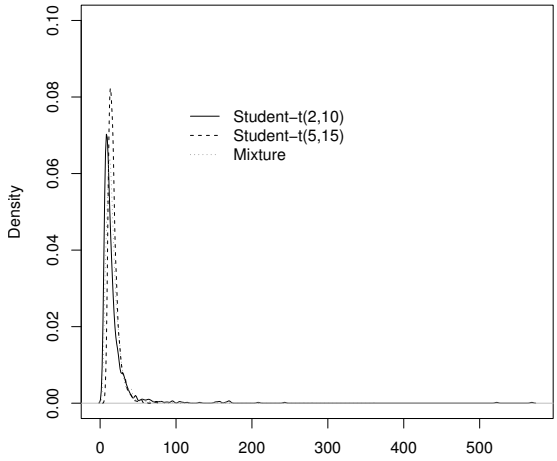

Figure 6.2: *Mixture of Student-t distributions (2,000 observations).*

6.2.1.3 High quantiles estimation

In this section we explore the behaviour of high quantiles when taking into account the mixture of components instead of a single distribution. To do so, we consider a mixture of Generalized Pareto distributions:

$$w_1 \text{GPD}\left(\sigma_1, \xi_1\right) + w_2 \text{GPD}\left(\sigma_2, \xi_2\right), \qquad w_2 = 1 - w_1.$$

Our aim is to estimate the scale and location parameters and the corresponding weights. By doing so, we will be able to compute the Value-at-Risk for different parameter sets and

Distribution	$u = q_{0.7}$	σ	ξ	$u = q_{0.8}$	σ	ξ	$u = q_{0.9}$	σ	ξ
Mixture	18.611	7.276	0.384	22.213	7.187	0.484	27.979	9.792	0.512
$t_{(2,10)}$	16.504	9.913	0.589	21.117	11.803	0.636	30.910	16.495	0.742
$t_{(5,15)}$	18.943	6.393	0.108	21.659	6.502	0.124	26.465	7.411	0.082

Table 6.7: *Classic GPD estimation for the mixture of Student-t distributions (2,000 observations).*

	Mixture	$t_{(2,10)}$	$t_{(5,15)}$
Parameter	Mean, median, std	Mean, median, std	Mean, median, std
u	20.821, 20.801, 2.498	11.893, 11.722, 0.245	17.312, 17.270, 1.064
σ	7.342, 7.254, 0.624	7.912, 7.855, 0.633	6.002, 5.976, 0.509
ξ	0.431, 0.423, 0.082	0.555, 0.552, 0.072	0.130, 0.121, 0.060

Table 6.8: *Bayesian parameter estimates for the mixture of Student-t distributions (2,000 observations).*

sample sizes.

Before doing the estimation, note that inference for mixtures of the classical Pareto distributions is difficult because the EM algorithm cannot be applied; see Bee et al. (2013). These authors point out the theoretical and computational difficulties in carrying out the estimation using MLE. They show that when we have different shapes and scales, the likelihood is non-regular and the EM algorithm breaks down. That is, the estimator of the scale parameter is not asymptotically efficient since the distributions of $\hat{\sigma}_1$ and $\hat{\sigma}_2$ do not have common support, violating one of the regularity conditions for asymptotic efficiency of MLEs.

Whether similar results apply also to mixtures of the Generalized Pareto distribution remains to be investigated. Keeping this fact in mind, we decided to perform a simple MLE estimation using the function `optim` in R. The estimation was carried out for different data sets, keeping σ_1, σ_2, ξ_1 and ξ_2 fixed, and varying the weights w_i and sample size $n = 150$, 500 and 1000. The data used in this study come from the following distributions:

1. 0.2 GPD(1,0.5)+0.8 GPD(3,0.2)

2. 0.5 GPD(1,0.5)+0.5 GPD(3,0.2)

3. 0.7 GPD(1,0.5)+0.3 GPD(3,0.2)

Table 6.9 shows the parameter estimates. From the table, we can see that most of the time, values are not estimated correctly. Notwithstanding, the essential purpose of this work is to compare the performance of the classical POT method with respect to the mixture of GPDs. To do so, we compute the Value-at-Risk using both methods. In general, the Value-at-Risk of mixture distributions is no longer given by a simple formula and has to be calculated numerically.

Tables 6.10–6.18 display the estimates of the Value-at-Risk at different levels. We can see that in general, the mixture produces good estimates of $\text{VaR}_{0.9}$, $\text{VaR}_{0.95}$ and $\text{VaR}_{0.99}$, as the mean square error (MSE) is small; although, for higher quantiles (0.999 and 0.9999) neither the mixture nor the classical POT method give accurate estimates and the MSE increases considerably. However, the mixture estimates are slightly closer to the true values than the classical POT estimates in almost all cases. It is worth to mention that the estimates are very sensitive to the sample size and we have to be very careful with our conclusions.

This analysis does not pretend to be exhaustive, but may be considered as a tentative, to be extended to more general situations, in the aim to improve the application of extreme value theory methods to mixture distributions, as more sophisticated methods become available to estimate the parameters of the mixture of GPDs.

	0.2 GPD(1,0.5)+0.8 GPD(3,0.2)				0.5 GPD(1,0.5)+0.5 GPD(3,0.2)				0.7 GPD(1,0.5)+0.3 GPD(3,0.2)			
Par.	True value	Estimate			True value	Estimate			True value	Estimate		
		$n = 150$	$n = 500$	$n = 1000$		$n = 150$	$n = 500$	$n = 1000$		$n = 150$	$n = 500$	$n = 1000$
w_1	0.2	0.047	0.658	0.733	0.5	0.547	0.056	0.729	0.7	0.100	0.999	0.799
σ_1	1	0.088	2.492	2.073	1	1.173	0.207	1.057	1	0.103	1.167	1.105
ξ_1	0.5	-0.002	0.287	0.327	0.5	0.443	0.559	0.487	0.5	0.815	0.478	0.329
σ_2	3	2.542	3.545	5.931	3	3.852	1.964	4.036	3	1.744	2.348	5.211
ξ_2	0.2	0.269	-0.077	-0.092	0.2	-0.108	0.347	-0.047	0.2	0.422	0.348	0.172

Table 6.9: *Parameter estimates for a mixture of GPDs*

0.2 GPD(1,0.5)+0.8 GPD(3,0.2)					
		Mixture of GPD's		Classical POT	
γ	True value	VaR_γ	MSE	VaR_γ	MSE
0.9	8.109	7.877	0.053	7.756	0.125
0.95	11.594	11.430	0.027	11.594	6.671e-07
0.99	22.116	22.746	0.397	25.043	8.569
0.999	46.945	50.376	11.774	63.575	276.563
0.9999	102.650	101.721	0.862	149.458	2191.039

Table 6.10: *VaR for a mixture of GPDs,* $n = 150$

0.2 GPD(1,0.5)+0.8 GPD(3,0.2)					
		Mixture of GPD's		Classical POT	
γ	True value	VaR_γ	MSE	VaR_γ	MSE
0.9	8.109	7.847	0.068	8.124	2.091e-04
0.95	11.594	10.798	0.632	11.356	0.056
0.99	22.116	20.328	3.197	20.738	1.898
0.999	46.945	47.188	0.059	40.325	43.826
0.9999	102.650	99.468	10.123	70.738	1018.368

Table 6.11: *VaR for a mixture of GPDs,* $n = 500$

0.2 GPD(1,0.5)+0.8 GPD(3,0.2)					
		Mixture of GPD's		Classical POT	
γ	True value	VaR_γ	MSE	VaR_γ	MSE
0.9	8.109	9.068	0.919	8.697	0.345
0.95	11.594	12.706	1.236	12.571	0.955
0.99	22.116	22.272	0.024	24.950	8.030
0.999	46.945	48.608	2.765	55.325	70.223
0.9999	102.650	110.469	61.137	112.092	89.142

Table 6.12: *VaR for a mixture of GPDs,* $n = 1000$

0.5 GPD(1,0.5)+0.5 GPD(3,0.2)					
		Mixture of GPD's		Classical POT	
γ	True value	VaR_γ	MSE	VaR_γ	MSE
0.9	6.927	6.621	0.093	6.872	0.003
0.95	10.284	8.998	1.651	9.830	0.205
0.99	21.050	15.207	34.135	19.006	4.176
0.999	51.308	40.610	114.440	40.431	118.298
0.9999	142.295	117.351	622.179	78.261	4100

Table 6.13: *VaR for a mixture of GPDs,* $n = 150$

0.5 GPD(1,0.5)+0.5 GPD(3,0.2)					
		Mixture of GPD's		Classical POT	
γ	True value	VaR_γ	MSE	VaR_γ	MSE
0.9	6.927	6.689	0.0563	6.610	0.100
0.95	10.284	10.049	0.055	10.111	0.029
0.99	21.050	21.815	0.585	23.146	4.392
0.999	51.308	55.527	17.799	64.510	174.303
0.9999	142.295	130.846	131.071	167.985	659.994

Table 6.14: *VaR for a mixture of GPDs,* $n = 500$

0.5 GPD(1,0.5)+0.5 GPD(3,0.2)					
		Mixture of GPD's		Classical POT	
γ	True value	VaR_γ	MSE	VaR_γ	MSE
0.9	6.927	6.226	0.492	5.917	1.019
0.95	10.284	9.0813	1.446	9.176	1.228
0.99	21.050	17.514	12.506	21.751	0.491
0.999	51.308	51.604	0.087	64.117	164.084
0.9999	142.295	162.827	421.575	177.479	1237.956

Table 6.15: *VaR for a mixture of GPDs,* $n = 1000$

0.7 GPD(1,0.5)+0.3 GPD(3,0.2)					
		Mixture of GPD's		Classical POT	
γ	True value	VaR_γ	MSE	VaR_γ	MSE
0.9	5.981	6.344	0.132	6.249	0.072
0.95	9.176	9.910	0.539	10.294	1.250
0.99	20.108	23.639	12.471	28.481	70.116
0.999	54.934	70.086	229.608	107.564	2769.970
0.9999	166.313	198.979	1067.089	388.878	4.953e+04

Table 6.16: *VaR for a mixture of GPDs,* $n = 150$

0.7 GPD(1,0.5)+0.3 GPD(3,0.2)					
		Mixture of GPD's		Classical POT	
γ	True value	VaR_γ	MSE	VaR_γ	MSE
0.9	5.981	4.901	1.165	5.488	0.243
0.95	9.176	7.785	1.93	8.762	0.171
0.99	20.108	19.628	0.230	22.380	5.164
0.999	54.934	63.873	79.914	74.375	377.947
0.9999	166.313	196.823	930.841	234.413	4637.639

Table 6.17: *VaR for a mixture of GPDs,* $n = 500$

0.7 GPD(1,0.5)+0.3 GPD(3,0.2)					
		Mixture of GPD's		Classical POT	
γ	True value	VaR_γ	MSE	VaR_γ	MSE
0.9	5.981	6.253	0.074	5.921	0.003
0.95	9.176	10.022	0.715	9.562	0.148
0.99	20.108	22.089	3.927	25.145	25.372
0.999	54.934	48.067	47.143	87.557	1064.293
0.9999	166.313	89.764	5859.797	290.186	1.534e+04

Table 6.18: *VaR for a mixture of GPDs,* $n = 1000$

6.2.1.4 Mixtures of distributions with disjoint supports.

To complete this study, we consider the case of mixtures of distributions with disjoint supports. Castellacci (2012) provides some formulas for calculating high quantiles for this type of mixtures:

Theorem 6.3. *Consider a proper* ($w_i > 0$ *for all* i) *mixture distribution* F *of the form* (6.1) *with* $I = \{1, \ldots, n\}$ *and let* $A_i = \operatorname{supp}(f_i)$ *be the support of* f_i*, where* f_i *denotes the PDF of* F_i*. Assume* $\sup A_i \leq \inf A_{i+1}$ *for* $i = 1, ..., n$*. Then the quantile* $Q(p)$ *of* F *at level* p *equals*

$$Q(p) = F^{-1}(p) = F_1^{-1}\left(\frac{p}{w_1}\right) 1_{B_1}(p) + \cdots + F_n^{-1}\left(\frac{p - \sum_{i=1}^{n-1} w_i}{w_n}\right) 1_{B_n}(p)$$

for any $p \in [0, 1]$*, where*

$$B_i := \left[\sum_{j=0}^{i-1} w_j, \sum_{j=0}^{i} w_j\right) \quad \text{and} \quad B_n := \left[\sum_{j=0}^{n-1} w_j, 1\right]$$

and we have defined $w_0 = 0$*.*

Details of the proof can be found in Castellacci (2012).

If $X \sim F$ represents portfolio losses over a prescribed time horizon, and F is a mixture as in Theorem 6.3, the Value-at-Risk at level c is

$$\mathrm{VaR}_c(X) = \sum_{i=1}^{n} \mathrm{VaR}_{c_i}(X_i) 1_{C_i}(c),$$

where X_i is a loss random variable with distribution F_i corresponding to the ith component of the mixture, $i \in \{1, \ldots, n\}$ and

$$c_i = \frac{c-1+\sum_{j=1}^{i} w_j}{w_i},$$
$$C_i = \left(1 - \sum_{j=1}^{i} w_j,\ 1 - \sum_{j=1}^{i-1} w_j\right].$$

This reformulation of VaR implies that the higher the confidence level, the smaller the c_i of the component whose VaR is selected for the computation. The following example is provided. If $c \in C_1 = (1 - w_1, 1]$,

$$\mathrm{VaR}_c(X) = \mathrm{VaR}_{c_1}(X_1) = -F_1^{-1}\left(\frac{1-c}{w_1}\right).$$

Conversely, given a confidence level c one can pick a weight w_1 such that $1 - c < w_1$. In this case, only the leftmost risk contributes to the VaR of the mixture.

6.3 Concluding remarks

To conclude, we suggest some other directions for future work:

- Inclusion of covariates in the analysis. Extremes are non-stationary in general and they vary with respect to covariates. Several authors have shown the importance of including covariates in the analysis. For example, Davison and Smith (1990) characterize extreme value model parameters in terms of one or more covariates. Chavez-Demoulin and Davison (2005) and Coles (2001) describe a non-homogeneous Poisson model in which ocurrence rates and extremal properties are modelled as functions of covariates. Moreover, as we saw in Chapter 2, threshold selection is an important issue to handle, and it is even more difficult when covariate effects are present, since the threshold may itself be a function of covariates. Thus, it is essential to accommodate covariate effects in the model.

- Extension to the multivariate case. Modelling of multivariate extremes is increasingly important in different fields. Contrary to what happens in the univariate case,

the whole family of MEVD cannot be described parametrically. Due to their flexibility, nonparametric estimation has been proposed in the literature (see Beirlant et al., 2004), however, these models are difficult to implement due to the increasing number of parameters needed to characterize the joint dependence structure accurately. Hence, this is still an active area of research that requires new approaches.

- The use of more sophisticated methods to combine expert opinion. In Chapter 4 we have used basic methodologies for aggregating expert opinion. Nonetheless, aggregation methods range from the simple to the complex. Some authors favor the theoretical elegance of the more sophisticated approaches, while others advocate for the simple methodologies, on the basis of the greater probability of obtaining meaningful results (Jenkinson, 2005). Therefore, we leave open the possibility of using more sophisticated elicitation methods in the analysis of extremes.

- The elicitation of information in the new model framework. Models introduced in Chapter 5 provide different alternatives to handle the discontinuity problem in the model of Behrens et al. (2004); however, we have not supplied information from experts. Thus, it would be interesting to assess the effect of incorporating expert opinion into the analysis.

Appendix A

Markov Chain Monte Carlo methods

The recent explosion in Markov chain Monte Carlo (MCMC) techniques owes largely to their application in Bayesian inference. In MCMC simulation, one constructs a Markov chain long enough for the distribution of the elements to stabilize to a stationary distribution, which is the distribution of interest. By repeatedly simulating steps of the chain, the method simulates draws from the distribution of interest.

MCMC has been widely applied for exploring posterior distributions. That is, through MCMC, one can simulate the entire joint posterior distribution of the unknown quantities and obtain simulation-based estimates that can be used directly for parameter inference and prediction.

There are many ways of constructing the appropriate Markov chain, but all of them, including the Gibbs sampler are special cases of the general framework of Metropolis et al. (1953) and Hastings (1970), or the Metropolis-Hastings algorithm.

A detailed introduction to the Metropolis-Hastings and the Gibbs sampler algorithms can be found in Robert and Casella (1999) and Brooks et al. (2011), among others. In Sections A.1–A.3, we summarize the algorithms presented in those books.

A.1 Metropolis–Hastings sampling

Suppose we wish to draw $\theta = (\theta_1, ..., \theta_d)'$ from a density $\pi(\theta)$ (usually the posterior density). Further, suppose that we have some arbitrary transition kernel $p(\theta_{i+1}, \theta_i)$ (which is easy to simulate from) for iterative simulation of successive values. Then consider the following algorithm:

1. Initialize the iteration counter to $k = 1$, and initialize the chain to $\theta^{(0)}$ at some value;
2. Generate a proposed value θ' using the kernel $p\left(\theta^{(k-1)}, \theta'\right)$;
3. Evaluate the acceptance probability $A\left(\theta^{(k)}, \theta'\right)$ of the proposed move, where

$$A\left(\theta^{(k)}, \theta'\right) = \min\left\{1, \frac{\pi(\theta') L(\theta' \mid x) p(\theta', \theta)}{\pi(\theta) L(\theta \mid x) p(\theta, \theta')}\right\}; \tag{A.1}$$

4. Put $\theta^{(k)} = \theta'$ with probability $A\left(\theta^{(k-1)}, \theta'\right)$, and put $\theta^{(k)} = \theta^{(k-1)}$ otherwise;
5. Change the counter from k to $k + 1$ and return to step 2.

So at each stage, a new value is generated from the proposal distribution. This is either accepted, in which case the chain moves, or rejected, in which case the chain stays where it is. Whether or not the move is accepted or rejected depends on the acceptance probability which itself depends on the relationship between the density of interest and the proposal distribution.

A.2 The Gibbs sampler

Suppose again that $\pi(\theta)$ is the density of interest. If we can draw from various conditional distributions of $\pi(\theta)$, then we can construct a Markov chain that eventually converges to the joint distribution. This is, suppose that

$$\pi(\theta_i \mid \theta_1, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_d) = \pi(\theta_i \mid \theta_{-i}) = \pi_i(\theta_i), \quad i = 1, ..., d \tag{A.2}$$

are available for simulating from (θ_{-i} denotes the parameter vector excluding θ_i). The Gibbs sampler uses the following algorithm:

1. Initialize the iteration counter to $k = 1$. Initialize the state of the chain to $\theta^{(0)} = \left(\theta_1^{(0)}, ..., \theta_d^{(0)}\right)'$ at some arbitrary values;

2. Obtain a new value $\theta^{(k)}$ from $\theta^{(k-1)}$ by successive generation of values

$$\begin{aligned}
\theta_1^{(k)} &\sim \quad \pi\left(\theta_1 \mid \theta_2^{(k-1)}, ..., \theta_d^{(k-1)}\right), \\
\theta_2^{(k)} &\sim \quad \pi\left(\theta_2 \mid \theta_1^{(k)}, \theta_3^{(k-1)}, ..., \theta_d^{(k-1)}\right), \\
&\quad . \qquad\qquad\qquad . \\
&\quad . \qquad\qquad\qquad . \\
&\quad . \qquad\qquad\qquad . \\
\theta_d^{(k)} &\sim \quad \pi\left(\theta_d \mid \theta_1^{(k)}, ..., \theta_{d-1}^{(k)}\right).
\end{aligned}$$

3. Change counter k to $k + 1$, and return to step 2.

Each simulated value depends only on the previous simulated value, and not on any other previous values or the iteration counter k. The Gibbs sampler can be used in isolation if we can readily simulate from the full conditional distributions; however, this is not always the case. Fortunately, the Gibbs sampler can be combined with Metropolis-Hastings schemes when the full conditionals are difficult to simulate from.

A.3 Metropolis-within-Gibbs

The Metropolis-within-Gibbs idea retains the idea of sequential sampling, but uses a Metropolis step on some or all variables rather than attempting to sample from the exact conditional distribution.

That is, we propose a move of the parameter in question from its current position to a new position in the state space (keeping all the other variables fixed); calculate the acceptance probability using the conditional distribution of that variable; decide whether

to accept or reject, and then move onto the next variable. We can mix "pure" Gibbs steps where we sample from the desired conditional distribution of some variables, with Metropolis steps on other variables.

A.4 Convergence diagnostics

Convergence refers to the idea that eventually the simulated Markov chain reaches the stationary distribution. Although, it is never possible to say with certainty that a finite sample from an MCMC algorithm is representative of an underlying stationary distribution, convergence diagnostics offer a worthwhile check on the algorithm's progress. In this section we present some general methods for monitoring convergence. We refer the reader to Brooks and Gelman (1997), Cowles and Carlin (1996), Gelman and Rubin (1992) and Heidelberger and Welch (1983) for a detailed explanation of these methods.

A.4.1 Gelman and Rubin diagnostic

This statistic is based on the idea that after simulating multiple sequences for over-dispersed starting points, the behaviour of all of the chains should be basically the same (Gelman and Rubin, 1992). That is, the variance within the chains should be the same as the variance across the chains.

The procedure to calculate the Gelman-Rubin statistic is as follows:

1. Estimate the model with a variety of different initial values and iterate for an n-iteration burn-in and an n-iteration monitored period.

2. Take the n-monitored draws of m parameters and calculate the statistic $\sqrt{R} = \sqrt{\frac{\hat{V}(\theta)}{W}}$, where

   - $W = \frac{1}{m(n-1)} \sum_{j=1}^{m} \sum_{i=1}^{n} \left(\theta_j^{(i)} - \overline{\theta}_j\right)^2$ is the within-chain variance.

- $B = \frac{n}{m-1} \sum_{j=1}^{m} \left(\theta_j - \overline{\theta}\right)^2$ is the between-chain variance.
- $\hat{V}(\theta) = \left(1 - \frac{1}{n}\right) W + \frac{1}{n} B$ is the estimated variance.

If convergence has been achieved, W and $\hat{V}(\theta)$ should be almost equivalent, so R should approximately equal to 1.

A.4.2 Heidelberg and Welch diagnostic

The Heidelberg and Welch statistic (Heidelberger and Welch, 1983) is based on the Cramér-von Mises statistic and is used to test the hypothesis that the Markov chain is from a stationary distribution. The diagnostic consists of two parts.

- Part I

1. Generate a chain of N iterations and define an $\alpha \in (0, 1)$ level.
2. Calculate the test statistic on the whole chain to accept or reject the null hypothesis of stationarity.
3. If the null hypothesis is rejected, discard the first 10% of the chain and calculate the test statistic to accept or reject the null.
4. If the null hypothesis is rejected, discard the next 10% and calculate the test statistic.
5. Repeat until the null hypothesis is accepted or 50% of the chain is discarded. If the test still rejects the null hypothesis, then the chain fails the test and needs to be run longer.

- Part II

1. If the chain passes the first part of the diagnostic, then it takes the part of the chain not discarded from the first part to test the second part.

2. The halfwidth test calculates half the width of the $(1-\alpha)\%$ credible interval around the mean.

3. If the ratio of the halfwidth and the mean is lower than some ϵ, then the chain passes the test. Otherwise, the chain must be run out longer.

A.4.3 Effective Sample Size

For a series of N observations from a dependent stochastic process, the effective sample size is given by (Chapter 1, Brooks et al. 2011)

$$N_{\text{eff}} = \frac{N}{1 + 2\sum_{k=1}^{\infty} \rho(k)}, \tag{A.3}$$

where $\rho(k)$ is the true lag-k autocorrelation for the Markov chain.

This quantity gives an estimate of the equivalent number of independent iterations that the chain represents.

A.5 Reversible jump MCMC

The reversible jump MCMC (RJMCMC) algorithm was introduced by Green in 1995 (see Green (1995) and Damien et al. (2013, Chapter 7)), and enables us to get a handle on both model selection and parameter estimation in one single algorithm. This is, we are able to traverse the posterior model space, in terms of the models and corresponding parameters, in a single Markov chain.

This algorithm can be seen as an extension of the MH algorithm, with an additional step which moves between the different models.

A.5.1 Model formulation

The model and notation presented in this section are based on Damien et al. (2013, Chapter 7).

Denote the generic Markov transition kernel $Q\left(\theta, d\theta^{'}\right) = P\left(\theta \in d\theta^{'} \mid x, \theta\right)$ and the modified transition kernel $P\left(\theta, d\theta^{'} \mid x\right)$ which has π^* as its unique stationary distribution. Let π and q denote the densities of π^* and Q with respect to Lebesgue measure.

Consider a countable collection of Bayesian models, $\{M_k,\, k = 1, 2, ...\}$, where model M_k is parameterized by θ_k with parameter space $\Theta_k \subset \mathbb{R}^{d_k}$. For data x the full posterior can be written as

$$\pi\left(M_k, \theta_k \mid x\right) = \frac{f\left(x \mid \theta_k, M_k\right) p\left(\theta_k, \mid M_k\right) p\left(M_k\right)}{\sum_j \left\{\int f\left(x \mid \theta_j, M_j\right) p\left(\theta_j, \mid M_j\right) d\theta_j\right\} p\left(M_j\right)}. \tag{A.4}$$

The objective is to construct an aperiodic and irreducible Markov chain on the union parameter space $\Theta = \bigcup_k \Theta_k$.

We assume that at a specific iteration the chain is in model M with a d-dimensional parameter value θ, and the proposal is to move to model $M^{'}$ with a d'-dimensional parameter value $\theta^{'}$. Suppose that, when in model M, the move between the two models is selected with probability $r\left(M, M^{'}\right)$ and consider the introduction of collections of latent variables u and v of dimension d_u and d_v, respectively, so that $d + d_u = d^{'} + d_v$. The proposal density q is then considered for the extended parameter vectors (θ, u) and $(\theta^{'}, v)$ such that $q\left((\theta, u), (\theta^{'}, v) \mid x\right)$ is reversible. This can be seen as a bijective differentiable mapping

$$(\theta^{'}, v) = g\left(\theta, u\right) \Longleftrightarrow (\theta, u) = h(\theta^{'}, v). \tag{A.5}$$

Let ψ and $\psi^{'}$ be the latent-augmented parameter vectors and denote the augmented posterior density by

$$\tilde{\pi}_m\left(\psi \mid x\right) = \tilde{\pi}_m\left(M, \theta, u \mid x\right) = \pi\left(M, \theta \mid x\right) p_u\left(u\right). \tag{A.6}$$

Let $\alpha_m\left(\psi, \psi^{'} \mid x\right)$ denote the acceptance probability for move type m, so that for arbitrary sets A and B,

$$\int_A \tilde{\pi}_m^*\left(d\psi \mid x\right) \int_B \alpha_m\left(\psi, \psi^{'} \mid x\right) Q_m\left(\psi, d\psi^{'} \mid x\right) = \int_B \tilde{\pi}_m^*\left(d\psi' \mid x\right) \int_A \alpha_m\left(\psi, \psi^{'} \mid x\right) Q_m\left(\psi, d\psi^{'} \mid x\right) \tag{A.7}$$

which implies that

$$\alpha_m\left(\psi,\psi^{'}\mid x\right)f_m\left(\psi,\psi^{'}\mid x\right)=\alpha_m\left(\psi^{'},\psi\mid x\right)f_m\left(\psi^{'},\psi\mid x\right), \tag{A.8}$$

where f_m is defined with respect to a common, symmetric measure on the product space.

For the forward move, we have

$$f_m\left(\psi,\psi^{'}\mid x\right)=\tilde{\pi}_m\left(\psi\mid x\right)q_m\left(\psi,\psi^{'}\mid x\right)=\pi\left(M,\theta\mid x\right)p_u\left(u\right)\vec{r}_m, \tag{A.9}$$

where $\vec{r}_m$ represents the probability of choosing to make a move m from M to $M^{'}$.

For the reverse move, we must set

$$f_m\left(\psi^{'},\psi\mid x\right)=\tilde{\pi}_m\left(\psi^{'}\mid x\right)q_m\left(\psi^{'},\psi\mid x\right)=\pi\left(M^{'},\theta^{'}\mid x\right)p_v\left(v\right)\left|\frac{\partial\left(\theta^{'},v\right)}{\partial\left(\theta,u\right)}\right|\overleftarrow{r}_m, \tag{A.10}$$

where

$$\left|\frac{\partial\left(\theta^{'},v\right)}{\partial\left(\theta,u\right)}\right|=\left|\frac{\partial g\left(t_1,t_2\right)}{\partial g\left(\theta,u\right)}\right|_{t_1=\theta,t_2=u} \tag{A.11}$$

is the Jacobian associated with the bijection $g:(\theta,u)\mapsto\left(\theta^{'},v\right)$ and $\overleftarrow{r}_m$ represents the probability of choosing to make a move m from $M^{'}$ to M. So the augmented posterior becomes

$$\tilde{\pi}\left(M^{'},\theta^{'},v\mid x\right)=\tilde{\pi}\left(M,h\left(\theta^{'},v\right)\mid x\right)\left|J\left(\theta^{'},v\right)\right|=\tilde{\pi}\left(M,\theta,u\mid x\right)\left|J\left(\theta,u\right)\right|^{-1}. \tag{A.12}$$

The acceptance probability for move type m is

$$\alpha_m\left(\left(M,\theta\right),\left(M^{'},\theta^{'}\right)\mid x\right)=\min\left\{1,\frac{\pi\left(M^{'},\theta^{'}\mid x\right)p_v\left(v\right)\overleftarrow{r}_m}{\pi\left(M,\theta\mid x\right)p_u\left(u\right)\vec{r}_m}\right\}\left|\frac{\partial\left(\theta^{'},v\right)}{\partial\left(\theta,u\right)}\right|. \tag{A.13}$$

This probability can be simplified depending on whether $\dim\left(\theta\right)>\dim\left(\theta'\right)$ or $\dim\left(\theta\right)<\dim\left(\theta'\right)$.

In general, within each iteration of the Markov chain, the algorithm involves two steps:

1. Update the parameters, conditional on the model using the MH algorithm and
2. Update the model, conditional on the current parameter values by proposing to move to a different model with some given parameter values and accepting this proposed move with the probability above.

Further details on this algorithm can be found in Green (1995) and Tierney (1998).

Appendix B

Jeffreys prior for the GPD

Let $Y \sim GPD\,(\sigma, \xi)$. The Jeffreys prior for a GPD with parameter $\theta = (\sigma, \xi)$ is given by

$$J\,(\theta) \propto \sqrt{|I\,(\theta)|}, \tag{B.1}$$

where $I\,(\theta) = -E\,[\partial^2 \ln f\,(y \mid \theta)\,/\partial^2 \theta]$.

The density of the GPD(σ, ξ) is

$$f\,(y \mid \theta) = \frac{1}{\sigma}\left(1 + y\frac{\xi}{\sigma}\right)^{-\frac{1}{\xi}-1},$$

where the support is

$$\begin{cases} y > 0 & \text{if} \quad \xi \geq 0, \\ 0 \leq y \leq -\sigma/\xi & \text{if} \quad \xi < 0. \end{cases}$$

The log of $f\,(y \mid \theta)$ is thus given by

$$L\,(y \mid \theta) = \ln f\,(y \mid \theta) = -\ln \sigma - \left(\frac{1}{\xi} + 1\right) \ln\,\left(1 + \frac{\xi}{\sigma} y\right). \tag{B.2}$$

The first derivative with respect to ξ is

$$\frac{\partial L\,(y \mid \theta)}{\partial \xi} = \left(\frac{1}{\xi^2}\right) \ln\left(1 + \frac{\xi}{\sigma} y\right) + \left(-\frac{1}{\xi} - 1\right)\left(\frac{y}{\sigma}\right)\left(1 + \frac{\xi}{\sigma} y\right)^{-1}. \tag{B.3}$$

We can write

$$y = -\frac{\sigma}{\xi}\left[1 - \left(1 + \frac{\xi}{\sigma} y\right)\right]. \tag{B.4}$$

Substituting (B.4) in (B.3) we get:

$$\begin{aligned}
&\frac{\partial L\left(y \mid \theta\right)}{\partial \xi} \\
&= \left(\frac{1}{\xi^2}\right) \ln\left(1+\frac{\xi}{\sigma}y\right) + \left(-\frac{1}{\xi}-1\right)\left(-\frac{1}{\sigma}\right)\left(\frac{\sigma}{\xi}\right)\left[1-\left(1+\frac{\xi}{\sigma}y\right)\right]\left(1+\frac{\xi}{\sigma}y\right)^{-1} \\
&= \left(\frac{1}{\xi^2}\right) \ln\left(1+\frac{\xi}{\sigma}y\right) + \left(-\frac{1}{\xi}-1\right)\left[-\frac{1}{\xi}+\frac{1}{\xi}\left(1+\frac{\xi}{\sigma}y\right)\right]\left(1+\frac{\xi}{\sigma}y\right)^{-1} \\
&= \left(\frac{1}{\xi^2}\right) \ln\left(1+\frac{\xi}{\sigma}y\right) + \left(-\frac{1}{\xi}-1\right)\left(\frac{1}{\xi}\right)\left[1-\left(1+\frac{\xi}{\sigma}y\right)^{-1}\right].
\end{aligned} \tag{B.5}$$

Now the first derivative with respect to σ is, using (B.4),

$$\begin{aligned}
\frac{\partial L\left(y \mid \theta\right)}{\partial \sigma} &= -\frac{1}{\sigma} + \left(-\frac{1}{\xi}-1\right)\left(-\frac{\xi y}{\sigma^2}\right)\left(1+\frac{\xi}{\sigma}y\right)^{-1} \\
&= -\frac{1}{\sigma} + \left(-\frac{1}{\xi}-1\right)\left(\frac{\xi}{\sigma^2}\right)\left(\frac{\sigma}{\xi}\right)\left[1-\left(1+\frac{\xi}{\sigma}y\right)\right]\left(1+\frac{\xi}{\sigma}y\right)^{-1} \\
&= -\frac{1}{\sigma} + \left(\frac{1}{\sigma}\right)\left(-\frac{1}{\xi}-1\right)\left[1-\left(1+\frac{\xi}{\sigma}y\right)\right]\left(1+\frac{\xi}{\sigma}y\right)^{-1} \\
&= -\frac{1}{\sigma} + \left(-\frac{1}{\sigma\xi}-\frac{1}{\sigma}\right)\left[\left(1+\frac{\xi}{\sigma}y\right)^{-1}-1\right] \\
&= -\frac{1}{\sigma} - \frac{1}{\sigma\xi}\left(1+\frac{\xi}{\sigma}y\right)^{-1} - \frac{1}{\sigma}\left(1+\frac{\xi}{\sigma}y\right)^{-1} + \frac{1}{\sigma} + \frac{1}{\sigma\xi} \\
&= \frac{1}{\sigma\xi} + \frac{1}{\sigma}\left(-\frac{1}{\xi}-1\right)\left(1+\frac{\xi}{\sigma}y\right)^{-1}.
\end{aligned} \tag{B.6}$$

Setting $k = 1+\dfrac{\xi}{\sigma}y$, the second derivative with respect to ξ is as follows

$$\frac{\partial^2 L\left(y \mid \theta\right)}{\partial \xi^2} = \frac{\partial}{\partial \xi}\frac{1}{\xi^2}\ln k - \frac{1}{\xi^2} - \frac{1}{\xi} + \frac{1}{\xi^2}k^{-1} + \frac{1}{\xi}k^{-1}. \tag{B.7}$$

The derivative of k with respect to ξ is, using (B.4),

$$\frac{dk}{d\xi} = \frac{y}{\sigma} = \frac{1}{\sigma}\left[-\frac{\sigma}{\xi}\left(1-k\right)\right] = -\frac{1}{\xi}\left(1-k\right). \tag{B.8}$$

Using B.8,

$$
\begin{aligned}
&\frac{\partial^2 L\left(y \mid \theta\right)}{\partial \xi^2} \\
&= -\frac{2}{\xi^3}\ln k + \frac{1}{\xi^2}\frac{\left[-\frac{1}{\xi}\left(1-k\right)\right]}{k} + \frac{2}{\xi^3} + \frac{1}{\xi^2} - \frac{2}{\xi^3}k^{-1} - \frac{1}{\xi^2}k^{-2}\left(-\frac{1}{\xi}\left(1-k\right)\right) \\
&\quad - \frac{1}{\xi^2}k^{-1} - \frac{1}{\xi}k^{-2}\left(-\frac{1}{\xi}\left(1-k\right)\right) \\
&= -\frac{2}{\xi^3}\ln k - \frac{1}{\xi^3}k^{-1} + \frac{1}{\xi^3} + \frac{2}{\xi^3} + \frac{1}{\xi^2} - \frac{2}{\xi^3}k^{-1} + \frac{1}{\xi^3}k^{-2} - \frac{1}{\xi^3}k^{-1} - \frac{1}{\xi^2}k^{-1} \\
&\quad + \frac{1}{\xi^2}k^{-2} - \frac{1}{\xi^2}k^{-1} \\
&= -\frac{2}{\xi^3}\ln k + \frac{1}{\xi^2} + \frac{3}{\xi^3} - \frac{4}{\xi^3}k^{-1} - \frac{2}{\xi^2}k^{-1} + \frac{1}{\xi^2}k^{-2} + \frac{1}{\xi^3}k^{-2} \\
&= -\frac{2}{\xi^3}\ln k + \frac{3+\xi}{\xi^3} - \frac{2\left(2+\xi\right)}{\xi^3}k^{-1} + \frac{\left(1+\xi\right)}{\xi^3}k^{-2} \\
&= -\frac{2}{\xi^3}\ln\left(1+\frac{\xi}{\sigma}y\right) + \frac{3+\xi}{\xi^3} - \frac{2\left(2+\xi\right)}{\xi^3}\left(1+\frac{\xi}{\sigma}y\right)^{-1} + \frac{\left(1+\xi\right)}{\xi^3}\left(1+\frac{\xi}{\sigma}y\right)^{-2}. \qquad \text{(B.9)}
\end{aligned}
$$

The second derivative with respect to σ is

$$
\frac{\partial^2 L\left(y \mid \theta\right)}{\partial \sigma} = \frac{\partial}{\partial \sigma}\frac{1}{\sigma\xi} + \left(-\frac{1}{\xi} - 1\right)\frac{1}{\sigma}k^{-1}. \qquad \text{(B.10)}
$$

The derivative of k with respect to σ is, using (B.4),

$$
\frac{dk}{d\sigma} = -\frac{\xi y}{\sigma^2} = -\frac{\xi}{\sigma^2}\left[-\frac{\sigma}{\xi}\left(1-k\right)\right] = \frac{1}{\sigma}\left(1-k\right). \qquad \text{(B.11)}
$$

Using (B.11)

$$
\begin{aligned}
\frac{\partial^2 L\left(y \mid \theta\right)}{\partial \sigma} &= -\frac{1}{\sigma^2\xi} + \left(-\frac{1}{\xi} - 1\right)\left[-\frac{1}{\sigma^2}k^{-1} - \frac{1}{\sigma}k^{-2}\left(\frac{1}{\sigma}\left(1-k\right)\right)\right] \\
&= -\frac{1}{\sigma^2\xi} + \left(-\frac{1}{\xi} - 1\right)\left(-\frac{1}{\sigma^2}k^{-2}\right) \qquad \text{(B.12)} \\
&= -\frac{1}{\sigma^2\xi} - \frac{1}{\sigma^2}\left(-\frac{1}{\xi} - 1\right)\left(1+\frac{\xi}{\sigma}y\right)^{-2}.
\end{aligned}
$$

Furthermore,

$$\begin{aligned}
&\frac{\partial^2 L\left(y \mid \theta\right)}{\partial\sigma\partial\xi} \\
&= \frac{\partial}{\partial\xi}\left(\frac{1}{\sigma\xi} - \frac{1}{\sigma\xi}k^{-1} - \frac{1}{\sigma}k^{-1}\right) \\
&= \frac{1}{\sigma}\left(-\frac{1}{\xi^2}\right) + \frac{1}{\sigma}\frac{1}{\xi^2}k^{-1} - \frac{1}{\sigma\xi}\left[-k^{-2}\left(-\frac{1}{\xi}\left(1-k\right)\right)\right] - \frac{1}{\sigma}\left[-k^{-2}\left(-\frac{1}{\xi}\left(1-k\right)\right)\right] \\
&= -\frac{1}{\sigma\xi^2} + \frac{1}{\sigma\xi^2}k^{-1} - \frac{1}{\sigma\xi^2}k^{-2} + \frac{1}{\sigma\xi^2}k^{-1} - \frac{1}{\sigma\xi}k^{-2} + \frac{1}{\sigma\xi}k^{-1} \\
&= -\frac{1}{\sigma\xi^2} + \frac{1}{\sigma\xi^2}\left(2+\xi\right)k^{-1} - \frac{1}{\sigma\xi^2}\left(1+\xi\right)k^{-2} \\
&= -\frac{1}{\sigma\xi^2} + \frac{1}{\sigma\xi^2}\left(2+\xi\right)\left(1+\frac{\xi}{\sigma}y\right)^{-1} - \frac{1}{\sigma\xi^2}\left(1+\xi\right)\left(1+\frac{\xi}{\sigma}y\right)^{-2}.
\end{aligned} \tag{B.13}$$

Next, we calculate

$$I\left(\theta\right) = -E\left[\frac{\partial^2 \ln f\left(y \mid \theta\right)}{\partial^2\theta}\right].$$

For $\xi < 0.5$, we have the following results (Davison and Smith,1990)

$$E\left[\left(1+\frac{\xi}{\sigma}Y\right)^{r}\right] = \frac{1}{1-r\xi} \tag{B.14}$$

and

$$E\left[\ln\left(1+\frac{\xi}{\sigma}Y\right)\right] = \xi. \tag{B.15}$$

Then

$$
\begin{aligned}
E\left(\frac{\partial^2 L\left(Y \mid \theta\right)}{\partial^2 \xi}\right) &= -\frac{2}{\xi^3} E\left[\ln\left(1+\tfrac{\xi}{\sigma}Y\right)\right] + \frac{3+\xi}{\xi^3} - \frac{2\left(2+\xi\right)}{\xi^3} E\left[\left(1+\frac{\xi}{\sigma}Y\right)^{-1}\right] \\
&\quad + \frac{\left(1+\xi\right)}{\xi^3} E\left[\left(1+\frac{\xi}{\sigma}Y\right)^{-2}\right] \\
&= -\frac{2}{\xi^3}\left(\xi\right) + \frac{3+\xi}{\xi^3} - \frac{2\left(2+\xi\right)}{\xi^3}\frac{1}{1+\xi} + \frac{\left(1+\xi\right)}{\xi^3}\frac{1}{1+2\xi} \\
&= -\frac{2\xi}{\xi^3} + \frac{1}{\xi^3}\left[\frac{\left(3+\xi\right)\left(1+\xi\right)\left(1+2\xi\right) - 2\left(2+\xi\right)\left(1+2\xi\right) + \left(1+\xi\right)^2}{\left(1+\xi\right)\left(1+2\xi\right)}\right] \\
&= -\frac{2\xi}{\xi^3} + \frac{1}{\xi^3}\left[\frac{2\xi^3+9\xi^2+10\xi+3-4\xi^2-10\xi-4+\xi^2+2\xi+1}{\left(1+\xi\right)\left(1+2\xi\right)}\right] \\
&= -\frac{2\xi}{\xi^3} + \frac{1}{\xi^3}\left[\frac{2\xi^3+6\xi^2+2\xi}{\left(1+\xi\right)\left(1+2\xi\right)}\right] \\
&= \frac{1}{\xi^3}\left[\frac{-2\xi\left(1+\xi\right)\left(1+2\xi\right) + \left(2\xi^3+6\xi^2+2\xi\right)}{\left(1+\xi\right)\left(1+2\xi\right)}\right] \\
&= \frac{1}{\xi^3}\left[\frac{-4\xi^3-6\xi^2-2\xi+2\xi^3+6\xi^2+2\xi}{\left(1+\xi\right)\left(1+2\xi\right)}\right] \\
&= \frac{1}{\xi^3}\left[\frac{-2\xi^3}{\left(1+\xi\right)\left(1+2\xi\right)}\right] \\
&= -\frac{2}{\left(1+\xi\right)\left(1+2\xi\right)}
\end{aligned}
\tag{B.16}
$$

and

$$
\begin{aligned}
E\left(\frac{\partial^2 L\left(Y \mid \theta\right)}{\partial^2 \sigma}\right) &= -\frac{1}{\sigma^2\xi} - \frac{1}{\sigma^2}\left(-\frac{1}{\xi}-1\right) E\left[\left(1+\frac{\xi}{\sigma}Y\right)^{-2}\right] \\
&= -\frac{1}{\sigma^2\xi} + \left(\frac{1}{\sigma^2\xi}+\frac{1}{\sigma^2}\right)\left(\frac{1}{1+2\xi}\right) \\
&= -\frac{1}{\sigma^2\xi} + \frac{1}{\sigma^2\xi\left(1+2\xi\right)} + \frac{1}{\sigma^2\left(1+2\xi\right)} \\
&= \frac{1}{\sigma^2}\left[-\frac{1}{\xi} + \frac{1}{\xi\left(1+2\xi\right)} + \frac{1}{1+2\xi}\right] \\
&= \frac{1}{\sigma^2}\left[-\frac{1}{\xi} + \frac{1}{\xi\left(1+2\xi\right)} + \frac{1}{1+2\xi}\right] \\
&= \frac{1}{\sigma^2}\left[\frac{-1-2\xi+1+\xi}{\xi\left(1+2\xi\right)}\right] \\
&= \frac{1}{\sigma^2}\left[\frac{-\xi}{\xi\left(1+2\xi\right)}\right] \\
&= -\frac{1}{\sigma^2\left(1+2\xi\right)}.
\end{aligned}
\tag{B.17}
$$

$$\begin{aligned}
E\left(\frac{\partial^2 L\left(Y \mid \theta\right)}{\partial\xi\partial\sigma}\right) &= -\frac{1}{\sigma\xi^2} + \frac{1}{\sigma\xi^2}\left(2+\xi\right) E\left[\left(1+\frac{\xi}{\sigma}y\right)^{-1}\right] - \frac{1}{\sigma\xi^2}\left(1+\xi\right) E\left[\left(1+\frac{\xi}{\sigma}y\right)^{-2}\right] \\
&= -\frac{1}{\sigma\xi^2} + \frac{1}{\sigma\xi^2}\left(2+\xi\right)\left(\frac{1}{1+\xi}\right) - \frac{1}{\sigma\xi^2}\left(1+\xi\right)\left(\frac{1}{1+2\xi}\right) \\
&= -\frac{1}{\sigma\xi^2}\left[1 + \left(1+\xi\right)\left(\frac{1}{1+2\xi}\right) - \left(2+\xi\right)\left(\frac{1}{1+\xi}\right)\right] \\
&= -\frac{1}{\sigma\xi^2}\left[\frac{\left(1+2\xi\right)\left(1+\xi\right) + \left(1+\xi\right)^2 - \left(2+\xi\right)\left(1+2\xi\right)}{\left(1+2\xi\right)\left(1+\xi\right)}\right] \\
&= -\frac{1}{\sigma\xi^2}\left[\frac{2\xi^2+3\xi+1+\xi^2+2\xi+1-2\xi^2-5\xi-2}{\left(1+2\xi\right)\left(1+\xi\right)}\right] \\
&= -\frac{1}{\sigma\xi^2}\left[\frac{\xi^2}{\left(1+2\xi\right)\left(1+\xi\right)}\right] \\
&= -\frac{1}{\sigma\left(1+2\xi\right)\left(1+\xi\right)}.
\end{aligned} \tag{B.18}$$

From these equations,

$$I\left(\theta\right) = I\left(\sigma,\xi\right) = \begin{bmatrix} \dfrac{2}{\left(1+\xi\right)\left(1+2\xi\right)} & \dfrac{1}{\sigma\left(1+2\xi\right)\left(1+\xi\right)} \\ \dfrac{1}{\sigma\left(1+2\xi\right)\left(1+\xi\right)} & \dfrac{1}{\sigma^2\left(1+2\xi\right)} \end{bmatrix} \tag{B.19}$$

and

$$\begin{aligned}
\left|I\left(\sigma,\xi\right)\right| &= \frac{2}{\left(1+\xi\right)\left(1+2\xi\right)}\frac{1}{\sigma^2\left(1+2\xi\right)} - \left[\frac{1}{\sigma\left(1+2\xi\right)\left(1+\xi\right)}\right]^2 \\
&= \frac{2}{\sigma^2\left(1+\xi\right)\left(1+2\xi\right)^2} - \frac{1}{\sigma^2\left(1+\xi\right)^2\left(1+2\xi\right)^2} \\
&= \frac{1}{\sigma^2\left(1+\xi\right)^2\left(1+2\xi\right)^2}\left[2\left(1+\xi\right)-1\right] \\
&= \frac{1}{\sigma^2\left(1+\xi\right)^2\left(1+2\xi\right)}.
\end{aligned} \tag{B.20}$$

Hence, the Jeffreys prior is

$$\begin{aligned}
J\left(\theta\right) \propto \sqrt{\left|I\left(\theta\right)\right|} = \sqrt{\frac{1}{\sigma^2\left(1+\xi\right)^2\left(1+2\xi\right)}} &= \frac{1}{\sigma\left(1+\xi\right)\sqrt{1+2\xi}} \\
&= \sigma^{-1}\left(1+\xi\right)^{-1}\left(1+2\xi\right)^{-1/2}.
\end{aligned} \tag{B.21}$$

Appendix C

Algorithm for the Bayesian model[1]

Simulations are done via Metropolis-Hastings steps within blockwise MCMC. Suppose that at iteration $j-1$, the chain is positioned at $\theta^{(j-1)} = (\alpha^{(j-1)}, \beta^{(j-1)}, u^{(j-1)}, \sigma^{(j-1)}, \xi^{(j-1)})$.

Then, at iteration j, the algorithm cycles through the following steps:

- $\xi^{(j)}$

ξ^* is sampled from the following candidate distribution

$$\xi^{(j)} \sim N\left(\xi^{(j-1)}, V_{\xi^{(j-1)}}\right) I\left(\Xi\right),$$

where $V_{\xi^{(j-1)}}$ is an approximation based on the curvature at the conditional posterior mode and $\Xi = \left[-\sigma^{(j-1)} / \left(M - u^{(j-1)}\right), \infty\right]$ is given by the support of the GPD function, and $M = \max(x_1, ..., x_n)$, the maximum value of the data.

- $\sigma^{(j)}$

The candidate distribution of σ depends on the current value of ξ. If ξ is positive, no restrictions are imposed on σ. Otherwise, σ must be drawn in an appropriate region given by the support of the GPD. That is,

If $\xi^{(j)} \geq 0$: $\sigma^{(j)} \sim Ga\left(a_\sigma, b_\sigma\right)$, where $a_\sigma / b_\sigma = \sigma^{(j-1)}$ and $a_\sigma / b_\sigma^2 = V_{\sigma^{(j-1)}}$. So, $\sigma^{(j)}$ is centered around $\sigma^{(j-1)}$ with some variance $V_{\sigma^{(j-1)}}$;

[1]This algorithm has been taken from Behrens et al. (2004).

If $\xi^{(j)} < 0 :\ \sigma^{(j)} \sim N\left(\sigma^{(j-1)}, V_{\sigma^{(j-1)}}\right) I\left(\Sigma\right)$, a truncated normal distribution, where Σ is given by the support of the GPD with lower bound at $\Sigma = \left[-\xi^{(j-1)}\left(M - u^{(j-1)}, \infty\right)\right]$, $V_{\sigma^{(j-1)}}$ is an approximation for the concavity in the conditional posterior mode.

- $u^{(j)}$

As for σ, the candidate distribution used for u depends on the current value of ξ as well as on σ. When we consider a continuous prior, the candidate distributions are

If $\xi^{(j)} \geq 0 :\ u^{(j)} \sim N\left(u^{(j-1)}, V_{u^{(j-1)}}\right) I\left(A\right)$, a normal distribution truncated on $A = (m, M)$, with M being the largest, as before, and m the smallest observations, respectively.

If $\xi^{(j)} < 0 :\ u^{(j)} \sim N\left(u^{(j-1)}, V_{u^{(j-1)}}\right) I\left(A\right)$, a normal distribution truncated on $A = (a_u, M)$, with $a_u = M + \sigma^{(j-1)}/\xi^{(j-1)}$. Again, $V_{u^{(j-1)}}$ is a value for the variance which is tuned to allow appropriate chain movements.

If u has a discrete prior, we have to follow the same model restrictions as before. Therefore, the candidate distribution is:

If $\xi^{(j)} \geq 0 :\ u^{(j)} \sim U_d\left(q_1, q_2\right)$, a discrete uniform distribution on data quantiles from q_1 to q_2, where q_2 can be any high quantile, like the largest observation M, and q_1 can be any quantile below q_2. It is important to keep in mind that q_1 must be reasonable to prevent the model bias and to respect the asymptotic properties of the model.

If $\xi^{(j)} < 0 :\ u^{(j)} \sim U_d\left(q_1, q_2\right)$, a discrete uniform distribution with q_2 as before, but q_1 have to respect the model restrictions as shown above: $q_1 \geq M + \sigma^{(j-1)}/\xi^{(j-1)}$

- $\alpha^{(j)}, \beta^{(j)}$

$\alpha^{(j)} \sim \log N\left(\alpha^{(j-1)}, V_{\alpha^{(j-1)}}\right)$, with $V_{\alpha^{(j-1)}}$ given by an approximation for the curvature at the conditional posterior mode and $\beta^{(j)} \sim GI\left(a_\beta, b_\beta\right)$, an inverse Gamma distribution centered at $\beta^{(j-1)}$ and with variance $V_{\beta^{(j-1)}}$ given by the approximation for the curvature at the conditional posterior mode.

Appendix D

Non-informative prior plots

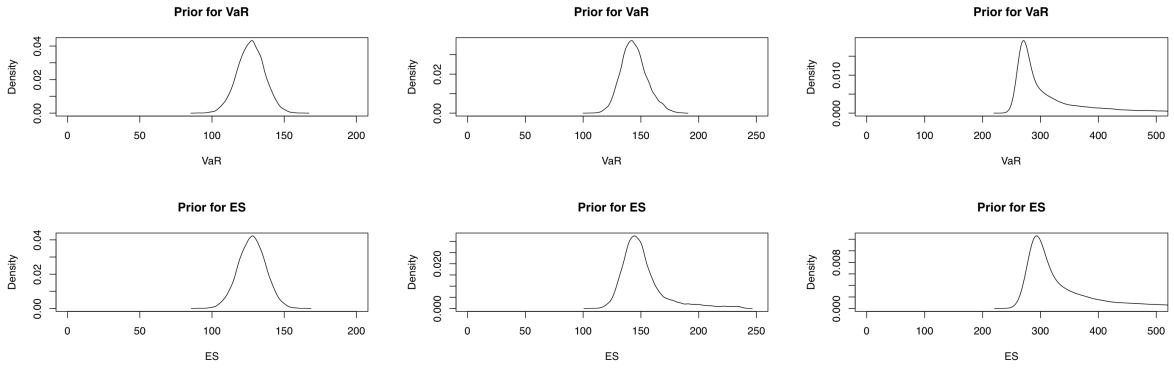

Figure D.1: *Fraud data:* $\sigma = 1,\ 10,\ 100$, *respectively;* $u \sim TN(\min(data))$; $\xi \sim$ *Jeffreys prior*

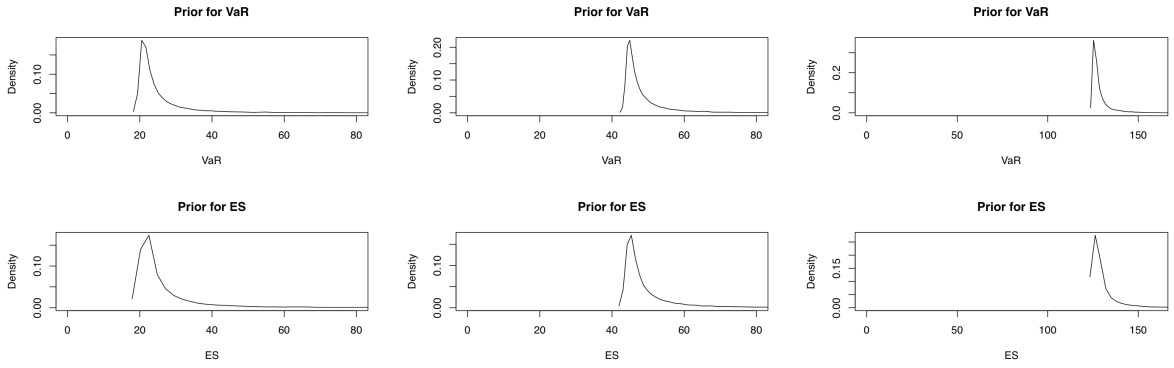

Figure D.2: *Fraud data:* $\sigma \sim LN(0, 1.25)$*;* $u = Q_{0.5}(19.89)$, $Q_{0.7}(43.77)$, $Q_{0.9}(124)$, *respectively;* $\xi \sim$ *Jeffreys prior*

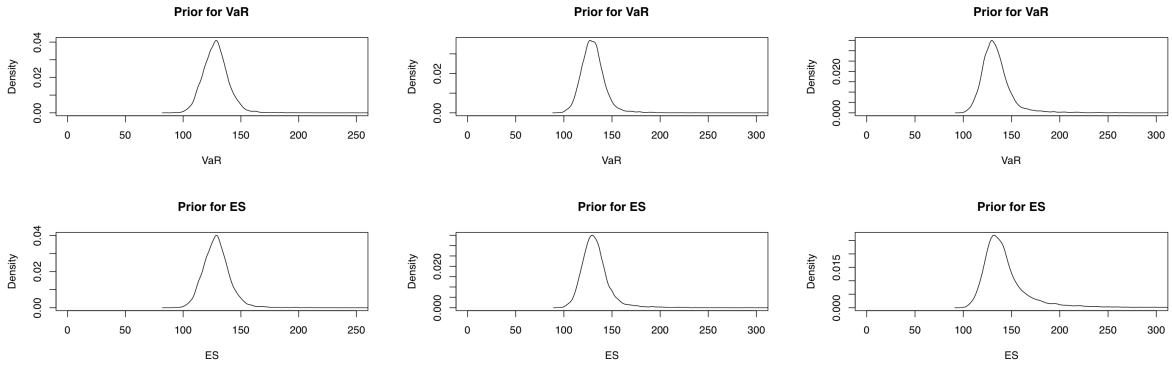

Figure D.3: *Fraud data:* $\sigma \sim LN(0, 1.25)$ *;* $u \sim TN(\min(data))$*;* $\xi = -0.3, 0.05, 0.45$, *respectively*

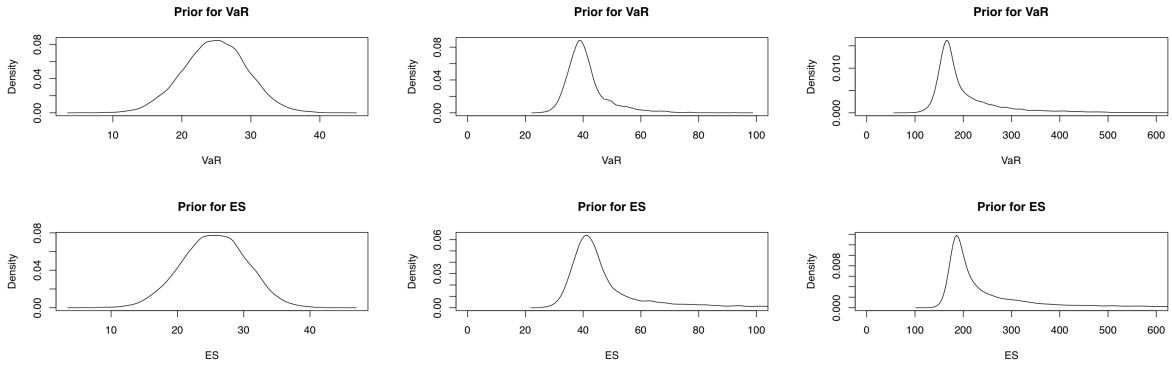

Figure D.4: *Exponential(0.1):* $\sigma = 1, 10, 100$, *respectively;* $u \sim TN(\min(data))$,$\xi \sim$ *Jeffreys prior*

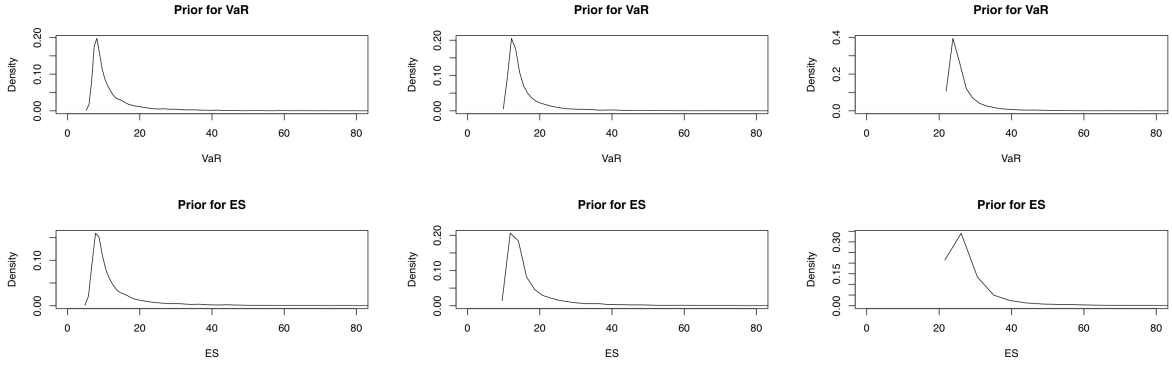

Figure D.5: *Exponential(0.1):* $\sigma \sim LN(0, 1.25)$*;* $u = Q_{0.5}(6.89)$, $Q_{0.7}(11.35)$, $Q_{0.9}(23.03)$, *respectively;* $\xi \sim$ *Jeffreys prior*

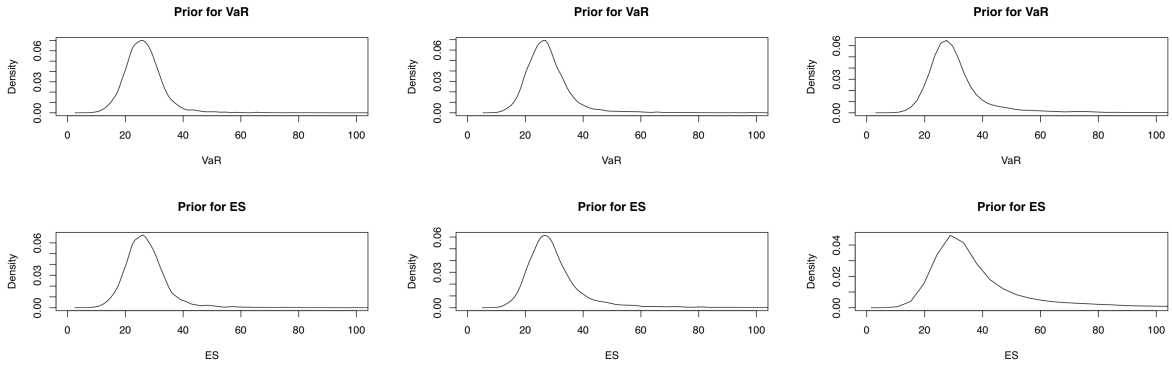

Figure D.6: *Exponential(0.1):* $\sigma \sim LN(0, 1.25)$ *;* $u \sim TN(\min(data))$*;* $\xi = -0.3,\ 0.05,\ 0.45$*, respectively*

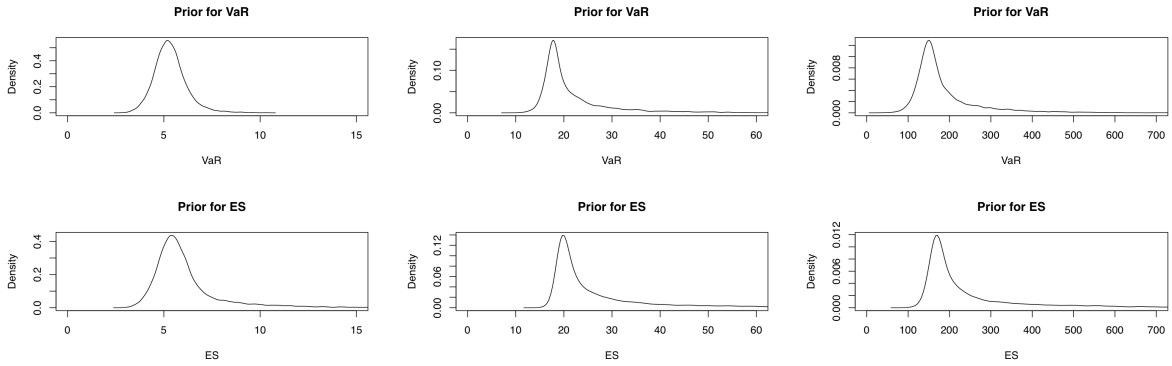

Figure D.7: *Log-Normal(0,1):* $\sigma = 1,\ 10,\ 100$*, respectively;* $u \sim TN(\min(data))$*,* $\xi \sim$ *Jeffreys prior*

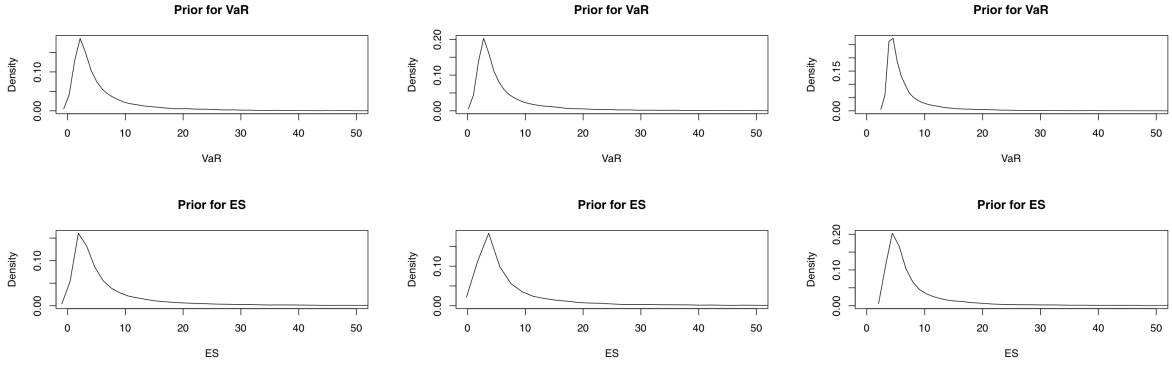

Figure D.8: *Log-Normal(0,1):* $\sigma \sim LN(0, 1.25)$*;* $u = Q_{0.5}(1.009)\,,\ Q_{0.7}(1.69)\,,\ Q_{0.9}(3.49)\,,$ *respectively;* $\xi \sim$ *Jeffreys prior*

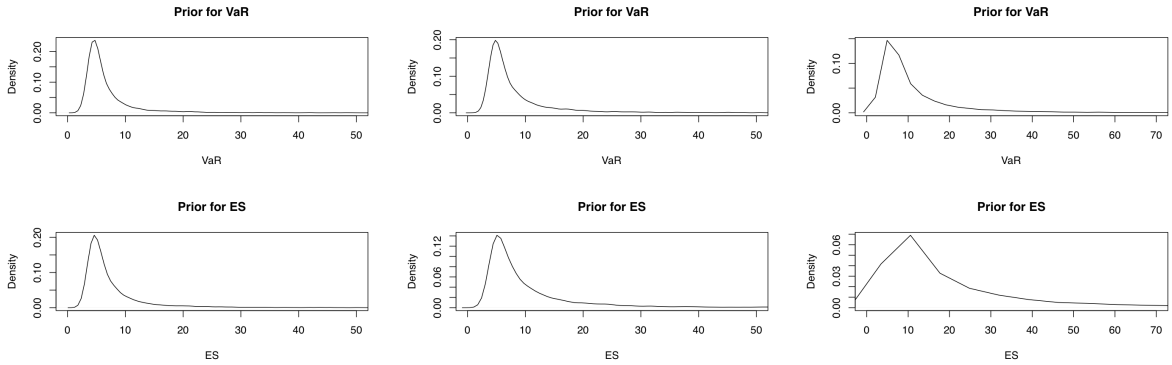

Figure D.9: *Log-Normal(0,1):* $\sigma \sim LN(0, 1.25)$ *;* $u \sim TN(\min(data))$*;* $\xi = -0.3,\ 0.05,\ 0.45$*, respectively*

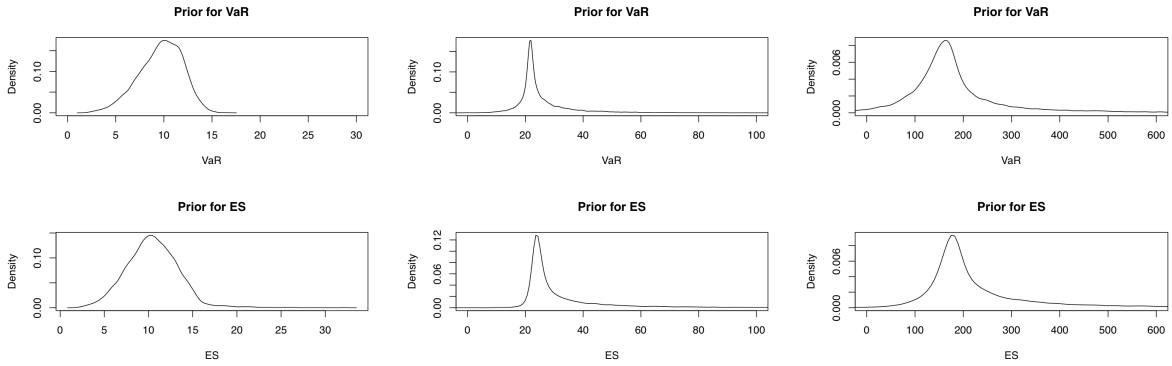

Figure D.10: *Gamma(2,0.5):* $\sigma = 1,\ 10,\ 100$*, respectively;* $u \sim TN(\min(data))$*,* $\xi \sim$ *Jeffreys prior*

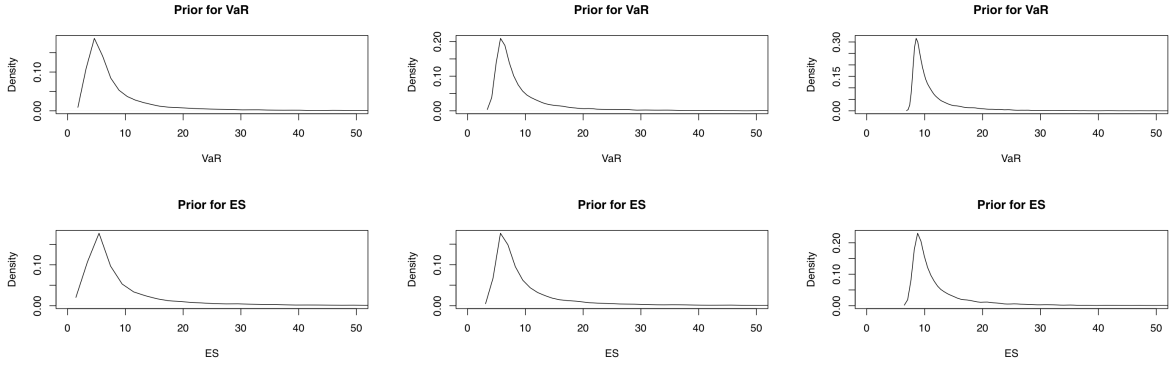

Figure D.11: *Gamma(2,0.5):* $\sigma \sim LN(0, 1.25)$*;* $u = Q_{0.5}(3.4)\,, Q_{0.7}(4.85)\,, Q_{0.9}(7.91)\,$*, respectively;* $\xi \sim$ *Jeffreys prior*

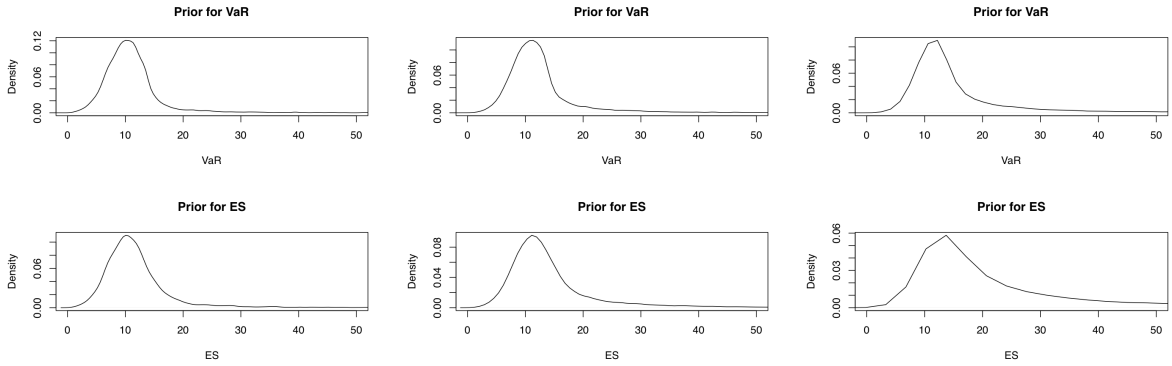

Figure D.12: *Gamma(2,0.5):* $\sigma \sim LN(0, 1.25)$ *;* $u \sim TN(\min(data))$*;* $\xi = -0.3,\ 0.05,\ 0.45,$ *respectively*

Bibliography

[1] Agostini, A., Talamo, P. and Vecchione, V. (2010): *"Combining operational loss data with expert opinions through advanced credibility theory"*. The Journal of Operational Risk, 5(1): 2–28.

[2] Antoniak, C. (1974): *"Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems"*. The Annals of Statistics 2: 1152–1174.

[3] Arora, S., Hazan, E. and Kale, S. (2012): *"The multiplicative weights update method: a meta-algorithm and applications"*. Theory of Computing, 8(1): 121–164.

[4] Basel Committee on Banking Supervision (2004): *"International convergence of capital measurement and capital standards"*. Available online: http://www.bis.org/publ/bcbs107.pdf.

[5] Bee, M., Espa, G. and Benedetti, R. (2013): *"On maximum likelihood estimation of a Pareto mixture"* . Computational Statistics and Data Analysis, 28(1): 161–178.

[6] Behrens, C. N., Lopes, H. F., and Gamerman, D. (2004): *"Bayesian analysis of extreme events with threshold estimation*". Statistical Modelling, 4: 227–244.

[7] Beirlant, J., Goegebeur, Y., Segers, J. and Teugels, J. (2004): *"Statistics of extremes: Theory and applications"*. Wiley, London.

[8] Beirlant, J., Vynckier, P. and Teugels, J. (1996): *"Excess functions and estimation of the extreme-value index"*. Bernoulli, 2(4): 293–318.

[9] Bermudez, P., Turkman, M.A. and Turkman, K.F. (2001): *"A predictive approach to tail probability estimation"*. Extremes, 4: 295–314.

[10] Bernardo, J.M. (1997): *"Noninformative priors do not exist: A discussion"*. Journal of Statistical Planning and Inference, 65: 159–189.

[11] Bortot, P. and Gaetan, C. (2013): *"A latent process model for temporal extremes"*. Scandinavian Journal of Statistics, forthcoming.

[12] Brooks, S. and Gelman, A. (1997): *"General methods for monitoring convergence of iterative simulations"*. Journal of Computational and Graphical Statistics, 7: 434–455.

[13] Brooks, S., Gelman, A., Jones, G. L., and Meng, X.L. (2011): *"Handbook of Markov Chain Monte Carlo"*. Chapman and Hall, London.

[14] Cabras, S. and Castellanos, M.E. (2010): *"An objective Bayesian approach for threshold estimation in the peaks over the threshold model"*. Improving Risk Management. Technical Report TR2010.10.

[15] Cabras, S., Castellanos, M. E., and Gamerman, D. (2011): *"A default Bayesian approach for regression on extremes"*. Statistical Modelling, 11 (6): 557–580.

[16] Carreau, J. and Bengio, Y. (2009): *"A hybrid Pareto model for asymmetric fat-tailed data: The univariate case"*. Extremes, 12: 53–76.

[17] Castellacci, G. (2012): *"A formula for the quantiles of mixtures of distributions with disjoint supports"*. Available online: http://ssrn.com/abstract=2055022.

[18] Castellanos, M.E. and Cabras, S. (2007): *"A default Bayesian procedure for the generalized Pareto distribution"*. Journal of Statistical Planning and Inference, 137(2): 473–483.

[19] Chavez-Demoulin, V. and Davison, A.C. (2005): *"Generalized additive modelling of sample extremes"*. Journal of the Royal Statistical Society C, 54: 207–222.

[20] Chavez-Demoulin, V., Embrechts, P. and Nešlehová, J. (2006): *"Quantitative models for operational risk: Extremes, dependence and aggregation"*. Journal of Banking and Finance, 30(9): 2635–2658.

[21] Clemen, R.T. and Winkler, R.L. (1997): *"Combining probability distributions from experts in risk analysis"*. Risk Analysis, 19: 187–203.

[22] Coles, S.G. (2001): *"An introduction to statistical modelling of extreme values"*. Springer, London.

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

[24] Coles, S.G. and Tawn, J.A. (1994): *"Statistical methods for multivariate extremes: An application to structural design"*. Applied Statistics, 43: 1–48.

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

[26] Cowles, M.K. and Carlin, B.P. (1994): *"Markov chain Monte Carlo convergence diagnostics: A comparative review"*. Technical report 94-008, Division of Biostatistics, School of Public Health, University of Minessota.

[27] Cruz, M.G. (2002): *"Modelling, measuring and hedging operational risk"*. Wiley, Chichester.

[28] Damien, P., Dellaportas, P., Polson, N. and Stephens, D.A. (2013): *"Bayesian theory and applications"*. Oxford University Press.

[29] Danielsson, J. and de Vries, C. (2002): *"Where do extremes matter?"*. Working Paper. Available online: http://www.riskresearch.org/files/CdV-JD-00-6-10-960592460-4.pdf.

[30] Davison, A.C. and Smith, R.L. (1990): *"Models for exceedances over high thresholds"*. Journal of the Royal Statistical Society B, 52: 393–442.

[31] Degen, M., Embrechts, P. and Lambrigger, D.D. (2007): *"The quantitative modelling of operational risk: Between g-and-h and EVT"*. Astin Bulletin, 37(2): 265–291.

[32] Diaconis, P. and Ylvisaker, D. (1985): *"Quantifying prior opinion (with discussion)"*. Bayesian Statistics 2 (J.-M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.): 133–156. North-Holland, Amsterdam.

[33] Do Nascimento, F.F., Gamerman, D. and Lopes, H.F. (2012): *"A semiparametric Bayesian approach to extreme value estimation"*. Statistics and Computing, 22(2): 661–675.

[34] Doucet, A., De Freitas, J.F.G. and Gordon, N.J. (2001): *"Sequential Monte Carlo Methods in Practice"*. Springer, New York.

[35] DuMouchel, W. H. (1983): *"Estimating the stable index in order to measure tail thickness: a critique"*. The Annals of Statistics, 11(4): 1019–1031.

[36] Dupuis, D.J. (2000)*: "Exceedances over high thresholds: A guide to threshold selection"*. Extremes, 1: 251–261.

[37] Dupuis, D. J. and Field, C. A. (1998): *"Robust estimation of extremes"*. The Canadian Journal of Statistics, 26(2): 199–215.

[38] Dutta, K. and Perry, J. (2006): *"A tale of tails: an empirical analysis of loss distribution models for estimating operational risk capital"*. Federal Reserve Bank of Boston. Working paper No. 06–13.

[39] Embrechts, P. (2000): *"Extreme Value Theory: Potential and limitations as an integrated risk management tool"*. Derivatives Use, Trading and Regulation, 6: 449–456.

[40] Embrechts, P. , Klüppelberg, C. and Mikosch, T. (1997): *"Modelling extremal events"*. Springer, Berlin.

[41] Embrechts, P. and Puccetti, G. (2006): *"Aggregating risk capital, with an application to operational risk"*. The Geneva Risk and Insurance Review, 31(2): 71–79.

[42] Embrechts, P. and Puccetti, G.(2008): *"Aggregation operational risk across matrix structured loss data"*. The Journal of Operational Risk, 3(2): 29–44.

[43] Escobar, M.D. and West, M. (1995): *"Bayesian density-estimation and inference using mixtures"*. Journal of the American Statistical Association, 90: 577–588.

[44] Falk, M., Hüsler, J. and Reiss, D. (2010): *"Laws of small numbers: Extremes and rare events"*. Springer, Basel.

[45] Ferson, S., Kreinovich, V., Ginzburg, L., Myers, D.S. and Sentz, K.(2003): *"Constructing probability boxes and Dempster-Shafer structures*". Sandia National Laboratories, Albuquerque, New Mexico 87185 and Livermore, California 94550. SAND report: SAND2002- 4015.

[46] Garthwaite, P.H., Kadane, J.B. and O'Hagan, A. (2005): *"Statistical methods for eliciting probability distributions"*. Journal of the American Statistical Association, 100(470): 680–701.

[47] Gelfand, A. E. and Kottas, A. (2002): *"A computational approach for full nonparametric Bayesian inference under Dirichlet process mixture models"*. Journal of Computational and Graphical Statistics, 11: 289–305.

[48] Gelman, A. and Rubin, D.B. (1992): *"Inference from iterative simulation using multiple sequences"*. Statistical Science, 7: 457–511.

[49] Genest, C. (1984a): *"A characterization theorem for externally Bayesian groups"*. The Annals of Statistics, 12: 1100–1105.

[50] Genest, C. (1984b): *"Pooling operators with the marginalization property"*. The Canadian Journal of Statistics, 12: 153–163.

[51] Genest, C. and McConway, K.J. (1990): *"Allocating the weights in the linear opinion pool"*. Journal of Forecasting, 9: 53–73.

[52] Genest, C. and Schervish, M.J. (1985): *"Modeling expert judgment for Bayesian updating"*. The Annals of Statistics, 13: 1198–1212.

[53] Genest, C. and Zidek, J.V. (1986): *"Combining probability distributions: A critique and an annotated bibliography (with discussion)"*. Statistical Science, 1: 114–148.

[54] Green, P.J. (1995): *"Reversible jump Markov chain Monte Carlo computation and Bayesian model determination"*. Biometrika, 82: 711–732.

[55] Hastings, W.K. (1970): *"Monte Carlo sampling methods using Markov chains, and their applications"*. Biometrika, 57: 97–109.

[56] Heidelberger, P. and Welch, P.D.(1983): *"Simulation run length control in the presence of an initial transient"*. Operations Research, 31: 1109–1144.

[57] Irony, T. Z. and Singpurwalla, N. D. (1997): *"Noninformative priors do not exist"*, Journal of Statistical Planning and Inference, 65: 159–89.

[58] Jenkinson, D. (2005): *"The elicitation of probabilities: A review of the statistical literature"*. Beep working paper, Department of Probability and Statistics, University of Sheffield.

[59] Kang, S., Serfozo, R.F. (1999): *"Extreme values of phase-type and mixed random variables with parallel processing examples"*. Journal of Applied Probability, 36: 194–210.

[60] Kiefer, N. M. (2010): *"Default estimation and expert information"*. Journal of Business and Economic Statistics, 28(2): 320–328.

[61] Lambrigger, D.D., Shevchenko and P.V., Wüthrich, M.V. (2007): *"The quantification of operational risk using internal data, relevant external data and expert opinions"*. The Journal of Operational Risk, 2(3): 3–27.

[62] MacDonald, A., Scarrott, C.J., Lee, D., Darlow, B., Reale, M. and Russell, G. (2011): *"A flexible extreme value mixture model"*. Computational Statistics and Data Analysis, 55: 2137–2157.

[63] Mahmoud, H. M. (2008): *"Pólya Urn Models"*. Chapman-Hall, Florida, USA.

[64] McConway, K. J. (1981): *"Marginalization and linear opinion pools"*. Journal of the American Statistical Association, 76: 410–414.

[65] McNeil, A., Frey, R. and Embrechts, P. (2005): *"Quantitative risk management"*. Princeton University Press, Princeton.

[66] Medova, E.A. (2007): *"Bayesian analysis and Markov chain Monte Carlo simulation"*. Judge Business School Working Papers, No.10/2007. University of Cambridge.

[67] Medova E.A. and Kyriacou M.N. (2002): *"Extremes in operational risk management"*. In Dempster, M.A.H. (ed.): Risk Management: Value-at-Risk and Beyond. Cambridge University Press, 247–274.

[68] Mendes, B. and Lopes, H.F. (2004): *"Data driven estimates for mixtures"*. Computational Statistics and Data Analysis, 47: 583–598.

[69] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953): *"Equations of state calculations by fast computing machines"*. Journal of Chemical Physics, 21(6): 1087–1092.

[70] Nešlehová, J., Embrechts, P. and Chavez-Demoulin, V. (2006): *"Infinite mean models and the LDA for operational risk"*. Journal of Operational Risk, 1(1): 3–25.

[71] Pickands III, J. (1994): *"Bayes quantile estimation and threshold selection for the generalized Pareto family"*. In Galambos, J., Leigh, S., and Simiu, E. (eds.), Extreme Value Theory and Applications, 123–138, Kluwer, Amsterdam.

[72] Reiss, R.D. and Thomas, M. (1997): *"Statistical analysis of extreme values with applications to insurance, finance, hydrology and other fields"*. Birkhauser Verlag, Basel.

[73] Resnick, S. I. (1987): *"Extreme values, regular variation and point processes"*. Springer, New York.

[74] Ribatet, M. A. (2006): "A user's guide to the POT package". http://cran.r-project.org/.

[75] Ripley B. D. (1987). *"Stochastic simulation"*. Wiley, New York.

[76] Robert, C. P., and Casella, G. (1999): *"Monte Carlo statistical methods"*. New York: Springer-Verlag.

[77] Scarrott, C. and MacDonald, A. (2012): *"A review of extreme value threshold estimation and uncertainty quantification"*. REVSTAT: Statistical Journal, 10(1): 33–60.

[78] Shevchenko, P.V. (2010): *"Implementing loss distribution approach for operational risk"*. Applied Stochastic Models in Business and Industry, 26(3): 277–307.

[79] Shevchenko P.V. (2011): *"Modelling operational risk using Bayesian inference"*. Springer, Berlin.

[80] Shevchenko, P.V. and Temnov, G. (2009) : *"Modelling operational risk data reported above a time-varying threshold"*. The Journal of Operational Risk, 4(2): 19–42.

[81] Shevchenko, P.V. and Wüthrich, M.V. (2006): *"The structural modelling of operational risk via Bayesian inference: combining loss data with expert opinions"*. The Journal of Operational Risk, 1(3): 3–26.

[82] Shi J., Samad-Khan, A. and Medapa P. (2000): *"Is the size of an operational loss related to firm size?"*. Operational Risk Magazine, 1(2): 22–25.

[83] Smith, R.L. (1987): *"Estimating tails of probability distributions"*. The Annals of. Statistics, 15: 1174–1207.

[84] Smith, R.L. (1985): *"Maximum likelihood estimation in a class of non-regular cases"*. Biometrika, 72: 67–90.

[85] Steinhoff, C., Baule, R.(2006): *"How to validate op risk distributions*". OpRisk and Compliance, 1(8): 36–39.

[86] Tancredi, A., Anderson, C. W. and O'Hagan, A. (2006): *"Accounting for threshold uncertainty in extreme value estimation"*. Extremes 9: 87–106.

[87] Teh, Y.W. (2010): *"Dirichlet Processes"*. Encyclopedia of Machine Learning. Springer, New York.

[88] Tierney, L. (1998): *"A note on Metropolis-Hastings kernels for general state spaces"*. The Annals of Applied Probability, 8(1): 1–9.

[89] Tversky, A. (1974): *"Assessing uncertainty"*. Journal of the Royal Statistical Society B, 36: 148–159.

[90] Venturini, S., Dominici, F. and Parmigiani, G. (2008): *"Gamma shape mixtures for heavy-tailed distributions*". The Annals of Applied Statistics, 2(2): 756–776.

[91] Wang, Z., Rodriguez, A. and Kottas, A. (2011): *"A nonparametric mixture modelling framework for extreme value analysis"*. SIGFIRM Working Paper No. 13.

[92] Zhao, X., Scarrott, C.J., Oxley, L. and Reale, M. (2009): *"Bayesian extreme value mixture modelling for estimating VaR"*. Working Papers in Economics 09/15, University of Canterbury, Department of Economics and Finance.