EXTREME VALUE THEORY WITH MARKOV CHAIN MONTE CARLO - AN AUTOMATED PROCESS FOR FINANCE

PHILIP BRAMSTÅNG & RICHARD HERMANSON

ROYAL INSTITUTE
OF TECHNOLOGY

Master's Thesis at the Department of Mathematics

Supervisor (KTH): Henrik Hult
Supervisor (Cinnober): Mikael Öhman
Examiner: Filip Lindskog

September 2015 – Stockholm, Sweden

ABSTRACT

The purpose of this thesis was to create an automated procedure for estimating financial risk using extreme value theory (EVT).

The "peaks over threshold" (POT) result from EVT was chosen for modelling the tails of the distribution of financial returns. The main difficulty with POT is choosing a convergence threshold above which the data points are regarded as extreme events and modelled using a limit distribution. It was investigated how risk measures are affected by variations in this threshold and it was deemed that fixed-threshold models are inadequate in the context of few relevant data points, as is often the case in EVT applications. A model for automatic threshold weighting was proposed and shows promise.

Moreover, the choice of Bayesian vs frequentist inference, with focus on Markov chain Monte Carlo (MCMC) vs maximum likelihood estimation (MLE), was investigated with regards to EVT applications, favoring Bayesian inference and MCMC. Two MCMC algorithms, independence Metropolis (IM) and automated factor slice sampler (AFSS), were analyzed and improved in order to increase performance of the final procedure.

Lastly, the effects of a reference prior and a prior based on expert opinion were compared and exemplified for practical applications in finance.

SAMMANFATTNING

Syftet med detta examensarbete var att utveckla en automatisk process för uppskattning av finansiell risk med hjälp av extremvärdesteori.

"Peaks over threshold" (POT) valdes som metod för att modellera extrempunkter i avkastningsdata. Den stora svårigheten med POT är att välja ett tröskelvärde för konvergens, över vilket alla datapunkter betraktas som extrema och modelleras med en gränsvärdesdistribution. Detta tröskelvärdes påverkan på olika riskmått undersöktes, med slutsatsen att modeller med fast tröskelvärde är olämpliga om datamängden är liten, vilket ofta är fallet i tillämpade extremvärdesmetoder. En modell för viktning av tröskelvärden presenterades och uppvisade lovande resultat.

Därtill undersöktes valet mellan Bayesiansk och frekventisk inferens, med fokus på skillnaden mellan Markov chain Monte Carlo (MCMC) och maximum likelihood estimation (MLE), när det kommer till applicerad extremvärdesteori. Bayesiansk inferens och MCMC bedömdes vara bättre, och två MCMC-algoritmer; independence Metropolis (IM) och automated factor slice sampler (AFSS), analyserades och förbättrades för användning i den automatiska processen.

Avslutningsvis jämfördes effekterna av olika apriori sannolikhetsfördelningar (priors) på processens slutresultat. En svagt informativ referensprior jämfördes med en starkt informativ prior baserad på expertutlåtanden.

The Reader may here observe the Force of Numbers, which can be successfully applied, even to those things, which one would imagine are subject to no Rules. There are very few things which we know, which are not capable of being reduc'd to a Mathematical Reasoning; and when they cannot it's a sign our knowledge of them is very small and confus'd; and when a Mathematical Reasoning can be had it's as great a folly to make use of any other, as to grope for a thing in the dark, when you have a Candle standing by you.

— John Arbuthnot
Of the Laws of Chance (1692)

ACKNOWLEDGMENTS

We would like to express our gratitude to our supervisor at the Royal Institute of Technology, Henrik Hult, for his valuable ideas and guidance. Furthermore, we would like to thank Mikael Öhman at Cinnober Financial Technology for his support, feedback, and advice throughout the process of this thesis work.

Philip Bramstång & Richard Hermanson
September 2015 – Stockholm, Sweden

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

AFSS	Automated Factor Slice Sampler
BI	Bayesian Inference
BM	Block Maxima
CDF	Cumulative Distribution Function
ES	Expected Shortfall
EVT	Extreme Value Theory
GARCH	Generalized Autoregressive Conditional Heteroskedasticity
GEV	Generalized Extreme Value (distribution)
GH	Generalized Hyperbolic (distribution)
GL	Generalized Lambda (distribution)
GP	Generalized Pareto (distribution)
i.i.d.	Independent and identically distributed
IM	Independence Metropolis
LA	Laplace Approximation
MCMC	Markov Chain Monte Carlo
MH	Metropolis-Hastings
MLE	Maximum Likelihood Estimation
MVC	Multivariate Cauchy (distribution)
MVN	Multivariate Normal (distribution)
PDF	Probability Density Function
POT	Peaks Over Threshold
SS	Slice Sampler
VaR	Value at Risk

NOMENCLATURE

θ	Parameter set
X	Data set (sample) of losses (negative returns)
$Pr(A)$	Probability of A
π	Target density
q	Proposal (conditional) density
K	Transition kernel or Markov kernel
P, p	CDF, PDF of generalized Pareto distribution
H, h	CDF, PDF of generalized hyperbolic distribution
V, v	CDF, PDF of generalized extreme value distribution
B, b	CDF, PDF of body distribution
T, t	CDF, PDF of tail distribution

1

INTRODUCTION

1.1 BACKGROUND

In the aftermath of the last financial crisis culminating in 2008, financial institutions face an increasing level of regulation on how they should measure and manage their exposure to risk. Banks are now required to hold more capital, covering risk at more extreme levels such as the 99% or 99.9% quantiles of their estimated loss distribution.

Previously, financial returns were often modelled using distributions such as the normal. Many of them are unable to properly describe the tails at these extreme levels. As a result, extreme value theory (EVT), containing results about limiting distributions of extreme values, has seen an increase in popularity as a template for statistical modelling.

However, there is an inherent difficulty with extreme risk, which is the scarcity of data, leading to substantial uncertainty when estimating parameters. Therefore, it is attractive to use some method that takes this uncertainty into account.

There are countless situations where investigating the behaviour of the tails of a distribution might be useful though this thesis focuses on, albeit is not limited to, financial applications.

1.2 PREVIOUS WORK

The "peaks over threshold" (POT) result from EVT requires selection of a threshold above which all data points are regarded as extreme events. The standard procedure is to choose this threshold graphically, by looking at a plot, see [15], or simply setting it to some high percentile of the data, see [14].

After selecting the threshold, it is assumed to be known and the other parameters are estimated. However, there is a lot of uncertainty about the selection of the threshold and previous works agree that it has

a significant effect on parameter estimates, see [44], [12], [11], and [17].

Many approaches have been suggested to improve on this, such as:

- selecting an optimal threshold by minimizing bias-variance, see [3].
- having a dynamic mixture model where one term was generalized Pareto (GP) and the other was a light-tailed density function as in [17], though they do not explicitly consider threshold selection.
- performing maximum likelihood estimation (MLE) on a mixture model where tails are GP and the center was normally distributed, see [34].
- choosing the number of upper order statistics and calculating a weighted average over several thresholds, as demonstrated in [5].
- having a model with Gamma as the center and GP as the tail where the threshold was simply considered another model parameter, see [2].

An another topic, to be able to use Markov chain Monte Carlo (MCMC), it is necessary to specify prior distributions for parameters. There have been many previous works on priors of different levels of subjectivity. Some aim to minimize the the subjective content and let the data speak for itself, see [4], while others attempt to augment the data with the help of subjective information from an expert, see [12].

1.3 PURPOSE

The purpose of this thesis was to develop an automatic procedure for estimating extreme risk from financial returns using EVT. Issues that were encountered and investigated include:

- Selecting an EVT limit result, i.e. block maxima (BM) vs "peaks over threshold" (POT).
- Threshold sensitivity and automatic threshold selection, eventually becoming automatic threshold weighting.
- Bayesian vs frequentist inference, with extra focus on Markov chain Monte Carlo (MCMC) vs maximum likelihood estimation (MLE) for EVT applications. This led to including a framework that allows financial experts to input their expertise into prior distributions.

The choices during development and the performance of the final procedure were evaluated.

1.4 DELIMITATIONS

In real world applications, risk analysis is often done on portfolios and it is well known that there often exist some inter-dependencies between financial instruments. This has been deemed outside the scope of this thesis but could very well be a future extension.

Due to the nature of EVT, there is often very little data available and, as such, standard back testing is virtually useless. Instead, simulations were used to get an idea for the effectiveness of the models. Along the same lines, the more extreme the risk, the fewer data points are available and the more uncertainty is incurred.

At some point, with fewer and fewer relevant samples, the estimates approach educated guesses.

The described transformation of real data is only provided as a standard example and there might be better ways to do it, which would yield better results.

Due to consensus in literature that financial data is heavy-tailed, see [15, p.38], and that there is not really any need for EVT otherwise, the tests and models focus on heavy-tailed data.

1.5 THESIS OUTLINE

The mathematical background theory necessary for understanding the models and methods used in this thesis is presented in Chapter 2. The reader is introduced to the core concepts of EVT, Bayesian inference (BI), MCMC, and certain finance-specific theory, such as volatility adjustment and risk measures.

Chapter 3 describes the complete process from financial data to risk measures and the decisions involved in arriving at said process. This entails data transformation, automatic threshold selection for POT, improvements on specific MCMC algorithms, and the use of different prior distributions in BI.

Chapter 4 presents the results, consisting of tables and plots highlighting different aspects of the process. This includes sensitivity of the risk measures to threshold choice, parameter estimation stability (for both MLE and MCMC), the effect of informative priors, and credible intervals or confidence intervals depending on the method and sample size. Moreover, an overview of the chosen data sets is given, both simulated and real-world.

The results are summarized and discussed in Chapter 5 and the thesis is concluded by Chapter 6, which briefly discusses ideas for future work.

2

BACKGROUND THEORY

This chapter will present the mathematical background of the problem. An outline of the presented theory can be found in Section 1.5.

2.1 EXTREME VALUE THEORY (EVT)

Two important results from EVT are the limit distributions of a series of (properly centered and normalized) block maxima (BM) and of excesses over a threshold, called "peaks over threshold" (POT), given that the distributions are non-degenerate and the sample is independent and identically distributed (i.i.d.).

> As a note of caution, it should be underlined that the existence of a non-degenerate limit distribution ... is a rather strong requirement. — *[33, Sornette p.47]*

Nonetheless, these results are commonly used as templates for statistical modelling and have displayed effectiveness in many applications.

2.1.1 *Block Maxima (BM)*

Consider a sample of N i.i.d. realizations $X_1, \ldots, X_N$ of a random variable, for example the daily returns of an index for one month. Let M_N denote the maximum of this sample, e.g. the monthly maximum of the returns:

$$M_N = max\{X_1, \ldots, X_N\}. \tag{1}$$

Then, the Fisher–Tippett–Gnedenko theorem states that, if there exist sequences of normalizing constants $\{a_N > 0\}$ and $\{b_N\}$;

$$M_N^* = \frac{M_N - b_N}{a_N}, \tag{2}$$

such that the distribution of M_N^* (e.g. the distribution of monthly maxima) converges to a non-degenerate distribution as N goes to

infinity, this limit distribution is then necessarily the generalized extreme value (GEV) distribution, see [10, p.46].

The main difficulty in using this result is often determining the optimal subsample size N, which comes down to a trade-off between bias and variance. For example, if one has 1000 data points, choosing $N = 10$ leads to many maxima, but each maximum is only informed by 10 data points, which leads to estimation bias, since approximation by the limit distribution (GEV) is likely poor. Choosing $N = 100$ leads to the opposite scenario: better convergence but few maxima and high variance.

2.1.1.1 *Generalized Extreme Value (GEV) Distribution*

The cumulative distribution function (CDF) of the GEV distribution is given by:

$$V(x) = exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} \qquad \xi \neq 0 \qquad (3)$$

$$V(x) = exp\left\{-exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right]\right\} \qquad \xi = 0 \qquad (4)$$

with support $x \in \{x : 1+\xi(x-\mu)/\sigma > 0\}$ when $\xi \neq 0$ and $x \in \mathbb{R}$ when $\xi = 0$. The three parameters are location μ, scale $\sigma > 0$ and shape ξ. The sign of the shape parameter determines the tail behaviour of the distribution. As $x \to \infty$ the probability density function (PDF) decays exponentially for $\xi > 0$, polynomially for $\xi = 0$, and is bounded above by $\mu - \sigma/\xi$ for $\xi < 0$, see [10, p.47].

2.1.2 *Peaks Over Threshold (POT)*

POT originates in the Pickands–Balkema–de Haan theorem, which continues from the earlier result from BM, see Section 2.1.1. Suppose the Fisher–Tippett–Gnedenko theorem from BM is satisfied, so that for large sample sizes N;

$$Pr\{M_N \leq x\} \approx V(x), \qquad (5)$$

where $V(x)$ is the GEV CDF. Let X be any term in the X_i sequence. Then, for a large enough threshold u, $X - u \mid X > u$, i.e. the threshold excesses, is approximately generalized Pareto (GP) distributed, see [10, p.75].

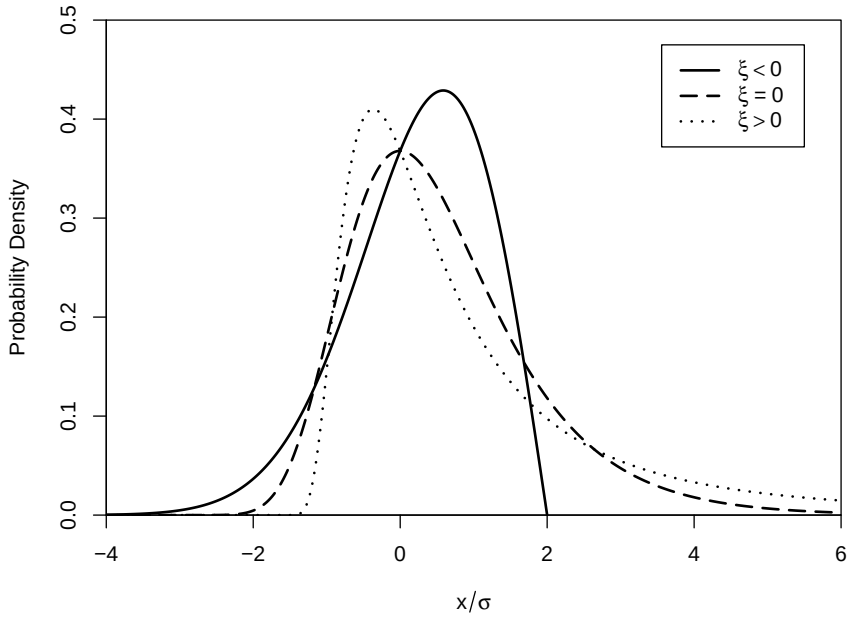


Figure 1: GEV PDF v(x) for different values of shape parameter ξ, all with $(\mu, \sigma) = (0, 1)$.

2.1.2.1 *Generalized Pareto (GP) Distribution*

The GP distribution has CDF

$$P(x) = 1 - \left(1 + \frac{\xi(x-u)}{\sigma}\right)^{-1/\xi} \qquad \xi \neq 0 \qquad (6)$$

$$P(x) = 1 - exp\left(-\frac{x-u}{\sigma}\right) \qquad \xi = 0 \qquad (7)$$

with support $x \geq u$ when $\xi \geq 0$, and $u \leq x \leq u - \sigma/\xi$ when $\xi < 0$. The three parameters are location u, scale $\sigma > 0$ and shape ξ. The shape parameter ξ plays the exact same role as for the GEV distribution, see Section 2.1.1.1, determining the tail behaviour as $x \to \infty$, refer to [10, p.75] for more details.

2.1.2.2 *Selecting Threshold*

Much like determining subsample size of BM, the biggest issue with using POT may be determining when the data has converged well enough and setting a corresponding threshold.

> ... determination of the optimal threshold ... is in fact related to the optimal determination of the subsamples size — *[33, Sornette p.48]*

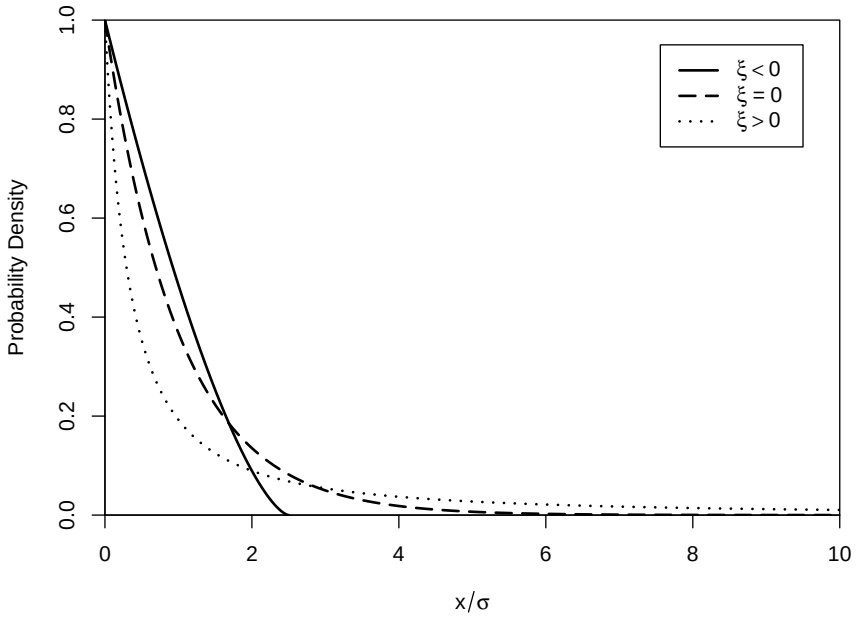


Figure 2: GP PDF p(x) for different values of shape parameter ξ, all with $(u, \sigma) = (0, 1)$.

The standard method for determining the threshold is the mean residual life (MRL) plot, described below, together with two alternative methods.

Mean Residual Life (MRL) Plot

The mean of a GP($u = 0, \sigma, \xi$) distributed variable X is

$$E[X] = \frac{\sigma}{1-\xi} \qquad \xi < 1. \tag{8}$$

When $\xi \geq 1$ the mean is infinite. Suppose this GP distribution is used to model excesses over a threshold u_0, then

$$E[X - u_0 \mid X > u_0] = \frac{\sigma_{u_0}}{1-\xi}, \tag{9}$$

where σ_{u_0} is the scale parameter corresponding to excesses of the threshold u_0. But if the GP is valid for threshold u_0 it is also viable for all thresholds $u > u_0$, only with a different σ given by

$$\sigma_u = \sigma_{u_0} + \xi u \tag{10}$$

as explained in [10, p.75]. So, for $u > u_0$

$$\begin{aligned} E[X - u \mid X > u] &= \frac{\sigma_u}{1-\xi} \\ &= \frac{\sigma_{u_0} + \xi u}{1-\xi} \end{aligned} \tag{11}$$

i.e. $E[X - u \mid X > u]$, which is the mean of the excesses, changes linearly with u if the GP model is appropriate. This means that the scatter plot of the points

$$\left\{ \left(u, \frac{1}{N_u} \sum_{i=1}^{N_u} (x_{(i)} - u) \right) : u < x_{max} \right\}, \tag{12}$$

where $x_{(1)} \dots x_{(N_u)}$ are the N_u excesses, should be linear in u. This plot is called the mean residual life (MRL) plot or the mean excess plot and is commonly used to determine an appropriate threshold for GP. However, it is very hard to read and doesn't give a definite answer, as shown in Section 3.3.2 and described in [10, p.78].

Fixed

In some papers, especially when focus is not on threshold selection, it is set to a high percentile as suggested by DuMouchel, see [14]. The 95th percentile is a common choice, see for example [26, p.312].

Stability of Parameters

Another technique is to fit the generalized Pareto distribution at a range of thresholds and look for stability in the parameter estimates, as described in [15, p.36].

> Above a level u_0 at which the asymptotic motivation for the generalized Pareto distribution is valid, estimates of the shape parameter ξ should be approximately constant, while estimates of σ should be linear in [threshold] u ...
> — *[10, Coles p.83]*

As with the MRL plot, deciding upon where the parameters are stable can be quite hard, especially as with higher thresholds, there are fewer and fewer data points which leads to decreasing accuracy and thus increasing changes in the parameter estimates.

2.2 RISK MEASURES FROM POT

2.2.1 *Value at Risk (VaR)*

VaR is a standard risk measure in finance that describes the worst loss over a horizon that will not be exceeded with a given level of confidence, see Figure 3.

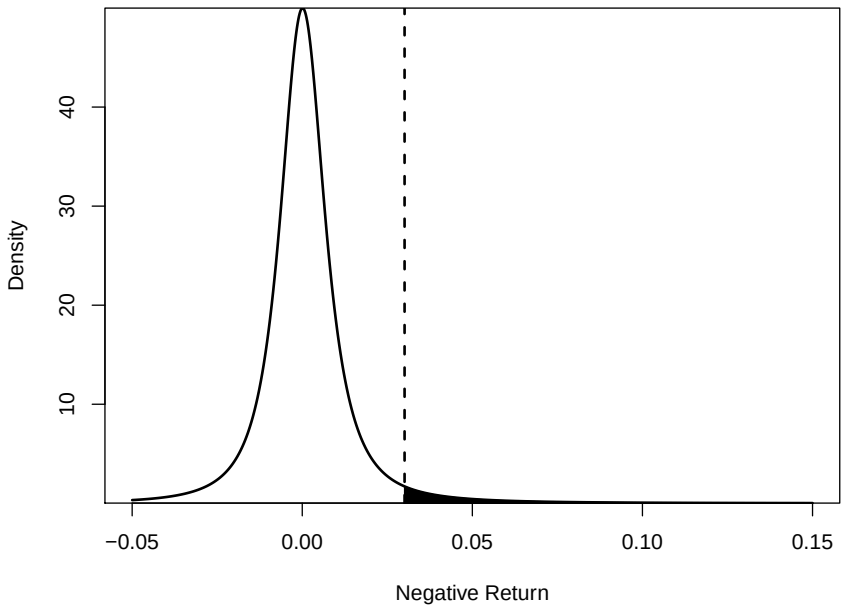


Figure 3: Example distribution of losses. The dashed line is the value at risk (VaR) at some level and the expected value of the filled area is the expected shortfall (ES) at the same level.

VaR at the confidence level α for a distribution X of losses is defined as:

$$VaR_\alpha(X) = F^{-1}(1-\alpha) \tag{13}$$

where F is the CDF of X, see [29, p.90].

Assuming that the data is in the form of losses, i.e. negative (log) returns, and that these losses, above some threshold u, are modeled by a GP distribution, the CDF for the full tail loss distribution is then

$$T(y) = B(u) + P(x)\Big[1 - B(u)\Big] \qquad y = x + u \qquad x > 0, \tag{14}$$

where $P(x)$ is the GP CDF and $B(x)$ is the CDF of the body distribution. The CDF value at the threshold, $B(u)$, can be approximated empirically. Let N be the total number of data points and N_u the number of data points exceeding the threshold. The standard method is to use the empirical CDF to approximate $B(u)$:

In the models presented later we use a body distribution to estimate $B(u)$ (instead of the empirical factor in Equation (15)).

$$B(u) \approx \frac{N - N_u}{N}, \tag{15}$$

which together with the expression for $P(x)$ from (6) and Equation (14) yields

$$T(y) \approx 1 - \frac{N_u}{N}\left[1 + \frac{\xi(y-u)}{\sigma}\right]^{-1/\xi}. \tag{16}$$

Solving for y gives an estimate of the $1-p$ quantile, which is the VaR at level p, see [47, p.26],

$$VaR_p \approx u - \frac{\sigma}{\xi}\left\{1 - \left[\frac{N}{N_u} \cdot p\right]^{-\xi}\right\}. \tag{17}$$

This expression for the VaR is only valid at quantiles above the threshold, i.e. in the area modelled by the GP distribution (small upper tail probability p).

2.2.2 *Expected Shortfall (ES)*

Also known as "average value at risk" or "conditional value at risk", ES is commonly used in financial literature and is very relevant for heavy tailed data. The ES at a certain level is the expected value of the loss, given that the loss exceeds the corresponding VaR, see Figure 3 and [29, p.91],

$$ES_p = E[X \mid X > VaR_p] = VaR_p + E[X - VaR_p \mid X > VaR_p]. \tag{18}$$

Using the properties of the GP distribution, it can be shown that

$$E[X - VaR_p \mid X > VaR_p] = \frac{\sigma + \xi(VaR_p - u)}{1-\xi} \tag{19}$$

for $0 < \xi < 1$, see [47, p.27]. Equation (18) then becomes

$$ES_p = \frac{VaR_p + \sigma - \xi u}{1-\xi}. \tag{20}$$

2.3 VOLATILITY ADJUSTMENT

Market circumstances may change significantly over time and, consequently, the historical returns from a period of a certain volatility (a volatility regime) may not be representative of the current market situation. For instance, if the market is currently very volatile and one tries to estimate today's 1-day VaR from historical returns from the last 3 years of low volatility, one will underestimate the risk.

Moreover, there is a known characteristic of financial time series, called volatility shocks. It is a tendency in the market for volatility to cluster. For example, large changes are often followed by large changes.

One way of trying to account for this is to model a time series of the historical volatility and adjust all the returns to today's estimated

volatility. This is done by dividing each return at time t by the estimated volatility at time t, and then multiplying it by today's volatility (time T). The standard method and a specialization for financial applications for modelling the volatility are presented below.

2.3.1 *Generalized Autoregressive Conditional Heteroskedasticity (GARCH)*

The standard GARCH model assumes that the dynamic behaviour of the conditional variance is given by

$$\sigma_t^2 = \omega + \alpha\epsilon_{t-1}^2 + \beta\sigma_{t-1}^2 \qquad \epsilon_t | I_{t-1} \sim N(0, \sigma_t^2) \tag{21}$$

where σ_t^2 is the conditional variance, ω is the intercept, and ϵ_t (called the market shock or unexpected return) is the mean deviation $(r_t - \bar{r})$ from the sample mean, i.e. the error term from ordinary linear regression, see [1, p.4].

The parameters are often estimated with maximum likelihood estimation (MLE). The model can be further improved by letting the ϵ_t terms be drawn from a distribution other than the normal and can thereby allow for non-zero skewness and excess kurtosis.

2.3.2 *Glosten-Jagannathan-Runkle GARCH (GJR-GARCH)*

Previous works suggest asymmetric GARCH models are often better when working with daily financial data. This is because of the so called leverage effect; that market volatility increases are larger following a large negative return than following a large positive return of equal size, see [6].

The GJR-GARCH model introduces a leverage parameter λ to model the asymmetric response from negative market shocks;

$$\sigma_t^2 = \omega + \alpha\epsilon_{t-1}^2 + \lambda I_{\{\epsilon_{t-1}<0\}}\epsilon_{t-1}^2 + \beta\sigma_{t-1}^2. \tag{22}$$

This time series of volatility estimates $\{\hat{\sigma}_t\}_{t=1}^T$ can then be used on historical returns $\{r_t\}_{t=1}^T$ to produce the volatility adjusted returns, as described in [6],

$$\tilde{r}_{t,T} = \left(\frac{\hat{\sigma}_T}{\hat{\sigma}_t}\right) r_t, \tag{23}$$

where T is the time at the end of the sample, e.g. today.

2.4 BAYESIAN INFERENCE (BI)

Statistical inference can be divided into two broad categories: Bayesian inference (BI) and frequentist inference. In a way, these two paradigms disagree on the fundamental nature of probability. The frequentist interpretation is that any given experiment can be considered as one of an infinite sequence of possible repetitions of the same experiment, each capable of producing statistically independent results. So the probability of an event is the limit of that event's relative frequency in an infinite number of trials. Many standard methods in statistics, such as statistical hypothesis testing and p-value confidence intervals are based on the frequentist framework.

BI, on the other hand, can assign probabilities to any statement, even in the absence of randomness, and updates knowledge about unknowns with information from data. In this framework, probability is a quantity representing a state of knowledge, or a state of belief. Merriam-Webster defines "Bayesian" as follows

> **Bayesian**: being, relating to, or involving statistical methods that assign probabilities or distributions to events (as rain tomorrow) or parameters (as a population mean) based on experience or best guesses before experimentation and data collection and that apply Bayes' theorem to revise the probabilities and distributions after obtaining experimental data.

There are also differing interpretations within BI, mainly objective vs subjective BI. As the names suggest, they differ in the degree that subjective information, as opposed to data, is allowed to influence the end result. Generally, objective Bayesians favor uninformative priors, while subjective Bayesians favor informative priors , see Section 2.4.2. For a more in-depth and formal overview, the reader is referred to [38].

2.4.1 *Bayes' Theorem*

The centerpiece of Bayesian inference (BI) is Bayes' theorem, which gives an expression for the conditional probability, or posterior probability, of an event A after the event B is observed, $Pr(A|B)$. In other words, it gives an expression for the *updated* probability of A, updated with the information that B occurred. Hence the word posterior probability, as opposed to prior probability $Pr(A)$.

From the formula for conditional probability;

$$Pr(A|B) = \frac{Pr(A \bigcap B)}{Pr(B)}, \tag{24}$$

and simply $A \bigcap B = B \bigcap A$, Bayes' theorem follows:

$$Pr(A|B) = \frac{Pr(B|A)Pr(A)}{Pr(B)}. \tag{25}$$

From Bayes' theorem, replacing probabilities Pr with densities p, A with a parameter set θ and B with a data set X, we have the relation

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)} = \frac{p(X|\theta)p(\theta)}{\int p(X|\theta)p(\theta)d\theta'}, \tag{26}$$

where $p(\theta)$ is the prior distribution (of the parameter set), $p(X|\theta)$ is the sampling distribution (the likelihood of the data X under some model) and $p(X)$ is the marginal likelihood, or the prior predictive distribution of X, which indicates what X should look like, given the model, before it has been observed, see [27].

The result, $p(\theta|X)$, is called the joint posterior distribution of the parameter set θ. It expresses the updated beliefs about θ after taking both prior and data into account. Due to the integral in the denominator of (26), it is rarely possible to calculate $p(\theta|X)$ directly. Instead, Markov chain Monte Carlo (MCMC) is often used to simulate samples from it.

The prior predictive distribution $\int p(X|\theta)p(\theta)d\theta$ normalizes the joint posterior distribution $p(\theta|X)$. Removing it from Equation (26) yields

$$p(\theta|X) \propto p(X|\theta)p(\theta), \tag{27}$$

i.e. that the unnormalized joint posterior is proportional to the likelihood times the prior. There are many methods that make use of this result.

The value of interest is often a function f of the parameter set θ.

$$E[f(\theta)|X] = \frac{\int f(\theta)p(X|\theta)p(\theta)d\theta}{\int p(X|\theta)p(\theta)d\theta} = \frac{\int f(\theta)\pi(\theta)d\theta}{\int \pi(\theta)d\theta}, \tag{28}$$

where $\pi(\cdot)$ is the posterior distribution of θ. For example, let f be value at risk at some confidence level and θ be the parameters of a GP distribution, then, if MCMC was used to produce the posterior, calculation of Equation (28) is as simple as taking the mean of the thinned posterior samples, after discarding the burn-in samples, see Section 2.7.2.

2.4.2 *Priors*

A prior probability distribution, often shortened to prior, is a probability distribution that expresses prior beliefs about a parameter θ before the data is taken into account. The prior is an integral part of

Bayes' theorem, see Equation (26), and can greatly affect the posterior distribution. One should make sure that the prior is proper, i.e. that

$$\int p(\theta) d\theta \neq \infty. \tag{29}$$

An improper prior can lead to an improper posterior distribution, which makes inferences invalid. In order for the joint posterior distribution to be proper, the marginal likelihood, i.e. the denominator in the last expression in Equation (26), must be finite for all X.

The two main approaches to choosing a prior, informative versus uninformative, are outlined below.

Priors can also be useful for attaining numerical stability or handling parameter bounds.

2.4.2.1 *Uninformative Priors*

The purpose of an uninformative prior is to minimize the subjective information content and instead let the data speak for itself. However, truly uninformative priors do not exist, as discussed in [25, p.159-189], and all priors are informative in some way. One instead speaks of weakly informative priors (WIP) and least informative priors (LIP).

Reference Prior for GP

A commonly used subcategory of least informative priors (LIP) is the reference prior, which is designed to let the data dominate the prior and posterior. The idea is to maximize the expected intrinsic discrepancy between the posterior distribution and prior distribution. This, in turn, also maximizes the expected posterior information about X, see [4, p.905] for details. The reference priors for the GP parameters are

$$p_\sigma(\sigma, \xi) \propto \frac{1}{\sigma\sqrt{1+\xi}\sqrt{1+2\xi}} \tag{30}$$

$$p_\xi(\sigma, \xi) \propto \frac{1}{\sigma(1+\xi)\sqrt{1+2\xi}} \tag{31}$$

from [30, p.1525] and [31, p.174], which include proofs of propriety. See Figure 17 for a visualization of these priors.

2.4.2.2 *Informative Priors*

Informative priors are based on the idea that when prior information is available about a parameter θ, that information should be used. One example of this is to use the knowledge of an expert to create a prior distribution. The knowledge contained in the elicited prior will then help supplement the data. There is a multitude of methods

As mentioned earlier, this can be helpful in extreme value theory applications because data is often scarce.

for converting expert knowledge into actual parameters for the prior distribution and the one used in this thesis is described in detail in Section 3.5.3.

2.5 LAPLACE APPROXIMATION (LA)

This is a minor part of the method but is nonetheless useful for initial exploration of the parameter space and expected value.

LA is a method of approximating integrals. Under the assumption that $f(x)$ both has a unique maxima $f(x_0)$ such that $f''(x_0) < 0$ and is a twice differentiable function on $[a, b]$, then

$$\int_a^b e^{Mf(x)} dx \approx \sqrt{\frac{2\pi}{M|f''(x_0)|}} \text{ as } M \to \infty. \tag{32}$$

2.6 MARKOV CHAINS

A Markov chain is a memoryless random process in the sense that the next state only depends on the current state. If the sequence of random variables $X_1, X_2 \ldots$ is a Markov chain, then

$$Pr(X_{n+1} = x \mid X_1 = x_1, \ldots, X_n = x_n) = Pr(X_{n+1} = x \mid X_n = x_n), \tag{33}$$

assuming that the conditional probabilities are well defined, i.e. that

$$Pr(X_1 = x_1, \ldots, X_n = x_n) > 0. \tag{34}$$

The possible values of X_i form the state space of the Markov chain. Under certain regularity conditions, the chain will converge to a unique stationary distribution, independent of the starting point X_1, see [18, p.113].

An important part of the theory of Markov processes is the Markov kernel $K(a, b)$. It is a function describing the transition probability of the chain from state a to state b.

2.7 MARKOV CHAIN MONTE CARLO (MCMC)

MCMC methods are a type of sampling algorithms that construct a Markov chain θ_n with a desired equilibrium distribution (θ is a parameter set). They focus on obtaining a sequence of random samples from a probability distribution for which direct sampling is difficult. This is often the case with the posterior distribution in Bayesian Inference (BI), see Equation (26).

The first few sections introduce important theorems and concepts that are needed to understand MCMC. This is followed by a few specific

algorithms. For a deeper discussion of these concepts, the interested reader is referred to [18].

2.7.1 *Existence of a Stationary Distribution*

Known as the detailed balance equation or the reversibility condition;

This is an important condition that will be revisited when analyzing specific algorithms later.

$$\pi(\theta)K(\theta,\theta^*) = \pi(\theta^*)K(\theta^*,\theta) \qquad \forall(\theta,\theta^*), \tag{35}$$

where K is the Markov kernel (see Section 2.6 for an explanation and Section 2.7.7 for an example), is a sufficient condition for the target distribution π to be the equilibrium or stationary distribution of the chain, see [28, p.21].

2.7.2 *Ergodic Average*

The ergodic average is very important for output analysis and tells us that

$$E[f(\theta)] \approx \frac{1}{N-s}\sum_{i=s+1}^{N} f(\theta_i) \tag{36}$$

where stationarity was reached at s iterations and N is sufficiently large, see [28, p.23]. This is how the risk measures or other functions of the parameter set θ are calculated from the posterior samples.

2.7.3 *Markov Chain Standard Error (MCSE)*

MCSE is the standard deviation around the mean of the samples, due to the uncertainty from using an MCMC algorithm. As the number of independent posterior samples tends to infinity, it approaches zero.

The initial monotone positive sequence (IMPS) estimator is used to estimate MCSE. It is a variance estimator that is more specialized for MCMC. It is valid for Markov chains that are stationary, irreducible, and reversible. It relies on the property that even-lag autocovariances are nonnegative, let:

$$\Gamma_m = \gamma_{2m} + \gamma_{2m+1}, \tag{37}$$

where γ_t is the autocovariance with lag t, is a strictly positive and strictly decreasing function of m.

Firstly, the so called initial positive sequence estimator, is

$$\hat{\sigma}^2_{pos} = \hat{\gamma}_0 + 2\sum_{i=1}^{2m+1} \hat{\gamma}_i = -\hat{\gamma}_0 + 2\sum_{i=1}^{m} \hat{\Gamma}_i, \tag{38}$$

where $\hat{\gamma}_t$ and $\hat{\Gamma}_t$ are estimates of their respective quantities, and m is chosen to be the largest integer such that

$$\hat{\Gamma}_i > 0 \qquad i = 1, 2, ..., m. \tag{39}$$

Secondly, this estimator was improved on by eliminating some noise by forcing the sequence to be monotone. This is done by replacing $\hat{\Gamma}_t$ above with

$$\min\{\hat{\Gamma}_1, \hat{\Gamma}_2, ..., \hat{\Gamma}_t\} \tag{40}$$

It can be shown that, as the sample size tends to infinity, the true variance will be smaller than or equal to the estimated variance, see [9, p.72-73].

2.7.4 *Burn-in*

This is widely discussed in MCMC literature and is the number of iterations that should be discarded before calculating the ergodic average. It is directly related to determining if the Markov chain has converged to the target distribution, which is a difficult problem. Suggestions for determining convergence include running multiple chains and, when they converge, only one continues running while burn-in is set to that point. There have been arguments against this method saying that convergence is better based on estimating if stationarity has been reached. For details, see [18, p.159,166-167] and [21, p.13-15].

2.7.5 *Stopping time*

There has also been much debate regarding stopping time as it is difficult to determine and, as with burn-in, there have been suggestions of simply using multiple chains and letting them converge sufficiently. However, effort has been put into making statistical estimates and one such is to look at the Markov chain standard error (MCSE), ensuring that it is small enough before stopping, see [21, p.15] and [45].

2.7.6 *Effective Sample Size (ESS)*

ESS is the sample size after the autocorrelation of the posterior samples has been taken into account. The correlated samples are thinned

(only every x:th sample is kept) by a factor determined by the auto correlation function and the size of the resulting sample is the ESS. This is done so that the final samples will be approximately independent. The standard estimator for effective sample size is given by

$$ESS = \frac{N}{1 + 2\sum_{i=1}^{\infty} \rho_i} \tag{41}$$

where ρ_i is the auto correlation function at lag i and N is the sample size.

2.7.7 *Metropolis-Hastings (MH)*

The MH algorithm is very general and there are many algorithms that fall into this category. It works as follows:

Set initial parameter value θ_0, then repeat;

1. Draw candidate θ_n^* from the proposal density $q(\,\cdot \mid \theta_{n-1})$.
2. Accept candidate as θ_n with probability $\alpha(\theta_{n-1}, \theta_n^*)$ or, if rejected, use θ_{n-1} instead,

until convergence with satisfactory accuracy, see [39, p.171]. The steps described above constitute the Markov kernel K, also called transition kernel.

Algorithms with acceptance probability $\alpha(\theta, \theta^*)$ and Markov kernel K based on the following satisfy the detailed balance condition stated in Section 2.7.1 and are referred to as MH algorithms.

2.7.7.1 *Acceptance Probability α*

$$\alpha(\theta, \theta^*) = \min\left\{1, \frac{\pi(\theta^*)q(\theta|\theta^*)}{\pi(\theta)q(\theta^*|\theta)}\right\}. \tag{42}$$

where $\pi(\cdot)$ is the target distribution, i.e. the prior times the likelihood, and $q(\cdot|\cdot)$ is the proposal density.

2.7.7.2 *Proposal Density q and Markov Kernel K*

The proposal density q is used to generate new candidate parameter sets and can be fairly arbitrary but should satisfy:

If $\theta \neq \theta^*$, then

$$K(\theta, \theta^*) = q(\theta^*|\theta)\alpha(\theta, \theta^*), \tag{43}$$

otherwise

$$K(\theta, \theta) = 1 - \int q(\theta|\theta^*)\alpha(\theta, \theta^*)d\theta^*, \tag{44}$$

This is the mathematical version of the description of the MH algorithm earlier in Section 2.7.7.

where α is the acceptance probability, and K is the Markov kernel.

2.7.8 *Independence Metropolis (IM)*

The IM algorithm is a special case of MH and generates candidates independently of the chain, i.e. the proposal density q does not depend on the current state θ:

$$q(\theta^*|\theta) = q(\theta^*). \tag{45}$$

As a result of this simplification, IM generates samples quickly and is used effectively once stationarity has been reached.

Many techniques are used when sampling in the multivariate case. In theory, any sampling distribution with sufficient support works but often the multivariate normal (MVN) distribution is used to sample all parameters simultaneously.

2.7.9 *Slice Sampler (SS)*

The slice sampler tries to sample inside the function graph by the use of, so called, slices and has an acceptance probability of 1. However, it doesn't work well with multimodal distributions due to the problematic nature of determining the horizontal slice, described below.

As a side note, SS does satisfy the Metropolis-Hastings-Green generalization but so does every sound MCMC algorithm, see [7, p.4,35].

It behaves much like the MH algorithm with $K(\theta, \theta^*) = q(\theta^*|\theta)$ and $\alpha(\theta, \theta^*) = 1$, if θ^* is in the support of $q(\theta^*|\theta)$, but doesn't always fulfill the MH requirements.

As mentioned above, the Markov kernel K and the sampling distribution q are one and the same and works as follows, refer to Figure 4 for ease of understanding and [13, p.3-5] for more details:

1. Sample y uniformly from the vertical slice $[0, f(\theta)]$.
2. Sample θ^* uniformly from the horizontal slice $f^{-1}[y, +\infty)$.

Keep in mind that θ^* is always accepted.

The horizontal slice is often difficult to determine. Slice samplers often use a user-defined step size ω and some variation of the following method:

1. An initial interval of size ω (called the step size) is placed randomly such that it contains θ.
2. (Expansion) Increment $n_\pm \in \mathbb{N}$ in the following fashion
   a) Step out left until $f(\theta - (a + n_-)\omega) < y$
   b) Step out right until $f(\theta + (b + n_+)\omega) < y$

The keywords in parentheses will be referred to later.

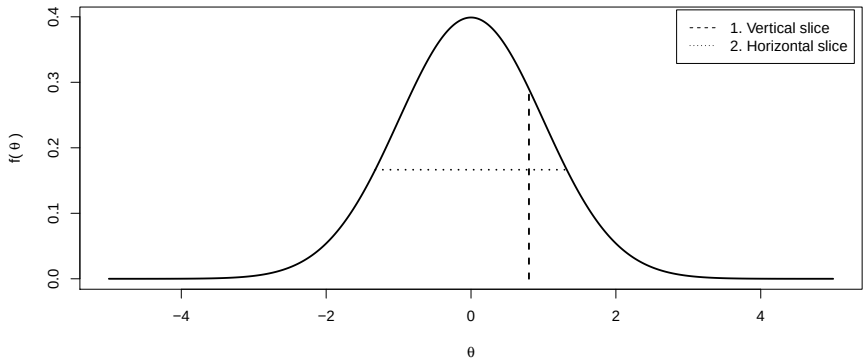


Figure 4: Example sampling for the slice sampler (SS).

where $a \in [0, 1]$, and $a + b = 1$ from step 1.

3. (Rejection sampling) Sample θ^* from the horizontal slice until $f(\theta^*) \geq y$. (Contraction) Decrease the size of the slice with each failed sampling (keeping θ within).

In the multivariate case, sampling often occurs with one parameter at a time which slows down convergence significantly in higher dimensions compared to some multivariate samplers.

2.7.10 *Automated Factor Slice Sampler (AFSS)*

AFSS is an extension of the slice sampler (SS), developed by Tibbits et al, see [46], that attempts to improve the rate of convergence in the multivariate case by reducing linear dependencies in sampling and tuning the step size ω sequentially. With diminishing tuning or if tuning is stopped, it can be used as a final algorithm for sampling from the posterior.

2.7.10.1 *Tuning Step Size*

The algorithm behaves exactly like SS but gathers information about how many expansions and rejections occur in each iteration. A Robbins-Monroe recursion is then used to tune the step size ω at certain intervals, aiming for a statistically and intuitively motivated target ratio.

The gathered statistics are used to tune ω for the i:th time at iteration $2^{(i-1)}$ after factors are recalculated. Tuning stops after a user-defined number of iterations A.

We define κ as the ratio of the number expansions to the total number of expansions and contractions.

$$\kappa = \frac{X}{X + C} \tag{46}$$

where X is the number of expansions and C is the number of contractions. The expected value of κ is estimated using the information gathered during the run and a target ratio $\alpha = 0.5$ is sought as motivated by Tibbits et al. [46].

The target ratio is achieved by setting the step size ω according to

$$\omega_{i+1} = \omega_i \frac{E[\kappa]}{\alpha} \tag{47}$$

Note that with an increased number of expansions the precision of the slice is likely to increase and fewer contractions occur. Vice versa, if there are many contractions, then the slice was likely imprecise originally as a result of few expansions. The interested reader is referred to the original article [46], which provides a more detailed explanation of the choices presented here.

2.7.10.2 *Factor Slice Sampling*

The covariance matrix of the parameters is estimated from the posterior samples. Its eigenvectors Γ_j are then used as a basis for constructing linearly independent updates. Where normally one would sample one parameter at a time, AFSS shifts all parameters according to the factors, sampling in one basis vector Γ_j at a time.

$$\theta^* = \theta + u_j \Gamma_j \tag{48}$$

where u_j is treated as the parameter that we are sampling, i.e. we need to find the vertical and horizontal slice w.r.t. u_j. Note that θ and θ^* are parameter sets.

It should also be noted that the factor sampling method will only lessen the impact of linear dependence among the parameters and will not help in the case of non-linear dependence.

2.8 GENERALIZED HYPERBOLIC (GH) DISTRIBUTION

The GH distribution is a normal variance-mean mixture with the mixture distribution set to the generalized inverse Gaussian (GIG). GH

is very general and is a superclass of the Student's t, Laplace, hyperbolic, normal-inverse Gaussian and the variance-gamma distributions. It possesses semi-heavy tails and has been claimed to model financial returns well, see [35].

With parameters μ = location, δ = peakness, α = tail, β = skewness and λ = shape, its PDF is

$$h(x) = \frac{(\gamma/\delta)^{\lambda}}{\sqrt{2\pi}\, K_{\lambda}(\delta\gamma)}\, e^{\beta(x-\mu)}\, \frac{K_{\lambda-1/2}(\alpha\sqrt{\delta^2+(x-\mu)^2})}{(\sqrt{\delta^2+(x-\mu)^2}/\alpha)^{1/2-\lambda}} \tag{49}$$

where $K_{\lambda}(\cdot)$ denotes the modified Bessel function of the second kind and $\gamma = \sqrt{\alpha^2-\beta^2}$. It is defined for all $x \in \mathbb{R}$. See Figure 5 for a visualization.

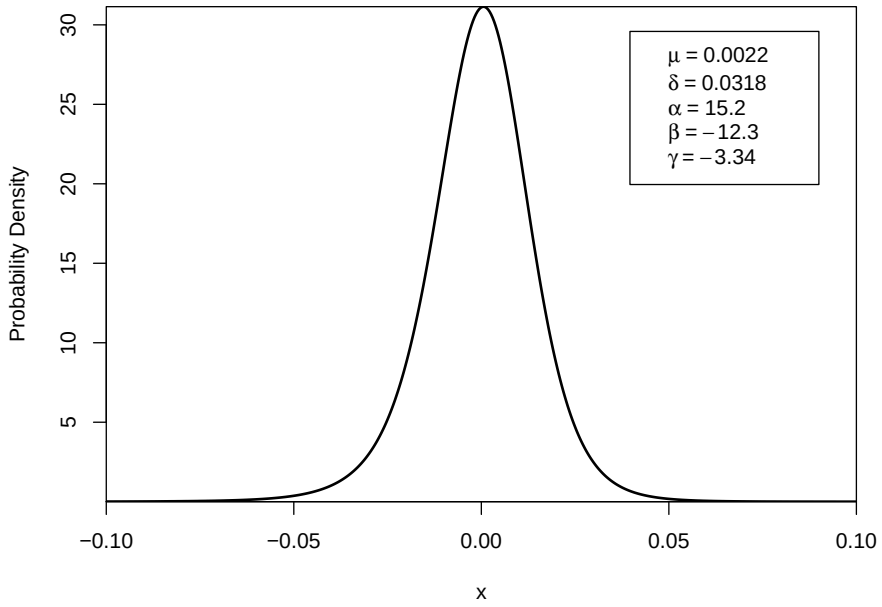


Figure 5: GH PDF h(x) with typical parameter values for modelling financial returns.

2.9 CONFIDENCE INTERVALS AND CREDIBLE INTERVALS FOR VAR AND ES

Several different methods are used in this thesis, each with its own procedure for computing intervals. The frequentist confidence interval and the Bayesian analogue, credible interval, are sometimes simply referred to as intervals. This section describes how to compute the 95% intervals for each method with the goal of being able to compare results from the different methods.

2.9.1 *Markov Chain Monte Carlo (MCMC)*

Since MCMC produces posterior samples for each parameter, VaR and ES can be calculated for each sample and the credible interval is simply the interval in which 95% of the samples fall, called the highest posterior density region. See Figure 6 for an illustration.

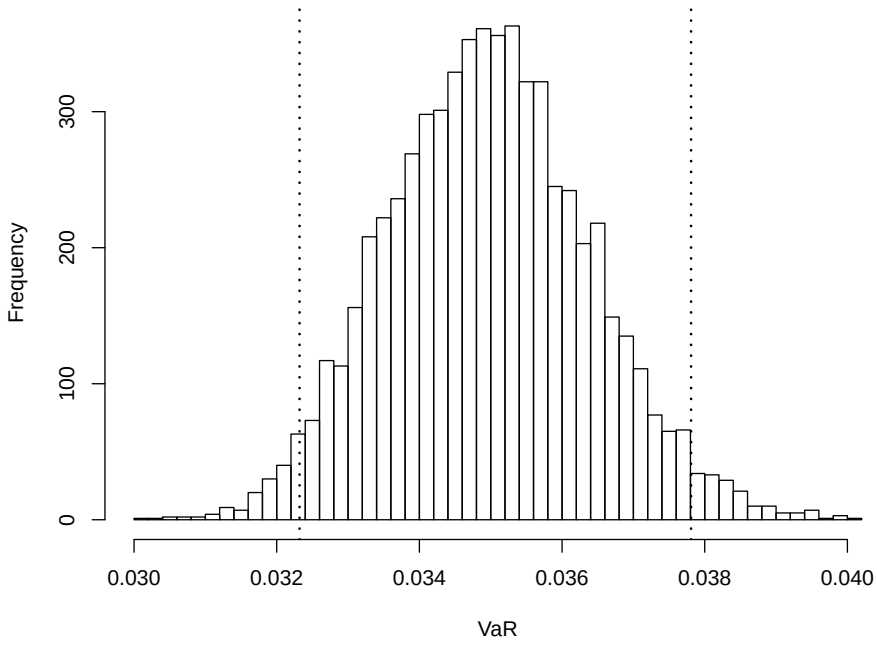


Figure 6: Posterior distribution of VaR (6250 posterior samples after thinning). The dashed lines mark the 95% credible interval.

2.9.2 *Historical*

Since *VaR* can be seen as a quantile of the empirical CDF, it is possible to compute confidence intervals for it. However, it is not possible for any desired confidence level. The procedure, described in [24, p.215], is based on the fact that the number of sample points exceeding VaR_p is $Bin(n,\ 1-p)$ distributed, where n is the sample size. One then tries to find $i > j$ and the smallest $q' \geq q$ such that

$$Pr(X_{i,n} < VaR_p < X_{j,n}) = q' \tag{50}$$

$$Pr(X_{i,n} \geq VaR_p) \leq (1-q)/2 \tag{51}$$

$$Pr(X_{j,n} \leq VaR_p) \leq (1-q)/2 \tag{52}$$

where $X_{1,n} \ldots X_{n,n}$ is the ordered sample. To be able to hit close to 2.5% probability in each direction, i.e. $q = 0.05$, there has to exist a fair amount of data points on either side of the target value. For instance, if the data set contains only 1000 points, it would not be

possible to compute a confidence interval for the historical $VaR_{0.1\%}$. The procedure is not applicable to ES, as it is not a quantile.

2.9.3 *Maximum Likelihood Estimation (MLE)*

The confidence intervals for MLE were calculated using the relative profile log-likelihood method described in [22, p.13]. If the parameter or function of interest is M (for example $M = VaR_{1\%}$), the profile log-likelihood function is defined as

$$L^*(M) = \max_{\xi} L(\sigma(M), \xi) \tag{53}$$

where L is the regular log-likelihood function for GP and $\sigma(M)$ means that σ is determined by the given M, so the maximization is only with respect to ξ. The relative profile log-likelihood function is then defined as

$$L^*(M) - L(\hat{\xi}, \hat{\sigma}) \tag{54}$$

where $\hat{\xi}$ and $\hat{\sigma}$ are the estimated parameters from MLE. So $L(\hat{\xi}, \hat{\sigma})$ is just the maximum log-likelihood. The sought confidence interval is given by all values of M satisfying

$$L^*(M) - L(\hat{\xi}, \hat{\sigma}) > -\frac{1}{2}\chi^2_{\alpha,1} \tag{55}$$

where $\chi^2_{\alpha,1}$ is the $(1-\alpha)$ quantile of the χ^2 distribution with 1 degree of freedom ($\alpha = 0.05$ if a 95% confidence interval is wanted). As can be seen in Figure 7, the interval is asymmetric, since there are less observations for the higher quantiles.

These intervals, unlike those based on standard errors, do not rely on asymptotic theory results and should therefore perform better with the small sample sizes in the tail, see [22, p.11]. Additionally, this method of calculating confidence intervals for a risk measure M directly (instead of for σ and ξ separately) captures the correlation between σ and ξ.

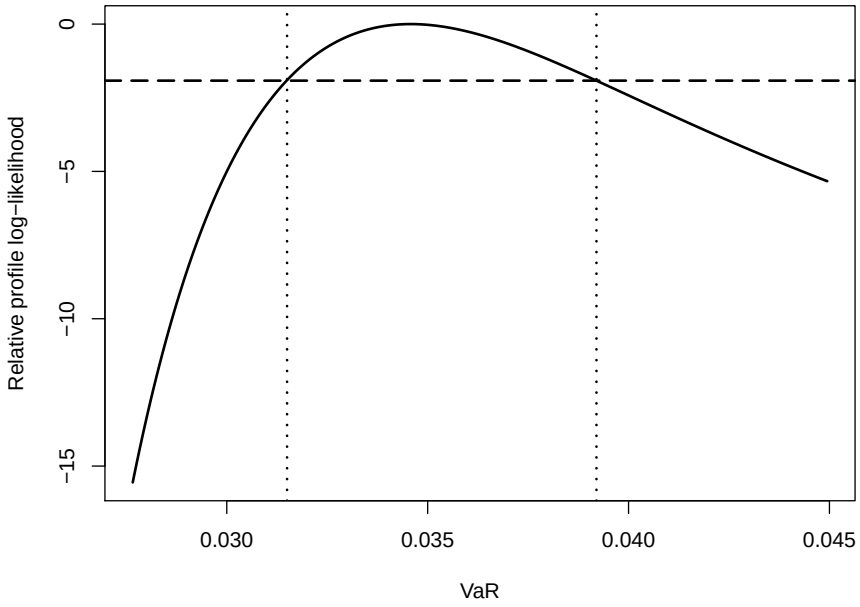


Figure 7: Relative profile log-likelihood for $VaR_{1\%}$. The dashed horizontal line is at $-\frac{1}{2}\chi^2_{0.05,1} = -1.92$ and the dotted vertical lines mark the calculated 95% confidence interval.

3

DEVELOPMENT

This chapter will focus on some decisions that were made in the process of constructing an automatic process that takes in transformed data and returns risk measures. Each section discusses one or two such crossroads. Results and discussions of discarded side-tracks are included so as to lend insight and possibly let others avoid these in the future.

3.1 SAMPLE INDEPENDENCE

A common assumption in the underlying theory is sample independence, yet it can be difficult to accomplish. The issue of independence becomes even more prominent if we want to take longer time-spans into account, which is very attractive in order to increase the small amount of data that is inherently available in extreme value theory (EVT) applications. There have been many theses on this issue alone and thus here we will focus on some standard methods that are applied in finance. Plots from this process are shown in Figure 8.

3.1.1 *Return Transformation*

The natural and standardized way is to transform the financial data into (log) returns[1], see Figure 8. This will make the time series approximately stationary. Additional modelling is rarely applied although more could possibly be done. For example, there could at times exist volatility regimes, see Section 2.3, where the instrument performs better or worse.

1 The mean of the time series should remain in the data and not be arbitrarily removed unless you have supporting information. You could, however, apply suitable additional modelling, such as some ARMA-model, on the realizations in order to remove some dependence between data points and then displace the remaining time series to a predicted expected value.

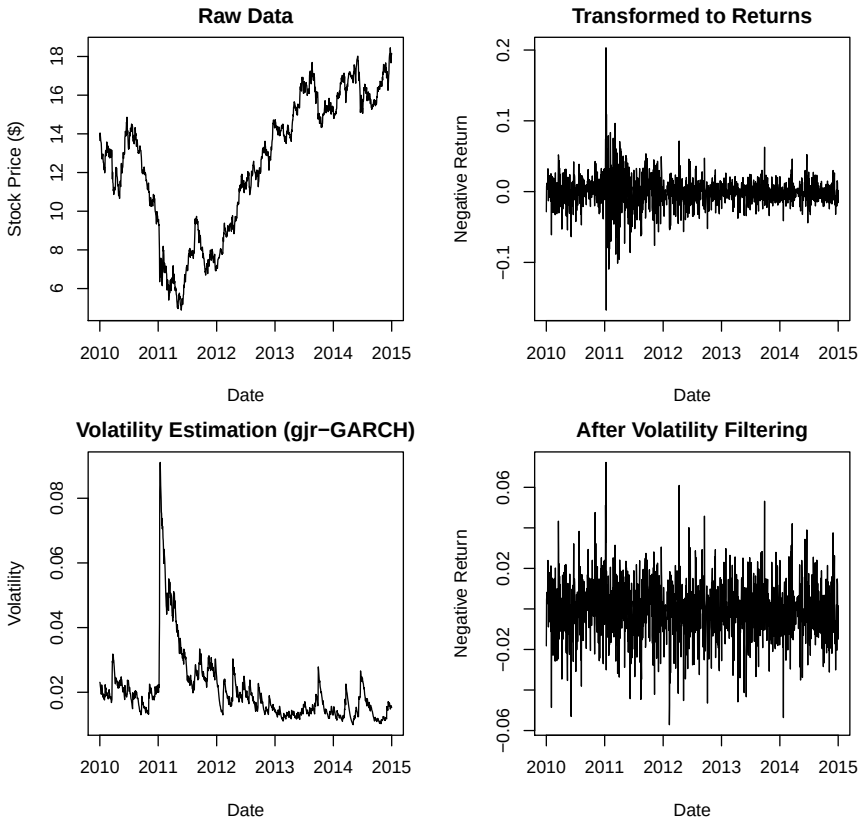


Figure 8: Data transformation and volatility filtering of the Bank of America data set.

3.1.2 *Volatility Filtering*

A well known characteristic of financial data are volatility shocks, which appear as volatility clusters, and its effects on EVT have been studied.

> ... results indicate that the dependence of the data does not constitute a major problem in the limit of large samples, so that volatility clustering of financial data does not prevent the reliability of EVT, we shall see that it can significantly bias standard statistical tools for samples of size commonly used in extreme tails studies — *[33, Sornette p.44]*

A standard method in finance for filtering and normalizing w.r.t. volatility is using an asymmetric GARCH model. For the real world data set, we use the GJR-GARCH procedure described in Section 2.3.2 and depicted in Figure 8, with the additional enhancement of letting the error terms be drawn from the generalized hyperbolic (GH) distribution instead of the normal distribution. This should affect the tails less, since GH can model semi-heavy tails. There is, however,

a flaw with this sequential modelling where, when filtering, it is assumed that the data is GH distributed. For future work it would be interesting to include modelling of volatility clusters in the main model.

On the subject of increasing reliability, it is attractive to look far back in time in order to increase the number of data points. However, the dangers of this should be carefully considered. For example, when doing this it is highly recommended to look for different volatility regimes. There is a myriad of techniques available related to this but they are outside the scope of this thesis.

3.1.3 *Further Modelling*

There is an infinite space of possible modelling solutions for data dependence available to us. Many depend heavily on the application at hand. The ones already presented are most common but there exists another, relevant only for POT, called peak declustering. It will not be covered here as a result of it being very subjective as it, much like the mean residual life (MRL) plot method, relies on interpreting graphs to decide upon a declustering threshold.

3.2 BLOCK MAXIMA (BM) VS PEAKS OVER THRESHOLD (POT)

This section will discuss the choice of BM vs POT for our problem formulation. The two major limit results of extreme value theory (EVT) both have their perks and the suitability of each for our problem formulation was investigated. Being able to automatize the process is one such criteria. However, the standard methods of both results is to graphically choose a block size or threshold.

There are some factors in favor of POT:

> ... modeling only block maxima is a wasteful approach to extreme value analysis if other data on extremes are available. — *[10, Coles p. 74]*

Moreover, related to the discussion of independence, there have been studies comparing the two:

> ... the standard generalized extreme value (GEV) estimators can be quite inefficient due to the possibly slow convergence toward the asymptotic theoretical distribution and the existence of biases in the presence of dependence between data. Thus, one cannot reliably distinguish between rapidly and regularly varying classes of distributions. The generalized Pareto distribution (GPD) estima-

> tors work better, but still lack power in the presence of strong dependence. — *[33, Sornette p.44]*

POT has been used extensively in financial applications. It has the same weakness as BM where a threshold is selected quite arbitrarily and the uncertainty in the selection process is not accounted for. However, a model presented by Gamerman, see [2], allows for automatic threshold selection and takes this uncertainty into account. Automatic threshold selection is one of the main issues considered in this thesis. Attempts were made to resolve the issues with the original model and improve performance.

In regards to BM, although it was ultimately discarded here in favor of POT, it does have its advantages yet has seen comparatively little research, at least within financial applications. See [16, p.1-3] for a more complete discussion.

3.3 THRESHOLD SELECTION

The problem of threshold selection is inherent in POT and, although a variation of Gamerman's model was finally chosen as aforementioned, all three methods discussed in the background theory, see Section 2.1.2.2, were investigated during the development process so as to provide a comparison and improve understanding. Lastly, improved models for threshold selection were iteratively developed.

3.3.1 *Fixed Threshold*

The fixed-threshold generalized Pareto (GP) model with a threshold corresponding to the 95th percentile of the data is illustrated in Figure 9. The model was evaluated for different thresholds, sample sizes, and using both MLE[2] and MCMC for comparison.

3.3.2 *Mean Residual Life (MRL) Plot*

Choosing threshold graphically from the MRL plot is highly subjective, exemplified in Figure 10 for the Bank of America data set. The plot should be approximately linear in u at values where the data is GP distributed. Determining at what threshold the plot becomes linear is obviously tricky. Arguments can be made for choosing u as low as 0.02 or perhaps as high as 0.045. This uncertainty makes a big

2 The Nelder-Mead algorithm performed well and was used for all MLE calculations.

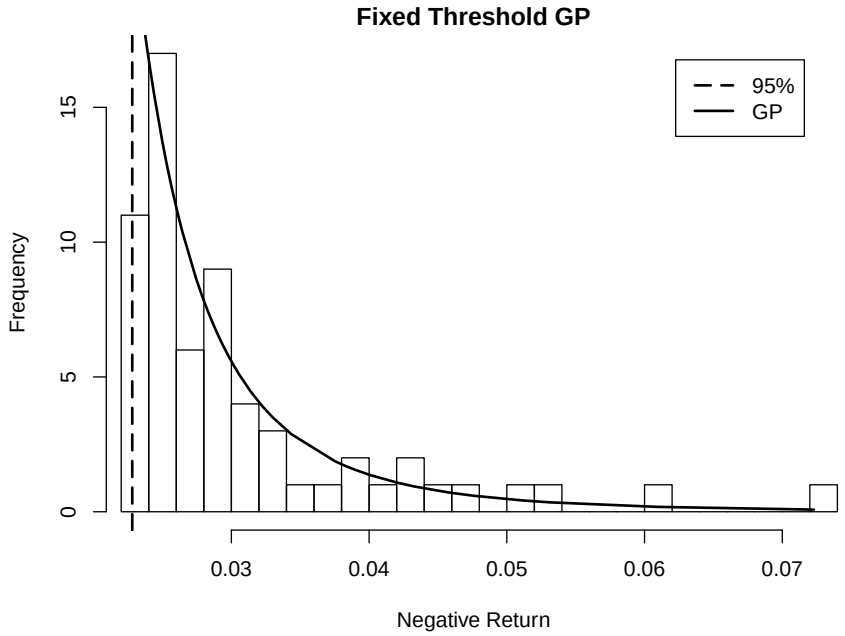


Figure 9: Plot of the fixed-threshold GP model fitted to example data.

difference since choosing $u = 0.02$ corresponds to modelling 12.4% of the data, while only 2.5% of the data is greater than $u = 0.045$.

3.3.3 *Stability of Parameters*

The effectiveness of using the stability of estimated parameters as a method for selecting threshold was investigated.

First, the effect of moving the threshold was analyzed analytically, refer to Section 2.1.2.1 for notation, by looking at the following equation:

$$p_{u=0}(x) = (1 - P_{u=0}(a)) \cdot p_{u=a}(x), \tag{56}$$

where $p_{u=0}$ is the original GP, $p_{u=a}$ is the GP with threshold $u = a$, and the factor $1 - P_{u=0}(a)$ is an adjustment such that the area of the two distributions below the graph from a to ∞ is equal.

This led to the result, also seen in Equation 10,

$$\sigma(u) = \sigma(0) + \xi u \tag{57}$$
$$\xi(u) = const, \tag{58}$$

which is in accordance with theory.

Note this linear dependence, it could be part of the reason why the AFSS algorithm behaved well.

With this result in mind, a fixed-threshold GP model was run while varying threshold using both MLE and MCMC on a simulated GP-distributed sample of size 10,000.

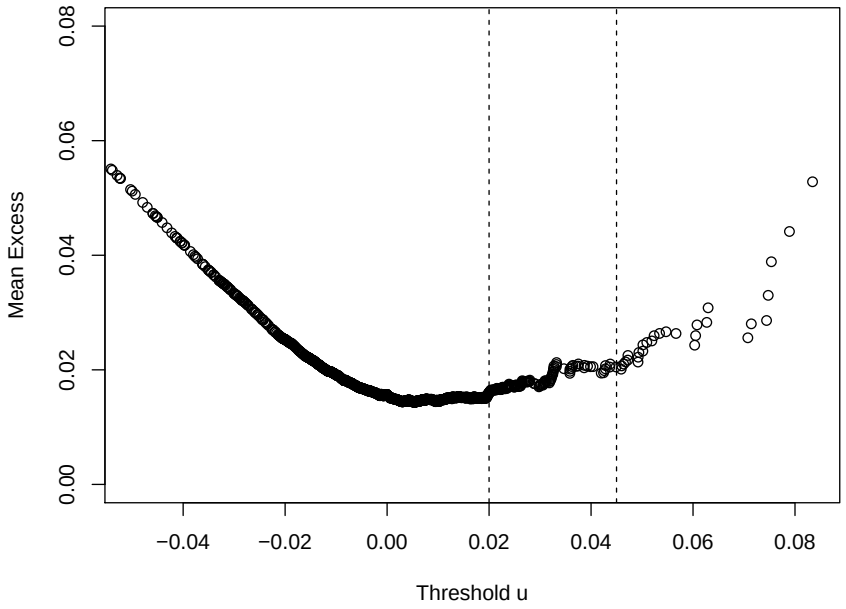


Figure 10: MRL plot for the Bank of America data set. The dashed vertical lines mark our lowest and highest estimate of the appropriate threshold.

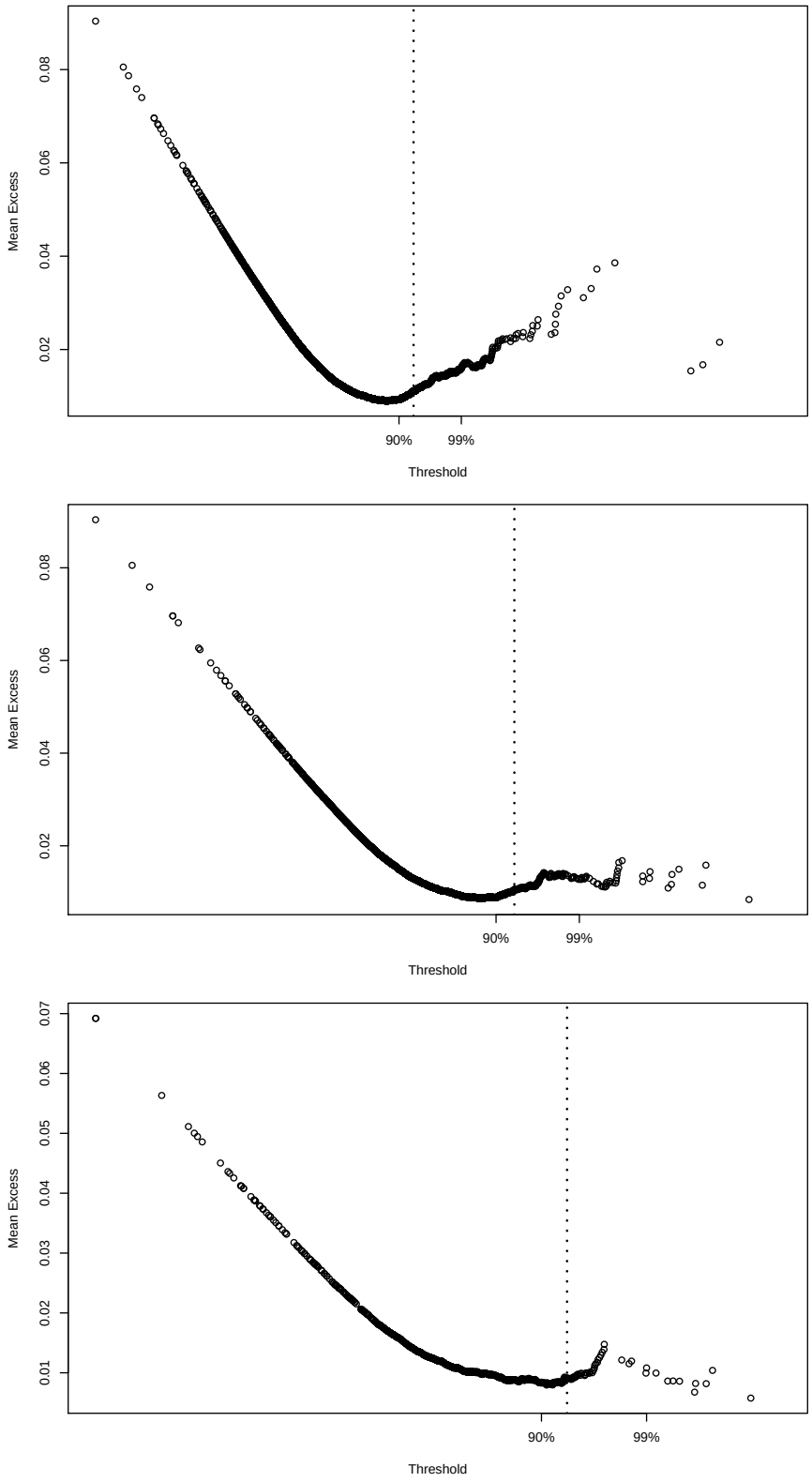


Figure 11: MRL plots for the simulated GHGP data with $N = (10000, 4000, 1000)$ (top, middle, bottom). The dashed line marks 95% of the data.

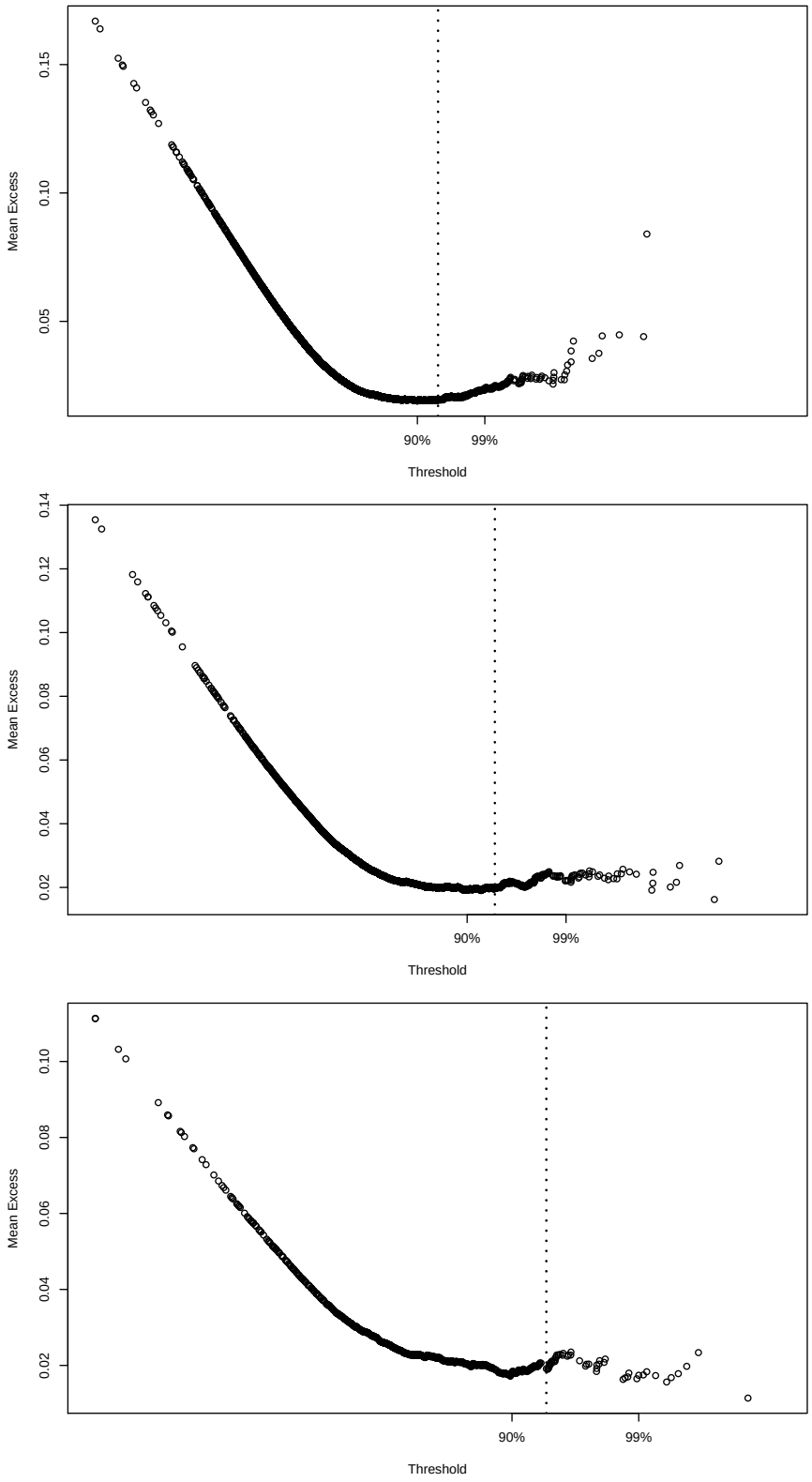


Figure 12: MRL plots for the simulated GL data with $N = (10000, 4000, 1000)$ (top, middle, bottom). The dashed line marks 95% of the data.

The MLE estimator for varying fixed thresholds on GP distributed data and σ behaved close to as predicted. However, ξ was not estimated to be constant at all, see Figure 21. Moreover, a large relative difference of as much as 15% was seen for $VaR_{1\%}$ when the threshold increased.

The test was repeated for MCMC with more stable results, see Figure 22. The changing threshold has much less effect on the quantiles compared to when using MLE though they are also not behaving optimally, compare with Equations (57) and (58).

Although "stability of threshold" was discarded, the results were useful for comparing MLE vs MCMC and were thus moved to the Results, see Figures 21 and 22.

Using stability of parameters to select threshold seems a difficult prospect, regardless of whether MLE or MCMC is used, even in the context of massive amounts of data. As a result, this method for threshold selection was discarded.

3.3.4 *Body-tail Models*

From the investigations into threshold selection techniques and the effect of choosing threshold, see Section 5.5, it became clear that threshold selection is quite arbitrary and fixed-threshold GP models are inadequate, at least when there are few relevant data points.

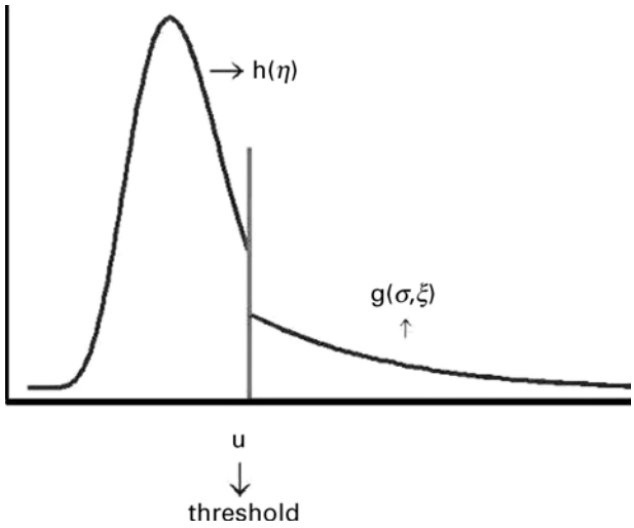


Figure 13: Gamerman's original model, from [2].

The model presented by Gamerman [2] used a Gamma distribution as the body and GP as the tail, shortened to Gamma-GP and illustrated in Figure 13. It was an attempt to both automatize threshold selection and take the threshold uncertainty into account. The transition between the distributions was very ill-fitted and may very well be the reason why the method hasn't readily reappeared. Attempts were made to resolve three serious problems with the original model:

An information criterion for choosing threshold is difficult to create, although, in essence, the body-tail models are heuristic attempts at this.

A. Threshold selection will depend on how well the body distribution models the data. The worse it models the data, the more data will fall to GP, and vice versa. Since Gamma was not able to model the data very well, the chosen threshold was much lower than expected from looking at an MRL plot. From the perspective of EVT it is better to have a threshold that is too high rather than the other way around, since convergence will be better.

B. Discrepancies in the data will affect the placement of the threshold greatly. Due to the lack of data in the tails there is often some "hole" in the histogram (great distance between closest observed variates) which might cause the body distribution to fall near zero at this point, and then the GP distribution takes over at a higher density where data is once again present.

C. The discontinuities of the original Gamma-GP disqualify many algorithms.

3.3.4.1 *GH-GP*

The first idea was to just replace the body distribution with GH which is very general and has been shown to accurately model the returns of many financial instruments, see [35, p.12].

The first attempt was using fixed GH parameters that were estimated at the start. This significantly improved threshold estimation to levels more in line with theoretical expectations. This was done in conjunction with a threshold relaxation which allowed the threshold parameter to be sampled as if continuous and then moved to a data point so as to avoid unnecessary multimodality.[3]

Seeing the seriousness of problem B stated above and since GH models the tail quite well on its own, it was used to, in a sense, smooth the discrete data set. This was done by making the two distributions match at the threshold and completely relaxing the threshold parameter. Note that this puts much more emphasis, compared to Gamerman's original model, on that the body distribution models the data quite well at least in the tails or, more accurately, the threshold parameter space. As a result, the model was changed to let the GH parameters vary as part of the MCMC process.

3 Changing the body distribution introduced another problem, namely, the threshold could, albeit rarely, exceed the quantile we were interested in. A boundary condition was placed upon the threshold so as to lie within $(90\%, min(99\%, quantile))$. This limitation applies to any application of POT. For larger quantiles ($\approx 5\%$) an approach such as historical risk should be sufficient.

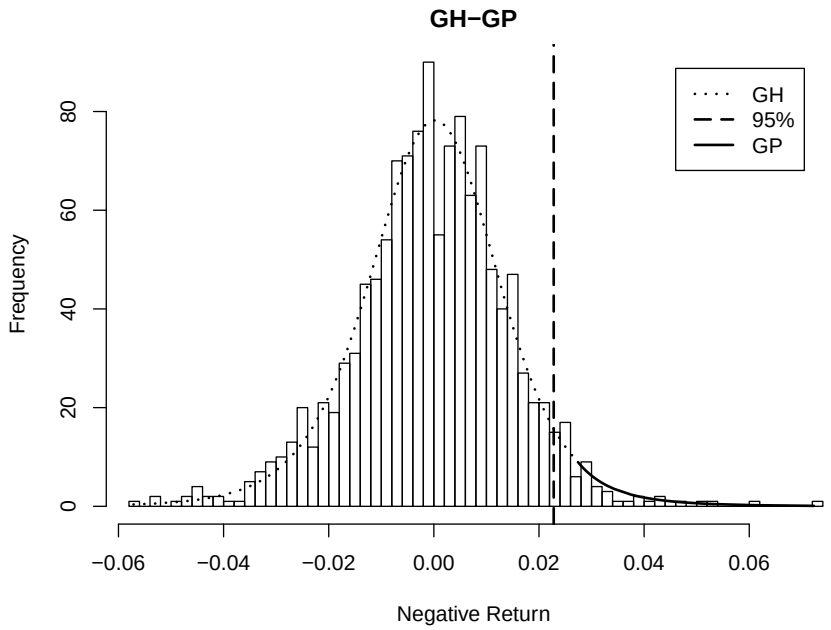


Figure 14: Plot of the GH-GP model fitted to example data.

The PDF $m(x)$ of the final model looks like this:

$$m(x) = \begin{cases} h(x), & \text{if } x \leq u \\ \frac{p(x)}{1-H(u)}, & \text{otherwise} \end{cases} \tag{59}$$

$$\sigma = \frac{1 - H(u)}{h(u)} \tag{60}$$

where H and h are the CDF and PDF of GH while p is the PDF of GP with parameter σ described by Equation (60).

Note that calculating the PDF of GH is quite computationally expensive and the CDF doesn't have a closed form, meaning that numerical integration has to be used. For these reasons we used an implementation of the PDF written in C.

If the body distribution doesn't fit the data well enough then the second improvement is not justified. In that case, the factor $1 - H(u)$ mentioned above is best replaced by factors based upon sample estimates. For example, weighting each distribution by the number of points it models as in Equation (15).

For financial returns, the second improvement also lead to approximately continuous first and second order derivatives over the threshold. This opens up a larger solution space and could also lead to faster convergence.

3.3.4.2 *GP-GP Distribution*

With the GH-GP model we experienced two problems:

A. GH can model the tails of financial data quite well but has a tendency to estimate the threshold over-confidently and very highly with consequently few data points above it. The latter is a classic case of bias vs variance. While GH has the advantage of ensuring better convergence, it may sometimes be better to include more data points in the tail.

B. Heavy computations.

   a) GH has a complex form and there is no analytical solution to its CDF so numerical methods have to be used. This can also lead to inaccuracy when normalizing the GP-tail.

   b) We are arguably modelling a much larger part of the sample than is needed or we are interested in. This also synergizes badly with the computation time per data point.

To resolve these issues, GP was used as a body distribution starting at 85%, letting the GP-tail take over when the data has converged sufficiently, see Section 2.1.2. This reduced the computation time immensely and it also estimated a wider range of thresholds with a lower mean which increased the average number of data points above the threshold.

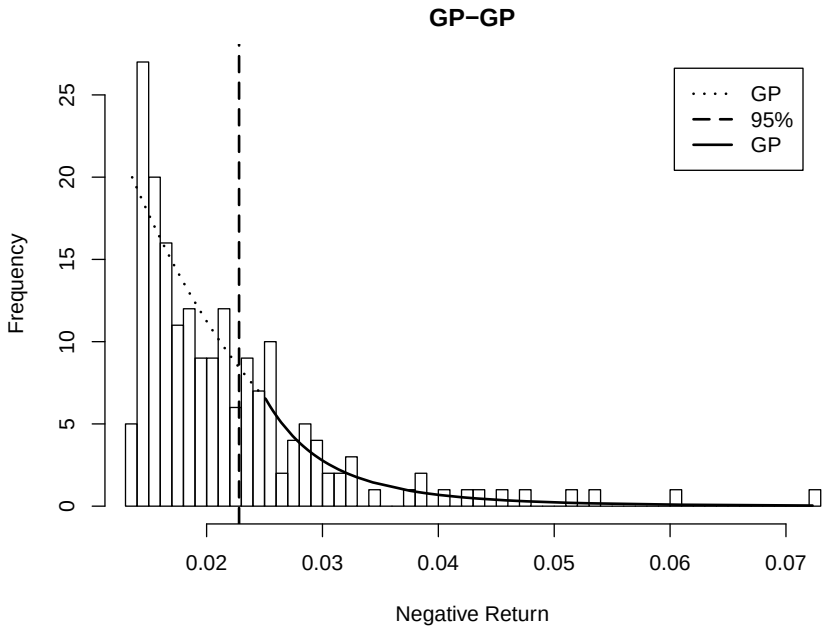


Figure 15: Plot of the GP-GP model fitted to example data.

3.4 BAYESIAN VS FREQUENTIST INFERENCE

Ignoring priors for a moment, then this is a comparison between taking the expected value (mean) vs. the most likely value (mode). The

value that we are interested in is the expected value but sometimes the mode is very close to the mean, which makes the frequentist approach interesting because it is often both easier to use and faster. However, if the distribution is, for example, multimodal or too asymmetric then the frequentist approach isn't viable, as illustrated in Figure 16.

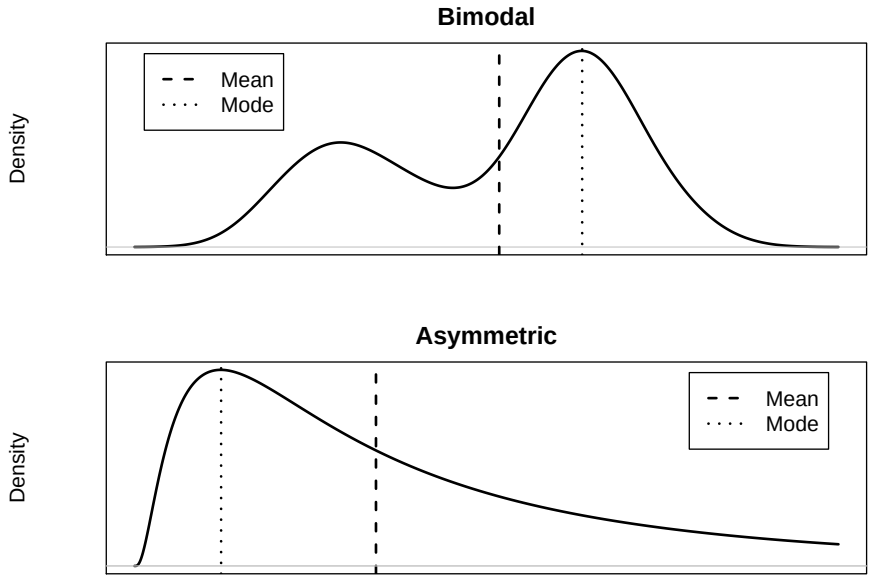


Figure 16: Example PDFs for a bimodal and an asymmetric distribution.

Additionally, calculating the expected value is not that simple. The two main drawbacks of Bayesian inference (BI) is the need to solve an integral and specify prior distributions for the parameters. In other words, prior (and often subjective) information modelled in a distribution which affects the end results. The integral causes BI to be much slower than the frequentist approach and the priors make the method harder to use, especially for non-experts.

In the case of EVT, very little data is available due to the very nature of the topic and, in such a context, using priors based on expert opinion could be powerful. For example, they can be used to avoid problems with model identification in complex models. We will compare a reference prior with an expert prior specialized for financial applications and see how much they influence our results.

A philosophical difference, which *in our opinion* favors BI, is that BI considers the data to be fixed and estimates the parameters while the frequentist approach considers the unknown parameters to be fixed, estimating based on the data at hand plus hypothetical repeated sampling in the future with similar data. Simplified, BI regards $p(hypothesis|data)$ while the frequentist approach looks at $p(data|hypothesis)$, see [41, p.15].

Moreover, unlike the frequentist approach, which only estimates the mode, BI estimates the full probability model. A confidence interval can still be calculated with the frequentist approach but previous results also indicate that MCMC provides smaller analogous intervals, see [45, p.25].

Another strong argument for BI, in the context of our problem set, is that via MCMC it is unbiased with respect to sample size unlike the frequentist approach which becomes more biased with smaller sample sizes. The effect of this in the context of heavy-tailed data was investigated, see Section 5.2.1.

A comparison of the MLE and MCMC estimators on large amounts of GP distributed heavy-tailed data was done, see Section 5.2, which favored MCMC.

In conclusion, BI and, more specifically, MCMC was chosen.

A more complete but slightly biased summary of the arguments for and against, non-specific to our purposes, can be found in [45, p.24-26] and an interesting related discussion can be found in [37]. Moreover, a deeper investigation into the effects for our problem set is provided in Section 5.2.

3.5 PRIORS

Priors are a necessary component of Bayesian inference about which there have been much debate over the years. This section will cover priors with different grades of information, from reference and weakly informative priors to expert priors. Priors have the possibility of being especially useful in EVT applications as they can be used to add information to the often meager data sets.

3.5.1 *Weakly Informative Prior for GH*

The priors for the GH parameters were all set to fairly flat distributions centered around the initial value for each parameter. For example, a normal distribution with very high variance. This allows the priors to have very little impact on the end result (compared to the data), while still being able to help numerical algorithms escape areas of flat density.

Less time has been spent on priors for the GH distribution, for two main reasons:

- There is a lot more data for the GH part of the distribution, usually around 95% of the total data. Unless overly informed, this reduces the impact of the priors massively.

- The impact of GH on the end result is quite small, rather it is mainly a tool to decide where the GP modeling should start and therefore generally less important.

3.5.2 *Reference Prior for GP*

In order to compare with frequentist inference, as uninformative priors as possible were used. These are the reference priors described in Section 2.4.2

$$p_\sigma(\sigma, \xi) \propto \frac{1}{\sigma\sqrt{1+\xi}\sqrt{1+2\xi}} \tag{61}$$

$$p_\xi(\sigma, \xi) \propto \frac{1}{\sigma(1+\xi)\sqrt{1+2\xi}} \tag{62}$$

These are used for all results, except for when specifically testing the informative priors. See Figure 17 for a visualization.

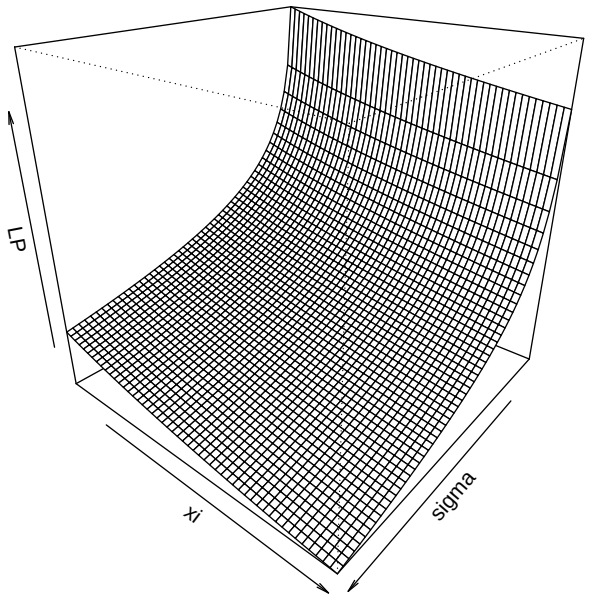


Figure 17: Log-probability surface of the combined reference prior from Equations (61) and (62) for $\sigma \in [0.001, 0.1]$ and $\xi \in [0.01, 1]$. Same view as for the informed prior in Figure 18.

3.5.3 *Prior Elicitation from Expert Opinion*

The goal here is to supplement the data with prior knowledge from an expert in the field. This so-called elicitation procedure is based on [12, p.467] but modified for our particular application. The idea is to ask the expert for estimates of VaR, which he or she is presumably very familiar with, and then convert these into a prior distribution for the GP parameters σ and ξ. Recall from Equation (17) that the VaR for a small tail probability p is given by

$$VaR_p = u - \frac{\sigma}{\xi}\left\{1 - \left[\frac{N}{N_u} \cdot p\right]^{-\xi}\right\} \tag{63}$$

The elicitation is done in terms of

$$d_1 = VaR_{1\%} \tag{64}$$
$$d_2 = VaR_{0.1\%} - VaR_{1\%} \tag{65}$$

and the expert is asked for the median and 90% quantile for d_1 and d_2. We then take marginal priors of the form

$$d_1 \sim Gamma(a_1, b_1) \tag{66}$$
$$d_2 \sim Gamma(a_2, b_2) \tag{67}$$

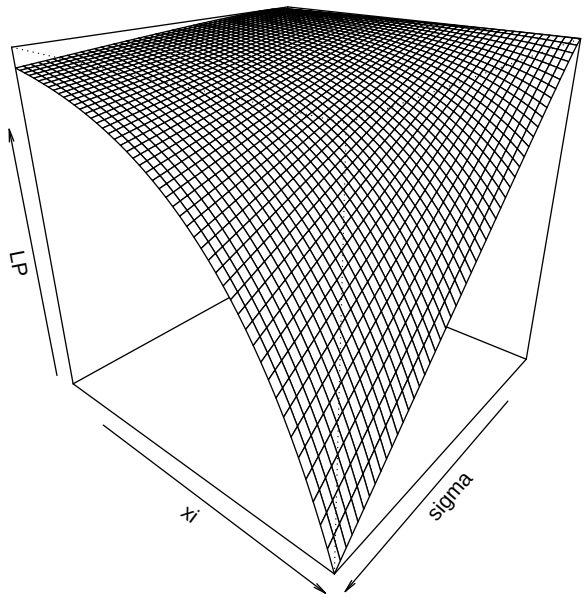
and calculate the hyperparameters a_1, b_1, a_2, b_2 from the information received from the expert. The joint prior is then

$$f(d_1, d_2) \propto {d_1}^{a_1 - 1} exp(-b_1 d_1)\ {d_2}^{a_2 - 1} exp(-b_2 d_2) \tag{68}$$

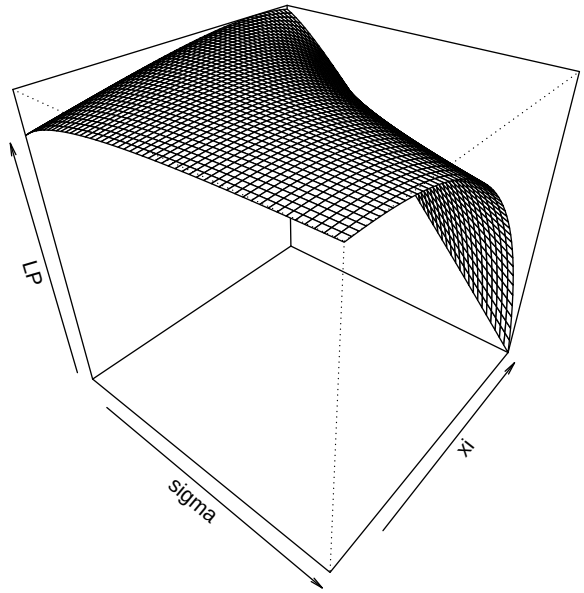
Substitution of the expression for VaR from Equation (63) and multiplication by the Jacobian of the transformation $(VaR_{1\%}, VaR_{0.1\%}) \rightarrow \theta = (\sigma, \xi)$ leads to the following prior for the GP parameters

$$\begin{aligned} p_{\sigma,\xi}(\sigma, \xi) \propto & \left[u^* - \frac{\sigma}{\xi}(1 - p_1^{-\xi})\right]^{a_1 - 1} exp\left[-b_1\left\{u^* - \frac{\sigma}{\xi}(1 - p_1^{-\xi})\right\}\right] \\ & \times \left[\frac{\sigma}{\xi}(p_2^{-\xi} - p_1^{-\xi})\right]^{a_2 - 1} exp\left[-b_2\left\{\frac{\sigma}{\xi}(p_2^{-\xi} - p_1^{-\xi})\right\}\right] \\ & \times \left|-\frac{\sigma}{\xi^2}\left[(p_1 p_2)^{-\xi}(log\ p_2 - log\ p_1) - p_2^{-\xi} log\ p_2 + p_1^{-\xi} log\ p_1\right]\right| \end{aligned} \tag{69}$$

where $p_1 = 0.01 \cdot N/N_u$, $p_2 = 0.001 \cdot N/N_u$ and u^* is the prior mean of the threshold (should in theory be the actual current threshold value, but we don't want the prior to change during runtime). See Figure 18 for a visualization.



(a) Same view as the reference prior in Figure 17.



(b) Rotated and zoomed in on smaller σ. The peak is at $\sigma = 0.004673469$ and $\xi = 0.3268$.

Figure 18: Log-probability-surface of the informed prior from Equation (69) with the expert's opinion equal to historical VaR. The upper plot has parameters $\sigma \in [0.001, 0.1]$ and $\xi \in [0.01, 1]$, and the lower plot has $\sigma \in [0.001, 0.01]$ with the same ξ.

3.6 BAYESIAN METHODS

There are many methods available, including iterative quadrature, variational Bayesian methods, Laplace approximation (LA), and MCMC. Iterative quadrature tries to solve the integral by numerical methods, both Laplace approximation and variational Bayesian methods are deterministic approximations, and MCMC is a stochastic sampler that is very general with many algorithms to choose from.

Initially, MCMC was chosen because it is unbiased for smaller sample sizes, see Section 3.4, and supports discrete parameters well but as this latter restriction was lifted, Laplace approximation (equivalent to MLE with priors for this purpose) was used to attain better starting parameters. Finally, iterative quadrature was tried but was found to be too computationally heavy.

3.7 MCMC ALGORITHMS

Two out of many MCMC algorithms were selected after showing good performance on the problem set. This section will discuss their uses and some improvements that were made to them in order to improve performance or to satisfy underlying theoretical conditions.

3.7.1 *Independence Metropolis (IM)*

The independence Metropolis (IM) sampler, see Section 2.7.8, is very useful for sampling quickly if the proposal distribution isn't misinformed. A useful technique is therefore to use a more advanced algorithm first and then switch to IM in order to get more samples and decrease the Markov chain standard error (MCSE) to an acceptable level.

Another use, for example if the Hessian estimation of the covariance fails, is to sample during a short run after the Laplace approximation (LA) and use these samples to estimate the covariance instead, see Section 3.8.

As long as the support of the proposal distribution is large enough then IM will behave very well, even in the context of multimodal distributions. As a result, the covariance supplied to IM is multiplied by 1.1 as recommended by [45].

For our purposes the proposal distribution will most often be multivariate normal (MVN) but for some cases the multivariate Cauchy (MVC) distribution may appear.

3.7.1.1 *Handling Bounds*

There are often parameter bounds that must be satisfied, e.g. $\sigma \in (0, \infty)$, and the two most intuitive ways of handling this is to set the likelihood outside the bounds to zero, with the help of priors, or use a proposal distribution with support equal to the bounds.

The former is an easy solution but, with difficult bounds, it can cause low acceptance rates and high correlation between posterior samples. The latter is better but can be a hard task to achieve at times. We will discuss some alternatives. It will be seen that it is very important to consider the effects of the method you use to handle bounds as it can easily affect the entire procedure.

The ideas presented here are true in general but choices w.r.t. optimization are made with MVN in mind, as proposal distribution.

Interval

The idea of interval is to mirror proposals that fall outside against the bounds until they are inside. If the bounds are semi-infinite then this is accomplished by a single mirroring but if the bounds are finite it may take many.

This method, recommended by [45], ensures that only one sample has to be generated, which may be very useful for some algorithms.

Looking back at the definition of the acceptance probability, see Equation (42), it is seen that is must be possible to both sample from q and calculate the density. As a result of the interval method there are many ways of ending up at a specific value x when sampling from a distribution q^*. For instance, assume a univariate case with a semi-infinite bound $(0, \infty)$ then

$$q(x) = q^*(x) + q^*(-x) \tag{70}$$

where q^* is the proposal distribution used to generate the sample but q is the real distribution after passing the interval function.

This was implemented and tried using recursion to calculate the possible combinations resulting in a value x and stopping once a certain tolerance is reached. The number of calculations increase heavily with the number of bounded parameters but can still be counted as viable. However, the main strength of IM is the ability to sample very quickly and therefore another method was sought.

Bijections

Another technique that was tried was using bijections. However, this enlarged the parameter space greatly, and as a result, using a normal distribution to sample was out of the question. Moreover, a very non-linear relation between the parameters was created. Hence, the bijection technique was rejected.

Resampling

This resampling improvement has been added to the LaplacesDemon package [45] under algorithms IM and CIM (the latter having MVC proposal distribution instead of MVN).

Under the assumption that the main part of the proposal distribution q is within the bounds, then resampling is an alternative. However, at face value, this would require an integration of the proposal distribution in order to normalize the distribution, setting the probability inside the bounds to one. For the univariate case it becomes:

$$q(x) = \frac{q^*(x)}{\int_a^b q^*(y)dy} \tag{71}$$

where a and b are the lower and upper bounds. Note that in the multivariate case they are allowed to depend on the parameters.

However, the integral factor is constant under IM since q is independent of the last parameter, see Equation (45), unlike for e.g. the popular algorithm "random walk Metropolis". As a result, the factor cancels out when calculating the acceptance probability and does not have to be calculated. The resampling method is therefore very well suited for IM but would require heavy calculations for most other common algorithms.

3.7.2 *Automated Factor Slice Sampler (AFSS)*

The automated factor slice sampler (AFSS), see Section 2.7.10, is a powerful general-purpose sampler that performs well in many cases. It is suitable for an automatic process because it requires hardly any manual tuning if provided with good enough initial values and covariance estimates.

In an attempt to keep the notation of the original article intact and not cause confusion, X is still defined as the number of expansions not including the initial randomly placed interval but X^* does include it. Moreover, the number of contractions C has a parallel S, which is the number of samplings and includes both rejections and acceptances.

Also note that although everything is multivariate in practice, it will not be explicitly stated but should be quite easily understood nonethe-

less. For example, ω is a vector, since the step size is different for each parameter.

3.7.2.1 *Handling Bounds*

The interval method discussed earlier in Section 3.7.1.1 is not applicable to AFSS although it has seen use. The reason for this is that we get a malformed sampling probability that is no longer uniform. This goes to further show that seemingly small changes may affect the internal workings differently depending on the algorithm.

Using bijections to solve the issue is also a bad idea since they are often very far from linear, which makes the algorithm behave badly.

The simple and intuitive way, to set the probability of the parameter to occurring outside the bounds to zero is equivalent to resampling and works well.

3.7.2.2 *Improvements to the Original Algorithm*

A few problems were discovered when the original algorithm was inspected and used. The following sections seek to discuss and remedy these issues.

Normalized Factors

The initial configuration of the step size vector ω, see Sections 2.7.9 and 2.7.10.1, held virtually no meaning in the original algorithm and issues between early tunings could occur as ω did not relate to the parameter space in any way.

A change was made so that each factor Γ_j (eigenvector of the covariance matrix of the parameters, see Section 2.7.10.2), is normalized with respect to its j:th element (the j:th element becomes 1). This results in the vector ω relating directly to the parameter space, which makes ω easier to understand and initialize, as well as making ω relevant between tuning regimes in the early stages of the run when the covariance estimate can shift a lot.

These improvements have been added to the LaplacesDemon package [45] under the algorithm AFSS.

Starting Expansion Heuristic

Implementations of AFSS normally have a user-defined maximum number of expansions before they give up on finding the horizontal slice. This isn't a problem in the long run as the problem often resolves itself as tuning takes place. However, in some cases it can take a long time for this to occur and, of course, the obtained samples are

erroneous. An alternative approach to abandoning ship is to tune the step size ω at this user-defined maximum m.[4]

If it is the first sampling since tuning reset, then every time the slice is expanded a multiple of m times, ω is doubled and X^* is halved.

The intuitive motivation for this is that if ω is doubled, then, with slightly improper notation,

$$E[X^*|2\omega] \geq X^*/2 \tag{72}$$

and if m is large enough, then they are approximately equal. It should be noted that doubling ω is a very modest increase when considering large values of m.

If we wanted to rid the implementation of this limitation then we would have to analyze how the total number of samplings S is affected by the change in ω.

That it is only done at the first sampling isn't a real limitation because if the step size ω is badly chosen then problems will manifest immediately.

Starting Sampling Heuristic

The other side of the coin is when the step size ω is much too large and expansion stops immediately. This results in many contractions of the horizontal slice before finding an acceptable slice. The idea here is to let these contractions also apply to ω under certain conditions.

If it is the first sampling since tuning reset and $X^* = 1$, then every time we contract a multiple of m times we set ω to be length of the current horizontal slice and reset the total number of samplings S to zero.

Expansion-Sampling Ratio

The original AFSS algorithm used κ defined in Equation (46) but this technique has a problem. It works well for when the number of contractions grow large but fails to tune sufficiently when the number of expansions grow. If C grows then $\kappa \to 0$ but as X becomes very large then $\kappa \to 1$. This means that with a target ratio $\alpha = 0.5$, ω is multiplied by a factor $\frac{\kappa}{\alpha} \in [0, 2]$. We would like the upper bound to be significantly greater than 2.

Therefore, the following ratio is instead proposed;

$$\kappa^* = \frac{X^*}{S}, \tag{73}$$

4 The results were created using $m = 50$.

with the target ratio $\alpha^* = 1$, equivalent of the old $\alpha = 0.5$ in accordance with the results in [46].

The effect is a ratio that can tune quickly in both directions, see Figure 19. Additionally, by counting all samplings and including the initial expansion, we get a better statistic, as a good sample run, such as 1-expansion-1-sampling, will be counted. Lastly, the step size ω is never set to zero.

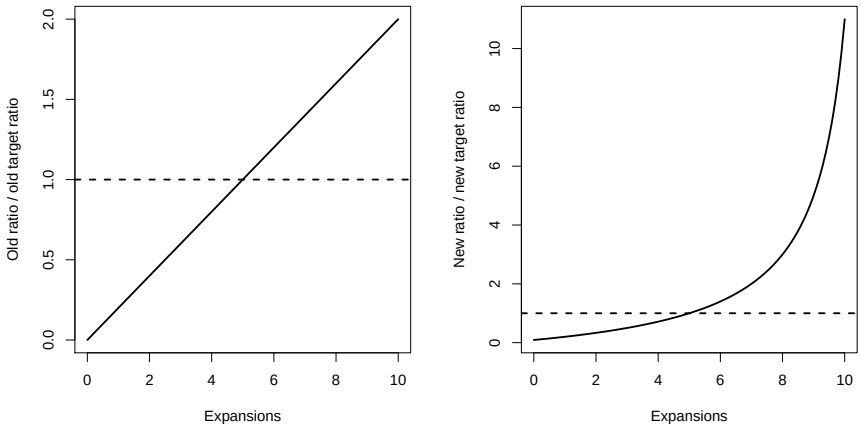


Figure 19: Comparison of old and new AFSS ratios while varying X with $X + C = 10$.

Note that with this improvement, the rough starting heuristics mentioned above become less important but can still be useful so as to always get proper samples, which is important since they are used to estimate the factors.

3.8 INITIAL VALUES AND COVARIANCE

The first step is to generate initial values based on prior knowledge about financial data. If many acceptable values are generated then the set of parameters which result in the best log-likelihood is chosen.

From here on there is a multitude of ways to explore the parameter space (to calculate the covariance) such as variational Bayesian methods, MCMC, or even iterative quadrature. However, a quite strong and fast initial method is Laplace approximation (LA) and a Hessian estimate of the covariance. In the automatic procedure much of the information provided by LA was ignored and it basically just finds the parameters that maximize the joint posterior distribution.

Note that the Hessian estimation should be multiplied by $\frac{2.38^2}{L}$, where L is the number of parameters, before being supplied to MCMC, see [40, p.113].

3.8.1 *Contingent Covariance Sampling*

If the Hessian cannot be inverted, then the process reverts to IM with an educated guess for the proposal covariance and the generated start values as mean. The obtained samples are then used to estimate the covariance matrix.

The contingent covariance sampling is not expected to be stationary and is just a method for exploring the parameter space so as to allow AFSS to work better.

3.9 STATIONARITY

AFSS is a powerful algorithm that converges quickly when applied to many problems. It is therefore run first, for 1,000 iterations, in order to achieve stationarity and find good parameter values for future IM runs. If stationary was not reached, then it continues for another 5,000 iterations.

Many other algorithms were tried as well but AFSS performed better than most with less configuration, especially after the improvements in Section 3.7.2.2. However, it is likely that there are other algorithms that could be configured to perform better given the task at hand.

3.10 ACCEPTANCE RATE

Previous research has shown that the theoretical ideal acceptance rate should be approximately in the range $(0.15,\ 0.5)$, recommended by [45], dependent on the target distribution. For example, it has been shown that the univariate normal target distribution has an ideal acceptance rate of 50% while for MVN it was approximately 23%, see [40].

The result of a suboptimal acceptance rate is slower convergence to the target distribution. However, since we already have stationarity at this point and our target distribution isn't univariate normal, the main concern is making sure the support of the proposal distribution is large enough to cover the entire parameter space properly. With this in mind we seek an acceptance rate in the range $(0.15, 0.4)$. In order to achieve this acceptance rate, we use an adaptive version of IM which estimates covariance from the samples until an acceptable rate is reached.

Another problem that can occur due to a low acceptance rate can be more correlation between samples but this is resolved in the next part.

3.11 MCSE

As a final step, in order to lower MCSE to an acceptable level, IM is run for 50,000 iterations and samples are thinned depending on how correlated samples were in previous runs, see Section 2.7.6.

Two terms commonly used in literature about MCMC are burn-in and stopping time which in this process would be, respectively, all iterations up until the final sampling run and when MCSE has reached an acceptable level, with some margin.

Posterior predictive checks were performed at this point but with so little data they were unable to diagnose if there were any problems. Instead simulations took the role of scrutinizing the correctness of the models.

4

RESULTS

This chapter contains results from the main thread from the development and some information on how the results were generated. Results are from the comparison of Bayesian vs frequentist inference with focus on MCMC vs MLE, the effect of variations in threshold selection, model comparisons, and the effect of priors.

In each simulation, a sample of 10,000 data points was generated and reused in every run. Parts of the sample were used when testing was performed with smaller data sets.

Moreover, in accordance with our problem formulation, all the data samples used were heavy-tailed.

The first sample was generated from a GH body distribution merged with a GP distribution, the threshold at 95%.[1]

The second simulation was done using the generalized lambda (GL) distribution which has been shown to be a good alternative for modelling and simulating financial data, see [8].[2]

The third sample is real data from the stock Bank of America (1258 data points or 5 years) that was transformed and filtered using the methods described in Section 3.1. Histograms of the data sets described above are shown in Figure 20

Although 10,000 data points ($\approx$ 40 years) is rarely realistic, it is useful for giving a reference point and showing the effect of a decreasing sample size.

The reference prior described in Section 3.5.2 was used, unless otherwise stated, in order to have a parameter estimate that was as unbiased as possible to allow for comparisons with MLE.

In the tables, "Avg." corresponds to either the mean or the mode (depending on the algorithm) while LB and UB are the lower and upper

1 The GHGP parameters were $\mu = 0.002200946$, $\delta = 0.031815232$, $\alpha = 15.241766213$, $\beta = -12.325859594$, $\lambda = -3.336424312$, $u = 0.022438637$, $\sigma = 0.007726189$, and $\xi = 0.3$.

2 The GL parameters were $\lambda_1 = -0.0003$, $\lambda_2 = -4$, $\lambda_3 = -0.06$, and $\lambda_4 = -0.06$

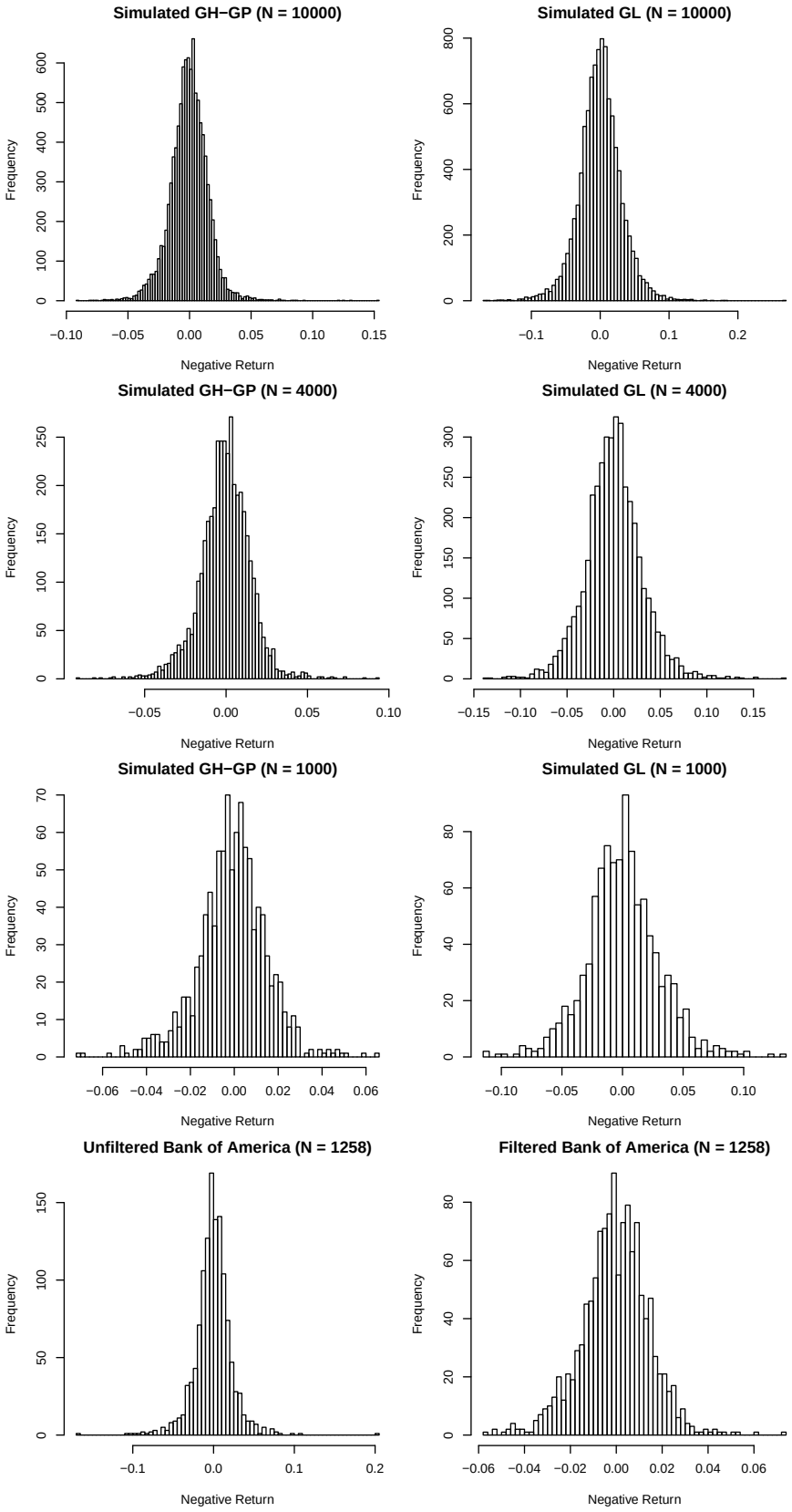


Figure 20: Histograms of the data sets used for testing.

bounds for the 95% confidence interval, or credible interval in the case of MCMC.

4.1 BAYESIAN VS FREQUENTIST INFERENCE

A comparison of the MLE and MCMC estimators was performed for a large amount of GP distributed data, see Figures 21 and 22.[3] This is related to the stability of parameters that was discussed earlier in Section 3.3.3.

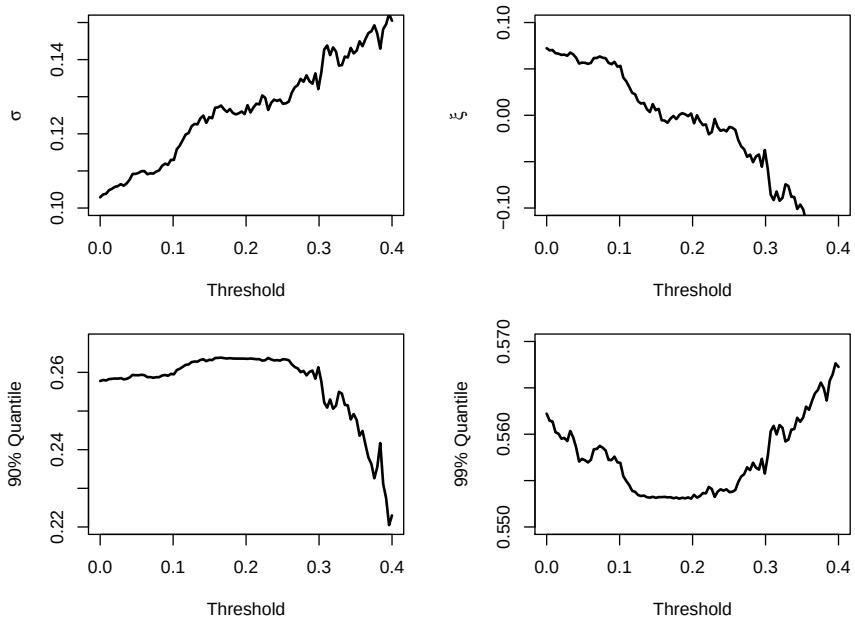


Figure 21: Effect of threshold on parameters and quantiles using MLE for 10,000 GP data points generated using $u = 0, \sigma = \xi = 0.1$.

Theory predicts MCMC is less biased than MLE for smaller sample sizes. This was investigated in Figure 23.

4.2 EFFECT OF THRESHOLD

The mean of the GP model (MCMC) was plotted in Figure 24 while varying the threshold and the posterior samples of GP-GP were plotted in Figure 25 for inference. In the prior case, the GH-GP sample was chosen in order to have a case where we know where the data has absolutely converged to GP.

3 The large GP data set was simulated using parameters $u = 0$ and $\sigma = \xi = 0.1$.

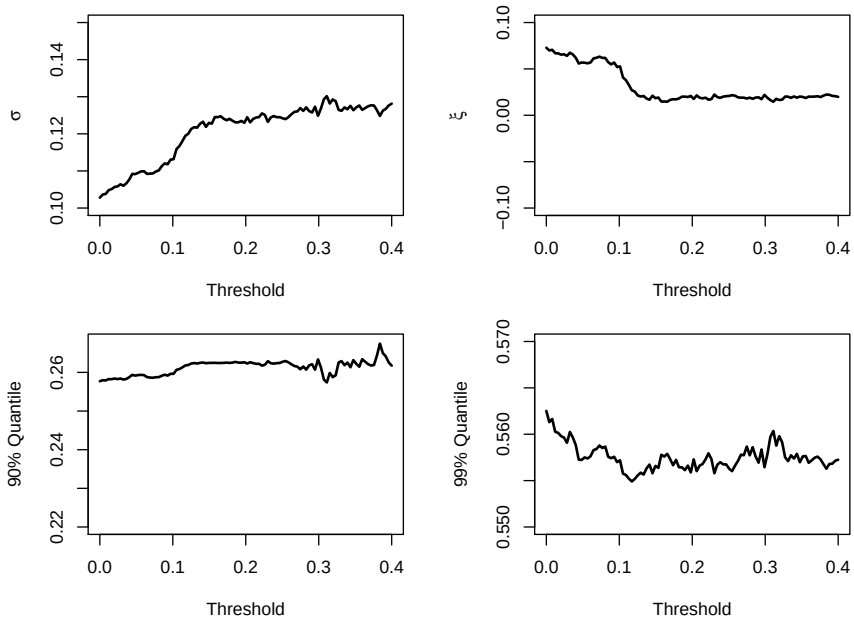


Figure 22: Effect of threshold on parameters and quantiles using MCMC for 10,000 GP data points generated using $u = 0, \sigma = \xi = 0.1$.

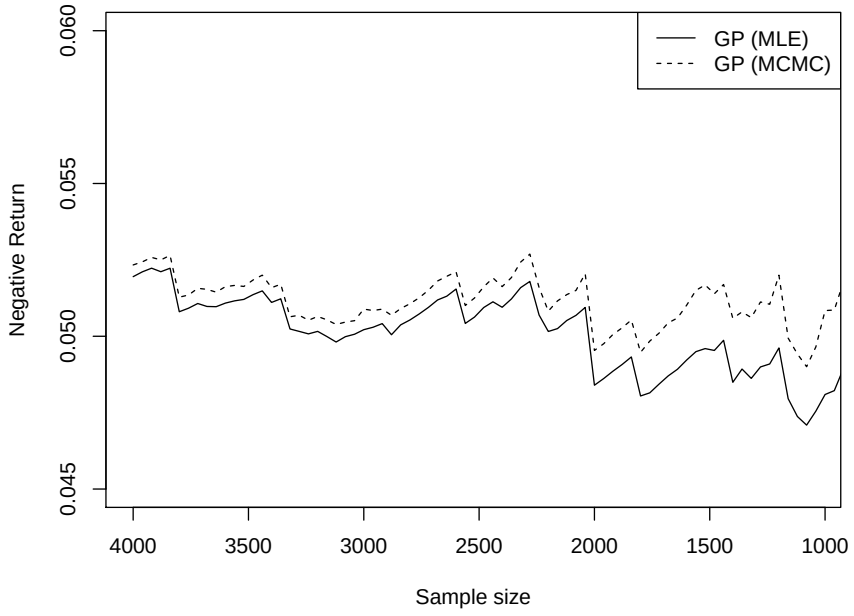


Figure 23: Mean of $ES_{1\%}$ for the GP model at 95% fixed threshold as GH-GP sample size decreases.

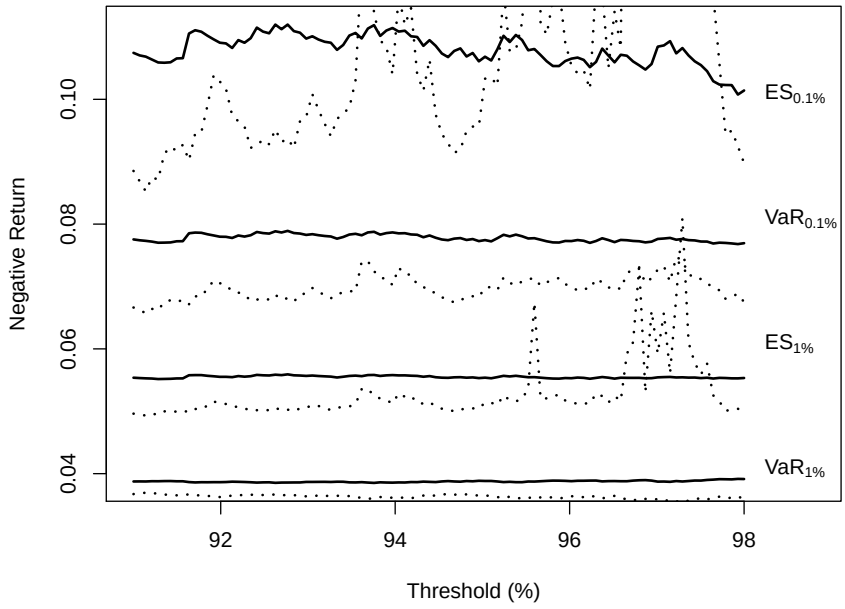


Figure 24: Mean of risk measures from the GP model for varying thresholds. The solid and dotted lines indicate GH-GP samples of size 10,000 and 1,000, respectively.

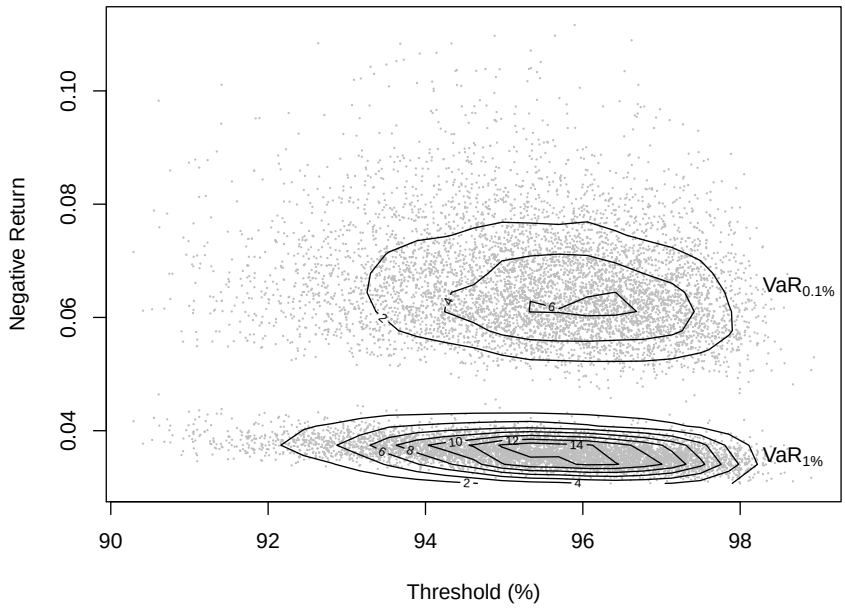


Figure 25: Posterior samples from the GP-GP model for 1,000 GH-GP samples.

4.3 MODEL COMPARISON

The GH-GP simulation was done so as to show the best-case scenario for all algorithms with the exception of GP-GP. Note that nothing should beat the fixed-threshold GP models in this data set, not even GH-GP. If GH-GP and especially GP-GP can perform nearly as well then it is very good. This gives us a reference point for comparisons, see Tables 2, 3, and 4.

Furthermore, results are presented for GL, see Tables 5, 6, and 7, and a real data set, see Table 8.

Although the models attempt to estimate the analytical value, too much emphasis shouldn't be put on it as, especially in the context of smaller sample sizes, the random variates may indicate something slightly different.

4.4 PRIORS

The procedure of prior elicitation, see Section 3.5.3, was tested on the Bank of America data set, using GP-GP as model. Recall that the expert is asked for the median and 90% quantile of

$$d_1 = VaR_{1\%} \tag{74}$$

$$d_2 = VaR_{0.1\%} - VaR_{1\%}. \tag{75}$$

Three different opinions and two different levels of certainty combine to produce six fictive elicitation scenarios, intended to test the impact of our informative prior on the end result. The opinions and certainty levels are described in Table 1.

Table 9 contains the final risk measure results when considering each of the six scenarios as well as the usual reference prior, for comparison. Figure 26 shows the fitted GP-GP model for the scenarios "Historical Certain", "Above Certain", and "Spread Certain", as well as for the reference prior.

Table 1: Explanation of prior elicitation scenarios

Historical (H)	The expert's opinion/estimate of both $VaR_{1\%}$ and $VaR_{0.1\%}$ coincide with the historical VaR.
Above (A)	Both $VaR_{1\%}$ and $VaR_{0.1\%}$ is 1.5 times their historical counterparts, i.e. the expert believes that the actual risk is substantially higher than the historical data suggests.
Spread (S)	$VaR_{1\%}$ is equal to historical but $VaR_{0.1\%}$ is twice its historical value. This would mean that the expert does not believe that the (extreme) 0.1% risk is captured in the historical data but rather is much higher.
Certain (C)	When asked for the 90% quantile of a given answer, the value (median) is multiplied by a factor 1.2.
Uncertain (U)	When asked for the 90% quantile of a given answer, the value (median) is multiplied by a factor 2.

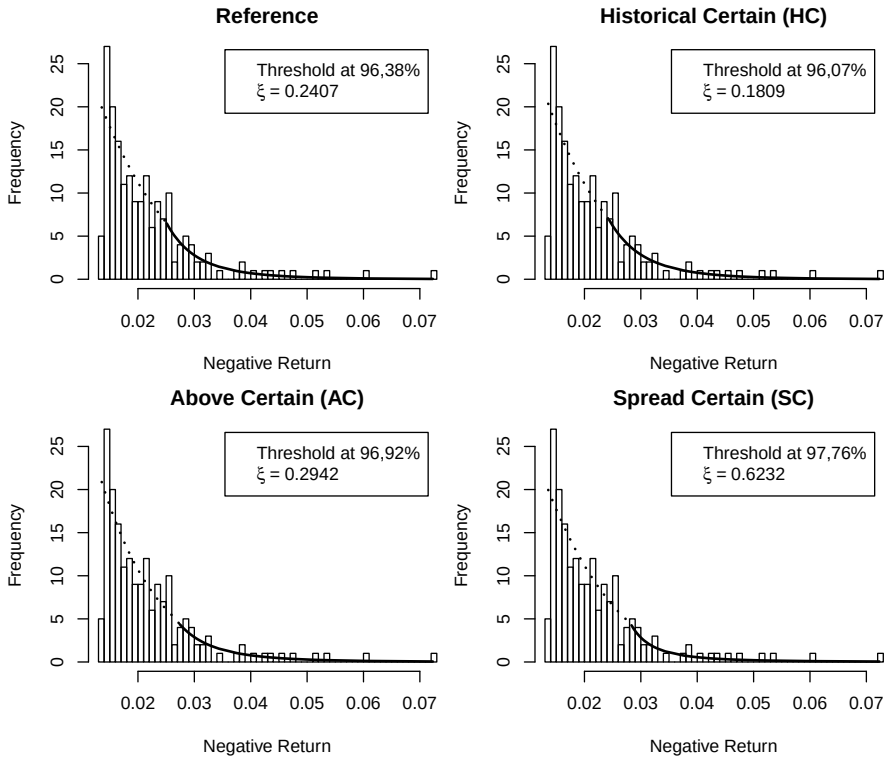


Figure 26: GP-GP model fitted to the Bank of America dataset. Informed priors are used, based on different elicitation scenarios, see table 1.

Table 2: GH-GP (10,000 samples)

	$VaR_{1\%}$ (%)			$ES_{1\%}$ (%)			$VaR_{0.1\%}$ (%)			$ES_{0.1\%}$ (%)			Threshold (%)		
Model	Avg.	LB	UB	Avg.	LB	UB	Avg.	LB	UB	Avg.	LB	UB	Avg.	LB	UB
Analytical	3.842			5.631			7.996			11.57					
Historical	3.919	3.736	4.150	5.502			7.373	6.617	9.376	10.30					
GP (MLE)	3.885	3.719	4.075	5.517	5.087	6.186	7.701	6.851	9.040	10.52	8.705	13.68	95.00		
GP	3.880	3.756	4.013	5.537	5.183	5.961	7.745	7.045	8.588	10.68	9.126	12.73	95.00		
GH-GP	3.879	3.798	3.961	5.548	5.358	5.754	7.773	7.410	8.155	10.82	10.02	11.68	94.36	93.88	94.82
GP-GP	3.849	3.694	4.019	5.521	5.158	5.963	7.745	7.071	8.586	10.83	9.396	12.65	93.42	91.27	95.40

Table 3: GH-GP (4,000 samples)

	$VaR_{1\%}$ (%)			$ES_{1\%}$ (%)			$VaR_{0.1\%}$ (%)			$ES_{0.1\%}$ (%)			Threshold (%)		
Model	Avg.	LB	UB	Avg.	LB	UB	Avg.	LB	UB	Avg.	LB	UB	Avg.	LB	UB
Analytical	3.842			5.631			7.996			11.57					
Historical	3.961	3.515	4.562	5.246			6.646			8.140					
GP (MLE)	3.848	3.610	4.134	5.198	4.700	6.151	7.004	6.052	8.959	8.916	7.137	11.59	95.00		
GP	3.837	3.652	4.032	5.226	4.806	5.791	7.080	6.248	8.232	9.171	7.503	11.79	95.00		
GH-GP	3.782	3.677	3.895	5.202	4.918	5.510	7.101	6.539	7.736	9.452	8.260	10.94	95.17	94.15	96.05
GP-GP	3.783	3.570	4.018	5.255	4.773	5.869	7.219	6.314	8.377	9.704	7.926	12.19	93.08	90.61	96.00

Table 4: GH-GP (1,000 samples)

	$VaR_{1\%}$ (%)			$ES_{1\%}$ (%)			$VaR_{0.1\%}$ (%)			$ES_{0.1\%}$ (%)			Threshold (%)		
Model	Avg.	LB	UB	Avg.	LB	UB	Avg.	LB	UB	Avg.	LB	UB	Avg.	LB	UB
Analytical	3.842			5.631			7.996			11.57					
Historical	3.818	2.982	4.781	4.900			5.866			6.443					
GP (MLE)	3.693	3.307	4.253	4.806	4.109	6.247	6.293	5.037	8.181	7.752	6.202	10.08	95.00		
GP	3.648	3.327	4.016	4.998	4.236	6.345	6.765	5.351	9.367	9.230	6.307	16.23	95.00		
GH-GP	3.860	3.565	4.233	5.141	4.568	6.052	6.839	5.794	8.623	8.812	6.781	13.07	97.14	96.45	97.78
GP-GP	3.633	3.278	4.063	4.929	4.168	6.153	6.633	5.258	8.912	8.988	6.233	14.87	95.23	92.02	97.64

Table 5: GL (10,000 samples)

	$VaR_{1\%}$ (%)			$ES_{1\%}$ (%)			$VaR_{0.1\%}$ (%)			$ES_{0.1\%}$ (%)			Threshold (%)		
Model	Avg.	LB	UB	Avg.	LB	UB	Avg.	LB	UB	Avg.	LB	UB	Avg.	LB	UB
Analytical	7.911			10.02			12.81			15.22					
Historical	7.913	7.605	8.425	10.31			13.14	12.27	15.33	16.44					
GP (MLE)	8.023	7.743	8.333	10.33	9.777	11.10	13.40	12.37	14.97	16.34	14.52	19.53	95.00		
GP	8.020	7.808	8.235	10.35	9.915	10.84	13.45	12.62	14.42	16.46	14.92	18.37	95.00		
GH-GP	8.214	8.064	8.364	10.60	10.34	10.85	13.76	13.29	14.22	16.69	15.84	17.56	97.37	96.89	97.80
GP-GP	7.976	7.716	8.240	10.28	9.826	10.79	13.35	12.54	14.29	16.39	14.88	18.26	96.30	93.29	98.33

Table 6: GL (4,000 samples)

	$VaR_{1\%}$ (%)			$ES_{1\%}$ (%)			$VaR_{0.1\%}$ (%)			$ES_{0.1\%}$ (%)			Threshold (%)		
Model	Avg.	LB	UB	Avg.	LB	UB	Avg.	LB	UB	Avg.	LB	UB	Avg.	LB	UB
Analytical	7.911			10.02			12.81			15.22					
Historical	8.257	7.435	8.876	10.48			13.36			15.53					
GP (MLE)	8.135	7.697	8.654	10.41	9.608	11.82	13.44	11.96	16.35	16.17	13.66	21.03	95.00		
GP	8.111	7.780	8.462	10.50	9.824	11.35	13.67	12.41	15.36	16.76	14.53	20.17	95.00		
GH-GP	8.389	8.104	8.677	10.80	10.33	11.33	14.00	13.17	14.97	16.91	15.43	18.68	97.13	96.59	97.64
GP-GP	8.084	7.677	8.525	10.40	9.693	11.27	13.49	12.25	15.02	16.44	14.37	19.16	94.41	91.13	97.30

Table 7: GL (1,000 samples)

	$VaR_{1\%}$ (%)			$ES_{1\%}$ (%)			$VaR_{0.1\%}$ (%)			$ES_{0.1\%}$ (%)			Threshold (%)		
Model	Avg.	LB	UB	Avg.	LB	UB	Avg.	LB	UB	Avg.	LB	UB	Avg.	LB	UB
Analytical	7.911			10.02			12.81			15.22					
Historical	8.324	7.128	9.737	10.06			12.22			13.36					
GP (MLE)	8.094	7.297	9.153	10.05	8.850	13.06	12.61	10.58	16.39	14.62	11.70	19.01	95.00		
GP	7.972	7.322	8.710	10.68	9.171	13.25	14.22	11.45	19.29	18.89	13.37	32.39	95.00		
GH-GP	8.773	8.055	9.618	11.47	10.27	13.03	15.03	13.08	17.93	18.40	15.29	23.75	96.57	95.60	97.52
GP-GP	7.976	7.212	8.860	10.40	9.021	12.26	13.62	11.28	17.04	17.07	13.20	23.52	93.07	90.56	96.10

Table 8: Bank of America (1258 samples)

	$VaR_{1\%}$ (%)			$ES_{1\%}$ (%)			$VaR_{0.1\%}$ (%)			$ES_{0.1\%}$ (%)			Threshold (%)		
Model	Avg.	LB	UB	Avg.	LB	UB	Avg.	LB	UB	Avg.	LB	UB	Avg.	LB	UB
Historical	3.434	3.017	4.330	4.699			6.094			6.996					
GP (MLE)	3.457	3.150	3.920	4.697	3.955	6.106	6.343	5.075	8.246	8.663	6.930	11.26	95.00		
GP	3.443	3.187	3.729	4.855	4.069	6.337	6.627	5.141	9.286	9.976	6.198	19.32	95.00		
GH-GP	3.641	3.431	3.893	4.893	4.355	5.742	6.543	5.489	8.202	8.741	6.389	13.56	97.23	96.46	97.93
GP-GP	3.476	3.191	3.794	4.689	4.067	5.637	6.277	5.132	8.076	8.581	6.113	13.34	96.38	93.43	98.24

Table 9: Use of informed priors on the Bank of America data set (1258 samples)

	$VaR_{1\%}$ (%)			$ES_{1\%}$ (%)			$VaR_{0.1\%}$ (%)			$ES_{0.1\%}$ (%)			Threshold (%)		
Prior	Avg.	LB	UB	Avg.	LB	UB	Avg.	LB	UB	Avg.	LB	UB	Avg.	LB	UB
Reference	3.476	3.191	3.794	4.689	4.067	5.637	6.277	5.132	8.076	8.581	6.113	13.34	96.38	93.43	98.24
Historical Certain	3.493	3.232	3.775	4.568	4.147	5.036	6.005	5.334	6.772	7.648	6.656	8.895	96.07	93.12	98.08
Historical Uncertain	3.523	3.212	3.885	4.778	4.153	5.596	6.437	5.322	7.946	8.757	6.540	12.21	96.33	93.40	98.29
Above Certain	3.652	3.321	4.012	5.130	4.516	5.797	7.080	6.023	8.242	10.02	8.264	12.31	96.92	94.71	98.45
Above Uncertain	3.532	3.211	3.888	5.015	4.251	6.139	6.913	5.493	8.973	10.29	6.906	16.38	96.78	94.28	98.45
Spread Certain	3.435	3.167	3.769	6.096	5.057	7.553	8.536	6.635	10.89	19.85	14.15	28.73	97.76	96.44	98.68
Spread Uncertain	3.527	3.182	3.950	5.724	4.481	7.984	8.033	5.936	11.23	15.83	8.317	33.18	97.42	95.51	98.73

5

DISCUSSION & CONCLUSIONS

This chapter will discuss the results from the main thread of the development process.

5.1 DATA TRANSFORMATION

Although not part of the process, return transformation and volatility filtering were attempted in order to test on real data and show an example procedure, see Section 3.1. This section seeks to discuss the effectiveness of this procedure.

The assumed heavy-tails of the data weren't always present after the financial data had been filtered using the standard example process presented. Although there is a general consensus that financial data is heavy-tailed, see [15, p.38], when modelling volatility clusters, even using a very general distribution such as the generalized hyperbolic (GH), the tails are sometimes affected to such an extent that they are no longer heavy-tailed.

This may be due to the limitation that GH is only able to model semi-heavy-tailed data, adjusting extreme values too harshly, or because the heavy-tails are related to the volatility clusters as indicated by [32, p.15-18]. Regardless, although the method can be extended to non-heavy-tailed data with a minimum of work, for the future it would be more valid to include modelling of volatility clusters in the main model.

In the context of non-heavy-tailed data, using EVT may be unnecessary and a simpler method may be appropriate.

It should be noted that the rest of the process works independently of this specific data transformation.

5.2 BAYESIAN VS FREQUENTIST INFERENCE

This section will compare and contrast Bayesian and frequentist inference for the different models.

5.2.1 *Fixed-threshold GP Model*

In Section 3.3.3 it was calculated how σ and ξ should behave, according to theory, as the threshold increased. To remind the reader: σ should increase linearly with increasing threshold while ξ remains constant. This is compared with how the MLE and MCMC estimators behaved in Figures 21 and 22. They both behave well to around threshold $u = 0.07$, which represents 50% of the data, but then grow troubled. MLE in particular underestimates the shape parameter ξ badly. MCMC also underestimates the shape but then stabilizes while σ is increasingly underestimated.

It isn't clear at this point if any of the estimators' behaviours are better. Therefore, to get a better understanding of the effect on our risk measures, the quantiles were calculated. In the same figures, it is seen that MCMC behaves much better when the number of data points decreases and otherwise they are interchangeable. Moreover, it should be noted that the expected shortfall is greatly affected by the shape ξ and this favors MCMC.

Note that though from this it may appear that just choosing a low threshold is good, the transition to GP is continuous and choosing threshold becomes a bias-variance trade-off.

The above observed behaviour of MCMC is in accordance with theory in that MLE becomes increasingly biased for smaller sample sizes while MCMC doesn't. Another experiment was performed with this in mind, the result can be seen in Figure 23, wherein the gap between MCMC and MLE grows larger as the number of samples decreases. All the risk measures showed similar behaviour.

Assuming the weakly informative priors (WIP), see Section 3.5.1, have little effect on the GP model, as intended, the results of MCMC and MLE are approximately the mean and mode, respectively. From the tables in Chapter 4 it is seen that they are quite close together and most often MLE estimates lower risk. This led us to investigate the likelihood function from which it was found that it is unimodal and becomes more asymmetric as uncertainty increases. Despite all the arguments against MLE, they are mainly significant in the context of fewer data points, and, from this perspective, MLE could still be considered a viable option for fixed-threshold GP models on large amounts of data.

Greater differences in estimates of intervals than these have been reported before but were possibly due to misuse of large sample theory. The confidence intervals reported here were instead calculated with the method described in 2.9.3.

Additionally, in accordance with prior knowledge, from the tables in Chapter 4 it is seen that MLE consistently estimates a larger confidence interval than MCMC's analogue, credible interval.

In conclusion, MCMC behaves slightly better than MLE when the number of samples decrease but they are otherwise very similar. On the other hand, it can be argued that the bias that MLE introduces becomes less relevant when the sample size decreases because the uncertainty is so great. Even with extreme amounts of data, both

estimators have a tendency to underestimate the heavy-tail and it could be beneficial to look into other estimators.

5.2.2 *Body-tail Models*

In Gamerman's original model, see Figure 13, there were many discontinuities which lead to multimodality and, hence, MLE was ill-suited. The body-tail models try to smooth out these discontinuities and, as a result, it is possible that MLE could be used. However, we would then lose the sought averaging effect discussed in Section 5.5 which was, in part, the purpose of the model.

5.3 MCMC ALGORITHMS

This section will discuss the final performance of the MCMC algorithms, some summarizing afterthoughts on the problems that were encountered during the development process, and hints at alternative solutions and possible improvements.

At a glance, AFSS is nearly automated and thus should require very little configuration. However, it was found that providing a good estimate of the covariance matrix was very important due to the factoring explained in Section 2.7.10. The covariance matrix is eventually updated using the obtained samples but this can take a long time. This was solved using a Hessian estimation of the covariance matrix with independence Metropolis (IM) sampling as a back-up.

However, the improvements made to AFSS lessened the problem described above significantly so that it is much less sensitive. There are likely further possible improvements to be made with regards to the timing of tunings and covariance updates which would make AFSS even more formidable.

The use of IM as a final sampler, once stationarity was reached, was very successful when using resampling. It samples very quickly while also having few limitations. The main thing to look out for with IM is that the support of the proposal distribution is large enough. Although such problems are forewarned by a very high acceptance rate, the CIM variation (IM using multivariate Cauchy instead of normal multivariate normal) was implemented and gave the same results.

The acceptance rate becomes high when the support is too small because, if stationarity has been reached, then proposals are mostly gathered in a small range with high likelihood.

The discontinuities of Gamerman's model didn't allow the use of many algorithms expected to have faster convergence, since they often rely on well defined derivatives.

However, since the removal of many discontinuities, there is a large number of algorithms available. This is interesting when using a body distribution for all the data, such as GH, and the execution of the model is slower. It can then be useful to look for an algorithm with faster convergence, e.g. with much weight on gradients. For this purpose, it is likely that there are better algorithms than AFSS. However, many algorithms require expertise to configure properly and the configuration might change from one data set to another.

5.4 PRIORS

In this section, the results from the testing of informative priors based on expert opinion are discussed and compared to the least informative reference priors. The Bank of America data set was analyzed using the GP-GP model and the results can be seen in Table 9.

For a reminder of the scenario definitions, see Table 1.

As one might expect, "Above" increases all risk estimates significantly while "Spread" increases all risk estimates except $VaR_{1\%}$. They both push the threshold upwards because as the tail GP tries to conform to the contradicting information contained in the prior, it fits the data worse. This results in more of the data being modelled by the body distribution. Although not shown in Table 9, the converse of "Above" and "Spread" were also tested (i.e. guessing below historical *VaR* and guessing $VaR_{1\%}$ and $VaR_{0.1\%}$ closer to each other, respectively) and behaved as expected after seeing these results.

The estimation of the shape parameter ξ varies from 0.1809 to 0.6232 between Historical and Spread, which has a large effect on the risk measures, especially ES. This is related to the worse fit of the tail GP mentioned above, pushing the threshold upwards, as seen in Figure 26. An example of the impact: in the, perhaps most extreme, scenario "Spread Certain", $VaR_{0.1\%}$ was estimated by the expert as 12.2% (twice the historical *VaR*). The end result was that the $VaR_{0.1\%}$ increased from 6.277% (Reference prior) to 8.536%.

The difference between "Certain" and "Uncertain" is that the former results in smaller credible intervals, representing the higher stated certainty. There is also a decrease in credible intervals when comparing the Reference prior to "Historical Certain", the only informative prior agreeing with the data. This reflects the confidence in the prior information added by the expert.

In conclusion, the informed priors can have a significant impact on the end result, and it might therefore be good to always compare with the reference prior. This is in accordance with theory, saying that the effect of priors increases with smaller samples and the Bank of America sample was only 1258 data points. It might also be argued

that the example elicitation scenarios are on the extreme side and that in reality an expert's opinion is going to align better with historical data. However, for the purposes of testing, the impact is easily seen like this even if it is slightly exaggerated.

5.5 EFFECT OF THRESHOLD

Even in the unrealistic case with 10,000 data points, it was seen that the threshold affects the risk measures in two ways, see Figure 24, which both increase in magnitude as the number of relevant data points decrease (smaller sample, greater quantile, or higher threshold). This is supported by the findings of most literature that discuss how the threshold selection influences the parameter estimation, see [44], [12], [11], and [17].

First is a fluctuation which causes small variations in the threshold to significantly influence the results, also seen in previous works such as [43, p.12-15] and [15, p.37]. The second is a slight trend.

The equivalent of Figure 24 was also tried for MLE and looked very similar in this regard

For the fluctuations it makes sense to not just pick a random point but calculate some average. The trend is more difficult to handle since it could be converging towards a value at first but then accuracy is eventually lost as the number of data points decrease. Therefore, it makes sense to account for the uncertainty caused by this trend.

It could be argued that the problems with using a fixed threshold become less relevant when the sample size decreases because the uncertainty is so great. However, if the uncertainty of selecting the threshold isn't taken into account then the estimation of intervals may become unsound.

Finally, it all comes back to the problem of choosing a threshold, but it has changed slightly to: "how much weight do we put on each threshold?" The body-tail models present one possible solution where the body distribution both smooths the discrete data set and decides the weighting, see Figure 25. At the very least, the results in Figure 24 indicate that, in the context of heavy-tailed and meager data, picking a single point threshold might be a bad idea.

5.6 MODEL COMPARISON

This section will seek to discuss and compare the effectiveness of the two body-tail models. Firstly, a more abstract view of the problem of threshold selection will be presented followed by comparisons based on gathered data.

In a sense, the body-tail models seek to establish heuristic information criteria (IC) for selecting a threshold. For the sake of being easier to understand, the IC can be divided into two parts;

$$IC(u, A) = f(u, A) + g(u, A), \tag{76}$$

which for GP distributed data should accomplish this;

$$LL(u, A) + f(u, A) = C(A) = \textit{ constant w.r.t. } u, \tag{77}$$

where LL is the log-likelihood, A is a set of parameters excluding u, and g is a function deciding the bias-variance trade-off.

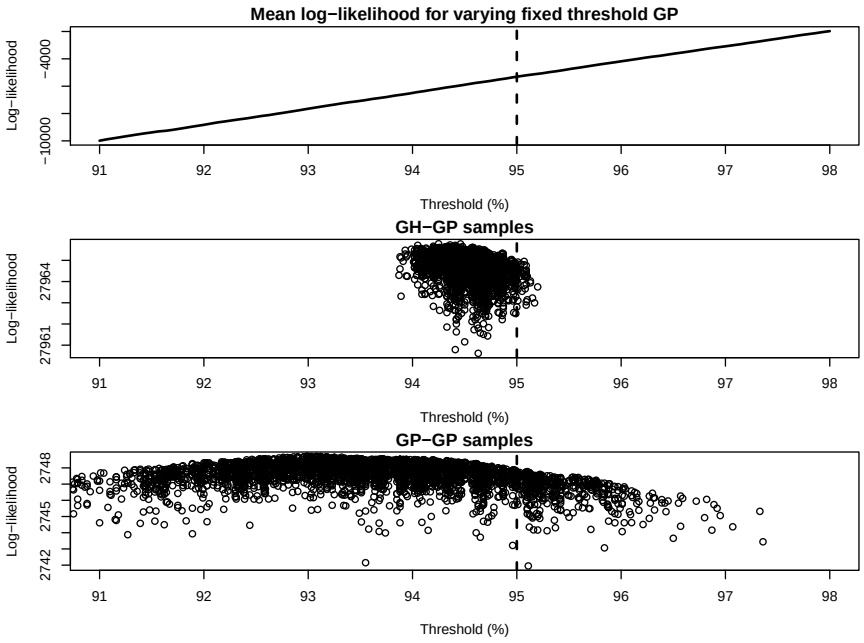


Figure 27: Comparing log-likelihood of different thresholds using different models on the GH-GP data set.

Looking at Figure 27, the first plot shows pure GP data to the right of the dashed line at 95%. The idea is that the function f should make all the thresholds equally likely for GP distributed data, i.e. transform the line above 95% into a horizontal line. Then, g decides the bias-variance trade-off by making modelling more data points more attractive, i.e. having higher log-likelihood.

From the posterior samples in the latter two plots, in Figure 27, and Figure 28 we can infer how the two models tackle this problem. The sample plots give an idea of what the distribution of thresholds looks like and how they assign higher probabilities to certain thresholds. The main test case is in Figure 27 where both models assign reasonable thresholds around what we know is true. The reason that the threshold isn't more confidently at 95% might be because the GH data merges very well into GP and it is possible that the sample can

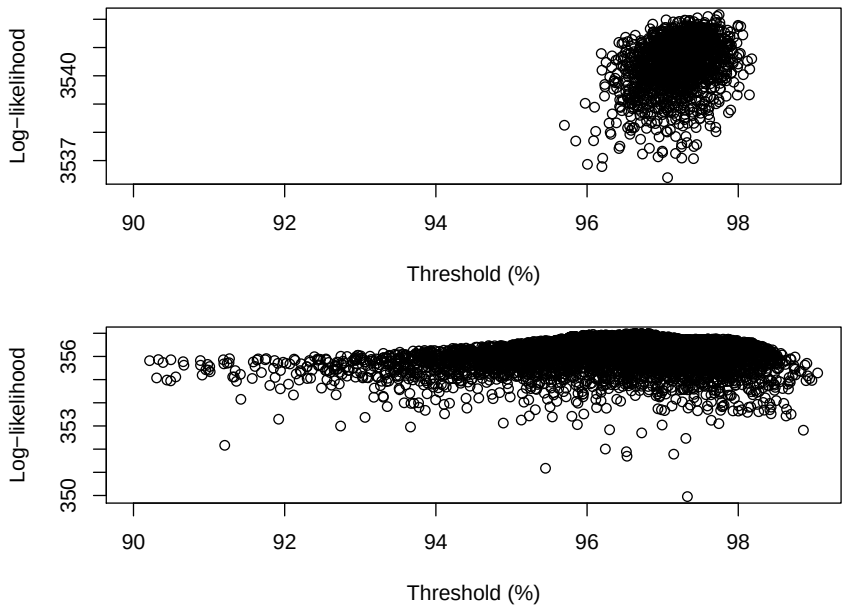


Figure 28: Comparing log-likelihood of different thresholds using the GH-GP model (top) and GP-GP model (bottom) on the Bank of America data set.

be modelled well with GP at a lower threshold than the 95th percentile.

Starting to compare the models, in Figure 27 the GH-GP model is very sure while GP-GP prefers a lower threshold on average and is uncertain. This behaviour makes sense for the GH-GP data but the GH-GP model shows very similar behaviour in Figure 28 and for the GL data set as seen in Table 5. It seems GH-GP models a lot of the data better than GP up to a point where it suddenly is much worse. This would explain why GH-GP is so certain of the threshold. GP-GP on the other hand is more easily interpreted and has a wider averaging effect which is more in line with the uncertainty of the MRL plots. The threshold is estimated to be higher/lower if the body distribution fits the data better/worse.

Comparing Figure 27 with the MRL plots in Figure 11 or Figure 28 with the range of estimates in the MRL plot in Figure 10, where the bounds on the threshold were estimated to be approximately (88%, 98%), we can see that GP-GP reflects the uncertainty of the MRL plots better.

All the models behave quite similarly but as discussed in Section 5.5, slight changes in threshold can have a great effect on the risk measures in the context of a small samples. This invalidates the use of the fixed-threshold GP model for our problem set.

In the tables, the most obvious difference is that GP-GP tends to estimate a lower threshold with more uncertainty than GH-GP. The latter is unsurprising when GH-GP is modelling GH-GP data but the phenomenon remains in the other data sets. The higher threshold leads to GH-GP modelling the tail using fewer data points, which causes increased uncertainty in the risk measures and higher mean because of the asymmetric nature discussed in Section 5.2.

Since the body-tail models include the uncertainty of the threshold parameter, it was expected that the credible interval would be larger. This seems to hold true for the GP-GP model but the GH-GP model shows inconsistencies, like in Table 5, where it seems overly confident in its threshold and deviates from the expected value and historically calculated risk measures. The sought averaging effect, discussed earlier in Section 5.5, is weaker in the GH-GP model.

Excluding the GH-GP data set, on average GP-GP outperforms the other models, especially in context of smaller sample sizes. Even in the GH-GP data set it competes well. This is unsurprising considering the earlier discussion about thresholds, see Section 5.5. However, there are few competing models and more work in the area of threshold weighting would be highly interesting. Notably, there is at least a philosophical design flaw in using GP as a body distribution, which is that it could compete with the tail for the GP data points because they both model these points well. This is avoided by choosing an appropriately low body GP threshold[1] but is still a concern and there are most likely more improvements to be made.

Lastly, the GP-GP model is much less computationally expensive than GH-GP which opens up for further modelling of, for example, volatility clusters or inter-dependencies between instruments, especially since many MCMC algorithms scale well with more parameters.

5.7 SUMMARY

The objective of the thesis was to develop an automatic procedure that takes in transformed data and returns risk measures. A few issues were encountered and worked on during this process, most notably threshold selection and sensitivity.

Possibly the most important result is that small variations in the chosen threshold have a significant effect on risk measures, at least in the context of heavy tails and few relevant data points. In this setting, the fixed-threshold GP model displayed possibly unsound behaviour in situations such as:

1 The body GP of GP-GP had at threshold at 85%.

- Very large sample (10,000), threshold 95%, and estimating $ES_{0.1\%}$
- Small sample (1,000), threshold 95%, and estimating $VaR_{0.1\%}$ or $ES_{1\%}$.

Therefore, taking the uncertainty of the threshold into account and calculating some average is a given, at least when estimating intervals.

This makes Bayesian inference, especially MCMC since it is unbiased for small sample sizes, and weighted threshold models a natural choice although MLE could still be argued to perform well enough. The developed GP-GP model demonstrated both consistency and effectiveness, showing that, in this regard, body-tail models are a viable option.

In order to improve performance of the final procedure, the AFSS and IM algorithms were analyzed and improved on. Notably, AFSS converges to optimal slice sampling faster and parameter bounds are handled better in AFSS and IM.

Lastly, the effect of a reference prior versus a prior based on expert opinion was investigated and exemplified for practical applications in finance. It was found that the expert prior could have a significant effect on the results and it is probably best to always compare with the reference prior.

Although an apparent focus was placed on financial applications, the method is in no way limited to financial data. Especially the GP-GP model is very general and, with slight variations, is applicable for any satisfactorily transformed data.

6

RECOMMENDATIONS & FUTURE WORK

A minimum of work would be required to extend the algorithm to allow for non-heavy-tailed data. From there on it would be interesting to continue the investigation into the effect of threshold on risk measures in a non-heavy-tailed context.

Following such an extension, a comparison with other models, such as [5], would be interesting.

Moreover, comparing different expert priors for financial applications could provide useful insight.

Portfolio interdependence was outside the scope of this paper but with the increased performance of GP-GP there is a possibility of modelling interdependence in a portfolio, possibly using regular vines. As a result of this, the data sets might grow very large and an investigation into using gradient MLE or MCMC could help.

Other possible model improvements include modelling volatility clusters and peak clusters. A deeper investigation into the effect of these would also be interesting and, if possible, a measure of clustering would be very insightful.

An alternative to the suggested body-tail composition is to use a convolution to fit the distributions together. The possible effects are a little unclear and depend on the length of the convolution but it could be worthwhile (at least using second order derivatives would be more sound).

Creating another information criterion for changing threshold could be most useful. For example, it could be heuristically based on simulated GP data, trying to cancel the effect of letting the threshold vary. In regards to this, Section 3.3.3 and looking at $E[p(X)] = \int_u^\infty p^2(x)dx = \frac{1}{\sigma(2+\xi)}$ or $E[p(X)|X > a]$ may be helpful.

Related to this is to look for a method for estimating convergence rate to GP.

Additionally, there are further improvements to be made to AFSS with regards to the timing of tuning and covariance updates.

Alternatives to AFSS could be tried, using a more specialized algorithm and proposal distribution. Making use of derivatives is attractive and component-wise algorithms could be effective especially if model complexity increases.

However, it could be better to look for another estimator that is more suited to heavy-tailed data and doesn't underestimate the shape parameter ξ as much. Possibly in combination with a better estimator, a simpler averaging method might be necessary. The downside will likely be that there is less rich theory behind it as MCMC is quite well developed.

BIBLIOGRAPHY

[1] Barone-Adesi, G. & Vosper, L. (1999). "VaR without correlations for portfolios of derivative securities". *Journal of Futures Markets*. **doi**:10.1002/(SICI)1096-9934(199908)19:5<583::AID-FUT5>3.0.CO;2-S

[2] Behrens, C. N., Lopes, H. F., & Gamerman, D. (2004). "Bayesian analysis of extreme events with threshold estimation". *Statistical Modelling*. **doi**:10.1191/1471082X04st075oa

[3] Beirlant, J., Vynckier, P. & Teugels, J. L. (1996). "Excess functions and estimation of the extreme-value index", *Bernoulli*, **2**, 293-318.

[4] Berger, J., Bernardo, J. & Dongchu, S. (2009). "The Formal Definition of Reference Priors." *Annals of Statistics*, **37(2)**, 905-938.

[5] Bermudez, P. Z., Turkman M. A. A. & Turkman K. F. (2001). "A predictive approach to tail probability estimation", *Extremes*, **4**, 295-314.

[6] Bollerslev, T. (1986). "Generalized Autoregressive Conditional Heteroskedasticity", *Journal of Econometrics*, **31**, 307-237.

[7] Brooks, S. (2011). *Handbook of Markov chain Monte Carlo*. Boca Raton: CRC Press/Taylor & Francis.

[8] Chalabi, Y., Scott, D. J., Würtz, D. "The Generalized Lambda Distribution as an Alternative to Model Financial Returns". Institut für Theoretische Physik, Zürich. University of Auckland, New Zealand. URL https://www.rmetrics.org/sites/default/files/glambda.pdf

[9] Chen, M. H., Shao, Q. M. & Ibrahim, J. G. (2000). *Monte Carlo methods in Bayesian computation*, Springer Series in Statistics.

[10] Coles, S. (2001). *An introduction to statistical modeling of extreme values*. London New York: Springer.

[11] Coles, S. G. & Powell, E. A. (1996). "Bayesian methods in extreme value modelling: a review and new developments", *International Statistical Review*, **64**, 119-36.

[12] Coles, S. G., & Tawn, J. A. (1995). "A Bayesian analysis of extreme rainfall data". *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **45(4)**, 463-478.

[13] Dittmar, D. (2013). "Slice Sampling". TU Darmstadt. URL http://www.ausy.informatik.tu-darmstadt.de/uploads/Teaching/RobotLearningSeminar/Dittmar_RLS_2013.pdf

[14] DuMouchel, W. H. (1983). "Estimating the stable index α in order to measure tail thickness: a critique", *The Annals of Statistics*, **11**, 1019-31.

[15] Embrechts, P., Resnick, S. I. & Samorodnitsky, G. (1999). "Extreme Value Theory as a Risk Management Tool", *North American Actuarial Journal*, **3(2)**.

[16] Ferreira, A. & De Haan, L. (2015). "On the Block Maxima Method in Extreme Value Theory: PWM Estimators", *The Annals of Statistics*, **43(1)**, 276-298.

[17] Frigessi, A., Haug, O. & Rue, H. (2002). "A dynamic mixture model for unsupervised tail estimation without threshold selection", *Extremes*, **5**, 219-235.

[18] Gamerman, D. & Lopes, H. (2006). *Markov chain Monte Carlo : stochastic simulation for Bayesian inference*. Boca Raton: Taylor & Francis.

[19] Gelman, A., Carlin, J., Stern, H. & Rubin, D. (2004). *Bayesian Data Analysis*. 2nd edition. Chapman & Hall, Boca Raton, FL.

[20] Ghalanos, A. (2014). **rugarch**: *Univariate GARCH models*. **R** package version 1.3-4. URL http://cran.r-project.org/web/packages/rugarch

[21] Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. (1996). Markov chain Monte Carlo in practice. London: Chapman & Hall.

[22] Gilli, M., & Këllezi, E. (2006). "An Application of Extreme Value Theory for Measuring Financial Risk". *Computational Economics*. **doi**:10.1007/s10614-006-9025-7

[23] Hu, W. & Kercheval, A. (2007). "Risk Management with Generalized Hyperbolic Distributions". Florida State University. Bell Trading, Chicago. URL http://www.math.fsu.edu/~aluffi/archive/paper321.pdf

[24] Hult, H. (2012). *Risk and portfolio analysis : principles and methods*. New York: Springer Science+Business Media.

[25] Irony, T. & Singpurwalla, N. (1997). "Noninformative Priors Do Not Exist: a Discussion with Jose M. Bernardo." *Journal of Statistical Inference and Planning*, **65**.

[26] Jaruskova, D. & Hanek, M. (2006). "Peaks over Threshold Method in Comparison with Block-Maxima Method for Estimating High Return Levels of Several Northern Moravia Precipita-

tion and Discharge Sites", *Journal of Hydrology and Hydromechanics*, **54(4)**, 309-319.

[27] Jeffreys, H. (1961). *Theory of Probability*. Third edition. Oxford University Press, Oxford, England.

[28] Johannes, M. S. & Polson, N. (2003). "MCMC Methods for Continuous-Time Financial Econometrics", *SSRN Electronic Journal*. **doi**:10.2139/ssrn.480461

[29] Jorion, P. (2007). *Value at risk: The new benchmark for managing financial risk*. New York: McGraw-Hill.

[30] Kang, S. G. (2013). "Noninformative priors for the scale parameter in the generalized Pareto distribution", *Journal of the Korean Data & Information Science Society* **24(6)**, 1521-1529

[31] Kang, S. G., Kim, D. H. & Lee, W. D. (2013). "Noninformative priors for the shape parameter in the generalized Pareto distribution", *Journal of the Korean Data & Information Science Society* **24(1)**, 171-178

[32] Kirchler, M. & Huber, J. (2007). "Fat tails and volatility clustering in experimental asset markets", *Journal of Economic Dynamics and Control*, **31(6)**, 1844-1874

[33] Malevergne, Y. & Sornette, D. (2006). *Extreme financial risks from dependence to risk management*. Berlin New York: Springer-Verlag.

[34] Mendes, B. V., & Lopes, H. F. (2004). "Data driven estimates for mixtures", *Computational Statistics & Data Analysis*, **doi**:10.1016/j.csda.2003.12.006

[35] Prause, K. (1999). "The Generalized Hyperbolic Model: Estimation, Financial Derivatives, and Risk Measures". Institut für Mathematische Stochastik, Albert-Ludwigs-Universität, Freiburg. URL http://bfi.cl/papers/Prause%20Eberlein%201999%20-%20The%20generalized%20hyperbolic%20distribution%20in%20finance.pdf

[36] Raoult, J. & Worms, R. (2003). "Rate of convergence for the generalized Pareto approximation of the excesses", *Advances in Applied Probability*. **doi**:10.1239/aap/1067436332

[37] Rasmussen, C. E. & Ghahramani, Z. (2002). "Bayesian Monte Carlo", University College, London. URL http://machinelearning.wustl.edu/mlpapers/paper_files/AA01.pdf

[38] Robert, C. (2007). *The Bayesian Choice*. 2nd edition. Springer, Paris, France.

[39] Robert, C. P. & Casella, G. (2010). *Introducing Monte Carlo methods with R*. New York: Springer.

[40] Roberts, G. O., Gelman, A. & Gilks, W. R. (1997). "Weak convergence and optimal scaling of random walk Metropolis algorithms". *Annals of Applied Probability*. **doi**:10.1214/aoap/1034625254

[41] Rossi, P. E. & Allenby, G. M. (2003). "Bayesian Statistics and Marketing", *Marketing Science*. **doi**:10.1287/mksc.22.3.304.17739

[42] Sawyer, S. (2006). "The Metropolis-Hastings Algorithm and Extensions". Washington University. URL http://www.math.wustl.edu/~sawyer/hmhandouts/MetropHastingsEtc.pdf

[43] Simiu, E. & Heckert, N. A. (1996). "Extreme Wind Distribution Tails: A 'Peaks Over Threshold' Approach", *Journal of Structural Engineering*, **547**.

[44] Smith, R. S. (2000). "Measuring risk with extreme value theory", *Extremes and Integrated Risk Management*. London: Risk Book, 19-35.

[45] Statisticat, LLC. (2015). **Laplaces Demon:** *Complete Environment for Bayesian Inference*. **R** package version 15.03.01,
URL http://www.bayesian-inference.com/software
Github Repo http://github.com/samedii/LaplacesDemon

[46] Tibbits, M. M. et al. (2014). "Automated Factor Slice Sampling." *Journal of computational and graphical statistics: a joint publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America*, 543–563. PMC.

[47] Tsay, R. S. (2009) "Extreme Values and Their Applications in Finance". University of Chicago. National University of Singapore. URL http://rmi.nus.edu.sg/_files/events/paper/Extreme%20Values%20and%20their%20Applications%20in%20Finance.pdf