Tail Estimators Using Weighted Quantiles

Mathematical optimisation for threshold exceedances in financial loss modelling using the generalised Pareto distribution

Minh H. Nguyễn

A thesis presented for the degree of

Master of Science
in
Computational Finance

University College London

Department of Computer Science

3 September 2019

Abstract

Extreme value theory has found interest in many applied sciences but some empirical issues stand as impediments to consistent results. The central theorems in extreme value theory are especially interesting for the financial sector due to the historical "fat-tail" characteristic of many asset classes. Our aim was to derive a method which can be applied to operational losses within a financial institution over a fixed period of time, thus modelling the greatest losses that the institution stands to lose. Most methods to estimate polynomial tail distributions regard all observations as equally important in estimating the tail which yields highly implausible estimators for the true tail-thickness. Our work comprises a two-stage procedure that corrects for the tail bias obtained from the first stage through iterative numerical optimisation on tail-overweighted quantiles. The procedure introduced proves to be stable throughout a wide range of tail cutoffs and highly applicable to loss modelling as well as asset returns analysis.

Acknowledgements

I would like to thank everyone who has contributed to the completion of this thesis. The list includes Peter Mitic for going over the details and talking me through the problems and Simone Righi for taking the time to be my academic supervisor and reading the manuscripts.

I am obligated to thank Dermot Fortune, John Magill, Edward Stockton, Sophie McGowan, Raj, Daniel White, Manuel Wallner, Rebecca Fryer and Winnifred Taylor who have all been instrumental to the process. I would also like to thank StackExhange, StackOverflow and GitHub for they have had a not-unsubstantial amount of contribution .

Dành cho mẹ, bố và chị.

"When I wrote this, only God and I understood what I was doing. Now, God only knows."

Karl Weierstrass

Declaration

I, Minh Hoàng Nguyễn, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Contents

Chapter 1

Introduction

Every day, financial institutions manage operations through internal processes that identify risk sources. The common thread among these risk sources is the capital requirement to "cushion" any shortfall that can damage the financial stability of the institution. Operational risk concerns less with conventional adverse market movements that result from excessive exposure but more with qualitative sources from both inside and outside the institution. Per the definition given in the Bank of International Settlement's Basel II framework [3]:

"Operational risk is defined as the risk of loss resulting from inadequate or failed internal processes, people and systems or from external events. This definition includes legal risk, but excludes strategic and reputational risk."

These risk sources pose a significant challenge for financial institutions due to their outsized effects on business and reputation, the latter of which has become increasingly more relevant since the great financial crisis. By definition, the majority of theses losses are preventable and are therefore an obvious target for operational discipline for large banks that have faced difficulties adjusting to the new economic environment.

Mathematically, operational risk is reduced down to a vector of losses accumulated throughout a time period. A lot of emphasis is put on the greatest empirical losses, which are indicative of the theoretical losses that the institution faces going forward. The job of the risk practitioner is to construct a *reasonable* probability measure for a given loss greater than a threshold. This thesis concerns the quantification of the empirically-consistent statistical properties of such losses and the empirical difficulties of applying theoretical results to actual data.

We know that the normalised sum of independent and identically distributed random variables with finite mean and variance is asymptotically normal. In the possibly infinite variance and mean case, common encounters of financial loss models revolve around the Lévy α-stable distribution [29] which can be written as a superposition of two other independent and identically distributed copies of another random variable, up to location and scale. The class of stable distributions allows practitioners to study density functions with "thicker" tails and thus would produce more extreme values. In a way, the stable distribution is a generalisation of the normal, given its similar characteristic function is of the same form as a the normal.

The stability index α in the parameterisation of any Lévy α-stable distribution is the central ingredient of excess modelling because it controls the thickness of the tail and the stability of the distribution. A special edge case is when $\alpha = 2$ which yields the standard normal distribution. In the general case, we can only retrieve the density through the non-trivial numerically inversion the characteristic function which is a significant practical drawback.

Extreme value theory explores the distribution of the extremal data points, possibly with infinite moments, in the sums described using a different class of thick-tailed distributions. Our work concerns the generalised Pareto distribution, which also has a tail index but is parameterised by a polynomial tail in most forms. Polynomial decay yields a non-vanishing tail, thus is appropriate for modelling losses which concentrate around small values but have frequent very large values that are not accurately represented in exponential decaying tails. The distribution of interest is the generalised Pareto distribution, which scales and shifts the famous tail distribution Pareto depending on the best fit for a given data set.

Previous works [2, 7, 13, 26, 27, 36, 39, 45] have been focused on analytical and moment-based methods to estimate the distribution but they have produced highly unfeasible tail estimates which predict extremely unlikely losses. Bayesian and quasi-Bayesian methods, such as the one found [13], have been found to yield good estimators which are somewhat related to our method. Our work focuses the application of the theorems introduced in the next section regarding the convergence of the tail to extremal distributions, therefore we overweigh penalisation of tail deviations to formulate better tail estimates. We divide the procedure up into two phases. The first phase aims to produce the best fit o the distribution function curve and a bias coefficient of the tail relative to the sample (and where possible the population) counterpart. The second phase employs an appropriate weigh vector to reduce this bias and match the fitted tail statistics to their sample counterparts.

Our results show numerous powerful applications to different synthetically generated data sets as well as actual data from different asset classes, some much more volatile than others. We found that quantile-consistent estimates of the left tail for most asset classes yield finite variances thus facilitating the application of the central limit theorem. Furthermore, standard methods in literature such as the Hill [25] and Pickands [2] estimators for the tail index are highly volatile throughout various cutoffs for most data sets whereas as produce tail cutoffs that constitute most of the data set. Our procedure yields a credible range for the tail index that ranges throughout the entire data set.

Our estimators are empirically consistent throughout the tail yielding the same level of bias as the initial estimates compared to estimates of truncated data sets. This result provides a powerful tool for extrapolations outside of the sample data sets in order to estimate true tail events in different asset classes, where most of the estimates show finite variance except for extremely volatile assets such as cryptocurrencies. Since our procedure produces a unique tail estimator, the results on tail behaviour are also generally more informative than the Pickands and Hill functions where the asymptotically normal properties only apply with larger data sets.

In chapter two, we introduce the theoretical framework for the thesis with rigorous probability proofs. Chapter three details the literature review and current methods to es-

timate the generalised Pareto distribution. Chapter four introduces the optimisation problems in two different stages along with a primer on feasibility of the estimator. Chapter five goes over the optimisation procedures in detail with full mathematical proofs along with an efficient implementation guide. Chapter six presents our results and chapter seven concludes. We also provide a concise plain English summary of our work in chapter 8.

Chapter 2

Extreme Value Theory

Suppose one is given a sequence of n realisations of a strictly positive random variable $X_1, X_2, ..., X_n$. Extreme value theory (EVT) is the study of the distribution of the largest element observed in the sequence for each increasing n. In the limit as n is infinity, we are almost surely guaranteed to have observed the maximum value that the random variable can attain (with probability 1). Thus, the distribution of this maximum value is just a Dirac mass at the right end-point of the support of the support. For example, if we are interested in the distribution of (unbounded) losses less than a threshold $c > 0$ then in the limit we either observe that the largest loss will be c if $c = \infty$ or it will not be c if $c < \infty$. We therefore scale and shift our largest loss appropriately as we increase n to get the generalised extreme value (GEVD) distribution given by the Fisher-Tippett-Gnedenko theorem [41].

If instead of the largest loss one wants the distribution of k largest losses, they can reasonably see that the condition for a useful (non-binary) distribution still applies, i.e. we still need to scale and normalise. Another way to get this distribution is to pick a threshold $c > 0$ as before, but ensure that c is appropriately large, and observe the distribution of *all* losses greater than c. This distribution is given by the Pickands-Balkema-de Haan [2][27] theorem which is the generalised Pareto distribution (GPD) of interest. We will proceed to introduce and prove both theorems seeing as they are intimately connected and give an outline of methods to estimate the GPD in literature.

2.1 Generalised Extreme Value Distribution

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X be an $\mathcal{F}$-measurable random variable. From a set of n iid realisations of X given by $X_1, X_2, ..., X_n$, we denote its Borel sub-σ-algebra fully contained in $\mathcal{F}$ by $\mathcal{F}_n$ and the maximum of the set by M_n. To perform statistical procedures on M_n, we need a stability condition on its distribution as we let n tend to infinity.

Let G denote the cumulative distribution function (cdf) of X. If we denote the theoretical right endpoint of G as $\omega(G) = \sup\{x \mid G(x) < 1\}$, then $M_n \xrightarrow[n\to\infty]{(a.s.)} \omega(G)$ where

$\xrightarrow{(a.s.)}$ denotes almost sure convergence (with probability 1) [8]. We can deduce that

$$\mathbb{P}\big(M_n \leq x\big) \xrightarrow[n\to\infty]{(a.s.)} 0 \quad \forall \quad x \leq \omega(G)$$
$$\mathbb{P}\big(M_n \leq x\big) \xrightarrow[n\to\infty]{(a.s.)} 1 \quad \forall \quad x > \omega(G)$$

which is is a degenerate limiting distribution, i.e. the Dirac measure $\delta_X(x)$ with the property $\int_X f(x')\,\mathrm{d}\delta_X(x') = f(x)$. This is immediate from Kolmogorov's zero-one law concerning the Borel σ-algebra generated by the limiting behaviour of the tail events

$$\mathcal{T}_\infty = \lim_{n\to\infty}\Big\{\mathcal{T}_n = \sigma(X_{n+1}, X_{n+2}, ...)\Big\}$$

which stipulates that if $A \in \mathcal{T}$ then $\mathbb{P}(A) \in \{0, 1\}$. We also note that $\mathcal{T}_n$ is independent of $\mathcal{F}_n$ for all n therefore the event $M_n \leq x$ for $n \to \infty$ is not affected by $\mathcal{F}_n$ and is a valid tail event. Moreover, it is clear that the sequence $\{M_n\}$ is a sub-martingale, i.e.

$$\mathbb{E}[M_{n+1} \mid M_1, ..., M_n] \geq M_n \quad \forall\, n$$

and therefore an appropriately normalisation of the observations would give us a valid probabilistic view on $M_n \leq x$. Specifically, there must exist sequences $\{a_n \mid a \in \mathbb{R}, n \in \mathbb{N}\}$ and $\{s_n \mid s \in \mathbb{R}^+, n \in \mathbb{N}\}$ that satisfy

$$\mathbb{P}\left(\frac{M_n - a_n}{s_n} \leq x\right) = \Big[G(s_n x + a_n)\Big]^n \xrightarrow[n\to\infty]{(d)} F(x) \tag{2.1.1}$$

where $\xrightarrow{(d)}$ denotes convergence in distribution and F is not degenerate. If the normalisation in (2.1.1) exists and the convergence holds, then we say that G is max-stable and G is in the domain of attraction of F, written $G \in \mathcal{D}(F)$.

Theorem 2.1.1 (Fisher-Tippett-Gnedenko [41] (Proof in A.1))**.** *If* (2.1.1) *holds, then* F *is given by*

$$F(x \mid \xi) = \exp\big[-(1+\xi x)^{-1/\xi}\big] \ , \ 1 + \xi x > 0 \tag{2.1.2}$$

where changing $\xi \in \mathbb{R}$ *corresponds to switching between three distinct distributions:*

1. *If* $\xi > 0$, $F(x \mid \alpha)$ *is Frechét*

$$F(x) = \exp\big(-x^{-\alpha}\big) I_{x\in[0,\infty]}(x)$$

2. *If* $\xi < 0$, $F(x \mid \alpha)$ *is Weibull*

$$F(x) = \begin{cases} \exp\big[-(-x)^{\alpha}\big] & x \leq 0 \\ 1 & \textit{otherwise} \end{cases}$$

3. $\lim_{\xi\to 0} F(x \mid \xi) = F(x)$ *is Gumbel*

$$F(x) = \exp\big(-e^{-x}\big) \ , \ x \in \mathbb{R}$$

Our proof of theorem 2.1.1 follows that of de Haan and Ferreira [11] although we do make some minor changes to improve readability and connect to theorem 2.2.1. The Frechét distribution imposes a lower bound on x and is usually observed after fitting the GEV on data. The Weibull distribution is used to model rainfall in hydrology and survival analysis due to its large body (it tends to the Dirac measure as $\alpha \to \infty$). The Gumbel is usually used to model the largest values of a sequence of exponential random variables. Thus generally, to account for scaling and translations of the distributions, we add two more parameters μ and σ to G to get the full GEVD

$$F(x) = \exp\left\{ -\left[1 + \xi\frac{(x-\mu)}{\sigma}\right]^{-1/\xi}\right\} \tag{2.1.3}$$

where $x > \mu - \frac{\sigma}{\xi}, \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+$ and $\xi \in \mathbb{R}$.

Often times, of greater interest is not just the greatest attainable value but the greatest $100(1-c)\%$ values. If the values in question are of losses, then the value is called *value-at-risk* (VaR) at the threshold of consideration VaR(c). Thus, a VaR(c) is defined as the left-continuous inverse of the cdf at the threshold

$$\mathrm{VaR}_F(c) = \inf\left\{x \;\middle|\; \mathbb{P}\big(X > x\big) \le c\right\} \;\; \forall\, c \in (0,1) \tag{2.1.4}$$

We then make the distinction between an extreme distribution and a tail distribution. A commonplace tail distribution is the log-normal with parameters μ and σ, whose cdf is

$$F(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\int_0^x \exp\left[-\frac{(\log x' - \mu)^2}{2\sigma^2}\right]\frac{\mathrm{d}x'}{x'} = \Phi\left(\frac{\log x - \mu}{\sigma}\right) \tag{2.1.5}$$

which has the half-infinite interval $x \in (0, \infty]$ support as well as a fat right tail. It is well known that the log-normal distribution does not have a moment-generating function, but all moments exist and are given analytically through a straightforward function

$$\mathbb{E}\left[X^k\right] = \exp\left(k\mu + k^2\sigma^2/2\right) \tag{2.1.6}$$

Empirical estimation for μ and σ are straightforward due to their analytical maximum likelihood (ML) estimators. The estimators are derived from differentiating the log-likelihood with respect to the parameters and equating to zero:

$$\begin{aligned} \hat{\mu}_{ML} &= \frac{1}{n}\sum_{i=1}^{n}\log x_i \\ \hat{\sigma}^2_{ML} &= \frac{1}{n}\sum_{i=1}^{n}\big(\log x_i - \hat{\mu}\big)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\log x_i - \frac{1}{n}\sum_{i=1}^{n}\log x_i\right)^2 \end{aligned} \tag{2.1.7}$$

The empirical parametric log-normal VaR follows immediately, using the relation from (2.1.5)

$$\mathrm{VaR}_F(c \mid \hat{\mu}, \hat{\sigma}^2) = \exp\left[\hat{\sigma}\Phi^{-1}(c) + \hat{\mu}\right] \tag{2.1.8}$$

Another distribution of interest is the Pareto distribution, $X \sim Pareto(\alpha, x^*)$, where x^* is the smallest value of X considered and the tail index α controls the size of the right tail decay

$$F(x \mid \alpha, x^*) = 1 - \left(\frac{x^*}{x}\right)^{\alpha} \tag{2.1.9}$$

The fat tail is the direct result of the polynomial decay instead of the exponential decay of thin-tailed distributions like the Gaussian.

2.2 Generalised Pareto Distribution

The second central result in EVT concerns the distribution of the exceedances for the general case with translation and scaling. Usually called the Pickands, Balkema and de Haan theorem, the result is the threshold exceedance counterpart to the Fisher, Tippett and Gnedenko theorem with special applicability to operational risk. Our proof of the theorem follows Pickand's [2] although we do not prove elementary topology results to keep the proof concise.

Theorem 2.2.1 (Pickands, Balkema and de Haan [2][27] (Proof in A.2))**.** *Given a distribution function G with the right endpoint of its support $\omega(G)$, define the distribution function for the exceedances over a threshold c as $G(X \mid X > c)$. If for an iid sequence $X_1, X_2, ..., X_n$, $G(X \mid X > c)$ is not degenerate up to translation and scale, then there exist $c < \omega(G) \leq \infty$, $\xi \in \mathbb{R}$ and a strictly positive measurable function $\sigma : X \to (0, \infty)$ such that*

$$\mathbb{P}\left[\frac{X - a(c)}{s(c)} \leq x \;\middle|\; X > c\right] \xrightarrow[c \to \omega(G)]{(d)} F\big[x \mid c, \sigma(c), \xi\big]$$

where F is the generalised Pareto distribution (GPD) or the three-parameter Pareto. An equivalent statement is

$$\lim_{c \to \omega(G)} \sup_{c < \omega(G)} \Big| G\big[s(c)x + a(c) \mid X > c\big] - F\big[x \mid c, \sigma(c), \xi\big] \Big| = 0 \;\; \forall\, x \geq c$$

Given a threshold parameter μ such that $X \geq \mu$ (the inequality is equivalent to the strict inequality $X > c$ in theorem 2.2.1 because a point has zero measure) there are two cases for the GPD

$$F(x \mid \mu, \sigma, \xi) = \begin{cases} 1 - \left[1 + \frac{\xi(x-\mu)}{\sigma}\right]^{-1/\xi} & \xi \neq 0 \\ 1 - \exp\left[-\frac{(x-\mu)}{\sigma}\right] & \xi \to 0 \end{cases} \tag{2.2.1}$$

The support for the GPD is defined on two intervals for different values of ξ, given by

1. $\xi \geq 0$: $x \geq \mu$
2. $\xi < 0$: $x \in \big[\mu, \mu - \sigma/\xi\big]$

for $\mu \in \mathbb{R}, \sigma \in \mathbb{R}^+$. The density (pdf) follows straight from the cdf

$$f(x \mid \mu, \sigma, \xi) = \begin{cases} \frac{1}{\sigma}\Big[1 + \frac{\xi(x-\mu)}{\sigma}\Big]^{-\frac{\xi+1}{\xi}} & \xi \neq 0 \\ \frac{1}{\sigma}\exp\Big[-\frac{(x-\mu)}{\sigma}\Big] & \xi \to 0 \end{cases} \tag{2.2.2}$$

Remark. *There are four special cases of the GPD:*

1. $F(x \mid \mu, \sigma, \xi < 0)$ *corresponds to the two-parameter Pareto* (2.1.9) $X \sim Pareto(\alpha = 1/\xi, x^* = \sigma)$ *where* $X \in [\mu, \infty]$.
2. $\lim_{\xi \to 0} F(x \mid \mu, \sigma, \xi)$ *corresponds to the exponential* $X \sim Exp(1/\sigma)$ *where* $X \in [\mu, \infty]$.
3. $F(x \mid \mu, \sigma, \xi = -1)$ *corresponds to the uniform* $X \sim \mathcal{U}(\mu - \sigma/2, \mu - \sigma/2)$.
4. $F(x \mid \mu = 0, \sigma, \xi > 0)$ *corresponds to the type-2 GPD (or Lomax), which is the Pareto shifted to begin at 0, i.e.* $(X - x^*) \sim Lomax(\alpha = 1/\xi, \lambda = \sigma/\xi) \iff X \sim Pareto(\alpha = 1/\xi, x^*)$.

The moments for both the $\xi < 0$ and $\xi \geq 0$ cases are identical. The k^{th} moment exists if $\xi < 1/k$. If $\xi < 1/3$, the first three raw moments of the GPD are

$$\mathbb{E}[X] = \int_\mu^\infty \frac{x}{\sigma}\left[1 + \frac{\xi(x-\mu)}{\sigma}\right]^{\frac{-1}{\xi}-1} \mathrm{d}x = \mu + \frac{\sigma}{1-\xi}$$

$$\mathbb{E}\left[X^2\right] = \mu^2 + \frac{2\sigma\mu}{1-\xi} + \frac{2\sigma^2}{(1-\xi)(1-2\xi)}$$

$$\mathbb{E}\left[X^3\right] = \mu^3 + \frac{3\sigma\mu^2}{1-\xi} + \frac{6\sigma^2\mu}{(1-\xi)(1-2\xi)} + \frac{6\sigma^3}{(1-\xi)(1-2\xi)(1-3\xi)}$$

and from which we get the variance and skewness

$$\mathbb{V}\mathrm{ar}[X] = \frac{\sigma^2}{(1-\xi)^2(1-2\xi)} \tag{2.2.3}$$

$$\mathbb{S}[X] = \frac{2(1+\xi)\sqrt{1-2\xi}}{1-3\xi} \tag{2.2.4}$$

We can also use the characteristic function to obtain the pattern of the moments for $0 < \xi < 1/k$

$$\chi(t) = \int_\mu^\infty \frac{\mathrm{e}^{\mathrm{i}tx}}{\sigma}\left[1 + \frac{\xi}{\sigma}(x-\mu)\right]^{\frac{-1}{\xi}-1} \mathrm{d}x = \mathrm{e}^{\mathrm{i}t\mu}\sum_{j=0}^{\infty} \frac{(\mathrm{i}\sigma t)^j}{\prod_{k=0}^{j}(1-k\xi)}$$

Theorem 2.2.2. *The GPD is stable with respect to exceedances past a threshold* $\mu < t < \omega(F)$ *with the same shape parameter* ξ*, that is if* $X \sim GPD(\mu_1, \sigma_1, \xi_2)$ *and* $X \geq t > \mu_1$ *then there exist* $\mu_2 \in (\mu_1, \infty)$ *and* $\sigma_2 \in (0, \infty)$ *such that* $(X - t) \sim GPD(\mu_2, \sigma_2, \xi)$.

Proof. This is a special case of theorem A.2.5. The cdf for $X - t$ is given by the conditional probability

$$\begin{aligned}\mathbb{P}\big(X \leq x \;\big|\; X > t\big) &= \frac{\mathbb{P}\big(X \leq x\big) \cap \mathbb{P}\big(X > t\big)}{\mathbb{P}(X > t)} \\ &= F\big[x \;\big|\; t, \sigma + \xi(t - \mu), \xi\big]\end{aligned}$$

□

The application of theorem 2.2.2 is crucial to the fitting problem. For example, if we can fit a GPD to one set of data, then we can move the threshold as required and the fit should still apply. One empirical challenge is to find the appropriate threshold c, which is the estimator for μ: a large value will result in too few measurements for meaningful statistical procedures whereas a small value will account for some body values in conjunction with the tail, leading to biased results.

Another evident challenge is the fact that the parameters are not mutually independent. This observation leads to big complications in numerical solutions because the theorem implies some sort of combinatorial structure to the problem. In the next chapter we introduce some methods discussed in literature to fit the GPD. This list is by no means exhaustive and acts only to provide the background needed to understand the drawbacks of off-the-shelf procedures.

Chapter 3

Literature Review

The usual approach in literature is to use the two parameter GPD with $\mu = 0$. This coincides with taking positive values of a data set or values greater than a threshold then remove the threshold. From theorem A.2.1 we know that the GPD is uniquely determined by two points, corresponding to a re-parameterisation of σ and ξ given a threshold. Castillo and Hadi [7] estimate the parameters using a number of pairs of distinct empirical quantiles and selects the median values. This method provides an empirically consistent set of estimators but fails to identify the tail threshold necessary to build a two-parameter model. Furthermore we also need estimates consistent for the tail and not for the whole body.

Diebolt et al.[14] discards the frequentist, parametric approaches to GPD estimation because their unreliability and use a Gamma mixture representation of the pdf of a special case of GPD in which the assumed cdf is $G \in \mathcal{D}(Frechét)$. The representation transforms the GPD density into a weighted superposition of a Gammma density with the weights given by an exponential density

$$g(x \mid \xi, \sigma) = \frac{\sigma^{\xi}}{\Gamma(\xi)} \int_0^{\infty} z\mathrm{e}^{-xz} z^{\xi-1} \mathrm{e}^{-\sigma} \, \mathrm{d}z$$

The Gamma representation is then used as the basis for a Gamma quasi-conjugate Bayesian framework (since there is no GPD conjugate). Since all parameters of the Gamma are strictly non-negative, this means σ and ξ are also positive, hence the Frechét domain of attraction requirement. Although this is of the form we require, the method only works for the two-parameter version of the GPD, which again requires us to specify a threshold to eliminate μ.

In their seminal papers, both Pickands [2] and Balkema and de Haan [27] outline their proofs for the theorem using two distinct continuity points in the support. They proceed to employ quantile estimates for parameters using the uniqueness of the parameters, derived from th continuity points, given an appropriate threshold (see theorems A.1, A.2.4 and A.2.5). Because we are using the thee-parameter GPD, we can extend the theorem to three distinct quantiles and estimate as in (3.0.14). This method, however, calls into question the appropriate quantiles as well as the dependency of ξ and σ on the threshold μ, specifically if the threshold is too low then the GPD tail is not applicable and therefore

$\hat{\xi}$ and $\hat{\sigma}$ may not converge to unique values.

Hosking and Wallis [26] review a number of methods in estimation of the GPD in hydrology and they recommend a popular tool called probability-weighted moments (PWM) which estimates

$$\mathcal{M}_k = \mathbb{E}\big\{X^k F(X)^\alpha [1-F(X)]^\beta\big\} \quad \forall \quad k, \alpha, \beta \in \mathbb{R}$$

in place of conventional central moments. The estimators are asymptotically normal therefore one does not lose tractability as well as Greenwood's [1] proof that some PWM estimates are easier to relate to their distributions than the conventional counterparts. However, Hosking and Wallis restrict the range of $\xi \in [-0.5, 0.5]$ due to their specific GPD application in hydrology rarely requires "unusual" forms which yield infinite variance ($\xi > 1/2$) or finite upper endpoint ($\xi < -1/2$). Financial applications necessitate thick tails therefore we cannot make the same assumption.

Park and Kim [35] were also interested in financial risk modelling usng the GPD. They begin with the peaks over threshold condition and relax the unconditional GPD tail by also using a weighted non-linear least-square scheme like our work but the weights are given explicitly as the inverse covariance matrix and there is no coordinate parameter. Our method can be seen as a generalised version that derives from the Goodness-of-Fit test given in chapter 3, with special conditions on parameter space search (details in chapter 4). Furthermore, while they choose to optimise using the Nelder-Mead [32] simplex algorithm we use the more stable Levenberg-Marquadt method.

3.0.a Maximum Likelihood/Entropy

The ML estimators for the GPD are significantly more challenging to obtain compared to the log-normal's. We begin with the two conditions that can be obtained at the optimal parameter values for σ and ξ. For n iid random variables $\mathbf{x} \in (\mathbb{R}^+)^n$ where $x_i \sim GPD(\mu, \sigma, \xi)$ and $x_i \geq \mu \; \forall \; i$, the likelihood and log-likelihood are respectively given by

$$\mathrm{L}(\mu, \sigma, \xi \mid \mathbf{x}) = \prod_{i=1}^{n} f(x_i \mid \mu, \sigma, \xi) = \frac{1}{\sigma^n} \prod_{i=1}^{n} \left[1 + \frac{\xi(x_i - \mu)}{\sigma}\right]^{-1/\xi - 1} \tag{3.0.1}$$

$$\mathcal{L}(\mu, \sigma, \xi \mid \mathbf{x}) = -n \log \sigma - \big(1/\xi + 1\big) \sum_{i=1}^{n} \log \left[1 + \frac{\xi(x_i - \mu)}{\sigma}\right] \tag{3.0.2}$$

and from which we get the optimal conditions

$$\begin{aligned} \frac{\partial}{\partial \sigma}\mathcal{L} &= -\frac{n}{\sigma} - (1/\xi + 1) \sum_{i=1}^{n} \frac{-\xi(x_i - \mu)/\sigma^2}{1 + \xi(x_i - \mu)/\sigma} = 0 \\ \implies & \sum_{i=1}^{n} \frac{(x_i - \mu)/\sigma}{1 + \xi(x_i - \mu)/\sigma} = \frac{n}{1 + \xi} \end{aligned} \tag{3.0.3}$$

$$\begin{aligned}\frac{\partial}{\partial\xi}\mathcal{L} = \frac{1}{\xi^2}\sum_{i=1}^{n}\log\left[1+\frac{\xi(x_i-\mu)}{\sigma}\right] - \big(1/\xi+1\big)\sum_{i=1}^{n}\frac{(x_i-\mu)/\sigma}{1+\xi(x_i-\mu)/\sigma} = 0 \\ \implies \sum_{i=1}^{n}\log\left[1+\frac{\xi(x_i-\mu)}{\sigma}\right] = n\xi\end{aligned} \tag{3.0.4}$$

However, $\partial\mathcal{L}/\partial\mu$ is strictly unbounded in μ

$$\frac{\partial}{\partial\mu}\mathcal{L} = \sum_{i=1}^{n}\frac{1}{\xi/\sigma+(x-\mu)} \tag{3.0.5}$$

and therefore does not have a minimiser with respect to μ.

A common practice to set $\mu = \min\{x_i\}_{i=1}^n$ or to 0. The somewhat arbitrary choice for μ is not satisfactory in spite of the consistency of the ML estimator under the assumption of correct parameterisation. Another source of concern is for $\xi > 1$ the likelihood (and hence the log-likelihood) is unbounded from above, therefore no ML estimators exist [7]. A common approach is to numerically minimise the log-likelihood to avoid dealing with two optimality conditions. This approach can sometimes diverge for the reasons described and is generally unreliable even with well-conditioned initial values.

A similar procedure is to maximise the empirical Shannon entropy of the observations

$$H\big(f\big) = -\mathbb{E}\Big[\log f\big(X,\theta\big)\Big] = -\int_X \log f\big(x,\theta\big)\,\mathrm{d}F\big(x\big)$$

under the GPD, such as in [39]. Empirical information such as sample mean, variance and skewness can be matched with their theoretical counterparts to maximise entropy, given they exist. Given a set of q moment constraints

$$m_i\big(\theta\big) = \int_X w_i\big(x\big)\,\mathrm{d}F\big(x\big) \quad \forall \quad i = 1,...,q$$

and from the principle of maximum entropy we know that the density that maximises H subject to $\int_X \mathrm{d}F(x) = 1$ satisfies

$$f\big(X,\theta\big) \quad \propto \quad \exp\big(-\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{w}\big)$$

where $\boldsymbol{\lambda} \in \mathbb{R}^q$ are Lagrange multipliers. The problem becomes a linear programme

$$\hat{\theta} = \arg\max_{\theta}\ H\big(f\big) = \arg\max_{\theta}\ \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{w}$$

Since some densities are very sharp and concentrated around the body for financial losses, the maximisation problem can become unbounded in θ if $f\big(x_i\big) > 1$ for some i. A common approach to mitigating this "spiking" behaviour is to normalise by another density function.

3.0.b Method of Moments

We denote $\bar{x}$, V and S as the sample mean, variance and skewness

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$V = \frac{1}{n}\sum_{i=1}^{n} \big(x_i - \bar{x}\big)^2$$

$$S = \frac{1}{nV^{3/2}}\sum_{i=1}^{n} (x_i - \bar{x})^3$$

The method of moments (MM) estimator for ξ is obtained from solving (2.2.4) by substituting the theoretical moment with its sample counterpart

$$\hat{\xi}_{MM} = \arg\min_{\xi} \left\{ S - \frac{2(1+\xi)\sqrt{1-2\xi}}{1-3\xi} \right\} \tag{3.0.6}$$

We can also use a tail estimate for ξ using the standard Hill [25] estimator in literature. Sort the observations into their order statistics given, i.e. $X^{(1)} \geq X^{(2)} \geq ... \geq X^{(n)}$, the Hill estimator is given by

$$\hat{\xi}_{Hill}\big(k\big) = \frac{\sum_{i=1}^{k} \log X^{(i)} - \log X^{(k)}}{k} \tag{3.0.7}$$

which coincide with the ML estimate for the 1-parameter tail index $1/\alpha$ in (2.1.9) and is plotted against a range of k to find a stable value for most observations. Hill shows in the seminal paper that one can get $\hat{\xi}_{Hill} \xrightarrow[n\to\infty]{\mathbb{P}} \xi$ and more generally it can be shown that the estimator is asymptotically normal

$$\sqrt{k}\Big(\hat{\xi}_k^{Hill} - \xi\Big) \xrightarrow[n,k\to\infty]{(d)} \mathcal{N}\big(0, \xi^2\big)$$

if and only if $k/n \to 0$ [12], i.e. if our sample grows faster than the sub-sample considered. One immediate problem with this type of tail estimate is that $\hat{\xi}$ can be greater than 1, prohibiting analytical estimates for μ an σ which was the reason for using the estimate in first place.

Another estimate is to use the Pickands [2] estimator. From A.2.4 we know that the GPD is uniquely identified by two quantiles up to an appropriate threshold by inverting the generalised Pareto function introduced in the proof. By truncating the smallest $n-k$ observations, the Pickands ξ estimator can be shown to be

$$\hat{\xi}_{Pick}(k) = \frac{1}{\log 2} \log \frac{X^{(k/4)} - X^{(k/2)}}{X^{(k/2)}} \tag{3.0.8}$$

where $X^{(k/4)}$ and $X^{(k/2)}$ are the 75^{th} and 50^{th} percentiles of the truncated data set respectively. It has also been shown that this estimator is also asymptotically normal

$$\sqrt{k}\Big(\hat{\xi}_k^{Pick} - \xi\Big) \xrightarrow[k\to\infty]{(d)} \mathcal{N}\left(0, \frac{\xi^2\big(2^{\xi+1}+1\big)}{\big(2\big(2^{\xi}-1\big)\log 2\big)^2}\right)$$

but because of the necessary gaps between the lower quantiles introduced by letting $k \to \infty$ the estimate can be volatile.

From (3.0.6), we can derive the MM estimators for σ and μ, after obtaining $\hat{\xi}$ from one of the previous methods, by

$$\hat{\sigma}_{MM} = (1-\hat{\xi})\sqrt{V(1-2\hat{\xi})} \tag{3.0.9}$$

$$\hat{\mu}_{MM} = \bar{x} - \frac{\hat{\sigma}}{1-\hat{\xi}} \tag{3.0.10}$$

This approach is used Quandt [36] among all ML, PWM and quantile estimators. Quandt's recommendation is to empirically minimise quantile residuals with a bias correction and he uses the spectral density of the residuals to assess the fit. We use a different measure to assess the goodness-of-fit (explained in chapter 3) but we also make extensive use of the residuals in our construction of the weights in our procedure.

A different MM approach is to avoid solving (3.0.6) and set μ to start at either 0 or $\min\{x_i\}_{i=1}^n$ as previously suggested. The resulting MM estimators in the $\mu = 0$ case are

$$\hat{\sigma}_{MM2} = \frac{\bar{x}}{2}\left(1 - \frac{\bar{x}^2}{V}\right) \tag{3.0.11}$$

$$\hat{\xi}_{MM2} = \frac{1}{2}\left(1 - \frac{\bar{x}^2}{V}\right) \tag{3.0.12}$$

MM estimators are known to be biased in many circumstances where one requires extrapolation outside of the sample range. Specifically, empirical moments for unbounded distributions tend to severely underestimate the tails. In theory, one can get an estimate for μ from (3.0.10) and then use that estimate to calculate ML estimators for σ and ξ in (3.0.3) and (3.0.4). Empirical performance of this hybrid procedure depends on the data set's smallest values, which can have a significant condition number [1] given that small losses occur in much greater frequency than large losses by construction.

Zhao et. al [45] propose a weighted moments method that is somewhat similar to our proposal but moment-based methods generally perform poorly for financial losses. The biases of the empirical moments compared to true moments can sometimes be severe and affords the practitioner little control over most of the tail. Our quantile-based methods are much more direct in their penalisation to produce more consistent results.

[1] Condition number in numerical analysis is directly related to the degree to which the output changes given small changes in the input. Take the function $f : X \to Y$ where X and Y are normed vector spaces. The condition number is defined as $\chi = \lim_{\Delta\to 0}\sup_{\|\Delta x\|<\Delta}\frac{\|f(x)\|}{\|x\|}$.

3.0.c Quantile Estimation

Let us define a function $x_n : [0,1] \to \mathbb{R}^+$ which supplies the empirical percentile of the data set, i.e $x_n\big[F_n(\mathbf{x}_\alpha)\big] = x_n(\alpha) = \mathbf{x}_\alpha$ for all $\alpha \in [0,1]$ where F_n is the empirical cdf of X. We can specify three percentiles, for example $\alpha \in \{0.25, 0.5, 0.75\}$, to obtain a system of equations to solve for the parameter estimators:

$$\begin{aligned} \alpha &= 1 - \left\{ 1 + \frac{\xi\big[x_n(\alpha) - \mu\big]}{\sigma} \right\}^{-1/\xi} \\ \implies (1-\alpha)^{\psi/\sigma} &\Big\{ \sigma + \psi\big[x_n(\alpha) - \mu\big] \Big\} = \sigma \end{aligned} \tag{3.0.13}$$

where $\psi = \xi\sigma$. Stacking the equations, we get

$$\begin{bmatrix} (1-\alpha_1)^{\psi/\sigma}\Big\{ \sigma + \psi\big[x_n(\alpha_1) - \mu\big] \Big\} - \sigma \\ (1-\alpha_2)^{\psi/\sigma}\Big\{ \sigma + \psi\big[x_n(\alpha_2) - \mu\big] \Big\} - \sigma \\ (1-\alpha_3)^{\psi/\sigma}\Big\{ \sigma + \psi\big[x_n(\alpha_3) - \mu\big] \Big\} - \sigma \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \tag{3.0.14}$$

which gives us three equations for three unknowns. Since that any three values of α can produce a set of estimators, therefore we can aggregate a number of quantile solutions and take the average value for each estimator. Even with aggregation, the estimators fluctuate significantly with the choice of quantiles and therefore the estimated VaR can sometimes take on unreasonable values.

We can also use the Pickands estimator from (3.0.8) in conjunction with his estimator for σ for a given threshold, thus it is a form of quantile estimator or the two-parameter GPD, with the estimators given by

$$\begin{aligned} \hat{\xi}_k^{Pick} &\frac{1}{\log 2} \log \frac{X^{(k/4)} - X^{(k/2)}}{X^{(k/2)}} \\ \hat{\sigma}_k^{Pick} &= \frac{X^{(k/2)}}{\int_0^{\log 2} \mathrm{e}^{\hat{\xi} z}\, \mathrm{d}z} \end{aligned}$$

where Pickands shows that these estimators converge in probability to true values

$$\lim_{n\to\infty} \mathbb{P}\left[\sup_{x\in[0,\infty)} \left| \frac{1 - G(X^{(k/4)} + x)}{1 - G(X^{(k/4)})} - \Big[1 - \hat{G}_{Pick}(x)\Big] \right| > \epsilon \right] = 0 \quad \forall \quad \epsilon > 0$$

where $\hat{G}_{Pick}(x) = G\big(x \mid \hat{\xi}^{Pick}, \hat{\sigma}^{Pick}\big)$. The problem with using this standalone procedure is that we do not know how many observations is enough for the convergence to apply. Indeed, this problem is closely related to the problem of finding the optimal cutoff point seeing as n has to increase faster than k to find a set of stable estimators.

Chapter 4

Methodology I: Framework

4.1 Weighted Quantile Estimators

The contribution of this manuscript stems from quantile estimators from (3.0.14). Instead of using an exactly-determined system of p equations to solve for p estimators for the parameter set $\theta \in \Theta$, we formulate a set of m quantiles where $m \geq p$. The appeal of such over-determined systems is the inclusion of as many descriptive statistics of the empirical distribution as needed. From theorem A.2.4 we know that moving up the tail ensures that we converge to the GPD, we therefore overweight tail residuals to prioritise tail quantiles. The estimators produced by our procedure is henceforth referred to as weighted quantile estimators (WQE).

Pickands [2] uses the 50 and 75 percentiles to reparameterise the GPD and solve for the estimates σ and ξ analytically. One can reasonably extend these estimates to the three-parameter case by using the 1,2 and 3 quartiles (25, 50 and 75 percentile). The main empirical problem with using these quartiles is that because the density is highly concentrated around 0, almost all quantiles up to the last 10% usually have very similar values. Curve fitting algorithms almost always struggle with such a large discrepancy in the changes in the curve, leading to non-convergence. Instead, we use $m-1$ equally spaced increments for $\mathbf{x}_{\boldsymbol{\alpha}}$ and their theoretical cdf evaluations as quantiles $F(\mathbf{x}_{\boldsymbol{\alpha}}) = \boldsymbol{\alpha} \in \mathbb{R}^m$ to give us m quantiles (including the empirical endpoint). By doing this, we initially underweight the tail of the cdf and overweight the body. We then use subsequent weighted optimisation to re-balance the biases.

The procedure is two-fold. First, we get an initial estimate to obtain the empirical noise level around specific areas of the cdf. The initial estimate is then used in conjunction with a weight matrix to produce a Tikhonov-like regularisation of the tail estimates. We use two specialised least-square algorithms called Gauss-Newton (GN) [21] and Levenberg-Marquadt (LM) [30][28] to have more control over the residuals. LM also facilitates convergence much more efficiently compared to GN due to its eigenvalue damping procedure and proves to be highly reliable. We make some minor adjustments to tailor the algorithms to the GPD specification in the next chapter.

Given n strictly positive observations $\mathbf{x} \in \mathbb{R}^n$, our aim is to match a fitted set of quantiles $F(\mathbf{x})$ with a set of empirical quantiles of the data set $F_N(\mathbf{x})$. Let $\theta = \{\mu, \sigma, \xi\}$

be the set of parameters for the GPD, $F_n(\mathbf{x}) \in [0,1]^m$ be the vector of m quantiles and $x_n : [0,1] \to \mathbb{R}^+$ repeat its role as the function that supplies the empirical quantile. Using the empirical and theoretical distribution functions F_n and F respectively, the residual condition we aim to satisfy is

$$\begin{aligned}\rho_\alpha(\theta \mid \mathbf{x}) &\triangleq F_n\big[x_n(\alpha)\big] - F\big[\theta \mid x_n(\alpha)\big] \\ &= \alpha - 1 + \left\{1 + \frac{\xi\big[x_n(\alpha) - \mu\big]}{\xi}\right\}^{-1/\xi} = 0 \ \ \forall\, \alpha\end{aligned} \tag{4.1.1}$$

To achieve the concurrent estimation, we formulate the least-square optimisation problem

$$\begin{aligned}\hat{\theta}_1 &= \arg\min_\theta \ \varphi_1(\theta \mid \mathbf{x}) \\ &\triangleq \arg\min_\theta \ \frac{1}{2}\sum_{j=1}^{m} \rho_j(\theta \mid \mathbf{x})^2 \\ &= \arg\min_\theta \ \frac{1}{2}\big\|\boldsymbol{\rho}(\theta \mid \mathbf{x})\big\|_2^2\end{aligned} \tag{4.1.2}$$

where $\boldsymbol{\rho} \in \mathbb{R}^m$ is the vector of residuals $\boldsymbol{\rho} = [\rho_1(\theta), ..., \rho_m(\theta)]^{\mathrm{T}}$. We will proceed to drop the implicit condition on $\mathbf{x}$ in the objective functions to clean up the notations.

Lemma 4.1.1 (Proof in appendix B.1)**.** *The residual in* (4.1.1) *is not convex, therefore the least squares problem in* (4.1.2) *is not convex and produces local minima.*

Convex optimisation is not at our disposal due to the non-linearity in the residual function creating concavity for some values of X according to lemma 4.1.1. Most optimisation algorithms struggle with local minima as stopping conditions are set to be below a step size threshold, therefore convergence to a stationary point does not guarantee the best solution. We also included a stopping condition based on the goodness-of-fit test introduced in equation (4.2.1).

The second stage of fitting has a general form that allows for flexibility in weights and damping parameters. For a semi-positive definite (spd) weight matrix $\mathbf{W} \in \mathbb{R}^{m\times m}$, we formulate the second minimisation problem

$$\begin{aligned}\hat{\theta}_2 &= \arg\min_\theta \ \varphi_2(\theta \mid \mathbf{x}, \mathbf{W}) \\ &\triangleq \arg\min_\theta \ \frac{1}{2}\sum_{j=1}^{m} \Big[W_{jj}\rho_j\big(\theta \mid x_j\big)\Big]^2 \\ &= \arg\min_\theta \ \frac{1}{2}\Big\|\mathbf{W}\boldsymbol{\rho}\big(\theta \mid \mathbf{x}\big)\Big\|_2^2\end{aligned} \tag{4.1.3}$$

Because of the non-linearity in each ρ, ℓ_p norm regularisation is not possible. We have to adjust the weight matrix $\mathbf{W}$ to penalise the tail according to first stage results. We note the resemblance to the standard linear regularisation such as Tikhonov [40] of the results

as a function of the appropriate weights.

Upon retrieval of $\hat{\theta}_0$, we also get an estimate the vector of residuals in (4.1.2) denotes by $\hat{\boldsymbol{\rho}}_0$. We observe that a uniform reweighting scheme, $\mathbf{W} = \mathbf{I}_m$ where $\mathbf{I} \in \mathbb{R}^{m\times m}$ is the identity matrix, is biased towards the tail if we use equally spaced $\mathbf{x}$ intervals and quantiles because of the asymmetry in the distribution function. The bias can be accounted for by adding perturbations to the coordinates, $x_j + \epsilon_j$, as parameters to be estimated in the optimisation phase

$$\varphi_2(\theta, \boldsymbol{\epsilon} \mid \mathbf{x}, \mathbf{W}_b) \triangleq \frac{1}{2}\sum_{j=1}^{m}\Big\{W_{jj}\rho_j\big(\theta, \epsilon_j \mid \mathbf{x}\big)\Big\}^2 \tag{4.1.4}$$

where $\mathbf{W}_b \in \mathbb{R}^{m\times m}$ is diagonal and contains monotonically increasing diagonal elements. Another option is to weigh the quantiles by their deviation from their empirical counterparts, mathematically given by

$$\mathbf{W}_c \triangleq \gamma\left[\operatorname{diag}\Big(\big|\hat{\boldsymbol{\rho}}_0\big|\Big)\right]^{-1} \tag{4.1.5}$$

The $\gamma \in (0, 1)$ scaling factor ensures that small residuals do not inflate the weights to the point where they hinder the optimisation algorithm. Assessing the fit, however, is a trickier task if the quantiles' weights are highly disproportional. This gives rise to the next section.

4.2 The Mitic Goodness-of-Fit Test

The typical yearly loss data for a financial institution contains largely of relatively small losses which make up the bulk of frequencies. The remainder are the massive losses that the firm incurs throughout the year which, if modelled correctly, can be contained to a reasonable degree. Take figure 4.1 as a typical empirical cdf of losses throughout a 6.5 year period at a financial institution. For this particular example of $\log X \sim \mathcal{N}(8, 1.96)$, 70% of losses are less than 5,000 GBP and 95% are less than 40,000 GBP but the maximum observed loss is 2,230,329 GBP. Evidently we do not want an upper bound on the losses, which historically can sum to or more than the entire market capitalisation of the institution. Recalling from the previous chapter, this implies $\xi \in \mathbb{R}^+$.

We want a fit that models the tail to a high precision which can be accommodated with a less exact fit for the body if necessary. More importantly, by theorem 2.2.1, if a data set is generated by independent GPD losses then if we pick a higher tail cutoff $\hat{\mu}$ we should still be able to fit a GPD with the same shape parameter but positively scaled and translated. This latter point justifies the former to a degree, although finding the "right" cutoff point is a potential nuisance in much of applied statistics.

The flat portion of the cdf, corresponding to the tail, is not too problematic for most algorithms find a reasonable fit. The curvature that connects the vertical and the flat portions, however, magnify the error in the fit due to the rather extreme non-linearity. The Cramér-von Mises (CvM) test

$$\Omega_{CvM} = \mathbb{E}_F\big[F_n(x) - F(x)\big]^2 = \int_X \big[F_n(x) - F(x)\big]^2 \,\mathrm{d}F(x)$$

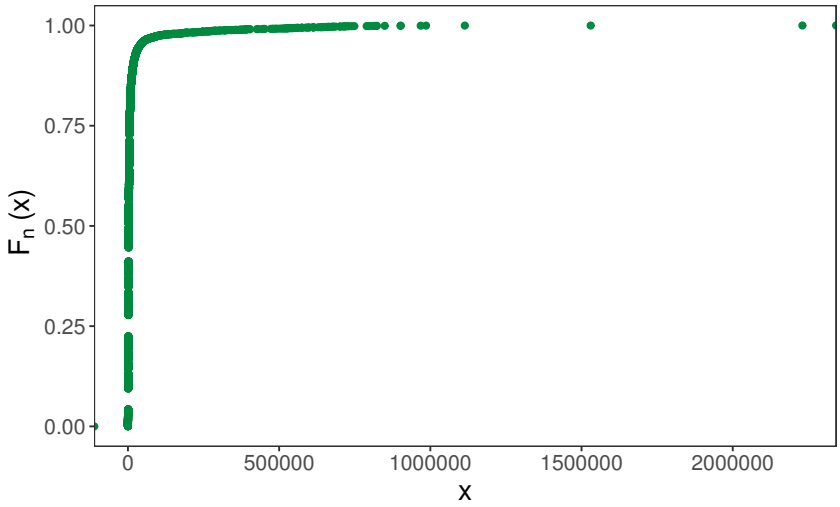


Figure 4.1: an example empirical financial loss cdf simulated from $\log X \sim \mathcal{N}(8, 2)$

along with the Anderson-Darling (AD) test

$$\Omega_{AD} = \int_X \frac{\big[F_n(x) - F(x)\big]^2}{F(x)\big[1 - F(x)\big]} \, \mathrm{d}F(x)$$

and the Kolmogorov-Smirnov (KS) test

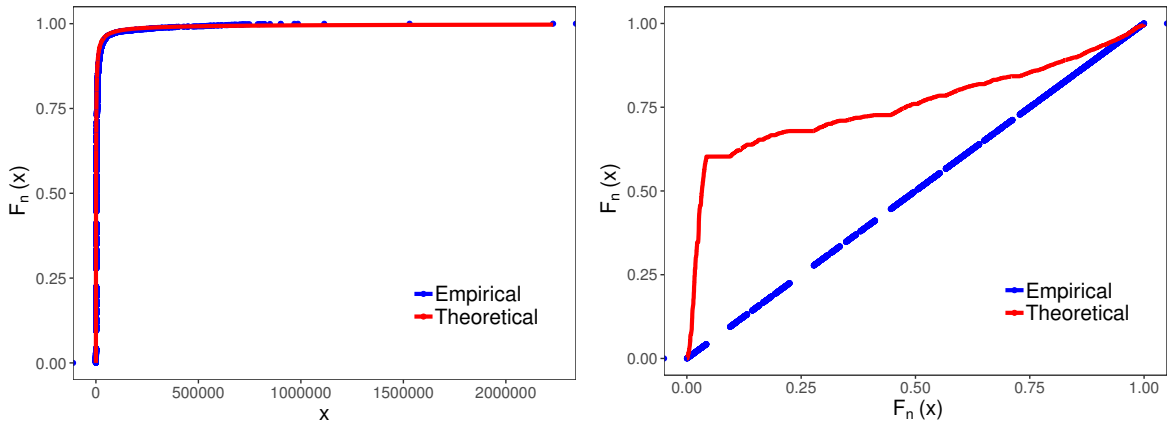
$$\Omega_{KS} = \Big\| F_n(x) - F(x) \Big\|_\infty = \sup_X \ \big| F_n(x) - F(x) \big|$$

are the most common one-sample theoretical distribution Goodness-of-Fit (GoF) tests and they are all very sensitive to outliers and prone to over-reject reasonable fits. The uniform weighting of both the CvM and KS tests exclude flexibility in less important regions as an option. The probability weighted AD test puts more weight on the tails evenly, which is the opposite of our requirement.

Mitic's [31] GoF test is fitting for the problem, especially the second stage. The main contribution is the mapping $T : X \times [0, 1] \to [0, 1] \times [0, 1]$, which takes the cdf space into a uniformly overlapping square of length 1 of the empirical and theoretical cdfs. The non-linearity in the cdf space leads to a poor fit in the transformed space that seemingly approximates the empirical cdf well in the cdf space. An example of this phenomenon is given in figure 4.2.

Mitic shows that the area off the 45° diagonal (which is between 0 and 1) corresponds to a GoF measure and he calculates the area in the transformed space directly on the diagonals by connecting endpoints of deviation vectors normal to the diagonal. We opted to project the absolute deviation onto one axis and numerically integrate. Due to the symmetry in the transformed space, the axis of projection is not important. The test statistic is given by

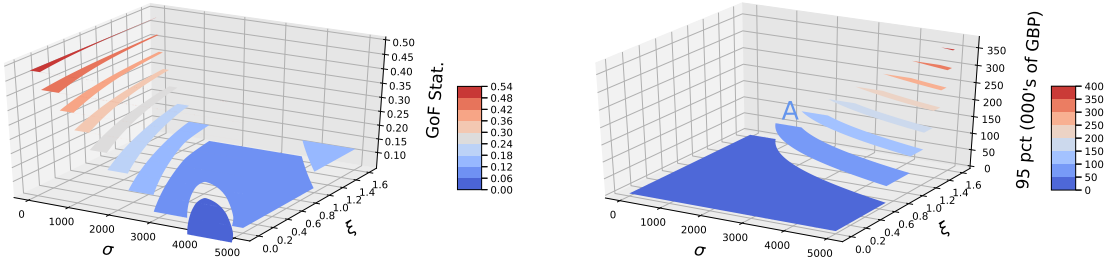
$$\Omega_{Mi} = \mathbb{E}_n \Big| F_n(x) - F(x) \Big| = \int_X \Big| F_n(x) - F(x) \Big| \, \mathrm{d}F_n(x) \tag{4.2.1}$$



(a) cdf space (b) transformed space

Figure 4.2: T mapping for log-normal data

To gauge the level of precision we require, the inclusion of two contour plots in figure 4.3 of the GoF statistic and the 95 percentile of the same data set seen above as a function of σ and ξ over their credible ranges while holding μ constant at 0.



(a) GoF statistic (b) 95 percentile

Figure 4.3: changes to the GoF statistic and 95 percentile as σ and ξ vary and $\mu = 0$

Proposition 4.2.1. *Holding μ constant, for an $x > \mu$ such that $F(x) \in (0, 1)$ from a given data set, $\sigma > 0$ and $\xi > 0$ necessarily have a positive relation when*

$$\xi \log\Big[1 - F(x)\Big] < 1 - \Big[1 - F(x)\Big]^{-\xi}$$

and negative relation otherwise.

Proof. By holding μ constant as well as imposing the condition that the cdf matches the empirical quantile, i.e. the pair $\{F(x), x\}$ must match for all x which in turns means holding them constant, the partial derivative of interest is

$$\frac{\partial \sigma}{\partial \xi} = \frac{x - \mu}{\big[1 - F(x)\big]^{-\xi} - 1}\left\{1 + \frac{\xi \log\big[1 - F(x)\big]}{\big[1 - F(x)\big]^{-\xi} - 1}\right\}$$

By construction

$$\frac{x-\mu}{\left[1-F(x)\right]^{-\xi}-1} > 0$$

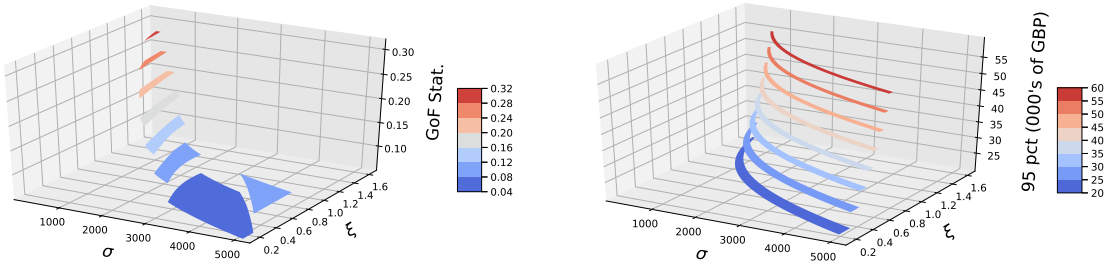
and

$$\log\left[1-F(x)\right] < 0$$

from which we can deduce the required conditions on the sum in the bracket. □

We see the graphical result of proposition 4.2.1 in figure 4.3a. However, not all "good" fits (per the GoF test statistic) are equally credible. The area corresponding to the upper ends of plotted ranges for σ and ξ with the reasonably low GoF statistics are associated with much greater percentiles than implied by the data, usually by orders magnitudes. The lowest achieved GoF statistics are found in the hollowed-out region with σ from about 4000 to 5000 and ξ near to 0 give virtually no predicted losses. Evidently, neither one of these cases is desirable and the optimisation problem has to be able to exclude them.

One credible region for the parameter space given the empirical quantiles and the GoF statistic is the strip running directly below A in the percentile plot where the 95-percentile is around the sample value of 40,000 GBP and the GoF statistic is low. The region can be isolated using the GoF threshold recommended by Mitic at 0.068. Graphically, we observe that the two variables are negatively related in the feasible region, thus contradicting (4.2.1). A closer view of the feasible region is provided in figure 4.4.



(a) GoF statistic

(b) 95 percentile

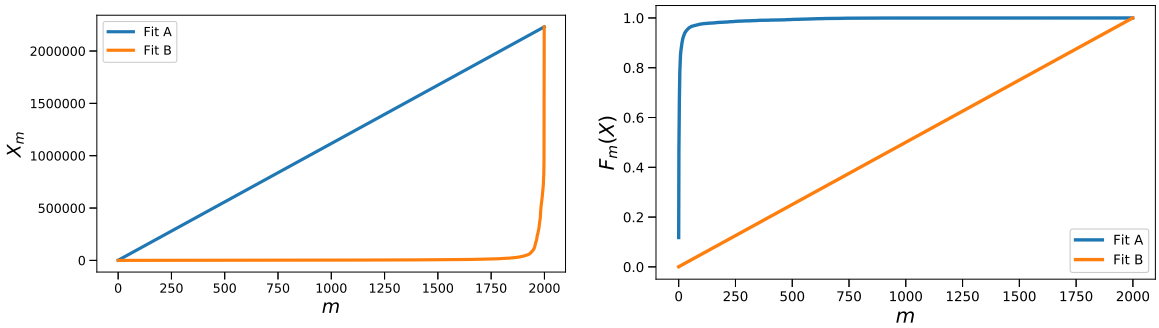
Figure 4.4: feasible region A from figure 4.3

4.3 The CDF Curve

There are two ways to obtain the empirical quantiles F_n.The first method is to create an empirical cdf out of the cumulative sum of the histogram out of the observations and extrapolate the required quantiles if needed, which we call fit A. The second way is to use the empirical quantiles themselves without alterations, and would include all extreme values, which we call fit B.

Under correct parameterisation, we can use the Glivenko-Cantelli theorem to justify using more quantiles to get a better fit because by theorem 2.2.1, we can use the GPD on the tail values. The Glivenko-Cantelli theorem is a stronger convergence condition

than the strong law of large number therefore we can just anchor the pointwise empirical quantiles to those of a GPD with an appropriate threshold to ensure that our estimates are uniformly consistent for observations greater that threshold.



(a) sample quantiles by their indexes

(b) sample ecdf at by the quantiles

Figure 4.5: example A and B fits from $\log X \sim \mathcal{N}(8, 1.96)$ with 2,000 quantiles

Theorem 4.3.1 (Glivenko-Cantelli [22] (Proof in B.2)). *If X_1, X_2, ..., X_n are iid and have the common distribution function F, then*

$$\sup_X \left| F_n(X) - F(X) \right| \xrightarrow[n\to\infty]{(a.s.)} 0$$

Fit A essentially "smoothens" the optimisation problem compared to fit B if the number of quantiles used is less than the number of observations, which we will discuss later in chapter 5. Fit A is also concentrated more at the tail if the data set is body-heavy as visualised in figure 4.5. The two method are equivalent when the number of quantiles used is the number of observations available.

A naïve observation can already reveal the large discrepancy between the two fits. For such a skewed data set as the one we are considering in this chapter, one would need many more quantile intervals to account for tail values with fit B the way fit A does. One can also deduce that fit A does not work as well as fit B on data sets that contain many outliers seeing as the quantiles will be greatly distorted whereas fit B would yield empirically-consistent quantiles.

Chapter 5

Methodology II: Optimisation Details

In this section, we outline the theoretical basis for obtaining numerical solutions to the WQE problem using two different algorithms. We will proceed to present in details the step-by-step outline of each of the two algorithms, complete with pseudo-codes as well as full convergence proofs. When necessary, we will indicate the adjustments we have made to the algorithms to tailor to adapt them to the GPD curve using some heuristics we have described in earlier chapters.

Because it is easy to lose track of the shapes of mathematical objects if one is not familiar with the context, we opted to present our results using Dirac's bra-ket notation borrowed from quantum mechanics. Simple expressions may seem cumbersome in the beginning but we believe many element indexes and operations are significantly clearer compared to standard notations.

A column vector $\mathbf{a}$ is represented by the "ket" $|a\rangle = \mathbf{a}$ and its conjugate (Hermitian) transpose, or dual (row) vector, is represented by the "bra" $\langle a| = \mathbf{a}^*$. Since all variables are real, the conjugate transpose is just the transpose $\mathbf{a}^* = \mathbf{a}^{\mathrm{T}}$. Matrix-vector multiplication is given in standard operator form: $\mathbf{Ba} = \mathbf{B}\,|a\rangle$ and $\mathbf{a}^{\mathrm{T}}\mathbf{Ba} = \langle a|\,\mathbf{B}\,|a\rangle$. Finally, inner product is given by $\mathbf{a}^{\mathrm{T}}\mathbf{a} = \langle \mathbf{a}, \mathbf{a}\rangle = \langle a \mid a\rangle$ along with the elegant expression for the a linear operator $\sum |a\rangle\langle a|$. All vectors and matrices evaluated by functionals in brackets are denoted without without the bra-kets to improve readability[1].

Using the standard orthonormal basis vectors $\{\mathbf{e}_i\}_{i=1}^n$ in $\mathbb{R}^n$, we label the bras and the kets by their indexes. For example, $\mathbf{e}_i = |e_i\rangle = |i\rangle$. The operation that evaluates a function at a point v is given by the an inner product with the relevant basis vector $f(v) = \langle v \mid f\rangle$ such that $|v\rangle = \int f(v)\,|v\rangle\,\mathrm{d}v$. In the finite dimensional case, the action that supplies a vector element is thus[2] $x_i = \langle i \mid x\rangle$. Similarly, the action that supplies the element in row i and column j of a matrix is $A_{ij} = \langle i|\,\mathbf{A}\,|j\rangle$. We make use of this notation when there are more than one index.

An important distinction in our usage of the notation is that we make no comment on the duality of the bra object, and hence the dual space in which it is conventionally defined in quantum mechanics. Our notation is purely for linear algebraic convenience and because we work only with finite dimensional objects everything is well defined (e.g.

[1] For example $\varphi(\theta)$.

[2] This is due to orthonormality of the basis vectors $\langle i \mid j\rangle = \delta_{i,j}$ where $\delta_{i,j}$ the Kronecker delta function.

there is no need to invoke Hilbert space). The bra object should *not* be understood as an abstract linear functional $\langle| : V \to \mathbb{C}$ that is an element of the dual space V^* but should be taken simply as the transpose of the corresponding column vector $|\rangle$, i.e. simply a row vector.

5.1 Backgound Information

Suppose we are given a vector containing the initial values for the parameters $|\theta_0\rangle = [\mu_0, \sigma_0, \xi_0]^{\mathrm{T}}$ for the problem in (4.1.2). Take the p-dimensional Taylor expansion of φ around $|\theta_0\rangle$ in direction $|\delta\rangle \in \mathbb{R}^p$, we get

$$\begin{aligned}\varphi(\theta_0 + \delta) \triangleq \varphi(\theta_0) + \langle \nabla\varphi(\theta_0) \mid \delta\rangle + \frac{1}{2}\langle\delta|\, \boldsymbol{\nabla}^2\varphi(\theta_0)\,|\delta\rangle \\ + \frac{1}{2}\int_0^1 \langle\delta|\, \boldsymbol{\nabla}^2\varphi(\theta_0 + t\delta) - \boldsymbol{\nabla}^2\varphi(\theta_0)\,|\delta\rangle\,(1-t)^2\,\mathrm{d}t \quad (5.1.1)\end{aligned}$$

where $|\nabla\rangle : \mathbb{R} \to \mathbb{R}^p$, $\boldsymbol{\nabla}^2 : \mathbb{R}^p \to \mathbb{R}^{p\times p}$ are the gradient and Hessian operators respectively. We wish to find the direction that minimises $\varphi(\theta_0 + \delta)$ at every step with respect to (wrt) the current value $\varphi(\theta_0)$ which corresponds to the standard quadratic Newton direction

$$\left|\delta^{GN}\right\rangle \triangleq \arg\min_{|\delta\rangle}\left\{\varphi(\theta_0 + \delta) - \varphi(\theta_0)\right\} \quad (5.1.2)$$

Ignoring the small remainder term in (5.1.1), we get the direction explicitly by equating the gradient of the function with zero

$$\begin{aligned}|\nabla\varphi(\theta_0 + \delta)\rangle &= 0 \\ \implies \left|\delta^{N}\right\rangle &= -\left[\boldsymbol{\nabla}^2\varphi(\theta_0)\right]^{-1}|\nabla\varphi(\theta_0)\rangle\end{aligned} \quad (5.1.3)$$

Newton-type algorithms have quadratic convergence, hence their superior empirical performance compared to basic algorithms such as steepest (gradient) descent. This characteristic is the direct result of the Hessian matrix that accounts for curvature in φ.

The one drawback of the Newton direction is that the calculation of the Hessian $\boldsymbol{\nabla}^2$ can be highly complicated and takes up a substantial amount of memory storage that scales up as $\mathcal{O}(p^2)$. Several Quasi-Newton methods construct matrices that are similar to the true Hessian: the Broyden-Fletcher-Goldfarb-Shanno (BFGS) [5][15][23][38] approach that uses successive gradients to construct approximate Hessians and the SR-1 [10] algorithm uses a rank-1 approximation of the true Hessian. Algorithms used in this manuscript are very similar to these approximations, because they all use linear approximations that progressively match the true Hessians near a feasible solution.

Least squares optimisation has a special structure that exploits the gradient of each residual to form an acceptable approximation of the true Hessian. Because the GPD cdf for most data sets is mostly flat for the majority of the domain, linear approximations are very convenient for efficient approximations. To facilitate computation, we follow Nocedal

and Wright [33] in defining the Jacobian $\mathbf{J} : \mathbb{R}^p \to \mathbb{R}^{m\times p}$ which is a mapping comprising of first-order partial derivatives

$$\mathbf{J} \triangleq \begin{bmatrix} \frac{\partial \rho_1}{\partial \mu} & \frac{\partial \rho_1}{\partial \sigma} & \frac{\partial \rho_1}{\partial \xi} \\ \vdots & \vdots & \vdots \\ \frac{\partial \rho_m}{\partial \mu} & \frac{\partial \rho_m}{\partial \sigma} & \frac{\partial \rho_m}{\partial \xi} \end{bmatrix} = \begin{bmatrix} \langle \nabla \rho_1 | \\ \vdots \\ \langle \nabla \rho_m | \end{bmatrix} \tag{5.1.4}$$

and the first order partial derivatives of the residual function of x_j are given by

$$\frac{\partial \rho_j}{\partial \mu} = \frac{1}{\sigma}\left[1 + \frac{\xi}{\sigma}\big(x_j - \mu\big)\right]^{-1/\xi - 1}$$

$$\frac{\partial \rho_j}{\partial \sigma} = \frac{\big(x_j - \mu\big)}{\sigma^2}\left[1 + \frac{\xi}{\sigma}\big(x_j - \mu\big)\right]^{-1/\xi - 1}$$

$$\frac{\partial \rho_j}{\partial \xi} = \left\{\frac{1}{\xi^2}\log\left[1 + \frac{\xi}{\sigma}\big(x_j - \mu\big)\right] - \frac{\big(x_j - \mu\big)}{\xi\sigma + \xi^2\big(x_j - \mu\big)}\right\}\left[1 + \frac{\xi}{\sigma}\big(x_j - \mu\big)\right]^{-1/\xi}$$

The Jacobian then allows us to concisely express the gradient and Hessian matrix of φ. Specifically, we note that the gradient $|\nabla\varphi\rangle$ and Hessian $\boldsymbol{\nabla}^2\varphi$ can be derived from each residual's gradient and Hessian respectively. Let $\varphi = \varphi(\theta)$ and $|\rho_j\rangle = |\rho_j(\theta)\rangle$, the operators read

$$\begin{aligned} |\nabla\varphi\rangle &= \sum_{j=1}^{m} \rho_j \, |\nabla\rho_j\rangle = \mathbf{J}^{\mathrm{T}} \, |\rho\rangle \\ \boldsymbol{\nabla}^2\varphi &= \sum_{j=1}^{m} |\nabla\rho_j\rangle \, \langle\nabla\rho_j| + \sum_{j=1}^{m} \rho_j \boldsymbol{\nabla}^2\rho_j \\ &= \mathbf{J}^{\mathrm{T}}\mathbf{J} + + \sum_{j=1}^{m} \rho_j \boldsymbol{\nabla}^2\rho_j \end{aligned}$$

All algorithms used to produce the results seen in this thesis use the approximation

$$\boldsymbol{\nabla}^2\varphi \approx \sum_{j=1}^{m} |\nabla\rho_j\rangle \, \langle\nabla\rho_j| = \mathbf{J}^{\mathrm{T}}\mathbf{J} \tag{5.1.5}$$

Discarding the second sum can save a significant amount of computation and storage and if the steps are small and almost linear then the approximation is very close to the true Hessian.

5.2 WQE First Stage

5.2.a Gauss-Newton

Perhaps the most straight forward algorithm given the Newton direction and approximate Hessian is the Gauss-Newton (GN) step. The algorithm is a essentially a quasi-Newton approach paired with an appropriate step size obtained from a line search algorithm to ensure sufficient decrease in φ. We will not go into great detail about the line search algorithm, which can be found in [33], but it is not too complicated to find a step size $\alpha \in \mathbb{R}^+$ given direction $|\delta\rangle$ that satisfy the strong Wolfe [44] conditions:

$$\begin{aligned} \varphi(\theta + \alpha\delta) &\leq \varphi(\theta) + c_1\alpha \langle \nabla\varphi(\theta) \mid \delta\rangle \\ -\langle \nabla\varphi(\theta + \alpha\delta) \mid \delta\rangle &\leq -c_2 \langle \nabla\varphi(\theta) \mid \delta\rangle \\ \big|\langle \nabla\varphi(\theta + \alpha\delta) \mid \delta\rangle\big| &\leq \big|c_2 \langle \nabla\varphi(\theta) \mid \delta\rangle\big| \end{aligned} \tag{5.2.1}$$

for $c1, c2 \in \mathbb{R}^+$. c_1 guarantees sufficient decrease in φ and c_2 bounds the step away from an ascent direction prevents too small an increment.

We shorten dependent matrices and vectors to $\mathbf{A}\big(\theta_k\big) = \mathbf{A}_k$. The GN step at iteration k is

$$\big|\delta_k^{GN}\big\rangle = -\Big(\mathbf{J}_k^{\mathrm{T}}\mathbf{J}_k\Big)^{-1}\mathbf{J}_k^{\mathrm{T}}\,|\rho_k\rangle \tag{5.2.2}$$

with the corresponding update rule

$$|\theta_{k+1}\rangle = |\theta_k\rangle - \alpha_k\Big(\mathbf{J}_k^{\mathrm{T}}\mathbf{J}_k\Big)^{-1}\mathbf{J}_k^{\mathrm{T}}\,|\rho_k\rangle \tag{5.2.3}$$

which is the solution vector that makes the residual normal to the column space of $\mathbf{J}$

$$\min_{|\delta\rangle} \ \Big\|\mathbf{J}_k\,|\delta\rangle + |\rho_k\rangle\Big\|_2^2 \tag{5.2.4}$$

Every step in the GN algorithm can be seen as solving for linear normal equations that yield the standard ordinary least-square (OLS) estimators. Explicit inversion of the approximate Hessian $\mathbf{J}_k^{\mathrm{T}}\mathbf{J}_k$ can be a demanding task but if p is small then the empirical boost to performance when a factorisation and back or forward substitution instead is only marginal.

GN directions should uniformly share common elements with the steepest descent direction for guaranteed convergence. These common elements aid convergence because they ensure that the GN directions are *generally* a descent direction (which the steepest descent direction is obviously one). One can conceivably see that a uniformly descent sequence of directions would lead the algorithm to a local minimum should there exist one. Specifically, we will use theorem 5.2.1 to show that the algorithm forces residuals to vanish if the directions are well conditioned.

Theorem 5.2.1. [33] *Let* $\mathbf{J}$ *be Lipschitz continuous in the neighbourhood* $\mathcal{N}$ *of the level set* $\mathcal{L} = \big\{\,|\rho\rangle \;\big|\; \varphi(\rho) \leq \varphi(\rho_0)\big\}$, φ *is bounded in* $\mathbb{R}$ *and continuously differentiable in* $\mathcal{N}$.

Algorithm 1: Gauss-Newton

Initialise $|\theta_k\rangle = |\theta_0\rangle$
while *not converged* **do**
  $\mathbf{J}_k = \mathbf{J}(\theta_k)$
  $|\rho_k\rangle = \big|\rho\big(\theta_k\big)\big\rangle$
  $|\delta_k\rangle = -\big(\mathbf{J_k}^{\mathrm{T}}\mathbf{J}_k\big)^{-1}\mathbf{J}_k\,|\rho_k\rangle$
  $\alpha_k = \mathrm{LineSearch}\big(\delta_k\big)$
  $|\theta_{k+1}\rangle = |\theta_k\rangle + \alpha_k\,|\delta_k\rangle$
end while
return $|\theta_{k+1}\rangle$;

Also let the sequence of directions $\big\{\,|\delta_k\rangle\,\big\}_{k=1}^{\infty}$ *satisfies the strong Wolfe conditions* (5.2.1). *Define* ϕ_k *as the angle between* $-|\nabla\varphi_k\rangle$ *and* $|\delta\rangle$, *as such we get*

$$\cos\phi_k = \frac{-\,\langle\nabla\varphi_k \mid \delta\rangle}{\Big\|\,|\nabla\varphi_k\rangle\Big\|_2\Big\||\delta\rangle\Big\|_2}$$

and the following condition (called Zountendijk's condition) holds

$$\sum_{k\in\mathbb{N}}\big(\cos\phi_k\big)^2\Big\|\mathbf{J}_k^{\mathrm{T}}\,|\rho_k\rangle\Big\|_2^2 < \infty$$

Proof. From (5.2.1) we know that at iteration k

$$\langle\nabla\varphi_{k+1} \mid \delta\rangle \geq c_2\,\langle\nabla\varphi_k \mid \delta\rangle$$

where

$$\varphi_{k+1} = \varphi\big(\theta_k + \alpha_k\delta\big)$$
$$\varphi_k = \varphi\big(\theta_k\big)$$

By linearity of the inner product

$$\langle\nabla\varphi_{k+1} - \nabla\varphi_k \mid \delta\rangle \geq \big(c_2 - 1\big)\,\langle\nabla\varphi_k \mid \delta\rangle$$

Since $\mathbf{J}$ is Lipschitz in $\mathcal{N}$, the individual gradients are also Lipschitz in $\mathcal{N}$, which implies there exists $M > 0$ such that

$$\Big\|\,|\nabla\varphi(x)\rangle - |\nabla\varphi(y)\rangle\Big\|_2 \leq M\Big\||x\rangle - |y\rangle\Big\|_2 \quad \forall\ |x\rangle, |y\rangle \in \mathcal{N}$$

and therefore

$$\begin{aligned}\big\langle\nabla\varphi_{k+1} - \nabla\varphi_k \;\big|\; \delta\big\rangle &\leq M\,\big\langle\theta_{k+1} - \theta_k \;\big|\; \delta\big\rangle \\ &= \alpha_k M\Big\||\delta\rangle\Big\|_2^2 \\ \implies \alpha_k &\geq \frac{\big(c_2 - 1\big)\,\langle\nabla\varphi_k \mid \delta\rangle}{M\big\||\delta\rangle\big\|_2^2}\end{aligned}$$

We know from the first strong Wolfe condition

$$\begin{aligned}\varphi_{k+1} &\leq \varphi_k + c_1\alpha \left\langle \nabla\varphi_k \mid \delta \right\rangle \\ &\leq \varphi_k + \frac{c_1\big(c_2-1\big)\left\langle \nabla\varphi_k \mid \delta \right\rangle^2}{M\Big\| |\delta\rangle \Big\|_2^2} \\ &= \varphi_k + \frac{c_1\big(c_2-1\big)}{M}\Big\| |\nabla\varphi_k\rangle \Big\|_2^2 \Big(\cos\phi_k\Big)^2\end{aligned}$$

By the k^{th} iteration we have obtained an additive condition of the form

$$\varphi_k \leq \varphi_0 - D\sum_{i=1}^{k}\Big(\cos\phi_i\Big)^2\Big\| |\nabla\varphi_j\rangle \Big\|_2^2$$

where $D = c_1\big(c_2-1\big)/M$. Since φ is bounded below in $\mathbb{R}$, the sum must converge. □

Corollary 5.2.1.1. *If the sequence $\big\{\phi_k\big\}_{k=1}^{\infty}$ is uniformly bounded away from $\pi/2$ then the final sum in theorem 5.2.1 implies that*

$$\lim_{k\to\infty} \quad \Big\| \mathbf{J}_k^{\mathrm{T}} \left|\rho_k\right\rangle \Big\|_2 = 0$$

Proof. Convergence of the sum of a strictly positive sequence requires that the sequence eventually vanish. □

Corollary 5.2.1.1 is important to the GN iterations as we show in the next section that the angle ϕ_k is usually close to $\pi/2$ which fails to force the algorithm to converge. The convergence theory for GN depends the Jacobian. In particular, we need that the singular values of $\mathbf{J}$, hence the eigenvalues of $\mathbf{J}^{\mathrm{T}}\mathbf{J}$, to not vanish near the solution so that $\mathbf{J}^{\mathrm{T}}\mathbf{J}$ is invertible. The following convergence theorems and proofs can be found in more detail in [33].

Lemma 5.2.2. [33] *The directions generated by $\left|\delta^{GN}\right\rangle$ are descent.*

Proof.

$$\begin{aligned}\left\langle \nabla\varphi \mid \delta^{GN} \right\rangle &= \langle\rho|\, \mathbf{J}\left|\delta^{GN}\right\rangle \\ &= -\left\langle \delta^{GN}\right|\mathbf{J}^{\mathrm{T}}\mathbf{J}\left|\delta^{GN}\right\rangle \\ &= -\Big\| \mathbf{J}^{\mathrm{T}}\left|\delta^{GN}\right\rangle \Big\|_2^2 \leq 0\end{aligned}$$

where the last inequality implies, and by construction the gradient is always an ascent direction, that $\left|\delta^{GN}\right\rangle$ is in fact descent. □

Theorem 5.2.3 (Proof in C.1)**.** [33] *Given a residual vector $|\rho\rangle$ that is Lipschitz and differentiable in the neighbourhood $\mathcal{N}$ of the bounded level set $\mathcal{L} = \big\{\, |\rho\rangle \mid \varphi(\rho) \leq \varphi(\rho_0)\big\}$ with $|\rho_0\rangle$ being the initial residual vector. Let $\mathbf{J}$ be of full rank uniformly through the*

iterations when $|\rho_k\rangle$ *is in* $\mathcal{N}$*, implying the eigenvalues of* $\mathbf{J}^{\mathrm{T}}\mathbf{J}$ *are positively bounded from below, i.e. there exists* $\tau > 0$ *such that*

$$\left\|\mathbf{J}_k^{\mathrm{T}}\,|\rho_k\rangle\right\|_2 \geq \tau\left\||\rho_k\rangle\right\|_2$$

The Gauss-Newton steps then satisfy the following condition

$$\lim_{k\to\infty}|\nabla\varphi_k\rangle = \lim_{k\to\infty}\mathbf{J}_k{}^{\mathrm{T}}\,|\rho_k\rangle = |\mathbf{0}\rangle$$

As we show in the estimation chapter, the requirement for convergence is practically difficult to achieve in practice due to the extreme shape of the GPD cdf curve. This empirical difficulty leads us to the next algorithm.

5.2.b Levenberg-Marquadt

The Levenberg-Marquadt (LM) algorithm falls is a trust-region method. Instead of finding a descent direction and taking the appropriate step length, we instead fix a maximum radius of a search neighbourhood around the current iterate $|\theta_k\rangle$ and looking for a direction that yields the maximum decrease.

The general problem at an iteration is

$$\begin{gathered}\left|\delta^{LM}\right\rangle \triangleq \underset{|\delta\rangle}{\arg\min}\left\{q\big(\theta+\delta\big) = q\big(\theta\big) + \left\langle\nabla q\big(\theta\big)\;\middle|\;\delta\right\rangle + \frac{1}{2}\left\langle\delta\right|\mathbf{H}(\theta)\left|\delta\right\rangle\right\} \\ \text{s.t.}\;\;\left\|\left|\delta\right\rangle\right\|_2 \leq r\end{gathered} \tag{5.2.5}$$

where $\mathbf{H} \in \mathbb{R}^{p\times p}$ is symmetric and spd and $r \in \mathbb{R}^+$ is the maximum search radius. The quadratic model q in (5.2.5) represents the hypothetical decrease in the objective function and we decide on the value of r based on the ratio of change in q and change in actual function evaluation φ. Evidently the divergence between the q and φ up to second order is $\mathcal{O}\big(\|\mathbf{H}-\boldsymbol{\nabla}^2\|^2\big)$, i.e. the better the Hessian approximation the more similar the convergence is to quadratic.

Theorem 5.2.4. [33] *The quadratic model* (5.2.5) *has a feasible solution* $\left|\delta^{LM}\right\rangle$ *if and only if there exists* $\lambda \geq 0$ *such that, and thus is equivalent to,*

$$\big(\mathbf{H}+\lambda\mathbf{I}_p\big)\left|\delta^{LM}\right\rangle = -\left|\nabla q\right\rangle$$

where $\mathbf{I}_p$ *is a* $p\times p$ *identity matrix.*

Proof. There are two cases. When the unconstrained minimiser $|\delta\rangle$ of q is strictly feasible, i.e. $\left\||\delta\rangle\right\|_2 < r_M$, then $\left|\delta^{LM}\right\rangle = \left|\delta^{GN}\right\rangle$ and thus $\lambda = 0$. The second case is when $\left\||\delta\rangle\right\|_2 \geq r_M$. From the Karush-Kuhn-Tucker conditions, we have that for a feasible solution to the inequality constrained problem we must have that the Lagrange multiplier $\lambda > 0$ satisfies

$$\begin{aligned}|\nabla q\rangle &= \lambda\big(r_M - \left\||\delta\rangle\right\|_2\big) = 0 \\ \implies \frac{\lambda}{2}\big(r_M^2 - \left\||\delta\rangle\right\|_2^2\big) &= 0\end{aligned}$$

since λ can be any non-negative number. We can then set up the Lagrangian

$$\mathbb{L} = q\big(\theta + \delta\big) - \frac{\lambda}{2}\big(r_M^2 - \big\||\delta\rangle\big\|_2^2\big)$$

which has the unique minimiser sought. □

The LM step can be derived as a damped version of (5.2.4). Using the expressions for the gradient and Hessian approximation $\mathbf{H} = \mathbf{J}^{\mathrm{T}}\mathbf{J}$ of φ, from 5.2.4 the optimality condition for the step at a general iteration is

$$\big(\mathbf{J}^{\mathrm{T}}\mathbf{J} + \lambda\mathbf{I}_p\big)\,\big|\delta^{LM}\big\rangle = -\mathbf{J}^{\mathrm{T}}\,|\rho\rangle \tag{5.2.6}$$

which are the normal equations for the least squares problem

$$\min_{|\delta\rangle} \left\| \begin{bmatrix} \mathbf{J} \\ \sqrt{\lambda}\mathbf{I}_p \end{bmatrix} |\delta\rangle + \begin{bmatrix} |\rho\rangle \\ 0 \end{bmatrix} \right\|_2^2$$

The main ingredient to solve for $|\delta\rangle$ is the QR decomposition of $\mathbf{J}$, which allows for continuous insertion of the factor $\sqrt{\lambda}$ from the initial factorisation. This is done automatically through a series of Givens rotation to zero out lower diagonal elements in $\mathbf{R}$ in the factorisation

$$\begin{bmatrix} \mathbf{J} \\ \sqrt{\lambda}\mathbf{I}_p \end{bmatrix} = \mathbf{QR} = \begin{bmatrix} \mathbf{Q}_1 & \mathbf{Q}_2 \end{bmatrix} \begin{bmatrix} \mathbf{R} \\ \mathbf{0} \end{bmatrix}$$

where $\mathbf{Q}_1 \in \mathbb{R}^{(m+p)\times p}, \mathbf{Q}_2 \in \mathbb{R}^{(m+p)\times m}$ are orthogonal, i.e. $\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I}_m$, $\mathbf{R} \in \mathbb{R}^{p\times p}$ is right upper-triangular and $\mathbf{0}$ is a matrix of zeros of size $m \times p$. A simple algebra operation shows

$$\mathbf{R}^{\mathrm{T}}\mathbf{R} = \mathbf{J}^{\mathrm{T}}\mathbf{J} + \lambda\mathbf{I}_p$$

Solving for the direction $\big|\delta^{LM}\big\rangle$ relies on the pivots of $\mathbf{R}$, which correspond to the singular values of $\begin{bmatrix}\mathbf{J}^{\mathrm{T}} & \sqrt{\lambda}\mathbf{I}_p\end{bmatrix}^{\mathrm{T}}$. If the pivots are near zero, we damp them by a small factor λ until we find a direction that satisfies the radius constraint $\big\||\delta\rangle\big\|_2 \leq r_M$. We give an outline of the implementation in appendix C.2. For more details on the applications we recommend Golub and Van Loan's book [24].

This procedure of QR decomposition for least squares problems falls under the umbrella of LSQR algorithms first introduced by Paige and Saunders [34]. An alternative procedure is to perform a singular value decomposition (SVD) of $\mathbf{J}$, which is more expensive than QR but has no obvious computational advantage for this specific cdf curve-fitting problem.

Theorem 5.2.5. [33] *Let* $\mathbf{J}$ *be full rank,* ν_m *be the smallest eigenvalue of* $\mathbf{J}^{\mathrm{T}}\mathbf{J} + \lambda\mathbf{I}_p$ *and* $r_M > 0$ *be fixed. For a monotonically increasing sequence* $\{\lambda_i\}_{i=1}^{\infty}$ *where* $\lambda_i > -\nu_m$ *for all* i *there exists* $N \in \mathbb{N}$ *such that* $n \geq N$ *implies*

$$\Big\|\big|\delta\big(\lambda_n\big)\big\rangle\Big\|_2 \leq r_M$$

Proof. Because $\mathbf{J}$ is full rank, $\mathbf{J}^{\mathrm{T}}\mathbf{J}$ is symmetric and spd, therefore has an eigenvalue decomposition. Let $\mathbf{Q}\mathbf{D}\mathbf{Q}^{\mathrm{T}} = \mathbf{J}^{\mathrm{T}}\mathbf{J}$ where $\mathbf{D}$ is a diagonal matrix containing the eigenvalues $\{\nu_j\}_{j=1}^{p}$ of $\mathbf{J}^{\mathrm{T}}\mathbf{J}$ and the columns of $\mathbf{Q}$ their corresponding orthonormal eigenvectors $\left\{\, |q_j\rangle \,\right\}_{j=1}^{p}$. Damping $\mathbf{J}^{\mathrm{T}}\mathbf{J}$ by λ_i results in

$$\mathbf{J}^{\mathrm{T}}\mathbf{J} + \lambda_i \mathbf{I}_p = \mathbf{Q}\big(\mathbf{D} + \lambda_i \mathbf{I}_p\big)\mathbf{Q}^{\mathrm{T}}$$

which leads to the direction

$$\begin{aligned} \big|\delta\big(\lambda_i\big)\big\rangle &= -\Big[\mathbf{Q}\big(\mathbf{D} + \lambda_i \mathbf{I}_p\big)\mathbf{Q}^{\mathrm{T}}\Big]^{-1} |\nabla\varphi\rangle \\ &= -\sum_{j=1}^{p} \frac{|q_j\rangle \langle q_j \mid \nabla\varphi\rangle}{\nu_j + \lambda_i} \end{aligned}$$

Since this corresponds to a projection onto with orthonormal bases with a biased scaling factor, we get

$$\big\langle \delta\big(\lambda_i\big) \mid \delta(\lambda_i)\big\rangle = \Big\| |\delta(\lambda_i)\rangle \Big\|_2^2 = \sum_{j=1}^{p} \frac{\langle q_j \mid \nabla\varphi\rangle^2}{\big(\nu_j + \lambda_i\big)^2}$$

which vanishes as λ_i tend to infinity since $\nu_j \geq \nu_m$ for all j. □

To find $\big|\delta^{LM}\big\rangle$, we first do forward substitution with $\mathbf{R}^{\mathrm{T}}\,|x\rangle = -\mathbf{J}^{\mathrm{T}}\,|\rho\rangle$ and then backward substitution with $\mathbf{R}\,\big|\delta^{LM}\big\rangle = |x\rangle$. It can be shown that the linearisation of the feasibility condition on the norm of $|\delta\rangle$ when plugged into the root-finding procedure Newton-Raphson gives the λ updating rule

$$\lambda_{i+1} \triangleq \max\left\{0, \lambda_i + \frac{\big\| |\delta\rangle \big\|_2}{\big\| \mathbf{R}^{-\mathrm{T}}\,|\delta\rangle \big\|_2} \left(\frac{\big\| |\delta\rangle \big\|_2 - r_M}{r_M} \right) \right\} \tag{5.2.7}$$

where $\mathbf{R}^{-\mathrm{T}}$ is the inverse transpose of $\mathbf{R}$.

Because the cdf is essentially flat for the vast majority of the support for a typical tail-loss model the Jacobian's elements are usually very small, thus making the λ update even more crucial to the trust region problem because it regularises the quadratic approximation of the residual function, thus the local minimum that the algorithm converges to depends on the initial radius r_M. We define the following ratio to assess the quadratic approximation's accuracy

$$\begin{aligned} \tau_k &\triangleq \frac{\varphi\big(\theta_k\big) - \varphi\big(\theta_k + \delta^{LM}\big)}{q_k\big(0\big) - q_k\big(\delta^{LM}\big)} \\ q_k\big(\delta^{LM}\big) &\triangleq q\big(\theta_k + \delta^{LM}\big) \end{aligned} \tag{5.2.8}$$

and shrink the radius to $r_k < r_M$ when $\tau_k << 1$. The full details of the steps are given in algorithm 2.

A potential concern is that the search space for a solution includes F evaluations that

yield invalid values. Specifically, if for some μ, σ, ξ and x

$$1 + \frac{\sigma}{\xi}\big(x - \mu\big) < 0$$

then $F(x)$ is complex. We create a barrier to the step component $\langle i \mid \delta_k \rangle$ such that if the step component makes the new $\langle i \mid \theta_{k+1} \rangle = \langle i \mid \theta_k + \delta_k \rangle$ negative then we do not update component i. This is equivalent to restricting the search space to the relative interior of the domain of the remaining elements (the smallest affine hull that contains the remaining search space). This concept is borrowed from convex programming and even though convex procedures do not apply we can derive still take advantage of their useful properties due to the spd second order approximation.

The LM algorithm is usually seen as a damped procedure with a steepest descent initial step, which converges towards the usual GN step as it zeroes in on a solution. A summary of the proof of convergence for LM is that since we can bound the residuals from above and below and the step's radius by r_M, we can find a positive bound for step sizes which produce a sequence of gradients whose norms converge to 0. Our proofs, given in appendix C.3, follow Nocedal and Wright's [33].

Algorithm 2: Levenberg-Marquadt Trust Region [33]

Initialise $|\theta_k\rangle = |\theta_0\rangle$
Initialise $r_k = r_M > 0$
Initialise $0 < \eta << 1$
while *not converged* **do**
  $\mathbf{J}_k = \mathbf{J}(\theta_k)$
  $\mathbf{JJ}_k = \mathbf{J}_k^{\mathrm{T}}\mathbf{J}_k$
  $\mathbf{JJ}_k = \mathbf{Q}_k\mathbf{R}_k$
  Initialise $\lambda_i > 0$
  while $\big\||\delta_k\rangle\big\|_2 > r_M$ **and** $\lambda > 0$ **do**
    $|\delta_k\rangle =$ solve (5.2.6)
    $\lambda_{i+1} =$ solve (5.2.7)
    if $\lambda == 0$ **then**
      $|\delta_k\rangle = \big|\delta^{GN}\big\rangle$
  end while
  $\tau_k =$ solve (5.2.8)
  if $\tau_k < 0.25$ **then**
    $r_k = 0.25 r_k$
  else if $\tau_k > 0.75$ **and** $\Big|\big\||\delta_k\rangle\big\|_2^2 - (r_k)^2\Big| < 1\mathrm{e}^{-12}$ **then**
    $r_k = \min\{2r_k, r_M\}$
  if $\tau_k > \eta$ **then**
    $|\theta_{k+1}\rangle = |\theta_k\rangle + |\delta_k\rangle$
end while
return $|\theta_{k+1}\rangle$;

Theorem 5.2.6 (Proof in C.3). *For $0 \leq \eta < 1$ in algorithm 2 with the level set $\mathcal{L} = \big\{\, |\rho\rangle \mid \varphi(\rho) \leq \varphi(\rho_0) \big\}$ be bounded along with a Lipschitz differentiable residual sequence in a neighbourhood $\mathcal{N} \subset \mathcal{L}$ near the solution. Furthermore, let the Hessian approximation $\mathbf{J}^{\mathrm{T}}\mathbf{J}$ be Lipschitz continuous in the neighbourhood $\mathcal{N}$. If $|\delta_k\rangle$, for each k such that the residual is near the solution, satisfies*

$$q_k\big(0\big) - q\big(\delta_k\big) \geq c_1 \Big\| \mathbf{J}_k^{\mathrm{T}} \, |\rho_k\rangle \Big\|_2 \bigg(r_M \bigwedge \frac{\big\| \mathbf{J}_k^{\mathrm{T}} \, |\rho_k\rangle \big\|_2}{\big\| \mathbf{J}_k^{\mathrm{T}} \mathbf{J}_k \big\|_2} \bigg)$$

for $c_1 > 0$, where $\big(a \wedge b\big) = \min\big(a, b\big)$, and there exists $1 < \gamma < \infty$ such that

$$\Big\| |\delta_k\rangle \Big\|_2 \leq \gamma r_M$$

Then, for $\eta = 0$, the following is true

$$\liminf_{k \to \infty} \quad \Big\| \big| \nabla \varphi\big(\theta_k\big) \big\rangle \Big\|_2 = 0$$

and for $\eta \in (0, 0.25]$, we have

$$\lim_{k \to \infty} \quad \Big\| \big| \nabla \varphi\big(\theta_k\big) \big\rangle \Big\|_2 = 0$$

5.3 WQE Second Stage

For the second stage of the algorithm, we have some flexibility when it comes to weights and residual types. Suppose we allow for the $|x\rangle$ vector coordinates to perturb in our optimisation scheme as well as attach pre-calibrated weights for both the standard residuals and the new perturbations, we get the least squares problem

$$\min_{|\theta\rangle, |\epsilon\rangle} \varphi \triangleq \frac{1}{2} \sum_{j=1}^{m} \Big\{ \langle j \mid w_1 \rangle^2 \Big[F_n\big(x_j\big) - F\big(\theta, \epsilon \mid x_j\big) \Big]^2 + \langle j \mid w_2 \rangle^2 \, \epsilon_j^2 \Big\} \tag{5.3.1}$$

where

$$F(\theta, \epsilon \mid x_j) = 1 - \bigg[1 + \frac{\xi}{\sigma} \Big(x_j + \epsilon_j - \mu \Big) \bigg]^{-1/\xi} \tag{5.3.2}$$

If the weight vectors $|w_1\rangle, |w_2\rangle \in \mathbb{R}^m$ are identical, the problem is equivalent to an orthogonal projection of the residuals onto the normal vector to φ at every point. This is contrasted to the standard least squares problem which projects the output onto the span of $|x\rangle$, and is accordingly called *orthogonal distance regression* (ODR).

We follow Boggs, Byrd and Schnabel [4] in the next following steps, with the exception that our data vector $|x\rangle$ is one dimensional, whereas they work with a general multidimensional case. This allows our calculation to simplify greatly. Let $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{m \times m}$ be $\mathrm{diag}(w_1)$ and $\mathrm{diag}(w_2)$ respectively. We begin by defining two residual vectors

$$|\rho_1(\theta, \epsilon)\rangle \triangleq \mathbf{W}_1 \, |F_n(x) - F(\theta, \epsilon \mid x)\rangle$$
$$|\rho_2(\epsilon)\rangle \triangleq \mathbf{W}_2 \, |\epsilon\rangle$$

and splitting (5.3.1) into

$$\min_{|\theta\rangle,|\epsilon\rangle} \varphi \triangleq \frac{1}{2}\sum_{j=1}^{2m} \begin{cases} \langle j \mid w_1\rangle^2 \left[F_n\big(x_j\big) - F\big(\theta, \epsilon \mid x_j\big)\right]^2 & j = 1, ..., m \\ \langle j \mid w_2\rangle^2 \epsilon_j^2 & j = (m+1), ..., 2m \end{cases} \tag{5.3.3}$$

which is equivalent to

$$\min_{|\theta\rangle,|\epsilon\rangle} \varphi = \frac{1}{2}\Big\||\rho\rangle\Big\|_2^2 = \frac{1}{2}\left\|\begin{bmatrix} |\rho_1(\theta,\epsilon)\rangle \\ |\rho_2(\epsilon)\rangle \end{bmatrix}\right\|_2^2 \tag{5.3.4}$$

where $|\rho\rangle \in \mathbb{R}^{2m}$.

The gradient of φ can again be expressed as a linear combination of the individual gradients

$$|\nabla\varphi(\theta,\epsilon)\rangle = \sum_{j=1}^{2m} \begin{cases} \langle j \mid \rho_1\rangle \Big(|\nabla\rangle \langle j \mid \rho_1\rangle \Big) & j = 1, ..., m \\ \langle j \mid \rho_2\rangle \Big(|\nabla\rangle \langle j \mid \rho_2\rangle \Big) & j = (m+1), ..., 2m \end{cases}$$

where

$$|\nabla\rangle \langle j \mid \rho_1\rangle = \begin{bmatrix} \frac{\partial\langle j|\rho_1\rangle}{\partial\mu} & \frac{\partial\langle j|\rho_1\rangle}{\partial\sigma} & \frac{\partial\langle j|\rho_1\rangle}{\partial\xi} & 0 & \dots & 0 & \frac{\partial\langle j|\rho_1\rangle}{\partial\epsilon_j} & 0 & \dots & 0 \end{bmatrix}^{\mathrm{T}}$$

$$|\nabla\rangle \langle j \mid \rho_2\rangle = \begin{bmatrix} 0 & \dots & 0 & \langle j \mid w_2\rangle & 0 & \dots & 0 \end{bmatrix}^{\mathrm{T}}$$

The Jacobian is thus made up of sub-Jacobians being different block matrices

$$\mathbf{J} = \begin{bmatrix} \mathbf{J}_\theta & \mathbf{J}_\epsilon \\ \mathbf{0} & \mathbf{W}_2 \end{bmatrix}$$

such that $\mathbf{J}_\theta \in \mathbb{R}^{p\times p}$ has elements $\langle i|\,\mathbf{J}_\theta\,|j\rangle = \partial\langle j \mid \rho_1\rangle/\partial\theta_i$, $\mathbf{J}_\epsilon \in \mathbb{R}^{m\times m}$ is diagonal and has elements $\langle j|\,\mathbf{J}_\epsilon\,|j\rangle = \partial\langle j \mid \rho_1\rangle/\partial\epsilon_j$ and $\mathbf{0}$ contains all zeroes and has dimension $m \times p$. The gradient of φ can finally be expressed as

$$|\nabla\varphi(\theta,\epsilon)\rangle = \mathbf{J}^{\mathrm{T}}\,|\rho(\theta,\epsilon)\rangle$$

and the approximate Hessian is given by the expression

$$\boldsymbol{\nabla}^2\varphi(\theta,\epsilon) \approx \mathbf{J}^{\mathrm{T}}\mathbf{J} = \begin{bmatrix} \mathbf{J}_\theta^{\mathrm{T}}\mathbf{J}_\theta & \mathbf{J}_\theta^{\mathrm{T}}\mathbf{J}_\epsilon \\ \mathbf{J}_\epsilon^{\mathrm{T}}\mathbf{J}_\theta & \mathbf{J}_\epsilon^2 + \mathbf{W}_2^2 \end{bmatrix}$$

We reapply the LM algorithm at this stage to prevent the small singular values of $\mathbf{J}$ from stalling convergence. Similar to (5.2.5), our general step can be expressed as the solution to the least squares problem

$$\min_{|\delta_\theta\rangle,|\delta_\epsilon\rangle} \frac{1}{2}\left\|\begin{bmatrix} \mathbf{J}_\theta & \mathbf{J}_\epsilon \\ \mathbf{0} & \mathbf{W}_2 \\ \sqrt{\lambda}\mathbf{I}_p & \mathbf{0} \\ \mathbf{0} & \sqrt{\lambda}\mathbf{I}_m \end{bmatrix}\begin{bmatrix} |\delta_\theta\rangle \\ |\delta_\epsilon\rangle \end{bmatrix} + \begin{bmatrix} |\rho_1\rangle \\ |\rho_2\rangle \\ 0 \\ 0 \end{bmatrix}\right\|_2^2$$

which has the normal equations

$$\begin{bmatrix} \mathbf{J}_\theta^{\mathrm{T}}\mathbf{J}_\theta + \lambda \mathbf{I}_p & \mathbf{J}_\theta^{\mathrm{T}}\mathbf{J}_\epsilon \\ \mathbf{J}_\epsilon^{\mathrm{T}}\mathbf{J}_\theta & \mathbf{J}_\epsilon^2 + \mathbf{W}_2^2 + \lambda \mathbf{I}_m \end{bmatrix} \begin{bmatrix} |\delta_\theta\rangle \\ |\delta_\epsilon\rangle \end{bmatrix} = - \begin{bmatrix} \mathbf{J}_\theta^{\mathrm{T}} |\rho_1\rangle \\ \mathbf{J}_\epsilon^{\mathrm{T}} |\rho_1\rangle + w_2 |\rho_2\rangle \end{bmatrix}$$

Proposition 5.3.1. *The normal equations for $|\delta_\theta\rangle$ can be expressed as*

$$\left[\mathbf{J}_\theta^{\mathrm{T}}\big(\mathbf{I}_m - \mathbf{J}_\epsilon^{\mathrm{T}}\mathbf{E}^{-1}\mathbf{J}_\epsilon\big)\mathbf{J}_\theta + \lambda \mathbf{I}_p\right] |\delta_\theta\rangle = -\mathbf{J}_\theta^{\mathrm{T}}\left[\big(\mathbf{I}_m - \mathbf{J}_\epsilon^{\mathrm{T}}\mathbf{E}^{-1}\mathbf{J}_\epsilon\big) |\rho_1\rangle - \mathbf{J}_\epsilon\mathbf{E}^{-1}\mathbf{W}_2 |\rho_2\rangle\right]$$

where $\mathbf{E} = \mathbf{J}_\epsilon^2 + \mathbf{W}_2^2 + \lambda \mathbf{I}_m$.

Proof. Because of the block structure of the total Jacobian, we can get the following expression in terms of $|\delta_\epsilon\rangle$

$$\mathbf{J}_\epsilon^{\mathrm{T}}\mathbf{J}_\theta |\delta_\theta\rangle + \mathbf{E} |\delta_\epsilon\rangle = -\Big(\mathbf{J}_\epsilon^{\mathrm{T}} |\rho_1\rangle + \mathbf{W}_2 |\rho_2\rangle\Big)$$

which we can use to plug into the expression for $|\delta_\theta\rangle$ to eliminate $|\delta_\epsilon\rangle$. □

Algorithm 3: Maximum Radius Initiation

Initialise $|\theta_0\rangle, r_M > 0, B \in (0,1), tol > 0, C > 1$
$|\delta_0\rangle = \left|\delta^{LM}(\theta_0, r_M)\right\rangle$
if $\left\| |\delta_0\rangle \right\|_2 < B \times r_M$ **then**
  $r_M = C \left\| |\delta_0\rangle \right\|_2$
while $\left| \left\| |\delta_0\rangle \right\|_2 - r_0 \right| < tol$ **do**
  $r_M = C r_M$
  $|\delta_0\rangle = \left|\delta^{LM}(\theta_0, r_M)\right\rangle$
end while
return r_M;

The normal equations for $|\delta_\theta\rangle$ is thus the solution to the reduced least squares problem

$$\min_{|\delta\rangle} \frac{1}{2} \left\| \begin{bmatrix} \mathbf{H}^{1/2}\mathbf{J}_\theta \\ \sqrt{\lambda}\mathbf{I}_p \end{bmatrix} |\delta\rangle + \begin{bmatrix} \mathbf{H}^{-1/2}\Big(\mathbf{H}^{1/2} |\rho_1\rangle + \mathbf{J}_\epsilon\mathbf{E}^{-1}\mathbf{W}_2 |\rho_2\rangle\Big) \\ |0_p\rangle \end{bmatrix} \right\|_2^2$$

where $\mathbf{H} = \big(\mathbf{I}_m - \mathbf{J}_\epsilon^{\mathrm{T}}\mathbf{E}^{-1}\mathbf{J}_\epsilon\big)$ and $|0_p\rangle$ is a vector of zeroes length p. We observe that both $\mathbf{H}^{\frac{1}{2}}$ and $\mathbf{E}$ are diagonal matrices whose elements can be obtained through simple algebra as follows

$$\begin{aligned} e_{jj}^{-1} &= 1/\left[\left(\frac{\partial \langle j \mid \rho_1\rangle}{\partial \epsilon_j}\right)^2 + \langle j \mid w_2\rangle^2 + \lambda\right] \\ h_{jj}^{\frac{1}{2}} &= \sqrt{1 - \left(\frac{\partial \langle j \mid \rho_1\rangle}{\partial \epsilon_j}\right)^2 e_{jj}^{-1}} \end{aligned} \tag{5.3.5}$$

Evidently, increasing λ is equivalent to damping the singular values of $\mathbf{H}^{\frac{1}{2}}$. Due to the diagonal structure of most of the matrices, we never actually invert any matrices and $|\delta_\epsilon\rangle$ is given analytically after obtaining $|\delta_\theta\rangle$

$$|\delta_\epsilon\rangle = -\mathbf{E}^{-1}\Big(\mathbf{J}_\epsilon^{\mathrm{T}}\,|\rho_1\rangle + \mathbf{W}_2^2\,|\rho_2\rangle + \mathbf{J}_\epsilon^{\mathrm{T}}\mathbf{J}_\theta\,|\delta_\theta\rangle\Big) \tag{5.3.6}$$

The appeal of ODR in the second stage of the problem stems from the Mitic GoF test. Because the GoF statistic is obtained by the trapezium rule applied on the m-dimensional vector of normal vectors to the 45° line generated by the fitted cdf against the empirical cdf, if the two weight vectors $|w_1\rangle$ and $|w_2\rangle$ are equal then ODR minimises this orthogonal distance directly.

Because the initial maximum radius r_M is crucial to the efficiency of the steps, we also propose an initialisation algorithm to find the optimal maximum radius detailed in algorithm 3. Due to the special structure of the problem, the order of complexity is not significantly greater than that of the standard LM problem when implemented efficiently. We give a concrete framework for our ODR code in appendix C.4.

Chapter 6

Estimation

In this chapter we present some results using WQE on data. We will provide extensive statistical analysis of WQE on a simulated data set to provide some example background information on the characteristics of the estimates produced by the algorithms. After the extensive and detailed results on this simulated data set, we will provide a brief look at WQE results from some other simulated data sets as well as actual data from different asset classes.

All optimisation and mathematical calculations are done in Python using the standard scientific and data management packages NumPy, SciPy, Scikit, Pandas, Random and Matplotlib. We have also converted the optimisation algorithms as well as statistical procedures and estimators into an R script using the relevant packages such as DPLYR, TLMoments, MASS and GGPlot. Application code for the application for the curve fitting and test procedures are written in R using Shiny. The computing system used is a personal Apple MacbookPro with a 2.9 GHz Intel Core i5 (l5-6360U) (dual core which hyper-threads up to four cores which NumPy exploits) and 8 gigabytes of 2133 MHz LPDDR3 RAM.

The Monte-Carlo (MC) procedure is an aggregation of 6.5 years of losses occurring over the period with yearly frequencies drawn from a Poisson. The frequency being the length of the data set divided by the time span $freq = m/6.5$ which leads to loss occurrence $l \sim Poi\big[freq/(1 - F(t))\big]$ where t is the threshold for tail cutoff, i.e. the estimate for μ in the GPD. MC VaR is measured at the 99.9^{th} percentile of the cumulative losses.

To strictly adhere to feasible analysis, once we get a set of estimators the data set is truncated below $\hat{\mu}$. This gives us more flexibility in terms of feasibility constraints when optimising and if $\hat{\mu}$ is relatively small (say under 1000), then the truncated data set would still be large enough for statistical analysis.

The arbitrary starting vector for all data sets and algorithms is $\mu_0 = 0$, $\sigma_0 = 150$ and $\xi_0 = 1.1$. Unless otherwise stated, optimisation parameters are: $tol = 1\text{e}^{-10}$, $maxIter = 1000$, $r_M = 100$, $\eta = 0.02$. Line search parameter values are $c_1 = 1\text{e}^{-4}$ nd $c_2 = 0.01$.

Default stopping condition is through the step size

$$\frac{\left\| |\theta\rangle_k - |\theta_{k-1}\rangle \right\|_2}{\left\| |\theta_{k-1}\rangle \right\|_2} \leq tol$$

and when appropriate, we also use the gradient's evaluation

$$\left\| |\nabla \varphi_k\rangle \right\|_\infty \leq tol\big(1 + |\varphi_k|\big)$$

and the most lenient condition using the GoF statistic

$$\Omega_k^{Mi} \leq 0.068$$

Our results mainly focus on the statistical properties of the estimators and details on optimisation iterations will be kept light due to the fact that convergence can make up another manuscript on its own as a subject of discussion. We have opted to present results using reasonably optimal parameter values (e.g. number of iterations, stopping condition) even though they should not contribute great differences if one were to attempt a reproduction.

We employ a helpful measure of bias given by the ratio

$$\mathcal{B}(\alpha) \triangleq \frac{x(\alpha)}{x_n(\alpha)}$$

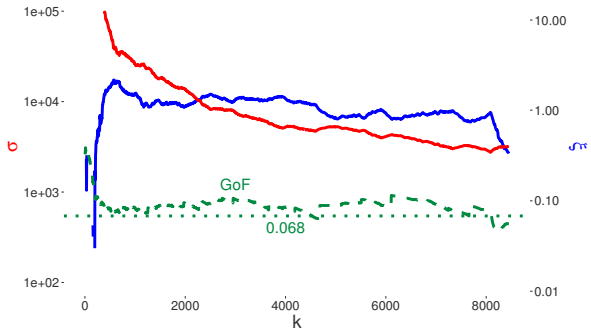
where x is the fitted quantile function, x_n is the empirical quantile function for $\alpha \in (0, 1)$. We work with two values for α at 0.95 and 0.999 to "anchor" the tail fitting process by aiming for $\mathcal{B} = 1$. ODR weights are thus calibrated using optimised parameters which yield the target bias of 1.

6.1 Baseline data set

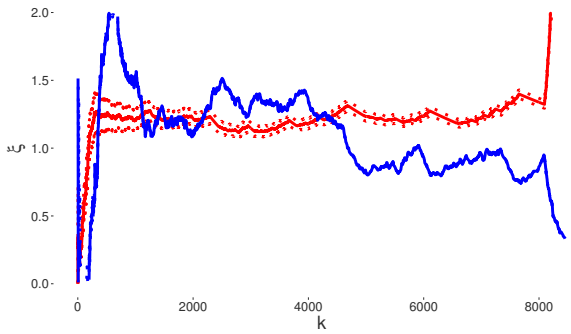
Recall the simulated log-normal data set introduced in 4.2 with mean 15,697, standard deviation 74,252, 95-percentile of 39,899 and 99.9-percentile of 821,045, all measured in GBP. The smallest value if 0.04 and largest value is 2,230,329. We will refer to this data set as the baseline data set. The true data generating process (dgp) is $\log X \sim \mathcal{N}(8, 1.96)$ with the 95 and of 99.9-percentile at 37,617 and 245,821 respectively. The empirical distribution's tail bias is 1.22 and 3.34 (as percentage of actual quantiles) respectively therefore the tail is significantly underestimated by the empirical quantiles.

The log-normal ML fit returns the values $\hat{\mu}_{ML} = 8.094$ and $\hat{\sigma}_{ML} = 1.953$, which actually fails the Mitic test significance level of 0.0882 but provides an accurate estimate for the tail, where the 95 and 99$^{\text{th}}$-percentiles are 81,333 and 1,368,270 respectively. MC calculation with 10 blocks of 10,000 trials each returns the cumulative 6.5-year 99.9% yearly VaR of 14m GBP. The mean-squared error (MSE) of the fit is reasonable at 1.8e^{-04}.

The Hill (3.0.7) function $\hat{\xi}_{Hill}(k)$ plateaus out at 1.25 for most cutoffs whereas $\hat{\xi}_{Pick}(k)$



(a) Pickands $\hat{\theta}$ with $\hat{\mu} = X^{(k)}$ and m = size of tail (GoF stat. shares right y-axis with ξ)



(b) Hill and Pickands tail estimators' 95% asymptotic normal confidence intervals

(3.0.8) maintains the 1.5 level until k approaches the body half of the data set as shown in figure 6.1b. The estimate for ξ implies infinite mean and variance leading to complex values for μ and σ, therefore no MM estimators exist. From figure 6.1a, one observes that Pickands' quantile estimates fail the GoF test at every cutoff point except the very first few observations. The results are informative about the range of the tail index but do not reveal much about σ or μ. The passing Pickands fits correspond to $\hat{\xi}$ values that are closer to 0.5 which we show using ODR is very close to target value which yields an empirically consistent MC VaR value.

We also observe that the asymptotic normality for the Pickands estimators yields a much narrower 95% confidence interval for the tail index at all cutoff points compared to the Hill estimator. Since the sample tail quantiles overestimate the true tail by over two-folds, one would need a large number of additional data points for the law of large numbers to deflate the tail appropriately using the Hill or Pickands tails.

6.1.a Fit A

We begin with fit A using m equal to a quarter of the number of observations, i.e $m = \lfloor n/4 \rfloor$. Most statistics, such as mean and variance, from both the GN and LM A fits match empirical counterparts with high accuracy except for the last tail quantiles. This is a major applicability problem for the GPD regarding loss data we outlines in chapters 2 and 3, where the tail's accuracy is usually neglected to achieve a good *overall* fit. Because of the large condition number at the tail, we get tail quantile values that can be up to an order of magnitude greater than their sample counterparts.

						95-pct		99.9-pct		
Algo.	tail thres.	GoF stat.	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	sam. bias	pop. bias	sam. bias	pop. bias	MC VaR
GN	0	0.067	-0.14	2196.39	1	1.05	1.4	2.69	9.79	2,742m
LM	0	0.068	0.04	2146.05	1.02	1.07	1.43	2.97	10.8	3,325m

Table 6.1: stage 1 WQE GPD A fits of the baseline data set

Prima facie from table 6.1, the sample lower quantiles fit their empirical counterparts reasonably well up to the 95-pct. The upper quantiles are significantly more biased, as

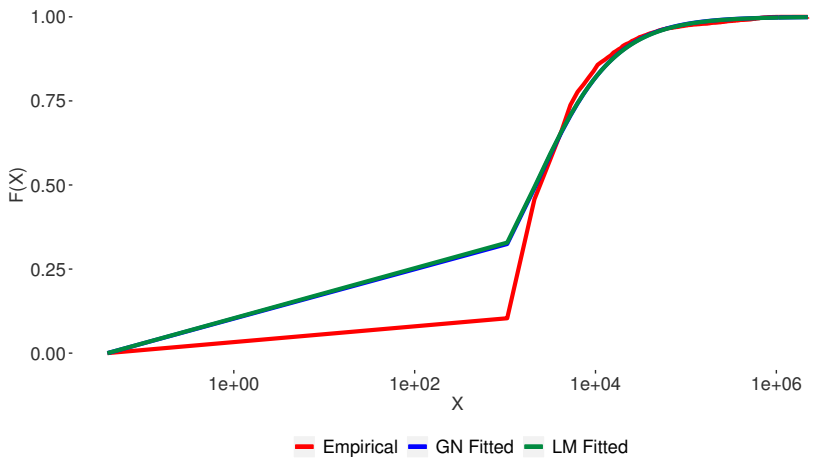
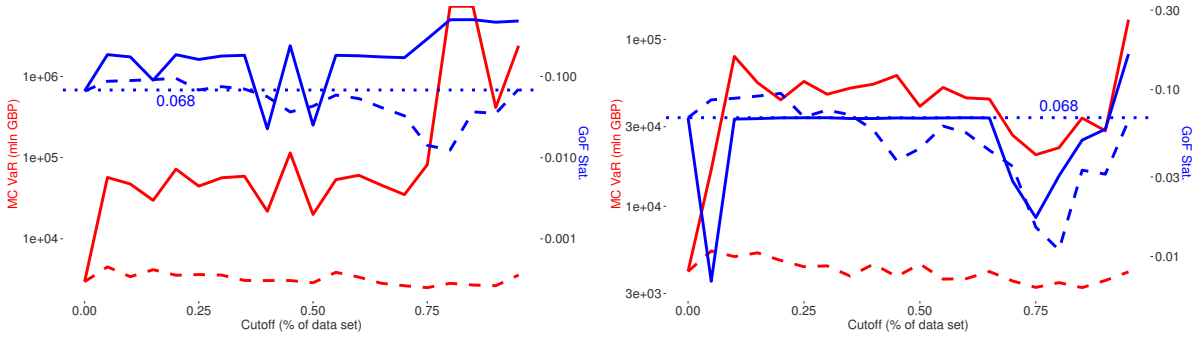


Figure 6.2: baseline data set stage 1 WQE A fits

they severely overestimate the sample tail three times and the population by ten to evelen times. To test the tail stability we check the algorithm's efficiency when applied to the sequentially truncated data set against the theoretical set of estimators. The hypothetical A fits are derived from the first (non-truncated) fit on the entire data set and the truncated as we compare its prediction using theorem 2.2.2 and the fits on the truncated data sets with normalised VaR levels [1] .

Figure 6.2 show that the truncated A fits and the theoretical A fits are very close to each other, which also implies that the MC VaR values are very high at all cutoff points except the last few. Also evident in figure 6.3a is the weakness of the GN algorithm as it struggles to produce estimators throughout the tail with fewer observations and variation in F (implying small singular values for $\mathbf{J}$).



(a) GN

(b) LM

Figure 6.3: truncated baseline data set stage 1 fit A MC VaR and GoF statistics with (dashed line is the predicted value based on the initial fit only, solid line is the value re-fitted at every cutoff)

[1] The renormalised quantile q with the data set truncated at $t\%$ is $q_t = x_n\big[1 - q/(1-t)\big]$.

The overall biases of the estimators produced by stage 1 algorithms fluctuate significantly with higher tail thresholds, indicating instability in the estimates regardless of the size of the data set or clustering of observations. The LM algorithm is able to produce passing fits for much of the cutoffs, whereas GN struggles to converge with the exception of the theoretical initial fit's prediction.

The residuals from the two fits have a vanishing trend towards the end of the tail. This trend suggests that the GPD is a progressively appropriate distribution for tail values and one can get good estimates of the tail by inflating tail residuals' gradients for ODR to attempt to minimise tail residuals first. We denote a weight vector containing logarithmically increasing weights with the final weight being v times the size of the first weight as $|\omega_v\rangle$. The results of $|\omega_v\rangle$ for several values for v are given in table 6.2.

Due to the monotonically increasing nature of the weight vector, truncating the body is equivalent to the relation [2]

$$\lim_{v\to\infty} \sup_{1\leq t<v} \frac{\langle t \mid \omega_v \odot \rho_1 \rangle}{\langle v \mid \omega_v \odot \rho_1 \rangle} = 0$$

where $|\rho_1\rangle$ is defined in (5.3.4). Thus finding the optimal weight vector is equivalent to finding the tail portion where the GPD fits the sample points well. There is always the risk of underestimating the tail if the weights cause the fitted quantiles to diverge far away from the theoretical implied tail by the initial fit, therefore heuristics and sound judgement must be made regarding the tolerance for deviation.

						95-pct		99.9-pct		
$\lvert w_1\rangle$	tail thres.	GoF stat.	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	sam. bias	pop. bias	sam. bias	pop. bias	MC VaR
$\lvert\omega_{10}\rangle$	0	0.068	0.04	2177.26	0.942	0.915	1.225	1.88	6.85	1,329m
$\lvert\omega_{50}\rangle$	0	0.054	0.042	2573.02	0.79	0.79	1.06	0.93	3.37	205m
$\lvert\omega_{300}\rangle$	0	0.05	0.035	2762.12	0.7	0.71	0.95	0.62	2.23	85m
$\lvert\omega_{1000}\rangle$	0	0.04	0.04	2773.09	0.7	0.7	0.93	0.58	2.11	80m

Table 6.2: stage 2 WQE GPD fit A of the baseline data set with $|w_2\rangle = |1_m\rangle$

By underestimating the 95-percentile to match the 99.9-percentile, we get a VaR value that is within an empirically consistent value and does not underestimate the sample tail quantiles too severely (In this case an arbitrary point of consideration is half the sample tail quantiles). The lower bound on the sample bias reduction appears to be around 0.58 since all weights greater than 1000 do not yield further reductions. Because for most data sets we would not have access to the population quantiles it is more relevant to target the sample quantiles. A series of plots for table 6.2 is given in appendix D.1.

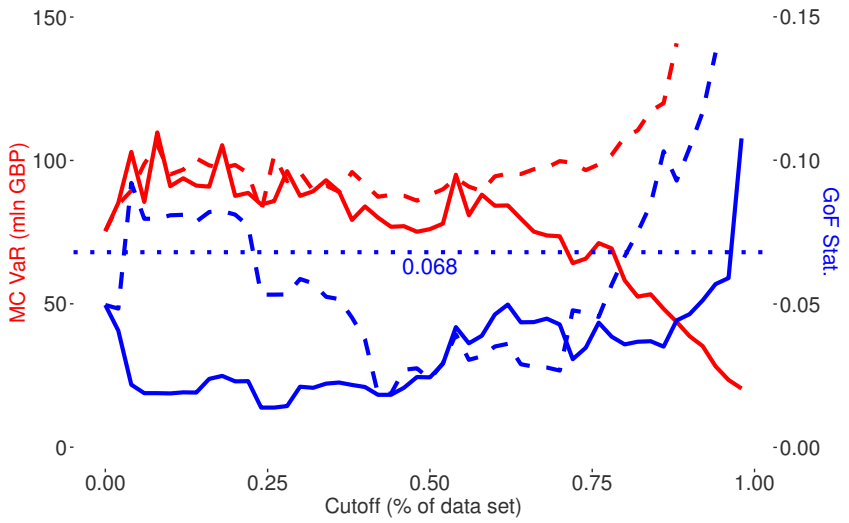
All the estimated thresholds $\hat{\mu}$ fitted are near 0, with the differences between the weighted fits lie in the ξ and σ trade-offs. The clear trend for fit A is that underweighting the sample tail quantiles involves reducing ξ while increasing σ appropriately to rescale the distribution's shape.

The $|\hat{\epsilon}\rangle$ vector contains values that are mainly close to 0. The implication of this result

[2] $\odot : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^m$ is the Hadamard element-wise vector multiplication introduced in C.4.2.

is that the predicted noise level in the measurements of the observations is low. Because the data set is synthetically generated this observation is expected. An important consequence from our expectation of the $|\hat{\epsilon}\rangle$ vector's elements is that if there are elements that are significantly different from 0, they need to be inspected for potential statistical manipulation which may lead to removal from the data set as an outlier.

For the truncated tail, we use the weight vector that produces the closest approximation to the sample tail quantiles $|\omega_{1000}\rangle$. The truncated fits match the theoretical fits up until the final 40% tail, where the re-fitted estimates begin to converge to the sample log-normal MC VaR whereas the initial theoretical fit's MC VaR begins to increase exponentially. This region of divergence is a possible cut-off point for the tail.



(a) GoF statistic

Figure 6.4: truncated baseline data set stage 2 fit A MC VaR and GoF statistics with $|w_1\rangle = |\omega_{1000}\rangle$ (dashed line is the predicted value based on the initial fit only, solid line is the value re-fitted at every cutoff)

From figure 6.4, we observe that the MC VaR values are significantly closer to the value implied by the log-normal fit for all tail cutoff values compared to stage 1 fits. All fits up until the final 4% of the tail pass the GoF test and the divergence between the theoretical tail and the fitted tail is approximately in line with the fitted tail's Gof statistics. The theoretical initial fit is a good approximation for the body but fails to capture the true tail, thus re-fitting at different cut-off points was standard procedure for our tests.

6.1.b Fit B

Switching over to fit B, the body of the distribution plays a much larger role. The optimisation algorithms never reach the level of flexibility that the A fits achieved, mostly because the very small values in the body have a disproportionate effect on the overall fit. The GN algorithm fails to produce a fit which passes the GoF test whereas the LM algorithm immediately yields parameters which corresponding to highly accurate MC VaR

values.

Comparing the initial LM fit for fit B and the weighted stage 2 fit A with increasing weights, we see a continuation of the trend characterised by inflating σ and deflating ξ to get better approximations for MC VaR. Even though the upper tail bias indicates a faily high underestimation of the sample tail quantiles, from table 6.4 we see that the fitted tail still overestimates the true dgp's tail quantiles hence the higher MC VaR level.

						95-pct		99.9-pct		
Algo.	tail thres.	GoF stat.	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	sam. bias	pop. bias	sam. bias	pop. bias	MC VaR
GN	*did not converge*									
LM	0	0.056	0.53	3205.59	0.567	0.63	0.85	0.32	1.236	25m

Table 6.3: stage 1 WQE GPD B fits of the baseline data set

The process of truncating the body for the B fits is more complex compared to A fits. Because F_n in fit B is not as smooth as the A fits, the importance of the initial fit from which the theoretical fit B can be used to compare against truncated fits is greatly magnified. If an initial fit is bad, the tail of the fit is almost certainly going to be a bad fit of the true tail and any MC simulation using the parameters is going to yield highly implausible results. Figures 6.5 reveal that the GN algorithm struggles with the tail again.

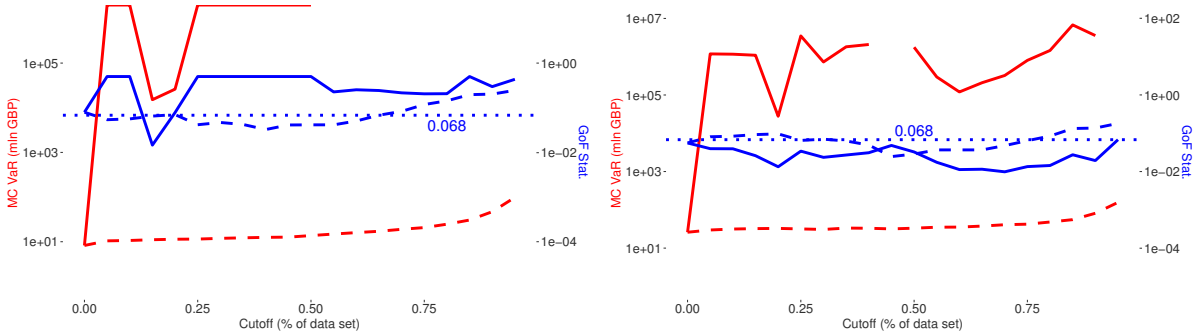


Figure 6.5: truncated baseline data set stage 1 B fits MC VaR and GoF statistics

The flatness of the tail hinders the crude GN algorithm greatly and it fails to converge for any cutoff points past 20% of the data set. LM performs relatively well, converging at all cutoff points with a stable albeit very high level of MC VaR throughout. The initial fit's MC VaR maintains its credibility along with passing GoF statistics until the final 25% tail. Recall that this 25% tail contains both the end quantiles as well as a large portion of the body due to the equidistant intervals in the cdf space, thus implying that there are significantly fewer tail quantiles available for the algorithm to fit. We observe the corresponding result as MC VaR increases exponentially along with GoF statistics.

The selected weights for tail truncation are $|\omega_8\rangle$. The fitted tails for most cutoff values have low Gof statistics, most are below 0.06, but the implied MC VaR values are massive at trillions of GBP. The theoretical tails give more reasonable MC VaR values with passing

GoF statistics for the majority of the body until the final quartile. The trend appears to be such that the final quartile poses significant challenges for fit B. The variability in the empirical quantiles biases the upper quantiles up so significantly that the weights do not have much correcting effect.

						95-pct		99.9-pct		
$\|w_v\rangle$	tail thres.	GoF stat.	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	sam. bias	pop. bias	sam. bias	pop. bias	MC VaR
$\|\omega_5\rangle$	0	0.053	0.3	3,289.01	0.452	0.524	0.701	0.192	0.7	12m
$\|\omega_8\rangle$	0	0.058	0.38	3,064.73	0.547	0.582	0.779	0.291	1.06	22m
$\|\omega_{10}\rangle$	0	0.056	0.43	2,967.56	0.587	0.609	0.0.815	0.349	1.271	27m
$\|\omega_{20}\rangle$	0	0.061	0.663	2,699.1	0.697	0.69	0.92	0.577	2.1	78m

Table 6.4: stage 2 WQE GPD B fits of the baseline data set with $|w_2\rangle = |1_m\rangle$

We also tested fit B with the weight vector $|\nu_\gamma\rangle$ which consists of the inverse of the squared-residuals scaled by an appropriate constant $\gamma \in (0, 1)$ as first described in (4.1.5). The ratio between the largest and smallest elements would be extremely high if the residual converges to 0 a the tail as is the case for the B fit.

The weight vector we tested was $|\nu_{1e^{-4}}\rangle$, which returns fitted MC VaR values from 10 to 20 mln GBP for the majority of cut off points and theoretical MC VaR values of 20 to 50 mln GBP until the final quartile. The biases introduced by the fitted parameters are low, at about 0.1 to 0.3 of the true tail quantiles which explains the low MC VaR values. Other values of γ also yield similar tail results therefore we have not reported them here due to the rigidity of the weights.

The weight vector $|\nu_\gamma\rangle$ can have a very steep gradient, given that the end result of the double-squaring of the reciprocal residuals is that the inverses of small residuals at the tail end are raised to the power of 4. This procedure ensures that tail observations are accounted for with greater importance but also reduces the flexibility of the modelling process. We observe these results in figure 6.6 where changing the scaling parameter yields very little changes to the figure plotted.

6.1.c Tail Estimation

We now proceed to estimate the tail of the data set. As we have demonstrated, the two methods to obtain the empirical cdf produce two sets of quantiles that are significantly different from each other. The differences disappear at the tail but because in stage 1 the quantiles are unweighted outliers in the body affect the B fits much more than A fits, in which they are "subsumed" into their corresponding bins.

Neither fit A nor B is "correct" as we demonstrate with other data sets, therefore we reduce the problem down to a credible range of parameters obtained from both fits. Because the A fits' empirical cdfs are smoother than the B fit's, optimisation algorithms do not struggle as much in the body and can achieve the best possible fits. Both methods have their advantages and drawbacks and our analysis has shown that, depending on the quantities of interest, there are compelling reasons to perform both procedures.

Suppose we are interested interested in estimated a feasible MC 99.9^{nth} yearly VaR range, using the estimated parameter sets for both fit types, we pick reasonable weight

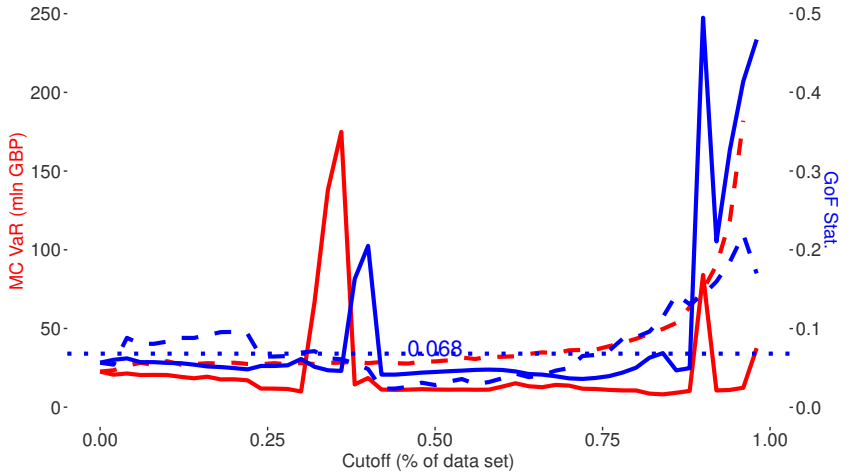


Figure 6.6: truncated baseline data set stage 2 B fits MC VaR and GoF statistics and $|w_1\rangle = |\nu_{1e^{-4}}\rangle$

ranges to narrow down to four anchor values. Denoting L_a and L_b as the lower bounds for A fits and B fits respectively, and the same idea for U_a and U_b for the upper bounds, we get table 6.5.

						95-pct		99.9-pct		
	tail thres.	GoF stat.	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	sam. bias	pop. bias	sam. bias	pop. bias	MC VaR
$L_a : \|\omega_{1000}\rangle$	0	0.04	0.04	2773.09	0.7	0.7	0.93	0.58	2.11	80m
$L_b : \|\omega_8\rangle$	0	0.058	0.38	3,064.73	0.547	0.582	0.779	0.291	1.06	22m
$U_a : \|\omega_{300}\rangle$	0	0.05	0.035	2762.12	0.7	0.71	0.95	0.62	2.23	85m
$U_b : \|\omega_{20}\rangle$	0	0.061	0.663	2,699.1	0.697	0.69	0.92	0.577	2.1	78m

Table 6.5: feasible range of parameter values for the baseline data set (using 14m $\leq$ MC VaR $\leq$ 90m GBP)

Recall figures 4.3 and 4.4. The feasible ranges for the parameters all lie inside the credible region A, up to μ which has virtually no effect on the curves' GoF statistics. We have demonstrated a numerical approach to validate the ad-hoc brute force approach of visually sectioning up the 3D-plots. To get a "reasonable" value for MC VaR, we have to underestimate the sample quantiles.

To convert the result into a conventional tail distribution, we recommend that one pick an MSE value as tolerance for the tail error and move down the body progressively from the largest observation until the MSE is greater than the tolerance level. This procedure is possible due to the fact that the residuals vanish towards the tail. One way to narrow down the appropriate tail level is to use the mean excess plot defined as the expectation with respect to the sub-σ-algebra generated by the restricted sample space, or conditional expectation, against the empirical counterpart. A similar function has been studied in literature [37], under the name *condition value-at-risk* (CVaR), due to its convexity wrt the threshold.

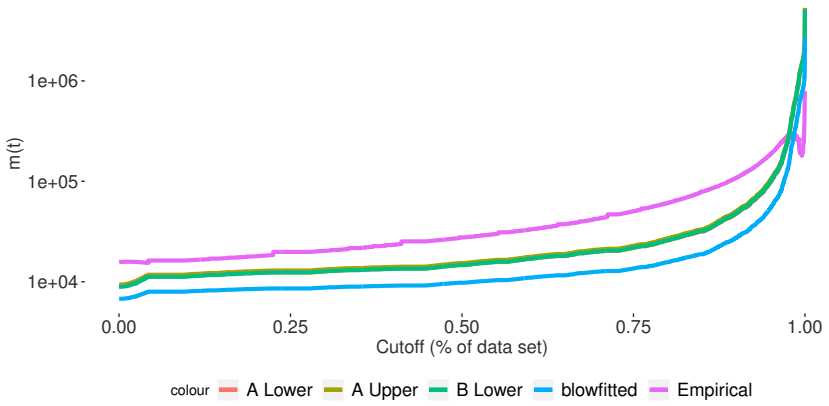


Figure 6.7: mean excess plots

For a given threshold $t \geq \mu$, let $\mathcal{T}$ denote the set $\{X \mid X > t\}$, we get the following theoretical and empirical mean excess functions

$$m(t) \triangleq \mathbb{E}[X - t \mid X > t] = \int_{\mathcal{T}} \frac{x - t}{1 - F(t)} \, \mathrm{d}F(x) = \frac{\sigma + \xi(t - \mu)}{1 - \xi}$$

$$\hat{m}(t) \triangleq \frac{\sum_{X \in \mathcal{T}} X - t}{\sum_{X \in \mathcal{T}} 1}$$

For all the narrowed-down fits, the cut-off point at which the mean excesses start to diverge from the a linear trend is around the 75-pct. This indicates the point at which tail observations dominate the data set's values, thus acts as an initial estimate for the tail threshold. Since the mean is not defined for $\xi \geq 1$, we cannot use this method for a number of fits that have thicker tails, for example the stage 1 A fits.

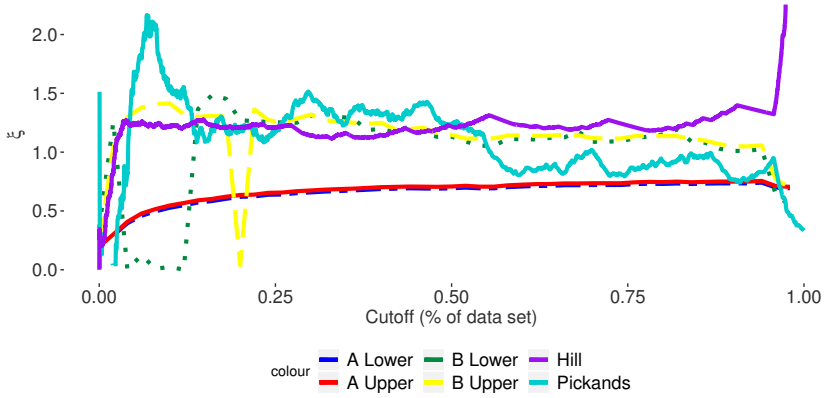


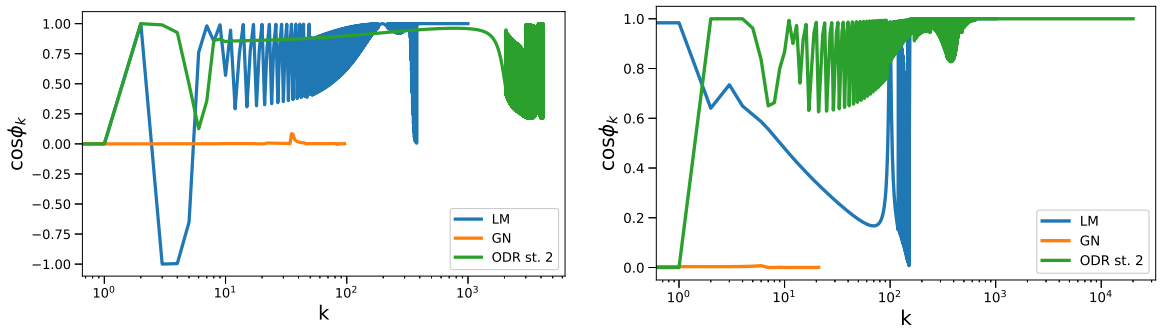
Figure 6.8: estimated $\hat{\xi}_{ODR}$ range and $\hat{\xi}_{Hill}$ and $\hat{\xi}_{Pick}$

Figure 6.8 shows that the B fits' estimated $\hat{\xi}$ are much more in-line with the Hill and

Pickands' estimates, slowly converging to the 1.25 level. This result is expected because the quantiles obtained from the ecdf fitting process are directly sampled from the data set and not interpolated like the A fits. We also see thetrend for the A fits' estimates to trend downwards as the data set is truncated to maintain the tail bias with an increasing $\hat{\sigma}$. We note the uniformity of the decrease in fit A's ξ as a way to visualise the aforementioned stability of the fit.

6.1.d Convergence

We now aim to provide a brief overview of the convergence of the algorithms. At all cut-off points the GN algorithm produces directions that are nearly orthogonal to $|\nabla\varphi\rangle$ whereas the LM algorithm gains ground to converge to $|\nabla\varphi\rangle$ after usual fluctuations in the beginning. Orthogonality implies that $\cos\phi$ is 0 and we show these results in figures 6.9. LM does not require that $\cos\phi$ be bounded away from $\pi/2$ to converge but it is helpful to differentiate the convergence trajectories to see the drawbacks of the GN algorithm.



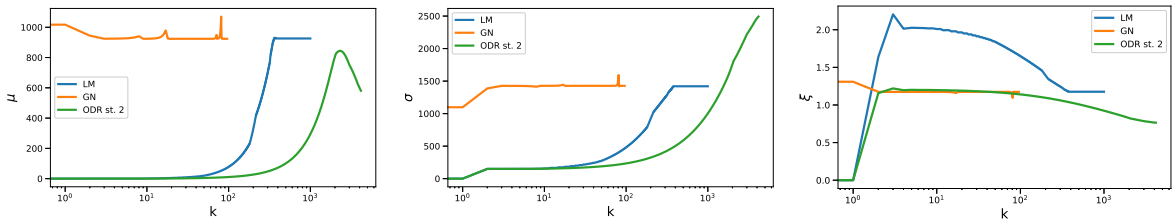
(a) fit A

(b) fit B

Figure 6.9: selected baseline data set's $\cos\phi_k$

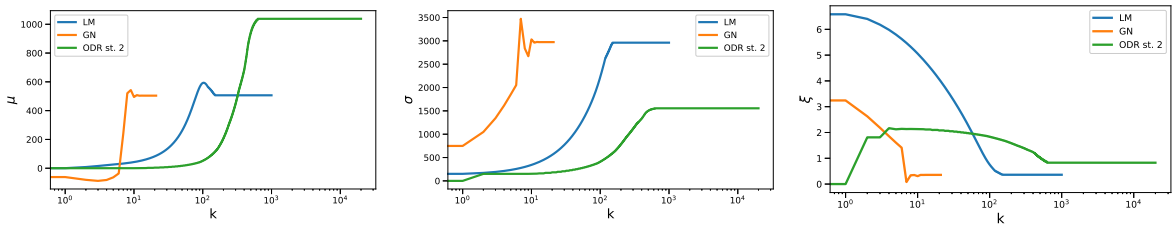
The most interesting observations from the convergence trajectories for stage 1 and stage 2 methods are the convergence paths for the parameter values. We picked an A and an B fit that yield similar tail quantiles to compare the trajectories. ODR convergence is very similar to LM, because they use the same underlying algorithm, but the weights significantly skew the optimal values. The result is very similar to what one would get with a Tikhonov [40] regularisation procedure in a linear model. The inverse relationship between σ and ξ can also be observed easily with σ tending upwards as $k \to \infty$ and ξ otherwise.

The shape parameter ξ characterises the tail just like in a Lévy α-stable distribution therefore its important role in the parameterisation commands more attention. We observe that weight penalisation applies downward pressure on ξ progressively to zero in on the true tail quantiles as expected but the different bias directions of the two quantile selection methods do create some uncertainty around the optimal weights. Overall, we find that optimal weights usually vary from around $v = 30$ to 100 for A fits and γ has a much larger range, from $1e^{-6}$ to 20 for B fits. These numbers are the result of balancing 95 and 99^{th} percentiles for the fits with much more extreme weights hindering fast convergence.



(a) μ (b) σ (c) ξ

Figure 6.10: baseline data set's A fits' parameter values with $|\omega_{80}\rangle$



(a) μ (b) σ (c) ξ

Figure 6.11: baseline data set's B fits' parameter values with $|\nu_{1e^{-5}}\rangle$

6.2 Other data sets

Among common non-negative distributions for the tail are those given by the Fisher-Tippett-Gnedenko theorem 2.1.1, we tested the WQE procedure on several generated data sets with true dgps following Frechét, Weibull as well as Pareto distributions and recorded their tail biases. Data sets with higher variances produce predictable more volatility in the tail estimates. fit B parameters are almost always significantly less mutable with several fits failing to converge at the tail despite acceptable body fits.

The tail thresholds, measured by $\hat{\mu}$'s quantile, are uniformly consistent with the minima of the respective data sets. Underestimating the lower tail quantiles almost always translates to overestimating higher tail quantiles, which is consistent with the baseline data set. The general implication is that for a more accurate tail estimate, a higher up the tail threshold should be considered. Since lower losses are not of great importance, the lack of a good body fit is not disqualifying of a set of estimates.

An important set of results tested is the Pareto tail distribution given in table 6.8. We observe that the A fits attempt to match the shape using the ξ parameter to compensate for the low σ whereas the B fits' ODR fit is able to recover the general shape of the distribution with a highly accurate tail index. This result highlights the B fits' accurate representation of the quantiles although both fits are unbiased at the tail after the stage 2 corrections. The tail results are particularly encouraging because we have the true population parameter values thus numerically proving the consistency of the estimators.

We observe that the algorithms struggle to find parameters using the B fits for some data sets such as the Weibull data set in table 6.6, which has a very dense body and

few extreme tail values. Some data sets are generally more challenging than others, such as the log-logistic simulated data set. This data set is a pathological case as the tail is never accurately recovered by stage 2 in either fit. The large distance between the tail observations and body creates a severity in spareness not observed in other data sets but ODR is still able to produce a range of estimates for the tail index.

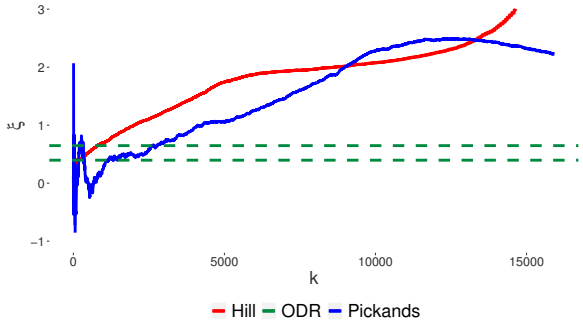
As observed throughout the manuscript, neither one of fit A nor fit B procedure yields accurate results for all dgps. If, however, the practitioner fits utilises both procedures then they are almost guaranteed *some* useful consensus information on the tail distribution. A difficult data set generated by a $Pareto\big(x^* = 150, \alpha = 1.93\big)$ has a very dense body and very few but extreme tail values proves somewhat difficult to optimise over at first but is ODR is still able to recover the parameters with extremely high accuracy. The density function itself almost resembles the Dirac delta mass, which poses significant challenges for all algorithms.

We also tested WQE on actual data ranging from swap rates, oil prices to cryptocurrencies. The data come from the St Louis Federal Reserve's Bureau of Economic Data (FRED) [16–19] and Kenneth French's asset returns database [20] as well as the cryptocurrency database CoinDesk [9]. Most tail estimates are below 0.3 thus implying finite mean and variance for all fits. The results show that not only can the strong law of large numbers be applied to these asset classes but also the central limit theorem, which may be unexpected for more volatile asset classes.

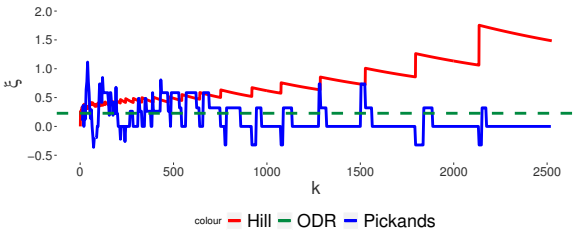
All tail estimates tested can be corrected in the second stage with logarithmic weights yielding superior results. Unlike the synthetically generated samples, both tail biases for all data sets can be made to converge reasonably close to 1 with the appropriate weight vectors. Large losses are likely but "black swan" events are still unlikely to occur if $\hat{\xi}$ is less than 1. We observe this property with all tested asset classes except for Bitcoin.

MC VaR is calculated as cumulative losses throughout a year-long period, without being offset by potential profit. The Fama-French momentum factor portfolio introduced by Carhart [6], known for its sensitivity to non-market forces (e.g. government intervention), performs particularly bad with all fits pointing to 850 to 930 percentage point yearly cumulative losses. The tail estimates also show thin tails, as such losses are mostly contained and concentrated indicating a systematic behaviour in the loss generating process. WQE therefore gives the investor a comprehensive picture of actual loss prospects of volatile portfolios as opposed to parametric methods with symmetric distributions both both profits and losses as well as retaining the majority of data points in most cases.

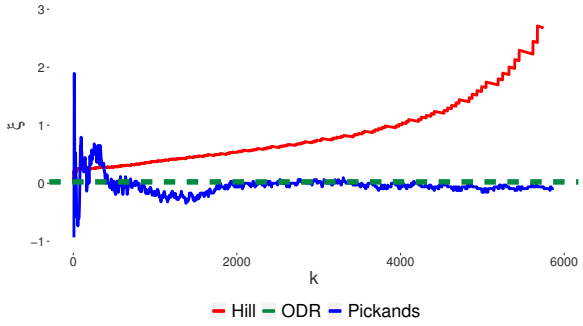
In addition, we have also included the median estimates for $\hat{\xi}_{Pickands}$ and $\hat{\xi}_{Hill}$ to frame the deflating effect of stage 2 fits and the tail threshold that is implied by the WQE estimate for ξ under the $\xi_{Hill}(k)$ and $\xi_{Pick}(k)$. For the majority of the data sets, we observe that correcting tail biases corresponds to moving up the tails as seen in the baseline data set. The Hill function produces systematically higher tail estimates for all cutoff points compared to the Pickands function, leading to very high thresholds for stage 2 estimates whereas Pickands tail estimates imply that the tail is perhaps lower down the body.



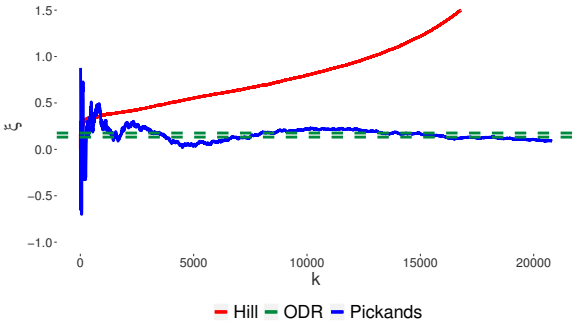
(a) Bitcoin hourly



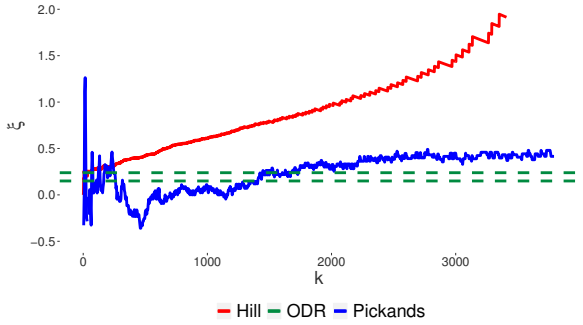
(b) BAML high-yield corp. bonds daily



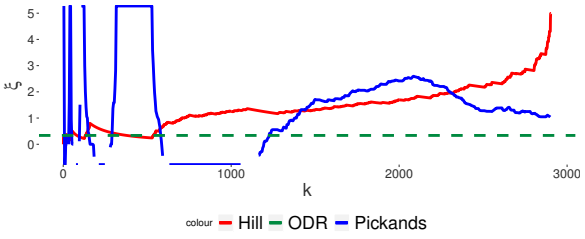
(c) GBP/USD daily



(d) Fama-French-Carhart momentum factor portfolio daily



(e) Brent crude daily



(f) Libor daily

Figure 6.12: comparison of between the WQE estimates vs Hill and Pickands estimates for $\hat{\xi}$ on loss data

The volatility in the Hill and Pickands functions' outputs also serve to highlight the stability of the strategy of anchoring the tail estimates through the stage 2 estimates. Their median estimates also tend to overestimate the tail quantiles significantly which implies that if a practitioner were to use just these estimators a significant number of observations would have to be discarded. An example of this phenomenon is given in figure 6.12(a) of the hourly USD losses of Bitcoin over the period 2010 to 2019. using the whole data set, we see that the estimates for ξ using GN fit A and B are 0.396 and 0.647 respectively. The second estimate implies infinite variance and a slight overestimate of the true tail with a 99-pct bias of 1.2 but is still within an acceptable range. From the

plot we see the behaviour with very high estimates for ξ using both the Hill and Pickands functions. ODR corrects the bias fairly effortlessly after some weight trials whereas to get a good fit using the Pickands or Hill functions one would need to truncate most of the data set.

Overall the A fits are more mutable, whereas B fits converge to true parameter values quicker. None of the data sets proved wholly impossible for WQE to provide estimates, even the most pathological 1-parameter Pareto case detailed in table 6.8. There is a non-trivial step to determine the optimal ecdf is A or B to pre-condition the optimisation algorithms but overall we see a lot of informational benefits with employing both fits. For example, if the data set is sparse and dense around a small area then the A fits would provide a set of quantiles that facilitate convergence. On the other hand, the B fits would suit data sets where there is a significant presence of tail values such that accurate quantiles are preferable.

	A fits		$\xi_{Hill} = 3.906$		$\xi_{Pick} = 2.84$	95-pct		99.9-pct	
Algo.	GoF stat.	tail thres.	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	sam. bias	pop. bias	sam. bias	pop. bias
GN St. 1	0.014	0	-0.435	908.64	1.46	0.749	0.679	3.226	3.239
LM St. 1	0.001	0	-0.015	553.337	1.608	0.642	0.583	4.842	4.862
ODR (v=13)	0.034	0.048	0	260.5	1.486	0.226	0.205	1.064	1.068
	B fits								
Algo.	GoF stat.	tail thres.	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	sam. bias	pop. bias	sam. bias	pop. bias
GN St. 1	*did not converge*								
LM St. 1	*did not converge*								
ODR (100)	*did not converge*								

Table 6.6: $Weibull\big(scale = 300, \alpha = 0.2\big)$ $(n = 8456, m = n/4)$

	A fits		$\xi_{Hill} = 1.275$		$\xi_{Pick} = 1.077$	95-pct		99.9-pct	
Algo.	GoF stat.	tail thres.	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	sam. bias	pop. bias	sam. bias	pop. bias
GN St. 1	0.026	0	0	60	1.098	1.025	1.041	1.436	1
LM St. 1	0.026	0	3.4	137.06	0.662	0.945	0.956	0.265	0.184
ODR (v=55)	0.044	0	3.035	126.27	0.923	1.489	1.512	1.087	0.757
	B fits								
Algo.	GoF stat.	tail thres.	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	sam. bias	pop. bias	sam. bias	pop. bias
GN St. 1	*did not converge*								
LM St. 1	0.023	0	4.13	86.49	0.678	0.617	1.512	0.183	0.757
ODR (100)	0.031	0	0	67	1.019	0.966	0.981	0.997	0.695

Table 6.7: $Frechét\big(scale = 50, \alpha = 0.9\big)$ $(n = 8456, m = n/4)$

	A fits		$\xi_{Hill} = 2.112$	$\xi_{Pick} = 1.929$		95-pct		99.9-pct	
Algo.	GoF stat.	tail thres.	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	sam. bias	pop. bias	sam. bias	pop. bias
GN St. 1	*did not converge*								
LM St. 1	0.068	0	1,246.33	1,448.48	2.557	0.203	0.179	1.907	2.079
ODR (v= 2.5)	0.067	0	1,254.03	1,452.92	2.46	0.157	0.139	0.996	1.086
	B fits								
Algo.	GoF stat.	tail thres.	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	sam. bias	pop. bias	sam. bias	pop. bias
GN St. 1	*did not converge*								
LM St. 1	0.068	0	1,475.078	13,034.94	3.63	32.29	28.489	19,976.95	22,180.777
ODR (v = 5)	0.005	0	1,506.715	35,650.42	1.969	1.114	1.05	0.983	1.146

Table 6.8: $GPD\big(\mu = 1500, \sigma = 40000, \xi = 1.93\big)$ $(n = 8456, m = n/4)$

	A fits		$\xi_{Hill} = 0.019$	$\xi_{Pick} = 0.552$		95-pct		99.9-pct	
Algo.	GoF stat.	tail thres.	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	sam. bias	pop. bias	sam. bias	pop. bias
GN St. 1	*did not converge*								
LM St. 1	0.002	0	145	0.98	0.5	1	1	1.019	0.968
ODR (v = 50)	0.001	0	150	0.972	0.546	1	1.001	1.039	1.039
	B fits								
Algo.	GoF stat.	tail thres.	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	sam. bias	pop. bias	sam. bias	pop. bias
GN St. 1	*did not converge*								
LM St. 1	0.001	0	150	0.96	0.562	1	1.002	1.032	1.064
ODR (v = 50)	0	0	149.993	0.978	0.541	1	1	1.036	0.985

Table 6.9: $Pareto\big(x^* = \mu = 150, \alpha = 1.93 \text{ or } \xi = 0.518\big)$ $(n = 8456, m = n/4)$

	A fits		$\xi_{Hill} = 2.71$	$\xi_{Pick} = 2.36$		95-pct		99.9-pct	
Algo.	GoF stat.	tail thres.	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	sam. bias	pop. bias	sam. bias	pop. bias
GN St. 1	*did not converge*								
LM St. 1	0.068	0	12.15	56,154.3	0.1	1.28	1.236	2.14	2.1
ODR ($v = 2.8$)	0.068	0	12.153	57,604	0	1.14	1.1	1.54	1.515
	B fits								
Algo.	GoF stat.	tail thres.	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	sam. bias	pop. bias	sam. bias	pop. bias
GN St. 1	*did not converge*								
LM St. 1	0.062	0	52	55,074	0.346	1.91	1.85	6.067	6.007
ODR (300)	0.067	0	52	57,773	0	1.14	1.1	1.54	1.52

Table 6.10: $Half - \mathcal{N}\big(\sigma = 80,000\big)$ $(n = 8456, m = n/4)$

	A fits		$\xi_{Hill} = 2.79$	$\xi_{Pick} = 2.26$		95-pct		99.9-pct	
Algo.	GoF stat.	tail thres.	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	sam. bias	pop. bias	sam. bias	pop. bias
GN St. 1	0	0	-0.385	3,305.53	1.969	1.789	1.127	0.892	0.905
LM St. 1	0.001	0	-0.018	150.009	2.43	0.172	0.165	0.863	0.798
ODR ($v = 0.8$)	0.001	0	-0.018	150.01	0.18	0.172	0.95	0.348	0.89
	B fits								
Algo.	GoF stat.	tail thres.	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	sam. bias	pop. bias	sam. bias	pop. bias
GN St. 1	0.018	0	-43.998	523.28	33.2	4.625	4.42	476.54	440.807
LM St. 1	0.018	0	-43.87	522.53	3.2	4.64	4.4.44	481.551	445.445
ODR (v = 1000)	0.068	0	-1,100.77	1,641.386	2.204	1.057	1.01	2.205	2.038

Table 6.11: $Log - Logistic\big(\alpha \text{ (scale)} = 1500, \beta \text{ (shape)} = 0.5\big)\text{(n = 8456, m = n/4)}$

A fits			$\xi_{Hill} = 0.826$		$\xi_{Pick} = 0.169$	95-pct	99.9-pct	
Algo.	GoF stat.	tail thres.	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	sam. bias	sam. bias	MC VaR
GN St. 1				*did not converge*				
LM St. 1	0.003	0	0	0.276	0.135	1.008	0.84	96
ODR ($v = 1.5$)	0.03	0	-0.136	0.302	0.176	1.043	1.052	111.5
B fits (Stop cond. = GoF)								
Algo.	GoF stat.	tail thres.	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	sam. bias	sam. bias	MC VaR
GN St. 1				*did not converge*				
LM St. 1				*did not converge*				
ODR ($v = 15$)	0.003	0	-0.02	0.277	0.132	1.001	0.829	956.2

Table 6.12: Fama-French Momentum factor 1926-2019 daily, re-weighted with equal weights in percentage point loss ($n = 10677, m = n/4$)

A fits			$\xi_{Hill} = 1.964$		$\xi_{Pick} = 1.718$	95-pct	99.9-pct	
Algo.	GoF stat.	tail thres.	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	sam. bias	sam. bias	MC VaR
GN St. 1				*did not converge*				
LM St. 1	0.067	0	-1.13	3.65	1.076	0.772	8.392	27m
ODR ($v = 40$)	0.045	0	-3.155	6.84	0.647	0.54	1.23	200,028
B fits (Stop cond. = GoF)								
Algo.	GoF stat.	tail thres.	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	sam. bias	sam. bias	MC VaR
GN St. 1				*did not converge*				
LM St. 1				*did not converge*				
ODR ($v = 0.4$)	0.051	0	-15.892	20.435	0.396	0.973	1.07	88,208

Table 6.13: Bitcoin 2010-2019 hourly losses in USD ($n = 17182, m = n/4$) [9]

A fits			$\xi_{Hill} = 0.716$		$\xi_{Pick} = -0.063$	95-pct	99.9-pct	
Algo.	GoF stat.	tail thres.	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	sam. bias	sam. bias	MC VaR
GN St. 1	0.013	0	0	0.007	0.07	1.2	1.29	1.31
LM St. 1	0.01	0	0	0.007	0.04	1.16	1.16	1.26
ODR	0.009	0	0	0.007	0	1.043	0.971	1.16
B fits (Stop cond. = GoF)								
Algo.	GoF stat.	tail thres.	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	sam. bias	sam. bias	MC VaR
GN St. 1				*did not converge*				
LM St. 1	0.059	0	0	0.007	0.54	2.74	11.31	9.6
ODR ($v = 50$)	0.012	0	0	0.006	0.046	1.068	1.08	1.2

Table 6.14: GBP/USD 1971-2019 daily nominal depreciation in USD ($n = 6103, m = n/4$) [19]

	A fits		$\xi_{Hill} = 0.679$		$\xi_{Pick} = 0.263$	95-pct	99.9-pct	
Algo.	GoF stat.	tail thres.	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	sam. bias	sam. bias	MC VaR
GN St. 1				*did not converge*				
LM St. 1	0.039	0	0.795	4.618	0.113	1.016	0.646	1064
ODR	0.043	0	0.952	3.519	0.288	1.04	1.02	1151
B fits (Stop cond. = GoF)								
Algo.	GoF stat.	tail thres.	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	sam. bias	sam. bias	MC VaR
GN St. 1	0.022	0	0.54	5.797	0.024	0.926	0.546	1,119
LM St. 1	0.022	0	0.54	5.798	0.025	0.929	0.547	1,116
ODR ($v = 70$)	0.019	0	0.825	5.021	0.15	0.99	0.762	1,151

Table 6.15: BAML options-adjusted digh-yield corp. bonds 1996-2019 daily positive spread wrt spot Treasury change in basis points ($n = 2833, m = n/4$) [18]

	A fits		$\xi_{Hill} = 1.318$		$\xi_{Pick} = 1.318$	95-pct	99.9-pct	
Algo.	GoF stat.	tail thres.	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	sam. bias	sam. bias	MC VaR
GN St. 1				*did not converge*				
LM St. 1				*did not converge*				
ODR ($v = 15$)	0.067	0	-0.511	1.818	0.337	0.992	1.578	541
	B fits							
Algo.	GoF stat.	tail thres.	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	sam. bias	sam. bias	MC VaR
GN St. 1				*did not converge*				
LM St. 1				*did not converge*				
ODR ($v = 6.8$)	0.001	0	-0.03	0.013	0.74	1.276	9.075	2,100

Table 6.16: 3-Month LIBOR Daily Positive Change in basis points ($n = 2622, m = n/4$) [16]

	A fits		$\xi_{Hill} = 0.921$		$\xi_{Pick} = 0.31$	95-pct	99.9-pct	
Algo.	GoF stat.	tail thres.	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	sam. bias	sam. bias	MC VaR
GN St. 1	0.008	0	-0.006	0.553	0.237	1	1.409	138.4
LM St. 1	0.008	0	-0.006	0.553	0.237	1	1.409	140.4
ODR ($v = 12$)	0.008	0	-0.037	0.626	0.15	0.968	1.105	133.7
B fits (Stop cond. = GoF)								
Algo.	GoF stat.	tail thres.	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	sam. bias	sam. bias	MC VaR
GN St. 1	0.008	0	0.003	0.492	0.361	1.103	2.21	191.3
LM St. 1	0.037	0	-0.032	0.492	0.361	1.086	2.2	198.4
ODR ($v = 6.8$)	0.001	0	-0.019	0.571	0.239	1.029	1.467	144.3

Table 6.17: Brent Crude Oil 1987-2019 Daily Losses in USD per barrel ($n = 4097, m = n/4$) [17]

Chapter 7

Conclusion

Throughout this thesis we have introduced the reader to some empirical problems probabilists and actuarial practitioners have had in estimating tail distributions using the generalised Pareto distribution. Along the way, we have introduced a new procedure to correct for the most egregious violations of tail biases using a set of weighted estimators divided into two stages. The author's contribution is merely but an adaptation of powerful tools from numerical optimisation that allows the practitioner the flexibility of concentrated penalisation whilst retaining full control over the acceptable values that WQE yields.

We have demonstrated correcting capabilities of WQE as well as the efficiency of its implementation due to the special structure of the problem. Conventional methods usually build on a prescribed threshold assumed by the practitioner and the results of these analytical methods are accurate to the degree of confidence in this threshold. We have shown that these methods can obfuscate a lot of the nuances in the body if one were to truncate it in search of a better tail fit. WQE resolves this issue by providing a good tail fit as well as retaining the divergences in the body so that there are both visual as well as numerical guidance towards an appropriate trade-off between tail unbiasedness and reasonable sample size.

Application of the method to real data from multiple asset classes also shows very encouraging results that hold very promising uses to interested parties. With applications in the broader financial sector, not just operational risk, we deem the estimator to have significant use in ad-hoc statistical contexts that require heuristics. We find that our studied procedure of anchoring tail quantiles and matching them to empirical quantiles provide feasible loss estimates as well as desirable finiteness of variances and usually means. Conventional methods to pick a threshold through a value at which the tail estimate plateaus proves unhelpful in some cases because they yield highly volatile estimates with less and less observation. Our procedure can provide similar tail estimates that obey the theoretical foundation of tail stability while allowing the practitioner to choose a convenient threshold.

We have not dealt with the optimal weight issues. Specifically, the author acknowledges the shortcomings regarding a systematic method to estimate weight vectors that correctly bias tail estimates to the right direction. We wish to explore this issue further

in another paper should the occasion arises. Specifically, we observe that the weights are somewhat proportional to the level of bias in the tail. One can feasibly construct an algorithm to find the link between tail bias and tail residual to account for the weight ratio differences among the fits and data sets. From the observation that we could recover reasonably accurate estimators for all data sets examined we believe the incorporation of a weight estimation step would not only be feasible but it would also enhance WQE's performance, allowing us to build a comprehensive framework for estimating tail values.

Another point we wish to consider in detail is the tail threshold. The common theme in literature thus far has been to estimate tail stability as a function of the threshold. But from theory we know that stability is asymptotic. We wish to explore further small sample estimates of the tail threshold that hold approximately, in the sense that its value can be bounded at least in distribution wrt to the asymptotic value. This consideration would necessarily differ from one dgp to the next and obviously from one real data set to another therefore poses a challenging but the results are potentially highly applicable.

Appendix A

Chapter 2

Before we begin the proofs, we wish to point out that this section was written under the assumption that the reader would examine both theorems 2.1.1 (Fisher-Tippett-Gnedenko) and 2.2.1 (Pickands-Balkema-de Haan) together. The proof for theorem 2.2.1 builds heavily on that of 2.1.1.

Definition A.0.1. *For a monotonically non-decreasing function $h\big(x\big)$, we define its left-continuous inverse as $h^* : h \to X$ as*

$$h^*\big(x\big) \triangleq \inf\Big\{u \;\Big|\; h\big(u\big) \geq x\Big\}$$

Lemma A.0.1. *If $g_n\big(x\big) \to f\big(x\big)$ for a monotonically non-decreasing functions g_n then we also have pointwise convergence for the left-continuous inverse at every point of continuity x*

$$g_n^*\big(x\big) \to f^*\big(x\big)$$

Proof. Because x is a continuity point, there must exist a limit point in any open neighbourhood around x. Since $g_n \to f$ for all points of continuity, $g_n^* \to f^*$ for all such points as well by definition of continuity. □

A.1 Theorem 2.1.1 (Fisher-Tippett-Gnedenko)

Proof. With proper normalisation for each $n \in \mathbb{N}$, the statement of the limiting distribution is

$$\begin{aligned} &\lim_{n\to\infty} \quad \Big[G\big(s_n x + a_n\big)\Big]^n = F\big(x\big) \\ \implies &\lim_{n\to\infty} \quad n \log G\big(s_n x + a_n\big) = \log F\big(x\big) \end{aligned}$$

and since $G(u) \in [0,1] \implies \log G(u) \in [-\infty, 0]$. By construction, the normalisation prevents the left hand side of the equality to diverge so we necessarily have $G \xrightarrow[n\to\infty]{} 1$. For $y \approx 1$, $\log y \approx y - 1$, we can deduce that

$$\begin{aligned} &\lim_{n\to\infty} \quad n\Big[G\big(s_n x + a_n\big) - 1\Big] = \log F\big(x\big) \\ &\lim_{n\to\infty} \quad \frac{1}{n\Big[1 - G\big(s_n x + a_n\big)\Big]} = \frac{1}{-\log F\big(x\big)} \end{aligned} \tag{A.1.1}$$

Now let $V\big(x\big) = \Big[1 - G\big(s_n x + a_n\big)\Big]^{-1}$, its left-continuous inverse is

$$\begin{aligned}
V^*\big(x\big) &= \inf\Big\{y \;\Big|\; V\big(y\big) \geq x\Big\} \\
&= \inf\left\{y \;\middle|\; \frac{1}{1 - G(y)} \geq x\right\} \\
\implies V^*\big(nx\big) &= \inf\left\{y \;\middle|\; \frac{1}{n\big[1 - G(y)\big]} \geq x\right\} \\
\implies \frac{V^*(nx) - a_n}{s_n} &= \inf\left\{\frac{y - a_n}{s_n} \;\middle|\; \frac{1}{n\big[1 - G(y)\big]} \geq x\right\} \\
&= \inf\left\{z \;\middle|\; \frac{1}{n\big[1 - G(s_n z + a_n)\big]} \geq x\right\}
\end{aligned}$$

By taking limits of both side, using continuity in probability to push the right-hand-side limit into the brackets, We then take the limit as $n \to \infty$ inside the brackets to arrive at the expression

$$\begin{aligned}
\lim_{n\to\infty} \quad \frac{V^*(nx) - a_n}{s_n} &= \inf\left\{z \;\middle|\; \frac{-1}{1 - F\big(z\big)} \geq x\right\} \\
&= \inf\Big\{z \;\Big|\; F(z) \geq \mathrm{e}^{-1/x}\Big\} \\
&= F^*\Big(\mathrm{e}^{-1/x}\Big)
\end{aligned}$$

Let $D(x) = F^*\Big(\mathrm{e}^{-1/x}\Big)$, x_0 be a continuity point of D and without loss of generality let $x_0 = 1$. We have

$$\lim_{t\to\infty} \quad \frac{V^*(tx) - V^*(tx_0)}{s(t)} = \lim_{t\to\infty} \quad \frac{V^*(tx) - V^*(t)}{s(t)} = D(x) - D(1) \tag{A.1.2}$$

where $s(t)$ is a generalisation of the discrete index of s_n to the real line such that $s(t) = s\big(\lfloor t \rfloor\big)$ and $a(t) = a\big(\lfloor t \rfloor\big)$ where $\lfloor u \rfloor$ is the floor function applied to u, i.e. the operation that supplies that nearest floor integer. Let $\Phi(x) = D(x) - D(1)$ and y be another continuity point in the domain of G_n, as such

$$\begin{aligned}
\Phi(xy) &= \lim_{t\to\infty} \quad \frac{V^*(txy) - V^*(tx_0)}{s(t)} \\
&= \lim_{t\to\infty} \quad \frac{V^*(txy) - V^*(ty)}{s(ty)} \frac{s(ty)}{s(t)} + \frac{V^*(ty) - V^*(t)}{s(t)}
\end{aligned}$$

We observe that the term $\big[V^*(txy) - V^*(ty)\big]/s(ty)$ is invariant wrt scaling by a common term, therefore $\big[V^*(txy) - V^*(ty)\big]/s(ty) = \Phi(x)$. Another claim is that both $s(ty)/s(t)$ and $\big[V^*(ty) - V^*(t)/\big]s(t)$ have unique limits. To see this, suppose there are at

least two limits for each of the two terms labelled by α_i and β_i respectively for $i = \{1, 2, ...\}$. Subtract one limit from the first to get

$$\Phi(xy) = \Phi(x)\alpha_i + \beta_i$$
$$\Phi(x)\big(\alpha_1 - \alpha_2\big) = \big(\beta_1 - \beta_2\big) = 0$$

and since we seek Φ such that it is non-constant (to prevent F being a singular probability measure), $\alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$. From equation (A.1.2) we know that these limits only depend on y, i.e. we can formulate a linear functional equation of the form

$$\Phi(xy) = \Phi(x)\alpha(y) + \Phi(y) \tag{A.1.3}$$

We make the following change of variables

$$\lambda = \log x$$
$$b = \log y$$
$$H(x) = \Phi\big(\mathrm{e}^x\big) \implies H(0) = \Phi(1) = 0$$

which allows us to re-express (A.1.3), in the limit as $t \to \infty$, as

$$\frac{H(\lambda + b) - H(b)}{\lambda} = \frac{H(\lambda) - H(0)}{\lambda}\alpha\big(\mathrm{e}^b\big) \tag{A.1.4}$$

which in the limit $\lambda \to 0$ becomes the derivative since b is continuous wrt y, which is a continuity point, and D is monotonic (via Lebesgue's theorem for differentiability of monotone functions on an open interval). More importantly, it is self-evident by identity (A.1.4) that H is differentiable for textitall b because Slutsky's theorem implies the right hand side converges to $\big[H(\lambda) - H(0)\big]C$ for a fixed $C \in \mathbb{R}$ and b as $t \to \infty$. We deduce that

$$\frac{\mathrm{d}}{\mathrm{d}b}H(b) = \alpha\big(\mathrm{e}^b\big)\frac{\mathrm{d}}{\mathrm{d}b}H(0)$$
$$\alpha\big(\mathrm{e}^b\big) = \frac{H'(b)}{H'(0)}$$

Define a new function Ψ given by

$$\Psi(b) = \frac{H(b)}{H'(0)}$$
$$\implies \Psi(0) = \frac{H(0)}{H'(0)} = 0 \tag{A.1.5}$$
$$\implies \Psi'(0) = \frac{H'(0)}{H'(0)} = 1$$

From equation (A.1.4)

$$\begin{aligned}\Psi(\lambda + b) - \Psi(b) &= \Psi(\lambda)\alpha\big(\mathrm{e}^b\big) \\ &= \Psi(\lambda)\frac{H'(b)}{H'(0)} \\ &= \Psi(\lambda)\Psi'(b)\end{aligned}$$

Recall that λ and b are two arbitrary points on the domain of G, we also get the symmetric identity by reversing the logic

$$\Psi(\lambda + b) - \Psi(\lambda) = \Psi(b)\Psi'(\lambda)$$

which also leads to the following result if we subtracted one from the other

$$\begin{aligned}
\Psi(\lambda)\big[1 - \Psi'(b)\big] &= \Psi(b)\big[1 - \Psi'(\lambda)\big] \\
\frac{\Psi(\lambda)}{\lambda}\big[\Psi'(b) - 1\big] &= \Psi(b)\frac{\big[\Psi'(\lambda) - 1\big]}{\lambda}
\end{aligned}$$

By L'Hôpital's theorem, we know that $\lim_{\lambda\to 0} \Psi(\lambda)/\lambda = 1$, which combines with (A.1.5) to give

$$\begin{aligned}
\lim_{\lambda\to 0} \quad \Psi'(b) - 1 &= \Psi(b)\frac{\big[\Psi'(\lambda) - \Psi'(0))\big]}{\lambda} \\
&= \Psi(b)\Psi''(0) \\
\implies \Psi''(b) &= \Psi'(b)\Psi''(0)
\end{aligned}$$

Let $\xi = \Psi''(0)$ and rearrange the terms to get a differential equation

$$\begin{aligned}
\frac{\Psi''(b)}{\Psi'(b)} &= \xi \\
\frac{\mathrm{d}}{\mathrm{d}b}\big(\log \Psi'\big)\big(b\big) &= \xi
\end{aligned}$$

which has the solution

$$\begin{aligned}
\Psi'\big(b\big) &= \mathrm{e}^{\xi b} \\
\Psi\big(b\big) &= \frac{1}{\xi}\Big(\mathrm{e}^{\xi b} - 1\Big)
\end{aligned}$$

which we can plug back into H to get

$$\begin{aligned}
H(x) &= \frac{H'(0)}{\xi}\Big(\mathrm{e}^{\xi x} - 1\Big) \\
&= \Phi\big(\mathrm{e}^{x}\big) \\
&= D\big(\mathrm{e}^{x}\big) - D\big(1\big) \\
D\big(x\big) &= D\big(1\big) + \frac{H'(0)}{\xi}\Big(x^{\xi} - 1\Big)
\end{aligned}$$

We now invert D to find D^* by

$$D^*(x) = \left\{1 + \frac{\xi\big[x - D(1)\big]}{H'(0)}\right\}^{1/\xi}$$

Recall that $D(x) = F^*\big(\mathrm{e}^{-1/x}\big)$, from which we can get the inversion

$$\begin{aligned} F(x) &= \exp\Big[-1/D^*(x)\Big] \\ \implies F(x) &= \exp\left\{ -\left[1 + \frac{\xi}{\sigma}\big(x-\mu\big)\right]^{-1/\xi}\right\} \end{aligned}$$

where $\mu = D(1)$ and $\sigma = H'(0)$. □

A.2 Theorem 2.2.1 (Pickands-Balkema-de Haan)

From the last proof, we know that $G(s_n x + a_n) \xrightarrow[n\to\infty]{(d)} F(x)$. We now examine the conditions that leads to this results, specifically that G lies in the domain of attraction of F or $G \in \mathcal{D}(F)$. First, we will derive the form of the limiting distribution for threshold exceedances, i.e. the GPD distribution function, using the proof from Balkema and de Haan [11]. Then we will show that the limiting distribution of G^n does indeed converge to the GPD using the approach employed by Pickands [2].

Theorem A.2.1. *The limiting distribution for threshold exceedances F given by the relation*

$$\mathbb{P}\left[\frac{X - a(c)}{s(c)} > x \;\middle|\; X > c\right] \quad \xrightarrow[c\to\omega(G)]{(d)} \quad F\big[x \mid c, \sigma(c), \xi\big] \tag{A.2.1}$$

satisfies the condition

$$F(x)F(y) = F\Big[A(y) + xS(y)\Big] \tag{A.2.2}$$

where $S \geq 1$, $A \in \mathbb{R}$, x, y as two continuity points and $F(x), F(y) < 1$.

Proof. The logic is similar to that of (A.1.3). Rewrite the conditional probability as

$$\mathbb{P}\left[\frac{X - a(c)}{s(c)} > x \;\middle|\; X > c\right] = \frac{1 - G\big[a(c) + xs(c)\big]}{1 - G(c)} \tag{A.2.3}$$

and define the ccdf's $\bar{Q} = 1 - G$ and $\bar{F} = 1 - F$. We note that one can rewrite functions of c as functions of $a(c) + ys(c)$. Wlog, we can assume that $y = 1$, because we can just appropriately shift the value, to get

$$\frac{\bar{Q}\Big[a(a(c)) + xs(a(c))\Big]}{\bar{Q}(a(c))}\frac{\bar{Q}(a(c))}{\bar{Q}(c)} = \frac{\bar{Q}\left\{a(c) + \left[\frac{a(a(c)) - a(c)}{s(c)} + x\frac{s(a(c)}{s(c)}\right]s(c)\right\}}{\bar{Q}(c)} \tag{A.2.4}$$

By Slutsky's theorem, the left hand side converges to $\bar{F}\big[a(c) + xs(c)\big]\bar{F}(0)$, in particular because $\bar{Q}(a(c))/\bar{Q}(c) \xrightarrow[c\to\omega(\bar{Q})]{} \bar{F}(0)$, which implies that $a(c) > c$ for all $\{y \mid \bar{F}(y) < 1\}$. Since the left hand side converges, the square bracket on the right hand side must be

finite. The only way for this to happen is if

$$\begin{aligned} \frac{a(a(c)) - a(c)}{s(c)} &\xrightarrow[c\to\omega(\bar{Q})]{} A(0) < \infty \\ \frac{s(a(c))}{s(c)} &\xrightarrow[c\to\omega(\bar{Q})]{} S(0) < \infty \end{aligned} \tag{A.2.5}$$

otherwise the limiting probability is always 0. If we take $x = y = 0$ then $0 < \bar{F}(S(0)) = \bar{F}(0)^2 < 1$, therefore thereare at least two unique continuity points for $\bar{F}$ at 0 and $S(0)$. This identity is proven backwards by Pickands [2] by taking the inverse cdf of the GPD and reparameterise with two distint points. If $A(0) = 0$, then we have that $S(0) = \omega(1-\bar{F}) = \omega(F)$, which is also the upperbound on the domain of $\bar{Q}$. Since $\bar{Q}(a(c))/\bar{Q}(c) \to \bar{F}(0) < 1$, the limit $S(0) > 0$. The limiting distribution satisfies

$$\bar{F}(x)\bar{F}(0) = \bar{F}\big[A(0) + xS(0)\big] \tag{A.2.6}$$

We now prove that $S(0) \geq 1$. It can be observed that $a(c)$ must approach an infinite limit, since if the limit is finite, say is $\alpha \in \mathbb{R}^+$, then $\bar{Q}(a \circ a(c))/\bar{Q}(a(c)) \xrightarrow[c\to\omega(\bar{F})]{} \bar{F}(\alpha)/\bar{F}(\alpha) = 1$ which contradicts $\bar{F}(x) < 1 \; \forall \; x$. It is clear that $a \circ a(c) - a(c) \sim \mathcal{O}(s(c))$ since by (A.2.5) the limit is $A(0)$, a constant. If $A(0) < 1$, then the sequence $a \circ a(c)$ vanishes and has a finite limit, which is a contradiction. $A(0) \geq 1$. □

Corollary A.2.1.1. *$\bar{F}(x) > 0$ for all x.*

Proof. Let A_{n+1} be the sequence given by $A_{n+1} = A(0) + A_n S$ and $A_0 = 0$ to ensure that 0 is a continuity point of $\bar{F}$ such that $\bar{F}(0) < 1$. Because $A(0) > 0$ and $S(0) \geq 1$, $\bar{F}(A_n) = \big[\bar{F}(0)\big]^{n+1} > 0$ for all n. □

Theorem A.2.2. *Let $G(x) < 1$ be the distribution function for X. If the limiting distribution for threshold exeedances in X is given by F defined in* (A.2.1) *then F is a form of the GPD.*

Proof. We again work with the ccdf's $\bar{Q}$ and $\bar{F}$. Let the set of continuity points for $\bar{F}$ be $\mathcal{C}$ where $\bar{F} < 1$. We prove the theorem by solving the general functional equation for (A.2.2). There are two cases: when $S(y) = 1$ and when $S(y) > 1$. The case when $A(y) = 0$ also leads to the same general family using a similar argument therefore we will not prove it.

Let $\Phi(x) = \log \bar{F}(x)$. When $S(y) = 1$ with a fixed $y \in \mathcal{C}$, we have

$$\begin{aligned} \Phi(x) + \Phi(y) &= \Phi\Big[A(y) + x\Big] \\ \Phi(p + x) - \Phi(x) &= cp \\ \Phi(z) &= cz - \Big[cx + \Phi(z - p)\Big] \end{aligned} \tag{A.2.7}$$

where $p = A(y)$, $c = \Phi(y)/A(y)$ and $z = x + p$. The the last square bracket term is periodic wrt z, therefore we can regroup the terms into

$$\Phi(z) = cz + \Phi_0(z) \tag{A.2.8}$$

The function Φ_0 is periodic modulo $A(y) = \Phi(y)/c$ because

$$\Phi_0(z) = -\Big\{c\big[z - A(y)\big] + \Phi\big[z - A(y)\big]\Big\} \tag{A.2.9}$$

Since c is the limiting behaviour of $\Phi(x)/x$ for $x \to \infty$, it does not depend on y. Another way to obtain the result is through the method seen in (A.1.4). Specifically, we can choose a y such that the limit for $A(y)$ can be necessarily small to obtain the differential equation

$$\begin{aligned}\lim_{A(y)\to 0} \quad \frac{\Phi(y)}{A(y)} &= \frac{\Phi\big[A(y) + x\big] - \Phi(x)}{A(y)} \\ &\approx \frac{\mathrm{d}}{\mathrm{d}x}\Phi(x)\end{aligned}$$

which has the solution given by (A.2.8). We know that Φ_0 and c must be negative because $\Phi(y) < 0$. Normalise the variable by the period p to arrive at

$$\bar{F}(z) = \exp\Big\{ -\frac{p}{\sigma}\Big[\big(z - \bar{F}_0\big)/p\Big]\Big\} \tag{A.2.10}$$

where $\sigma > 0$. Let $\xi = p$ and $\mu = F_0$. There are two special cases: when $p << \infty$, we use the logarithmic approximation

$$\bar{F}(z) = \Big[1 + \frac{\xi}{\sigma}\big(z - \mu\big)\Big]^{-1/\xi} \tag{A.2.11}$$

and when F_0 is constant, i.e. we do not normalise, which gives

$$\bar{F}(z) = \exp\Big\{ -\frac{1}{\sigma}\big(z - \mu\big)\Big\} \tag{A.2.12}$$

The second case is when $S(y) > 1$. Let $y_1, y_2 \in \mathcal{C}$ and $R_i = A(y_i) + xS(y_i)$ for $i = 1, 2$. It is clear that for any fixed x, $\Phi \circ R_1 \circ R_2(x) = \Phi \circ R_2 \circ R_1(x)$ because two continuity points uniquely determine the distribution. This identity in turn implies that $R_1 \circ R_2(x) = R_2 \circ R_1(x)$ and substituting the expressions into one another, we get

$$\begin{aligned}A(y_1) + S(y_1)A(y_2) &= A(y_2) + S(y_2)A(y_1) \\ A(y_1)\big[1 - S(y_2)\big] &= A(y_2)\big[1 - S(y_1)\big]\end{aligned} \tag{A.2.13}$$

which only holds if the two expressions are always constant, i.e. $c = A(y)\big[1 - S(y)\big]$ for all $y \in \mathcal{C}$. Let $c = 0$ and write the new form of (A.2.7) as

$$\Phi(x) + \Phi(y) = \Phi\Big[xS(y)\Big]$$

Transforming the variables to $z = \mathrm{e}^x, u = \mathrm{e}^y$ to get

$$\Psi(z) + \Psi(u) = \Psi\Big[z + \tau(u)\Big]$$

which is in the same form as (A.2.7). We then write

$$\bar{F}(x) = \exp\left\{ -\frac{p}{\sigma}\Big[\big(\log x\mathrm{e}^{S_0}\big)/p\Big]\right\}$$

which is of the form of (A.2.11) and we can deduce the same line of argument. □

We now turn the proof to the sufficient tail conditions for the GPD to apply using the same line of argument by Pickands [2]. Let us the generalised Pareto function (GPF), which generalises the GPD as introduced by von Mises [43] as the limiting extremal distribution for $G \in \mathcal{D}(F)$, defined as

$$\mathcal{P}(x) \triangleq \exp\left\{ -\int_0^{\frac{x}{\sigma}} \frac{\mathrm{d}u}{(1+\xi u)_+}\right\} \tag{A.2.14}$$

where $s > 0$. Since we can replace the scale σ and shape ξ parameters by two distinct continuity points as previously shown, $\mathcal{P}(x, \sigma, \xi) = \mathcal{P}(x, x_1, x_2)$ for $x_1, x_2 \in \mathcal{C}$. Define the metric

$$d(\mathcal{P}, \bar{\mathcal{P}}) \triangleq \sup_{x\in(0,\infty)} \big|\mathcal{P}(x) - \bar{\mathcal{P}}(x)\big| \tag{A.2.15}$$

Because it can be shown that x_1, x_2 are continuous functions of σ, ξ, we can scale the two GPFs by a common positive term, such as $\bar{x}_1$, and gather the expressions into one term as

$$d(\mathcal{P}, \bar{\mathcal{P}}) = \phi\left(\frac{x_1}{\bar{x}_1}, \frac{x_2}{\bar{x}_1}, \frac{\bar{x}_2}{\bar{x}_1}\right) = \phi(y, \alpha, \beta) \tag{A.2.16}$$

where

$$\lim_{y\to 1, \alpha\to\beta} \phi(y, \alpha, \beta) = 0$$

Theorem A.2.3. *Let $\bar{Q}(x)$ be a ccdf. There exist $x_1, x_2 \in (0, \infty)$ and $\delta > 0$ such that if*

$$d(\bar{Q}, \bar{\mathcal{P}}) \leq \delta$$

then

$$\big|x_i - \bar{x}_i\big| \leq h_i(\bar{x}_1, \bar{x}_2, \delta) \;\; ; \; i = 1, 2$$

where

$$\lim_{\delta\to 0} \;\; h(\bar{x}_1, \bar{x}_2, \delta) = 0$$

Proof. This is true by continuity, since one can always find a pair $(\bar{x}_1, \bar{x}_2)$ close enough to make this result hold. □

Definition A.2.1. *$\mathcal{P}$ is associated with a ccdf $\bar{Q}$ if there exist $x_1, x_2 \in (0, \infty)$ such that*

$$\mathcal{P}(x) = \mathcal{P}(x, x_1, x_2) \tag{A.2.17}$$

where x_1, x_2 are two distinct points in the domain of $\bar{Q}$.

Theorem A.2.4. *Let $\mathcal{P}(x), \bar{\mathcal{P}}(x)$ be two GPFs, $\bar{Q}$ a ccdf and let $\bar{\mathcal{P}}(x)$ be associated with $\bar{Q}$. For $\delta > 0$ if*

$$d(\mathcal{P}, \bar{Q}) \leq \delta$$

then

$$d(\mathcal{P}, \bar{\mathcal{P}}) \leq \phi\left(\frac{x_1}{\bar{x}_1}, \frac{x_2}{\bar{x}_1}, \frac{\bar{x}_2}{\bar{x}_1}\right)$$

such that

$$\left|x_i - \bar{x}_i\right| \leq h_i(\bar{x}_1, \bar{x}_2, \delta) \ ; \ i = 1, 2$$

then

$$\lim_{\delta \to 0} \ d(\bar{\mathcal{P}}, \bar{Q}) = 0$$

Proof. This can be seen as a consequence of (A.2.16) and theorem A.2.3. □

Corollary A.2.4.1. *This implies $d(\bar{\mathcal{P}}, \bar{Q}) \xrightarrow[\delta \to 0]{} 0$.*

Proof. Note that by the triangle inequality

$$\begin{aligned} d(\bar{\mathcal{P}}, \bar{Q}) \ &\leq \ d(\mathcal{P}, \bar{Q}) + d(\mathcal{P}, \bar{\mathcal{P}}) \\ &\leq \delta + \phi\left(\frac{x_1}{\bar{x}_1}, \frac{x_2}{\bar{x}_1}, \frac{\bar{x}_2}{\bar{x}_1}\right) \end{aligned}$$

□

Let $\bar{G}_c(x)$ be the tail conditional probability

$$\bar{G}_c(x) = \frac{\mathbb{P}(X > x + c)}{\mathbb{P}(X > c)} = \frac{\bar{G}(x + c)}{\bar{G}(x)}$$

For a ccdf $\bar{Q}$ and a corresponding GPF $\bar{\mathcal{P}}$, combined with a^* being the left continuous inverse of a given in A.0.1, define the quantity

$$D(y) \ \triangleq \ \inf_{y^* : \bar{G}(y^*) = y} d\Big(\bar{\mathcal{P}}_{y^*}, \bar{Q}_{y^*}\Big) \tag{A.2.18}$$

which is well-defined if $G(x)$ is continuous. Define the set containing strictly positive elements

$$\mathcal{S}_y \ \triangleq \ \Big\{\xi \ \Big| \ \bar{\mathcal{P}}(y^*, \sigma, \xi) = y\Big\}$$

and its bounds

$$\begin{aligned} \xi^+ &= \sup\left\{\xi \ \middle| \ \xi \in \mathcal{S}_y\right\} \\ \xi^- &= \inf\left\{\xi \ \middle| \ \xi \in \mathcal{S}_y\right\} \end{aligned}$$

Theorem A.2.5. *If there exists $y \in (0,1]$ such that $D(y) = 0$, then for all $u \leq y$ we have $D(u) = 0$ where*

$$\xi^+(y) = \xi^-(y) = \xi^+(u) = \xi^-(u)$$

Proof. Since we have established that there exists y^* such that $\bar{G}(y^*) = y$ and $d\big(\bar{\mathcal{P}}_{y^*}, \bar{Q}_{y^*}\big)$, let $\bar{Q}_{y^*} = \bar{\mathcal{P}}_{y^*}$, the GPF associated with $\bar{Q}_{y^*}$. For any $t > 0$, we plug in the expression for the GPF gives

$$\begin{aligned}
\bar{Q}_{y^*+t}(x) = \frac{\bar{G}(y^* + t + x)}{\bar{G}(y^* + t)} = \frac{\bar{Q}_{y^*}(x+t)}{\bar{Q}_{y^*}(t)} &= \exp\left\{ -\int_{\frac{t}{\sigma}}^{\frac{x+t}{\sigma}} \frac{\mathrm{d}x'}{1+\xi x'} \right\} \\
&= \exp\left\{ -\int_{0}^{\frac{x}{\sigma}} \frac{\mathrm{d}x'}{1+\xi(x'+t/\sigma)} \right\} \\
&= \exp\left\{ -\int_{0}^{\frac{x}{\sigma+\xi t}} \frac{\mathrm{d}x'}{1+\xi x'} \right\}
\end{aligned}$$

which implies that ξ is unique past the y^* threshold and σ is translated by ξt. □

The proof of A.2.5 establishes the stability of the GPF past a threshold which corresponds to a ccdf $\bar{G}$ converging to its associated GPF $\bar{\mathcal{P}}$. We can deduce that $\bar{G}$ follows a GPD if and only if $\lim_{y\to 0} D(y) = 0$ and there exists $\xi \in (-\infty, \infty)$ such that

$$\lim_{y\to 0} \ \xi^-(y) = \lim_{y\to 0} \ \xi^+(y) = \xi$$

Theorem A.2.6. *A distribution function $G(x)$ has a tail that follows the GPD if and only if $G \in \mathcal{D}(F)$ where F is the GEVD distribution given in* (A.1)*.*

Proof. We know that $G \in \mathcal{D}(F)$ if and only if

$$\lim_{n\to\infty} n \log G(A_n + xS_n) = \log F(x)$$

which we need for $\lim_{n\to\infty} F(A_n + xS_n) = 1$ for all x to hold, which leads to

$$\lim_{n\to\infty} n\Big[1 - G(A_n + xS_n)\Big] = -\log F(x)$$

Let $c < \omega(G)$ where $A_n \leq c \leq A_{n+1}$, which is possible since $\big\{A_n\big\}_{n=1}^{\infty}$ can be chosen to be monotonically non-decreasing. For a measurable functions $\sigma(c) \triangleq S_n$, we have

$$\begin{aligned}
&\lim_{c\to\omega(G)} \bar{G}_c\big[x\sigma(c)\big] = \lim_{c\to\omega(G)} \frac{1 - G\big[c + x\sigma(c)\big]}{1 - G(c)} = \frac{-\log F(x)}{-\log F(0)} \\
&\implies \lim_{c\to\omega(F)} \left| \frac{1 - G\big[c + x\sigma(c)\big]}{1 - G(c)} - \left[1 + \frac{\xi}{\sigma}(x-\mu)\right]^{\frac{-1}{\xi}} \right| = 0
\end{aligned}$$

Because of compactness, the upper bound of this expression must also go to zero. This

implies that if we fix an x, then for all x we have uniform convergence, i.e.

$$\lim_{c\to\omega(G)} \sup_{x<\omega(G)} \left| \frac{1-G\big[c+x\sigma(c)\big]}{1-G(c)} - \left[1+\frac{\xi}{\sigma}(x-c)\right]^{\frac{-1}{\xi}} \right| = 0$$

$$\implies \lim_{c\to\omega(G)} \sup_{x<\omega(G)} \left| \frac{1-G\big[c+x\big]}{1-G(c)} - \left[1+\frac{\xi}{\sigma(c)}(x-c)\right]^{\frac{-1}{\xi}} \right| = 0$$

The first equality is derived from the fact that this identity has to hold over the positive domain, we have to replace the fixed μ with c. Note that we are able to do this because by definition in (A.1.2) $\mu = D(1)$ was an arbitrary continuity point in the support of F. The second equality is obtained by scaling $\sigma(c)$ by a fixed constant σ, therefore yielding another function $\sigma(c)$. □

Appendix B

Chapter 4

B.1 Lemma 4.1.1

Proof. From (5.1.5) we know that the approximation of the Hessian matrix of the least-squares problem is given by

$$\nabla^2\varphi = \mathbf{J}^{\mathrm{T}}\mathbf{J} + + \sum_{i=1}^{m} \rho_i \nabla^2 \rho_i$$

and the first term on the right is always spd, thus it is enough to show that $\sum_{i=1}^{m} \rho_i \nabla^2 \rho_i$ is not spd for some i. We look closer at the individual Hessians $\nabla^2\rho_i$. The diagonal elements of $\nabla^2\rho_i = \nabla^2\rho(\theta \mid x_i)$ are made up of

$$\frac{\partial^2 \rho_i}{\partial \sigma^2} = \frac{(x_i - \mu)}{\sigma^3}\left[1 + \frac{\xi(x_i - \mu)}{\sigma}\right]^{-1/\xi - 1} - \frac{(x_i - \mu)^2(\xi + 1)}{\sigma^4}\left[1 + \frac{\xi(x_i - \mu)}{\sigma}\right]^{-1/\xi - 2}$$

It suffices to show that the sum of the individual Hessians of the residuals $\nabla^2\rho_i$ is not spd for some i. Let $\mathbf{A}_i = \nabla^2\rho_i$ and define the k^{th} pivot of $\mathbf{A}_i$

$$p_k \triangleq \frac{\det\big(\mathbf{A}_{i,k}\big)}{\det\big(\mathbf{A}_{i,k-1}\big)}$$

where $\mathbf{A}_{i,j}$ is the $j \times j$ sub-matrix making up the first j rows and columns of $\mathbf{A}_i$ [1]. Suppose $\mathbf{A}_i$ is spd. This implies that $p_k \geq 0 \ \forall$ k for some i as a consequence of Sylvester's criterion. The first pivot is just $\langle 1| \mathbf{A}_i |1\rangle$ and since the ordering of the parameters is arbitrary, we just need to show that all three of the diagonal elements of $\mathbf{A}_i$ are always positive. Rewriting $\partial^2\rho_i/\partial\sigma^2$, we get

$$\frac{\partial^2 \rho_i}{\partial \sigma^2} = \kappa\left\{1 - \frac{(x_i - \mu)(\xi + 1)}{\sigma}\left[1 + \frac{\xi(x_i - \mu)}{\sigma}\right]^{-1}\right\}$$

$$\kappa = \frac{(x_i - \mu)}{\sigma^3}\left[1 + \frac{\xi(x_i - \mu)}{\sigma}\right]^{-1/\xi - 1}$$

[1] This relation can be trivially checked by transforming the matrix into reduced row-echelon form.

Note that $\kappa > 0$ whenever $x_i > \mu$, which is the case we consider. We can reduce the terms down to

$$\frac{\partial^2 \rho_i}{\partial \sigma^2} = \kappa \left\{ 1 - \frac{(\xi + 1)}{\frac{\sigma}{x_i - \mu} + \xi} \right\}$$

For $x_i - \mu \in (0, 1]$

$$\frac{\partial^2 \rho_i}{\partial \sigma^2} \geq 0$$

and for $x_i - \mu > 0$

$$\frac{\partial^2 \rho_i}{\partial \sigma^2} < 0$$

therefore the Hessian $\mathbf{A}_i$ is not spd for $x_i > \mu + 1$. The optimisation problem therefore is not convex. □

B.2 Theorem 4.3.1 (Glivenko-Cantelli) [42]

Proof. First note the almot sure convergence condition is equivalent to

$$\mathbb{P}\left[\lim_{n \to \infty} \sup_X \left| F_n(X) - F(X) \right| = 0 \right] = 1 \tag{B.2.1}$$

which is a stronger condition compared to the law of large numbers because of uniformity. Let $\epsilon > 0$ and $k > \lceil 1/\epsilon \rceil$, the ceiling function applied to $1/\epsilon$, be an increasing index for a finite division of $\mathbb{R}$, i.e.

$$-\infty = k_0 < k_i < k_k = \infty \quad \forall\, i = 1, ..., k-1$$

which have the property

$$F\big(k_i^-\big) \leq \frac{i}{k} \leq F\big(k_i\big) \quad \forall i = 1, ..., k-1$$

where the superscript denotes the probability

$$F\big(k_i^-\big) = \mathbb{P}\big(X_i < k_i\big) = F\big(k_i\big) - \mathbb{P}\big(X = k_i\big)$$

We then have

$$F\big(k_i^-\big) - F\big(k_{i-1}\big) \leq \frac{i}{k} - \frac{i-1}{k} = \frac{1}{k} \leq \epsilon$$

By the strong law we have for each point

$$\left| F_n\big(k_i^-\big) - F\big(k_i^-\big) \right| \xrightarrow[n \to \infty]{(a.s.)} 0 \quad \text{and} \quad \left| F_n\big(k_i\big) - F\big(k_i\big) \right| \xrightarrow[n \to \infty]{(a.s.)} 0$$

Let the sequence $\{\delta_n\}_{n=1}^{\infty}$ be given by

$$\delta_n = \max_{i=1,\ldots,k-1} \left\{ \left| F_n\big(k_i^-\big) - F\big(k_i^-\big) \right|, \left| F_n\big(k_i\big) - F\big(k_i\big) \right| \right\} \xrightarrow[n\to\infty]{(a.s.)} 0$$

For each x, we can identify an interval it lies in given by $k_{i-1} \leq x < k_i$ which, by right continuity of the cdf, gives

$$\begin{aligned} F_n\big(x\big) - F\big(x\big) \;&\leq\; F_n\big(k_i^-\big) - F\big(k_{i-1}\big) \;\leq\; F_n\big(k_i^-\big) - F\big(k_i^-\big) + \epsilon \\ F_n\big(x\big) - F\big(x\big) \;&\geq\; F_n\big(k_{i-1}\big) - F\big(k_i^-\big) \;\geq\; F_n\big(k_{i-1}\big) - F\big(k_{i-1}\big) - \epsilon \\ \implies \left| F_n(X) - F(X) \right| &\leq \delta_n + \epsilon \xrightarrow[n\to\infty]{(a.s.)} \epsilon \end{aligned}$$

For this to hold for all x, we need that

$$\sup_X \quad \left| F_n(X) - F(X) \right| \xrightarrow[n\to\infty]{(a.s.)} \epsilon$$

Let A_ϵ be the set where this uniform convergence for ϵ occurs, such that $\mathbb{P}(A_\epsilon) = 1$, we define

$$A \triangleq \bigcap_{\epsilon} A_\epsilon \triangleq \lim_{\epsilon\to 0} A_\epsilon$$

then by continuity of probability

$$\begin{aligned} \mathbb{P}(A) \;&=\; \mathbb{P}\Big(\lim_{\epsilon\to 0} A_\epsilon \Big) \;=\; \lim_{\epsilon\to 0} \mathbb{P}(A_\epsilon) \;=\; 1 \\ \mathbb{P}\Big[\lim_{n\to\infty} \sup_X \left| F_n(X) - F(X) \right| = 0 \Big] &= 1 \end{aligned}$$

□

Appendix C

Chapter 5

C.1 Theorem 5.2.3

Proof. We drop the subscripts and take all norms to be ℓ_2 norm for notational convenience. From lemma 5.2.2 we know that the GN directions are descent. This implies that there exist $c_1, c_2 > 0$ such that we can bound the residuals

$$\left|\langle j \mid \rho_k \rangle\right| \leq c_1 \quad \forall \quad j,k$$
$$\left|\langle j| \, \mathbf{J}_k\right| \preccurlyeq c_1 \quad \forall \quad j,k$$

where the second inequality is directly from the Lipschitz condition, which also implies that a linear combination of the residuals for some k

$$\sum_{j=1}^{m} \rho_j \left|\nabla \rho_j\right\rangle = \left|\nabla \varphi\right\rangle$$

is also Lipschitz continuous. We also know that

$$\left|\nabla \varphi\right\rangle = \mathbf{J}^{\mathrm{T}} \left|\rho\right\rangle = -\mathbf{J}^{\mathrm{T}}\mathbf{J} \left|\delta\right\rangle$$

Take the cosine of the angle ϕ between the direction and the gradient at every iteration, which measures the linear correlation between the to vectors, we get

$$\begin{aligned}
\cos\phi &= \frac{\langle -\nabla \varphi_k \mid \delta \rangle}{\sqrt{\langle \nabla \varphi \mid \nabla \varphi \rangle}\sqrt{\langle \delta \mid \delta \rangle}} \\
&= \frac{\langle \delta| \, \mathbf{J}^{\mathrm{T}} \, \mathbf{J} \left|\delta\right\rangle}{\left\|\mathbf{J}^{\mathrm{T}} \, \mathbf{J} \left|\delta\right\rangle\right\| \left\||\delta\rangle\right\|} \\
&\geq \frac{\left\|\mathbf{J}^{\mathrm{T}} \, \left|\delta\right\rangle\right\|^2}{\left\|\mathbf{J}^{\mathrm{T}} \, \mathbf{J}\right\| \left\||\delta\rangle\right\|^2}
\end{aligned}$$

Because the eigenvalues of $\mathbf{J}^{\mathrm{T}}\mathbf{J}$ are positively bounded from below in $\mathcal{N}$, we also have that

$$\begin{aligned}\cos\phi &\geq \frac{\tau^2\left\||\delta\rangle\right\|^2}{\left\|\mathbf{J}^{\mathrm{T}}\ \mathbf{J}\right\|\left\||\delta\rangle\right\|^2} \\ &= \frac{\tau^2}{\left\|\mathbf{J}^{\mathrm{T}}\ \mathbf{J}\right\|}\end{aligned}$$

Since the residuals are Lipschitz continuous in $\mathcal{N}$, their gradients must also be continuous and differentiable, therefore $\left\|\mathbf{J}^{\mathrm{T}}\ \mathbf{J}\right\| < \infty$ and $\cos\phi > 0$. By corollary 5.2.1.1, we know $\left\||\nabla\varphi_k\rangle\right\| \to 0$. □

C.2 Givens Rotation for LSQR

For a given QR decomposition $\mathbf{J} = \mathbf{QR}$ of $\mathbf{J} \in \mathbb{R}^{m\times p}$ where $\mathbf{Q}$ is orthogonal and $\mathbf{R}$ is upper triangular, observe that for any dimension-conforming matrix $\mathbf{\Lambda}$ we have

$$\begin{bmatrix}\mathbf{J}\ \ \mathbf{\Lambda}\end{bmatrix}\begin{bmatrix}\mathbf{J}\\ \mathbf{\Lambda}\end{bmatrix} = \begin{bmatrix}\mathbf{R}\ \ \mathbf{\Lambda}\end{bmatrix}\begin{bmatrix}\mathbf{R}\\ \mathbf{\Lambda}\end{bmatrix}$$

Let $\mathbf{\Lambda}$ be $\sqrt{\lambda}\mathbf{I}_p$, the diagonal matrix containing the sqare root of the damping factor on its diagonal. Take an example of the first row $\langle 1|\,\mathbf{\Lambda}$ being appended to the bottom of $\mathbf{R}$

$$\mathbf{R} = \begin{bmatrix} R_{11} & R_{12} & \dots & R_{1p} \\ 0 & R_{22} & \dots & R_{2p} \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \dots & R_{pp} \\ \sqrt{\lambda} & 0 & \dots & 0 \end{bmatrix}$$

To zero out the last row of $\mathbf{R}$, we first consider the 2D subspace spanned by R_{11} and $\sqrt{\lambda}$. We can rotate the coordinates using an orthonormal rotation matrix which preserves the energy of the vector $\begin{bmatrix}R_{11}\ \ \sqrt{\lambda}\end{bmatrix}$. As such, there exists an angle ϕ such that

$$\begin{bmatrix}\cos\phi & -\sin\phi \\ \sin\phi & \cos\phi\end{bmatrix}\begin{bmatrix}R_{11}\\ \sqrt{\lambda}\end{bmatrix} = \begin{bmatrix}a\\ 0\end{bmatrix}$$

where we can then replace R_{11} with a in the new $\mathbf{R}$ matrix. Let We have a system of three equations with three unknowns

$$\begin{gathered}\cos\phi R_{11} - \sin\phi\sqrt{\lambda} = a \\ \sin\phi R_{11} + \cos\sqrt{\lambda} = 0 \\ \sin^2\phi + \cos^2\phi = 1\end{gathered}$$

with we can use to deduce that the solutions are

$$\cos\phi = \frac{R_{11}}{\sqrt{R_{11}^2 + \lambda}}$$
$$\sin\phi = -\frac{\sqrt{\lambda}}{\sqrt{R_{11}^2 + \lambda}}$$

Because of floating-point arithmetics, we can avoid overflow problems when summing two squares by defining the alternative set of solutions if $|R_{11}| \geq |\sqrt{\lambda}|$

$$\cos\phi = \frac{1}{\sqrt{1 + \tan^2\phi}}$$
$$\sin\phi = \cos\phi\tan\phi$$
$$\tan\phi = -\frac{\sqrt{\lambda}}{R_{11}}$$

and likewise if $|R_{11}| < |\sqrt{\lambda}|$

$$\sin\phi = \frac{1}{\sqrt{1 + \cot^2\phi}}$$
$$\cos\phi = \cos\phi\tan\phi$$
$$\cot\phi = -\frac{R_{11}}{\sqrt{\lambda}}$$

so that we only square elements strictly less than or equal to 1. Givens rotation works for any pair of elements, therefore to zero out an entire row we define the general Givens matrix

$$\mathbf{G}(i,j,\phi) \triangleq \begin{bmatrix} 1 & 0 & 0 & \dots & \dots & \dots & \dots & \vdots & \vdots \\ 0 & 1 & 0 & \dots & \dots & \dots & \dots & \vdots & \vdots \\ 0 & 0 & \ddots & \dots & \dots & \dots & \dots & \vdots & \vdots \\ \vdots & \vdots & \dots & \cos\phi & \vdots & -\sin\phi & \dots & \vdots & \vdots \\ \vdots & \vdots & \dots & \dots & \ddots & \dots & \dots & \vdots & \vdots \\ \vdots & \vdots & \dots & \sin\phi & \dots & \cos\phi & \dots & \vdots & \vdots \\ \vdots & \vdots & \dots & \dots & \dots & \dots & \ddots & 0 & 0 \\ \vdots & \vdots & \dots & \dots & \dots & \dots & 0 & 1 & 0 \\ \vdots & \vdots & \dots & \dots & \dots & \dots & 0 & 0 & 1 \end{bmatrix}$$

where $\cos\phi$ appears in $\langle i|\,\mathbf{G}\,|i\rangle$ and $\langle j|\,\mathbf{G}\,|j\rangle$, $-\sin\phi$ in $\langle i|\,\mathbf{G}\,|j\rangle$ and $\sin\phi$ in $\langle j|\,\mathbf{G}\,|i\rangle$. For our LM problem j corresponds to $m + l$ and where l is the l^{th} row of Λ being considered. The new $\mathbf{Q}$ matrix is obtained through repeated multiplication $\mathbf{Q} = \mathbf{G}(p, m+1)\mathbf{G}(p -$

$1, m+1)...\mathbf{G}(1, m+1)$. It should be self-evident that the new $\mathbf{Q}$ is orthogonal since all $\mathbf{G}$'s are orthogonal.

This process is repeated for all rows of Λ, where the order of complexity is $\mathcal{O}(n)$ because when a Givens matrix is applied to $\mathbf{R}$, each column's elements are given by

$$\mathbf{G}(i,j,\phi)\mathbf{R}\left|k\right\rangle = \begin{cases} \cos\phi \left\langle i\right| \mathbf{R} \left|k\right\rangle - \sin\phi \left\langle j\right| \mathbf{R} \left|k\right\rangle & k = i \\ \sin\phi \left\langle i\right| \mathbf{R} \left|k\right\rangle + \cos\phi \left\langle j\right| \mathbf{R} \left|k\right\rangle & k = j \\ \left\langle l\right| \mathbf{R} \left|k\right\rangle & k = l \neq i, j \end{cases}$$

It is not difficult to check that $\sin\phi \left\langle i\right| \mathbf{R} \left|k\right\rangle + \cos\phi \left\langle j\right| \mathbf{R} \left|k\right\rangle = 0$ for $k = j$ as needed.

C.3 Theorem 5.2.6

There are two cases we consider like in [33]. The first case is when $\eta = 0$, i.e. we accept the step if it yields any decrease in φ. All norms in the following two proofs are ℓ_2 norm.

Proof. The proof for the case $\eta = 0$ is by contradiction. Consider the second order Taylor expansion for φ

$$\varphi\big(\theta_k + \delta_k\big) = \varphi\big(\theta_k\big) + \big\langle \nabla\varphi\big(\theta_k\big) \mid \delta_k \big\rangle + \int_0^1 \big\langle \nabla\varphi\big(\theta_k + t\delta_k\big) - \nabla\varphi\big(\theta_k\big) \mid \delta_k \big\rangle \, \mathrm{d}t$$

The absolute difference between q and φ up to second order is given by

$$\Big| q\big(\delta_k\big) - \varphi\big(\theta_k + \delta_k\big) \Big| = \left| \frac{1}{2} \left\langle \delta_k\right| \mathbf{J}^{\mathrm{T}}\mathbf{J} \left|\delta_k\right\rangle - \int_0^1 \big\langle \nabla\varphi\big(\theta_k + t\delta_k\big) - \nabla\varphi\big(\theta_k\big) \mid \delta_k \big\rangle \, \mathrm{d}t \right|$$

Since $\mathbf{J}^{\mathrm{T}}\mathbf{J}$ is Lipschitz in $\mathcal{N}$, we can bound its Frobenius norm by $B > 0$, i.e. $\left\|\mathbf{J}^{\mathrm{T}}\mathbf{J}\right\|_2 \leq M$. By linearity of the inner product, we conbine with the Frobenius norm to get

$$\Big| q\big(\delta_k\big) - \varphi\big(\theta_k + \delta_k\big) \Big| \leq \left| \frac{1}{2} B \big\| \left|\delta\right\rangle \big\| - \left\langle \int_0^1 \nabla\varphi\big(\theta_k + t\delta_k\big) - \nabla\varphi\big(\theta_k\big) \, \mathrm{d}t \;\middle|\; \delta_k \right\rangle \right|$$

The integral term therefore is a a multiple of the norm of $\left|\delta\right\rangle$. Because this term vanishes as we reduce the norm of the direction $\big\| \left|\delta\right\rangle \big\|$, we can make the inner product arbitrarily small by bounding it from above with $\big\| \left|\delta\right\rangle \big\|$ and let this evaluate to $C_T\big(\delta\big) \big\| \left|\delta\right\rangle \big\|$.

$$\Big| q\big(\delta_k\big) - \varphi\big(\theta_k + \delta_k\big) \Big| \leq \frac{1}{2} B \big\| \left|\delta\right\rangle \big\| + C_T\big(\delta\big) \big\| \left|\delta\right\rangle \big\|$$

Suppose that there exist $K \in \mathbb{N}$ and $\epsilon > 0$ such that if $k \geq K$ then $\big\| \left|\nabla\varphi_k\right\rangle \big\| \geq \epsilon$, therefore

$$\begin{aligned} q\big(0\big) - q\big(\delta\big) &\geq c_1 \big\| \left|\nabla\varphi_k\right\rangle \big\| \left(r_k \bigwedge \frac{\big\| \left|\nabla\varphi_k\right\rangle \big\|}{\big\| \mathbf{J}^{\mathrm{T}}\mathbf{J} \big\|} \right) \\ &\geq c_1 \epsilon \left(r_k \bigwedge \frac{\epsilon}{B} \right) \end{aligned}$$

For conciseness let $\varphi_k = \varphi\big(\theta_k\big)$ and $\varphi_{k+1} = \varphi\big(\theta_k + \delta_k\big)$. Using the expression for τ_k defined in 5.2.8, we get the quantity

$$\begin{aligned}
\big|\tau_k - 1\big| &= \left|\frac{\varphi_k - \varphi_{k+1} - q\big(0\big) + q\big(\delta_k\big)}{q\big(0\big) - q\big(\delta_k\big)}\right| \\
&\leq \left|\frac{q\big(\delta_k\big) - \varphi_{k+1}}{q\big(0\big) - q\big(\delta_k\big)}\right| \\
&\leq \left|\frac{q\big(\delta_k\big) - \varphi_{k+1}}{c_1\epsilon\big(r_k \wedge \epsilon/B\big)}\right| \\
&= \frac{\frac{1}{2}B\big\||\delta\rangle\big\| + C_T\big(\delta\big)\big\||\delta\rangle\big\|}{c_1\epsilon\big(r_k \wedge \epsilon/B\big)} \\
&= \frac{\gamma r_k\Big[B\gamma r_k + 2C_T\big(\delta\big)\Big]}{2c_1\epsilon\big(r_k \wedge \epsilon/B\big)}
\end{aligned}$$

We can choose an r_k to satisfy an arbitrarily small value, therefore pick a threshold value r^* such that if $\big\||\delta\rangle\big\| \leq \gamma r_k \leq \gamma r^*$, the term in the square bracket satisfies

$$\begin{aligned}
B\gamma r_k + 2C_T\big(\delta\big) &\leq \frac{c_1\epsilon}{2\gamma} \\
\implies \big|\tau_k - 1\big| &\leq \frac{r_k}{4\big(r_k \wedge \epsilon/B\big)}
\end{aligned}$$

This also implies that we can pick a smaller r^* such that $r_k \leq r^* \leq \epsilon/B \;\; \forall\, k \geq K$, which yields

$$\big|\tau_k - 1\big| \leq \frac{1}{4} \tag{C.3.1}$$

When $\tau \geq 1$ because then the algorithm is performing at or better than the full second order approximation of the actual function, it is not a source for concern. We must get that for $\tau < 1$

$$\tau_k \geq \frac{3}{4}$$

From algorithm 2 we know that this is the condition for increasing r_k, thus $r_{k+1} > r_k$ when $r_k \leq r^*$, which in turn implies

$$r_k \geq \left(r_k \bigwedge \frac{r^*}{4}\right) \quad \forall \quad k \geq K \tag{C.3.2}$$

Conversely, r_k is only reduced by a quarter if $r_k \geq r^*$. Suppose that there exists an infinite subsequence $\mathcal{K}$ of the iterations such that if $k \geq K$ and $k \in \mathcal{K}$ then the value of τ produced by all iterations in $\mathcal{K}$ satisfies $\tau_k \geq 1/4$. By definition

$$\begin{aligned}
\varphi_k - \varphi_{k+1} &\geq \frac{1}{4}\Big[q\big(0\big) - q\big(\delta\big)\Big] \\
&\geq c_1\epsilon\Big(r_k \bigwedge \epsilon/B\Big)
\end{aligned}$$

Since $\varphi > -\infty$ and $\epsilon/B > 0$, it must be that

$$\lim_{k \in \mathcal{K}, k \to \infty} r_k = 0$$

but this contradicts (C.3.2) hence $\mathcal{K}$ does not exist and the converse must be true, i.e. $\tau_k < 1/4$ when $k \geq K$. Per the algorithm, the radius r_k will eventually vanish for such a sequence of τ, therefore contradicts (C.3.2) as well. We conclude that for $\eta = 0$,

$$\liminf_{k \to \infty} \quad \Big\| |\nabla \varphi_k\rangle \Big\|_2 = 0$$

□

The second case when $\eta \in \big(0, 0.25\big]$ is pertinent to our problem since it bounds the norm of the directions towards 0 and also because the η value tested ranges from $1\mathrm{e}^{-2}$ to 0.25.

Proof. We begin with $\eta \in \big(0, 0.25\big)$. Because $\mathbf{J}^{\mathrm{T}}\mathbf{J}$ is Lipschitz in $\mathcal{N}$, we know that for some $B_m > 0$

$$\Big\| |\nabla \varphi(\theta)\rangle - |\nabla \varphi(\theta_m)\rangle \Big\| \leq B_m \Big\| |\theta\rangle - |\theta_m\rangle \Big\|$$

Define a ball around a particular $|\theta_m\rangle$ in $\mathbb{R}^p$ with radius R by

$$\mathcal{B}\big(\theta_m, R\big) \triangleq \Big\{ \, |\theta\rangle \;\Big|\; \Big\| |\theta\rangle - |\theta_m\rangle \Big\| \leq R \Big\}$$

and

$$\epsilon = \frac{1}{2} \Big\| |\nabla \varphi\rangle \Big\|$$

$$R = \frac{\Big\| |\nabla \varphi\rangle \Big\|}{2B_m}$$

By Cauchy-Schwartz, we know

$$\begin{aligned}
\Big\| |\nabla \varphi_k\rangle \Big\|^2 &\geq \langle \nabla \varphi_k \mid \nabla \varphi_m \rangle \\
\implies \Big\| |\nabla \varphi_k\rangle \Big\| &\geq \Big\| |\nabla \varphi_k\rangle \Big\| - \Big\| |\nabla \varphi_k\rangle - |\nabla \varphi_m\rangle \Big\| \\
&\geq \Big\| |\nabla \varphi_k\rangle \Big\| - B_m R \\
&= \epsilon \quad \forall \quad \Big\{ \, |\theta_k\rangle \;\Big|\; |\theta_k\rangle \in \mathcal{B}\big(\theta_m, R\big) \Big\}
\end{aligned}$$

By the previous proof, we know this cannot happen for $k \to \infty$, therefore for some $\big\| |\theta_k\rangle \big\|$ must lie outside $\mathcal{B}\big(\theta_m, R\big)$. Let $s+1$ denote the index that $\big\| |\theta_k\rangle \big\|$ first leave the ball after iteration m. This implies

$$\Big\| |\nabla \varphi_k\rangle \Big\| \geq \epsilon \quad \forall \quad k = m, ..., s$$

By definition

$$\begin{aligned} q\big(0\big) - q\big(\delta\big) &\geq c_1 \Big\| |\nabla \varphi_k\rangle \Big\| \left(r_k \bigwedge \frac{\Big\| |\nabla \varphi_k\rangle \Big\|}{\Big\| \mathbf{J}^{\mathrm{T}} \mathbf{J} \Big\|} \right) \\ &\geq c_1 \epsilon \left(r_k \bigwedge \frac{\epsilon}{B} \right) \end{aligned}$$

where B is the overall Lipschitz constant for $\mathbf{J}^{\mathrm{T}}\mathbf{J}$. Define a subsequence $\mathcal{S}$ of $m, ..., s$ given by the iterations where the steps are actually taken, i.e. $|\theta_{k+1}\rangle \neq |\theta_k\rangle$ for alk $k \in \mathcal{S}$, we get

$$\begin{aligned} \varphi_m - \varphi_{s+1} &= \sum_{k=m}^{s} \varphi_k - \varphi_{k+1} \\ &\geq \sum_{k \in \mathcal{S}} \eta \Big[q\big(0\big) - q\big(\delta\big) \Big] \\ &\geq \sum_{k \in \mathcal{S}} \eta c_1 \epsilon \left(r_k \bigwedge \frac{\epsilon}{B} \right) \end{aligned}$$

The first case is where $r_k \leq \epsilon / B$ for all $k = m, ..., s$. Because $B \geq B_m$ and $\Big\| |\nabla \varphi_{s+1}\rangle \Big\| \geq R$, we have

$$\begin{aligned} \varphi_m - \varphi_{s+1} &\geq \eta c_1 \epsilon \sum_{k \in \mathcal{S}} r_k \\ &\geq \eta c_1 \epsilon R \\ &\geq \eta c_1 \epsilon^2 / B_m \end{aligned}$$

The second case is where $r_k > \epsilon / B$ for some $k = m, ..., s$. We instead know the general lower bound is

$$\varphi_m - \varphi_{s+1} \geq \eta c_1 \epsilon^2 / B$$

Since $\varphi > -\infty$, let $\varphi^* = \inf \varphi$. For a monotonically decreasing sequence of φ we have that

$$\varphi_k \underset{k \to \infty}{\downarrow} \varphi *$$

which implies

$$\begin{aligned} \varphi_m - \varphi^* &\geq \varphi_m - \varphi_{s+1} \\ &\geq \eta c_1 \epsilon^2 \left(\frac{1}{B} \bigwedge \frac{1}{B_m} \right) \\ &= \frac{1}{4} \Big\| |\nabla \varphi\rangle \Big\|^2 \eta c_1 \left(\frac{1}{B} \bigwedge \frac{1}{B_m} \right) \\ \implies \Big\| |\nabla \varphi\rangle \Big\| &\geq 2 \sqrt{\frac{\varphi_m - \varphi^*}{\eta c_1 \big(\frac{1}{B} \wedge \frac{1}{B_m} \big)}} \end{aligned}$$

therefore as $\varphi_k \underset{k\to\infty}{\downarrow} \varphi*$

$$\left\| |\nabla \varphi_k\rangle \right\| \to 0$$

□

C.4 ODR Efficient Implementation

Definition C.4.1. *$|1_n\rangle$ is a vector of ones length n. When pre-multiply by a diagonal matrix it compresses the matrix's diagonal elements into a vector, i.e. if* $\mathbf{A}$ *is given by*

$$\mathbf{A} \triangleq \begin{bmatrix} A_{11} & 0 & \dots & 0 \\ 0 & A_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & A_{yy} \end{bmatrix}$$

then $\mathbf{A}\,|1_n\rangle = |a\rangle = \begin{bmatrix} A_{11} & A_{22} & \dots & A_{yy} \end{bmatrix}^{\mathrm{T}}$.

Definition C.4.2. *$\odot : \mathbb{R}^y \times \mathbb{R}^y \to \mathbb{R}^y$ is the Hadamard element-wise product operator, i.e. if $|a\rangle$ and $|b\rangle$ are given by*

$$|a\rangle = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_y \end{bmatrix} \qquad |b\rangle = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_y \end{bmatrix}$$

then $|a\rangle \odot |b\rangle \triangleq \begin{bmatrix} a_1 b_1 & a_2 b_2 & \dots & a_y b_y \end{bmatrix}^{\mathrm{T}}$.

The trivial first step is to pre-multiply the weights at every iteration to form the $|\rho\rangle$ vector (5.3.4) to reuse at later stages. The naïve evaluation of the quadratic approximation

$$q\big(\delta\big) = \left\| |\rho\rangle \right\|_2^2 + \langle\rho|\,\mathbf{J}\,|\delta\rangle + \frac{1}{2}\,\langle\delta|\,\mathbf{J}^{\mathrm{T}}\mathbf{J}\,|\delta\rangle \tag{C.4.1}$$

is $\mathcal{O}\big[2m(p+m)^2 + 4m^2(p+m)^2\big]$. Recall that

$$\mathbf{J} = \begin{bmatrix} \mathbf{J}_\theta & \mathbf{J}_\epsilon \\ \mathbf{0} & \mathbf{W}_2 \end{bmatrix}$$

where $\mathbf{J}_\epsilon$ and $\mathbf{W}_2 = \text{diag}(\mathbf{w}_2)$ are diagonal. It can then be shown that (C.4.1) is equivalent to

$$q\big(\delta\big) = \big\||\rho\rangle\big\|_2^2 + \langle a \mid \delta\rangle + \big\||b\rangle\big\|_2^2 + 2\,\langle b \mid j_\epsilon \odot \delta_\epsilon\rangle + \langle c \mid \delta_\epsilon \odot \delta_\epsilon\rangle$$

where

$$|j_\epsilon\rangle = \mathbf{J}_\epsilon\,|1_n\rangle$$

$$|a\rangle = \Big[\langle\rho_1|\,\mathbf{J}_\theta \qquad \langle j_\epsilon| \odot \langle\rho_1| + \langle w_2| \odot \langle\rho_2|\Big]^{\mathrm{T}}$$

$$|b\rangle = \mathbf{J}_\theta\,|\delta_\theta\rangle$$

$$|c\rangle = |j_\epsilon\rangle \odot |j_\epsilon\rangle + |w\rangle \odot |w\rangle$$

which, if $p < m$, yields the (maximum) complexity $\mathcal{O}\big[2m(p+m) + pm^2\big]$. Running time savings can be very substantial if $m >> p$.

The bulk of the work lies in finding the direction $\big|\delta^{LM}\big\rangle$ for the total Jacobian. Because of the diagonal structure of the additional matrices, we get the following vectors without much effort from (5.3.5)

$$|e(\lambda)\rangle = \mathbf{E}^{-1}\,|1_n\rangle \quad \text{where} \quad \langle i|\,\mathbf{E}^{-1}\,|i\rangle = \frac{1}{\langle i|\,\mathbf{J}_\epsilon\,|i\rangle^2 + \langle i|\,\mathbf{W}_2\,|i\rangle^2 + \lambda} \;\forall\; i = 1, ..., m$$

$$|h(\lambda)\rangle = |1_n\rangle - |e(\lambda)\rangle \odot |j_\epsilon\rangle \odot |j_\epsilon\rangle$$

$$|\rho^*(\lambda)\rangle = |\rho_1\rangle - |e(\lambda)\rangle \odot |j_\epsilon\rangle \odot \Big(\,|j_\epsilon\rangle \odot |\rho_1\rangle + |w_2\rangle \odot |\rho_2\rangle\,\Big)$$

We define the the new matrix $\mathbf{J}^*$ with column i given by

$$\Big[\mathbf{J}^*(\lambda)\Big]\,|i\rangle \triangleq \Big(\mathbf{J}_\theta\,|i\rangle\,\Big) \odot \big|h(\lambda)^{1/2}\big\rangle$$

This definition allows us to solve for the initial QR factorisation of $\mathbf{J}$

$$\mathbf{Q}_0\mathbf{R}_0 = \mathbf{J}^*(0)$$

and the following damped matrices $\mathbf{Q}_k\mathbf{R}_k = \mathbf{Q}(\lambda_k)\mathbf{R}(\lambda_k)$ are obtained via Givens rotations applied on the original matrices and each subsequent additional row for each step.

The direction for step k is obtained by solving

$$\mathbf{R}_k^{\mathrm{T}}\mathbf{R}_k\,\big|\delta_\theta^k\big\rangle = -\mathbf{J}_\theta\,|\rho^*(\lambda_k)\rangle$$

via backward and forward substitution as in the standard LM case. Finally, the remaining direction is obtained through the expression

$$|\delta_\epsilon\rangle = -\,|e(\lambda_k)\rangle \odot \Big[\,|j_\epsilon\rangle \odot |\rho_1\rangle + |w_2\rangle \odot |\rho_2\rangle + |j_\epsilon\rangle \odot \Big(\mathbf{J}_\theta\,|\delta_\theta\rangle\,\Big)\,\Big]$$

Appendix D

Chapter 6

D.1 Baseline Stage 2 WQE Fit A Plots

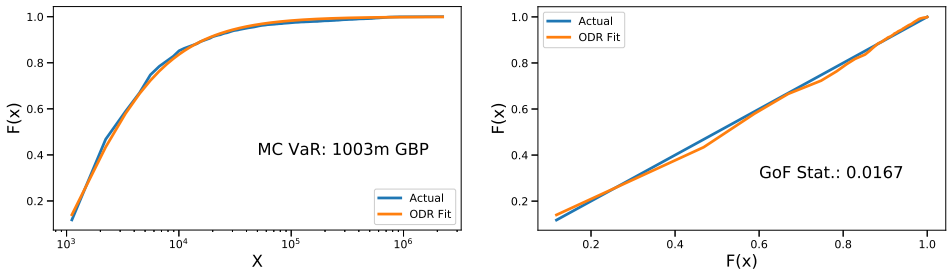


Figure D.2: $|w_1\rangle = |\omega_{10}\rangle$

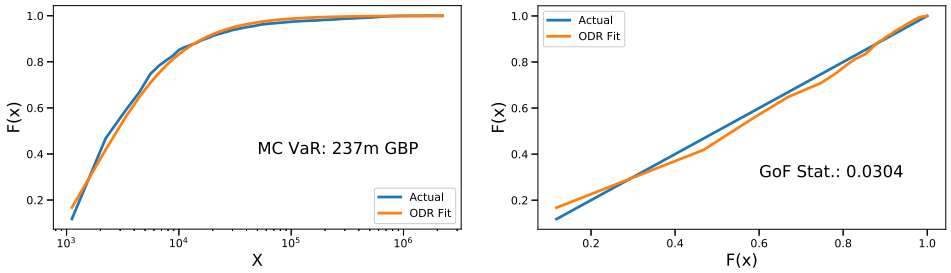


Figure D.4: $|w_1\rangle = |\omega_{40}\rangle$

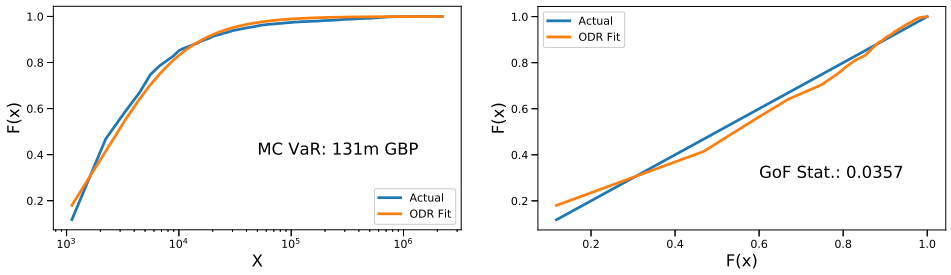


Figure D.6: $|w_1\rangle = |\omega_{80}\rangle$

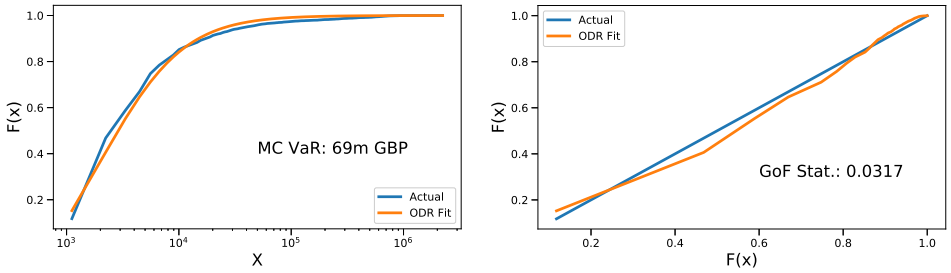


Figure D.8: $|w_1\rangle = |\omega_{500}\rangle$

D.2 Baseline Stage 2 WQE Fit B Plots

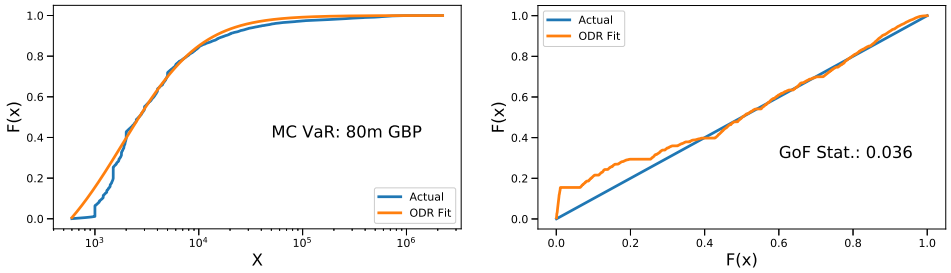


Figure D.10: $|w_1\rangle = |\omega_5\rangle$

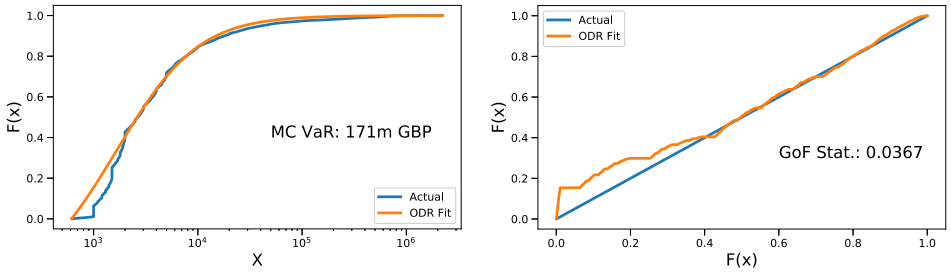


Figure D.12: $|w_1\rangle = |\omega_8\rangle$

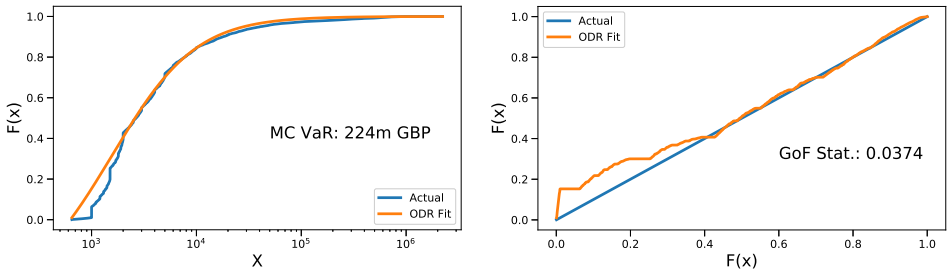


Figure D.14: $|w_1\rangle = |\omega_{10}\rangle$

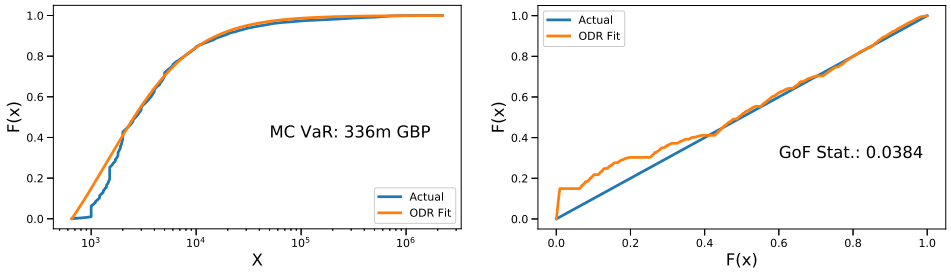


Figure D.16: $|w_1\rangle = |\omega_{15}\rangle$

Appendix E

Plain English Summary

Suppose we record losses over a period of a year. Our yearly losses will change from year to year with a different maximum loss for each year. Our work attempts to model the largest losses the institution can incur given all the yearly losses it has incurred and we make a statistically rigorous approximation of the probability of losses greater than a certain "pain" threshold.

Theorem 2.1.1 tells us that the more losses we incur, the probability distribution of the largest possible loss follows a particular distribution named the generalised extreme value distribution. Theorem 2.2.1 tells us that if we pick a cutoff point according to this "pain" threshold and forget about losses smaller than the threshold, then the probability distribution of the losses greater than this threshold follows the generalised Pareto distribution (GPD).

Aside from the statistical problems of estimating the distribution a major problem is the threshold to pick: too low and the theorem does not apply and too high leads to very few observations left. Previous works mainly focus on standard statistical techniques which yield closed-form solutions to the three parameters of the GPD, one method of which is to use three observations from the dataset (e.g. median, 25^{th}-percentile and 75^{th}-percentile) and solve for the three parameter values in terms these observations. Our work is similar to this, except we pick a very high number of losses that are equidistant from each other (fit A) or equidistant in an ordered list (e.g. the smallest, the 5^{th} smallest, the 10^{th} smallest,..., largest loss) (fit B). We then match the curve we get from the dataset and the estimated curve using numerical algorithms.

The naïve procedure does not yield a great outcome. We know for a fact that we need a higher threshold for the losses for theorem 2.2.1 to apply and the procedure matches all the observations to the estimated curve as if they are all *equally* important. Optimisation algorithms stop when the *overall* fit is good enough, which means that they all try to attack points that violate the dataset we have given it the most regardless of their sizes. This means that if the body of the estimated curve is far away from the body of the curve drawn out by the losses, the algorithms will attempt to reduce these points first. We see this result in 6.2 and table 6.1 where the 95^{th} largest loss is about the same as the one given by the estimated curve but the 99.9^{th} largest loss is only about a quarter of the estimated value. This discrepancy leads to predictions of possible future losses in

the billions of GBP that are not realistic given the dataset.

Our solution to this problem is to penalise the discrepancies between the largest losses in the dataset and their counterparts in the estimated curve. This inflated penalisation forces the algorithm to work on these "tail" values first then focus on the smaller values when there is nothing to gain from minimising the tail values anymore. This is achieved since optimisation algorithms always look for the most efficient reduction, thus if tail violations are large they attack those "low hanging fruit" values first. The result of this procedure does not explicitly give an optimal threshold. This number depends greatly on the dataset and can be prohibitively high seeing as the theorem holds approximately when there are an infinite number of losses. What the procedure does provide, however, is a curve that fits the larger losses very well, which one can use to compare with the actual losses to see where the two curves start to diverge significantly which is where the threshold needs to be at least at high as.

Bibliography

[1] J Arthur Greenwood, Jurate Landwehr, NC Matalas, and J.R.M. Wallis, *Probability weighted moments: Definition and relation to parameters of several distributions expressable in inverse form*, Water Resources Research (197905), 1049–1054.

[2] A. A. Balkema and L. de Haan, *Residual life time at great age*, Ann. Probab. **2** (197410), no. 5, 792–804.

[3] Bank of International Settlements, *Basel ii: International convergence of capital measurement and capital standards: a revised framework*, 2004.

[4] Paul T. Boggs, Richard H. Byrd, and Robert B. Schnabel, *A stable and efficient algorithm for nonlinear orthogonal distance regression*, SIAM J. Sci. Stat. Comput. **8** (November 1987), no. 6, 1052–1078.

[5] Charles G. Broyden, *The convergence of a class of double-rank minimization algorithms 1. General considerations*, IMA Journal of Applied Mathematics **6** (1970), no. 1, 76–90.

[6] Mark M. Carhart, *On persistence in mutual fund performance*, The Journal of Finance **52** (1997), no. 1, 57–82, available at `https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-6261.1997.tb03808.x`.

[7] Enrique Castillo and Ali S. Hadi, *Fitting the generalized pareto distribution to data*, Journal of the American Statistical Association **92** (1997), no. 440, 1609–1620, available at `https://doi.org/10.1080/01621459.1997.10473683`.

[8] Myriam Charras-Garrido and Pascal Lezaud, *Extreme Value Analysis : an Introduction*, Journal de la Societe Française de Statistique **154** (2013), no. 2, pp 66–97.

[9] Coindesk, *Bitcoin price (btc)*, 2019.

[10] A. R. Conn, N. I. M. Gould, and Ph. L. Toint, *Convergence of quasi-newton matrices generated by the symmetric rank one update*, Math. Program. **50** (March 1991), no. 2, 177–195.

[11] Laurens de Haan and Ana Ferreira, *Extreme value theory: An introduction (springer series in operations research and financial engineering)*, 1st Edition., Springer, 2010.

[12] Laurens de Haan and Sidney Resnick, *On asymptotic normality of the hill estimator*, Communications in Statistics. Stochastic Models **14** (1998), no. 4, 849–866, available at `https://doi.org/10.1080/15326349808807504`.

[13] P. de Zea Bermudez and M. A. Amaral Turkman, *Bayesian approach to parameter estimation of the generalized pareto distribution*, Test **12** (2003Jun), no. 1, 259–277.

[14] Jean Diebolt, Mhamed-Ali El-Aroui, Myriam Garrido, and Stéphane Girard, *Quasi-conjugate bayes estimates for gpd parameters and application to heavy tails modelling*, Extremes **8** (2005Jun), no. 1, 57–78.

[15] Roger Fletcher, *A new approach to variable metric algorithms*, The Computer Journal **13** (1970), no. 3, 317–322.

[16] FRED, *3-month london interbank offered rate (libor), based on u.s. dollar* (FRED, ed.), Board of Governors of the Federal Reserve System (US), 2019.

[17] ———, *Crude oil prices: Brent - europe* (FRED, ed.), Board of Governors of the Federal Reserve System (US), 2019.

[18] ———, *Ice bofaml us high yield master ii option-adjusted spread* (FRED, ed.), Board of Governors of the Federal Reserve System (US), 2019.

[19] ———, *U.s. / u.k. foreign exchange rate* (FRED, ed.), Board of Governors of the Federal Reserve System (US), 2019.

[20] K. French, *Momentum factor (mom) [daily]*, Vol. 4, 1933.

[21] Carl Friedrich Gauß, *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*, ETH-Bibliothek Zürich, Zurich, Switzerland, 1809.

[22] V. GLIVENKO, *Sulla determinazione empirica delle leggi di probabilita*, Gion. Ist. Ital. Attauri. **4** (1933), 92–99.

[23] Donald Goldfarb, *A family of variable metric updates derived by variational means*, Mathematics of Computation **24** (1970), no. 109, 23–26.

[24] Gene H. Golub and Charles F. van Loan, *Matrix computations*, Fourth, JHU Press, 2013.

[25] Bruce M. Hill, *A simple general approach to inference about the tail of a distribution*, Ann. Statist. **3** (197509), no. 5, 1163–1174.

[26] J. R. M. Hosking and J. F. Wallis, *Parameter and quantile estimation for the generalized pareto distribution*, Technometrics **29** (September 1987), no. 3, 339–349.

[27] James Pickands III, *Statistical inference using extreme order statistics*, Ann. Statist. **3** (197501), no. 1, 119–131.

[28] Kenneth Levenberg, *A method for the solution of certain non-linear problems in least squares*, Quarterly of applied mathematics **2** (1944), no. 2, 164–168.

[29] P. Lévy, *Calcul des probabilités*, PCMI collection, Gauthier-Villars, 1925.

[30] Donald W. Marquardt, *An algorithm for least-squares estimation of nonlinear parameters*, SIAM Journal on Applied Mathematics **11** (1963), no. 2, 431–441.

[31] Peter Mitic, *Improved goodness-of-fit measures*, Journal of Operational Risk **10** (2015), no. 1.

[32] J. A. Nelder and R. Mead, *A Simplex Method for Function Minimization*, The Computer Journal **7** (196501), no. 4, 308–313, available at `http://oup.prod.sis.lan/comjnl/article-pdf/7/4/308/1013182/7-4-308.pdf`.

[33] J. Nocedal and S. Wright, *Numerical optimization*, Springer Series in Operations Research and Financial Engineering, Springer New York, 2006.

[34] Christopher C. Paige and Michael A. Saunders, *Lsqr: An algorithm for sparse linear equations and sparse least squares*, ACM Trans. Math. Software (1982), 43–71.

[35] Myung Hyun Park and Joseph H.T. Kim, *Estimating extreme tail risk measures with generalized pareto distribution*, Computational Statistics Data Analysis **98** (2016), 91 –104.

[36] R. Quandt, *Old and new methods of estimation and the pareto distribution*, Metrika: International Journal for Theoretical and Applied Statistics **10** (1966), no. 1, 55–82.

[37] R. Tyrrell Rockafellar and Stanislav Uryasev, *Optimization of conditional value-at-risk*, Journal of Risk **2** (2000), 21–41.

[38] David F. Shanno, *Conditioning of quasi-Newton methods for function minimization*, Mathematics of Computation **24** (1970), no. 111, 647–656.

[39] Vijay Singh and H GUO, *Parameter estimation for 3-parameter generalized pareto distribution by the principle of maximum entropy (pome)*, Hydrological Sciences Journal **40** (199504).

[40] A. N. Tikhonov and V. Y. Arsenin, *Solution of ill-posed problems*, John Wiley & Sons, 1977.

[41] B V. Gnedenko, *On the limiting distribution of the maximum term in a random series*, 199201.

[42] Aad W. van der Vaart and Jon A. Wellner, *Glivenko-cantelli theorems*, Springer New York, New York, NY, 1996.

[43] R. Von Mises, *La distribution de la plus grandede nvaleurs*, Rev., Math, Union Interbalcanique **1** (1965), no. 1, 141–160.

[44] Philip Wolfe, *Convergence conditions for ascent methods. ii*, SIAM Rev. **13** (April 1971), no. 2, 185–188.

[45] Xu Zhao, Zhongxian Zhang, Weihu Cheng, and Pengyue Zhang, *A new parameter estimator for the generalized pareto distribution under the peaks over threshold framework*, Mathematics **7** (201905), 406.