# Practical likelihood-based inference for the univariate generalized hyperbolic distribution

By

Arnold van Wyk

Submitted in partial fulfillment of the requirements for the degree

MSc (Advanced Data Analytics)

in the

Faculty of Natural and Agricultural Sciences

at the

University of Pretoria

Supervisors: Prof A. Azzalini and Prof A. Bekker

November 2021

# Declaration

I, Arnold van Wyk declare that the dissertation, which I hereby submit for the degree MSc Advanced Data Analytics at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

Signature

Date

# Acknowledgements

- I would firstly like to thank everyone that played a role in this journey. My friends, family and loved ones, you have my sincerest and warmest gratitude for all your time, effort and patience.

- Professor Azzalini: It was an absolute honour and a privilege to work under your supervision. Thank you for not only affording me the opportunity to learn from your expertise, but for paving the road on what was such a fruitful journey both academically and otherwise.

- Professor Bekker: Thank you for embarking on this journey with me, and for all your support and guidance along the way. Through all the ups and downs of the past two years, our meetings were always such a pleasure, and I can confidently say that my love of statistics has only grown under your guidance and supervision.

- This work was based on research supported in part by the National Research Foundation (NRF) of South Africa, SARChI Research Chair UID: 71199; Ref.: SRUG190308422768 grant No. 120839, and STATOMET at the Department of Statistics at the University of Pretoria. The opinions expressed and conclusions arrived at are those of the authors and are not necessarily attributed to the NRF.

# Abstract

Maximum likelihood estimation is a powerful estimation tool that is widely used to fit models to data. In this study, the behaviour of the log-likelihood function, and the ensuing impact on the maximum likelihood estimation process is explored. This exploration is conducted using the univariate generalized hyperbolic distribution, a highly flexible distribution with tail properties making it desirable as a model for financial returns data. The study aims to explore potential issues that may present when estimating the parameters of such flexible distributions, especially those stemming from the behaviour of the log-likelihood function. Different numerical methods are applied to showcase the effect of not only the shape and behaviour of the log-likelihood function, but the structure of the parameters themselves on the outcome of the estimation process. Application to real-world financial data shows that the behaviour of the log-likelihood function has a significant impact on the estimation outcome, and that an understanding of these components is fundamental to the success of the process of estimation.

**Keywords**: Generalized hyperbolic distribution; maximum likelihood estimation; log-likelihood; EM algorithm; profile likelihood.

# Contents

# List of Figures

# List of Tables

# Nomenclature

| | |
|---|---|
| pdf | probability density function |
| cdf | cumulative distribution function |
| mgf | moment generating function |

| | |
|---|---|
| GH | Generalized hyperbolic distribution |
| VG | Variance-gamma distribution |
| ALap | Asymmetric Laplace distribution |
| *t* | Student's *t* distribution |
| AC | Asymmetric Cauchy distribution |
| Ca | Cauchy distribution |
| Hyp | Hyperbolic distribution |
| NIG | Normal inverse Gaussian distribution |
| HA*t* | Hyperbolic asymmetric Student's *t* distribution |

| | |
|---|---|
| $f_X(x)$ | pdf of random variable $X$ |

$M_X(t)$ moment generating function of random variable $X$

$E(X)$ expected value of random variable $X$

$V(X)$ variance of random variable $X$

$\mu_n(X)$ $n^{th}$ central moment of random variable $X$

$K_\nu(\cdot)$ Modified Bessel function of the third kind with index $\nu$

$\Gamma(\cdot)$ Gamma function

# Chapter 1

# Introduction

## 1.1 Motivation and background

Maximum likelihood estimation is one of the most predominantly used methods of parameter estimation. It involves finding estimates such that the observed data are most likely to occur under the predefined statistical model, and this is achieved by maximizing the likelihood function. The likelihood function measures the support provided by the data for each possible combination of the underlying parameters. The likelihood function is maximized as this maximum represents the parameter values that are most likely given the underlying dataset.

Although the method of maximum likelihood is a strong means of parameter estimation, it is not without its shortfalls. The susceptibility of the standard minimization algorithms to restricted parameter spaces and the flatness of the likelihood function are an example of this (see Prause, 1997). This problem is amplified when dealing with flexible distribution classes (see Ley, 2015), as near non-identifiability, especially that stemming from a flat log-likelihood function is a common issue in this case. This creates a situation where two vastly different parameter estimates can result in the same fitted distribution of the data. In a broad sense, exploration of these problems is a necessity, especially when the underlying distribution possesses properties that may hinder the overall quality of the estimation process, such as the the lack of a closed form derivative or a badly behaved likelihood function.

The generalized hyperbolic distribution was first developed by Barndorff-Nielsen (1977) to

model the mass-size distribution of sand particles. This model emanated from a geostatistical study, and results from a normal mean-variance mixture, where the mixture variable emanates from a generalized inverse Gaussian distribution.

The generalized hyperbolic distribution is considered a flexible distribution. By convention, this refers to distributions that allow substantial variation of their behaviour when the underlying parameters span their admissible range. The formal construction of these distributions are represented by the Pearson system of curves, whereby the pdf is regulated by four parameters, thus allowing for greater variation in terms of measures of skewness and of kurtosis. This naturally provides a greater flexibility, than for example the normal distribution, where only location and scale can be varied.

The generalized hyperbolic distribution is desirable for its semi-heavy tails, which are generally heavier than those of the normal distribution, allowing it to better accommodate extreme values. It is because of this property that the generalized hyperbolic distribution has become rather popular in the field of econometrics, particularly in the prediction of financial markets and in risk analysis (see Eberlein et al. 1995 and Puig and Stephens 2001).

What sets the generalized hyperbolic distribution apart from the hyperbolic distribution, and the flexible distribution framework as a whole, is the introduction of the index parameter $\lambda$, giving us five parameters in total. This results in this distribution being a superclass of numerous flexible distributions, often referred to as subfamilies, which include but are not limited to: the variance-gamma distribution, the Laplace distribution, the Student's $t$ distribution, and naturally, the hyperbolic distribution (see Paolella, 2007, pp. 317-326). This adds an extra layer of flexibility, and it is of interest whether this flexibility has an impact on the behaviour of the log-likelihood function, and possibly on the quality of the maximum likelihood estimation process as a whole.

There is clear evidence in literature of potential issues when estimating the parameters of the generalized hyperbolic distribution by means of the method of maximum likelihood. Challenges resulting from a flat likelihood function, which is believe is in large part caused by the index parameter $\lambda$ are discussed in Snoussi and Idier (2006), Prause (1999), and Barndorff-Nielsen and Blaesild (1981). As previously stated, the $\lambda$ parameter is largely responsible for the existence of the various subclasses, begging the question if there is possible over-parametrization, or if

the nature of flexible distributions has an inherent, negative impact on the maximum likelihood estimation method. They also report a possible identifiability issue, with reference to a specific case where the normal inverse Gaussian subfamily ($\lambda$ = -0.5) and the Hyperbolic subfamily ($\lambda$ = 1) are nearly identical. This is potentially alarming as it could greatly decrease the validity, as well as the inferential power of the resulting estimates.

## 1.2 Contributions

What is apparent from the literature, is that although there is some discussion, albeit brief, on the impacts of the above issues, there does not seem to be a sufficient, in-depth analysis on where or not the steps taken in any way ensure that the quality of the estimates are of an acceptable standard. This opens up the opportunity for further study into these behaviours and potential issues and forms the basis of the motivation for this study as a whole.

The primary aim of this study is thus, to explore the behaviour of the log-likelihood function of the generalized hyperbolic distribution, as well as investigate potential solutions to problems stemming from this particular log-likelihood function. The goal is to undertake a detailed exploratory analysis on the potential problems that may be present when using the maximum likelihood as a means of estimating parameters, especially with flexible distributions. The idea is to create a platform that highlights these issues, such that potential solutions and/or recommendations can be proposed, keeping in mind that the primary aim is to explore the potential behaviours of the log-likelihood function in the context of maximum likelihood estimation. That being said, it is still of interest to identify the existence as well as extent with which these issues may pervade the method and whether this warrants greater care, or even certain steps when performing the estimation. Although there are mentions of potential issues stemming from the flatness of the log-likelihood function, as of writing there is an apparent lack of active methods to deal with these issues.

## 1.3 Dissertation outline

- **Chapter 2**: provides an in depth overview of the generalized hyperbolic distribution. A derivation of the distribution is provided for the major parameterizations, as well as a com-

parison of the subclasses of the generalized hyperbolic distribution. Finally, the role of the parameters, some alternative parameterizations, and some useful properties of the generalized hyperbolic distribution are provided.

- **Chapter 3**: provides an overview of the method of maximum likelihood, with specific departure to the profile likelihood that will be used in the model fitting process.

- **Chapter 4**: gives a breakdown and description of the Numerical methods that will be used to fit the Generalized Hyperbolic distribution to the data.

- **Chapter 5**: contains a full discussion and commentary on the exploratory process of fitting the generalized hyperbolic distribution to data. Any and all findings pertaining to the behaviour of the log-likelihood function as investigated in the study, as well the estimation methods are also discussed. The estimation results are both discussed and analysed with some commonly used goodness-of-fit methods. A simulation study is included to allow for better discussion, as well as comparison of the findings.

- **Chapter 6**: includes a summary of the findings, as well as an outline of the significance of performing such a study and the challenges/limitations thereof. Also included are some concluding remarks and potential departures or expansion.

- **Appendix A**: gives a brief overview of some continuous normal mixture mechanics that form the basis of the derivation of the generalized hyperbolic distribution.

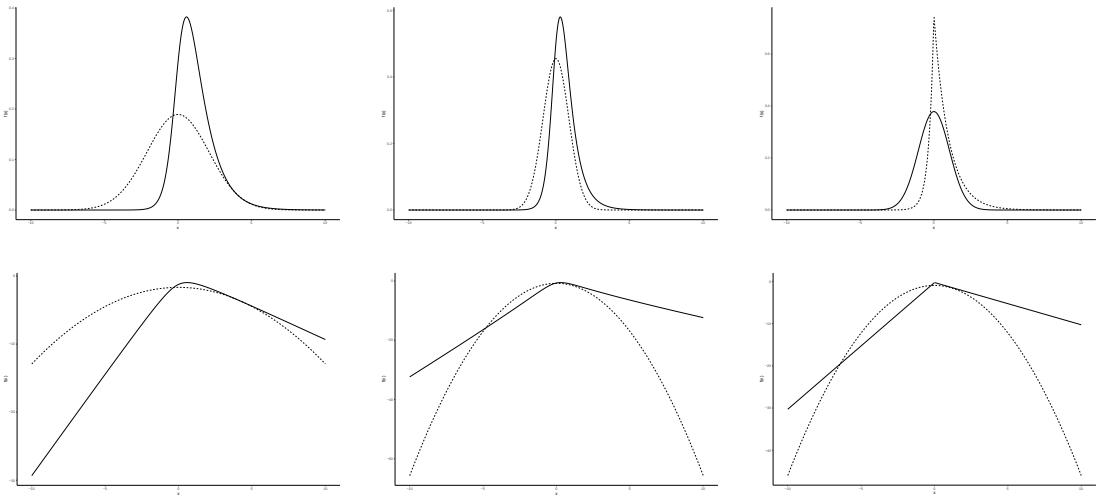
- **Appendix B**: defines the modified Bessel function of the third kind, and provides some crucial results needed for the derivation of the generalized hyperbolic distribution and its subclasses.

# Chapter 2

# The Generalized Hyperbolic Distribution

The generalized hyperbolic distribution was first introduced by Barndorff-Nielsen (1977), and was initially used to model the mass-size distributions of particles of wind blown sands from beaches and dunes (see Bagnold 1941). It has since gained traction as an alternative to the normal distribution when modelling financial data due to its desirable tail properties and flexibility (see Prause 1999; Bibby and Sørensen 2003; Eberlein et al. 1995; Küchler et al. 1999; Behr and Pötter 2009). Many of these papers deal with the multivariate case of the GH distribution, but for the work that follows we only consider the univariate case. The various subclasses of the GH distribution include: the hyperbolic distribution, the Student's $t$ distribution, the variance gamma distribution, and the Laplace distribution.

One of the primary appeals of the generalized hyperbolic distribution lies in its semi-heavy tails. This makes it rather useful when extreme values require a greater representation, and is one of the reasons for the generalized hyperbolic distributions popularity in modelling financial markets (see Eberlein et al. 1995 and Puig and Stephens 2001). Figure 2.1 depicts the pdfs as well as log-pdfs for the hyperbolic, normal inverse Gaussian, and variance-gamma subclasses.



(a) (b) (c)

Figure 2.1: Examples of GH distributions. (a) Hyperbolic subclass. (b) Normal inverse Gaussian subclass. (c)Variance-gamma subclass. In each instance the pdf is on the top row and the log-pdf on the bottom row. The dashed line corresponds with the normal distribution with the same mean and variance.

## 2.1 Definition, parameters and general formulae

**Definition 2.1.1.** (Jørgensen 1982, p. 1) The random variable $W$ is said to follow a generalized inverse Gaussian distribution (GIG), denoted $W \sim \text{GIG}(\lambda, \chi, \psi)$, if the pdf of $W$ is given by

$$f_W(w; \lambda, \chi, \psi) = \frac{\left(\frac{\psi}{\chi}\right)^{\frac{\lambda}{2}}}{2\, K_\lambda(\chi, \psi)} w^{\lambda-1} e^{-\frac{1}{2}(\chi w^{-1} + \psi w)} \qquad (w > 0) \qquad (2.1.1)$$

where $K_\lambda(.)$ is the modified Bessel function of the third kind with index $\lambda$ (see B.1), and $\lambda, \chi, \psi$ are the parameters with parameter space:

$$\begin{aligned} &\chi \geq 0, \quad \psi > 0, \qquad \text{if } \lambda > 0, \\ &\chi > 0, \quad \psi > 0, \qquad \text{if } \lambda = 0, \\ &\chi > 0, \quad \psi \geq 0, \qquad \text{if } \lambda < 0. \end{aligned}$$

**Definition 2.1.2.** Let $X$ be a random variable such that either of the equivalent representations (A.3) and (A.4) hold, i.e. $X$ has a normal variance-mean mixture distribution. If the mixing variable $W$ belongs to the generalized inverse Gaussian distribution, i.e. $W \sim \text{GIG}(\lambda, \chi, \psi)$, then $X$ is said to follow a generalized hyperbolic distribution.

**Theorem 2.1.1.** Let $X$ follow a generalized hyperbolic distribution as in Definition 2.1.2, then the pdf of $X$ is given by

$$\begin{aligned} f_X(x; \lambda, \psi, \beta, \chi, \mu) = a(\lambda, \psi, \beta, \chi)\, e^{\beta(x-\mu)} \\ \times K_{\lambda-\frac{1}{2}}\left(\sqrt{(\chi + (x-\mu)^2)(\psi+\beta^2)}\right) \\ \times \left((\chi + (x-\mu)^2)(\psi+\beta^2)\right)^{\frac{1}{2}(\lambda-\frac{1}{2})} \qquad (x \in \mathbb{R}) \end{aligned} \tag{2.1.2}$$

where $\lambda, \psi, \beta, \chi$ and $\mu$ are the parameters, the expression $K(.)$ denotes the modified Bessel function of the third kind as defined in (B.1), and $a(\lambda, \psi, \beta, \chi)$ is a norming constant given by

$$a(\lambda, \psi, \beta, \chi) = \frac{\left(\frac{\chi}{\psi}\right)^{-\frac{\lambda}{2}} (\psi+\beta^2)^{\frac{1}{2}-\lambda}}{\sqrt{2\pi}\, K_\lambda(\sqrt{\chi\psi})}. \tag{2.1.3}$$

The domain of variation of the parameter space is given by

$$\begin{aligned} &\chi \geq 0, \quad \psi > 0, \qquad \text{if } \lambda > 0, \\ &\chi > 0, \quad \psi > 0, \qquad \text{if } \lambda = 0, \\ &\chi > 0, \quad \psi \geq 0, \qquad \text{if } \lambda < 0 \end{aligned}$$

where $\beta, \mu \in \mathbb{R}$.

**Proof.** Let $X \in \mathbb{R}$ be a random variable such that (A.3) and (A.4) hold and assume $W \sim$

GIG$(\lambda, \chi, \psi)$. Then the pdf of $X$ is given by

$$
\begin{aligned}
& f_X(x; \lambda, \psi, \beta, \chi, \mu) \\
&= \int_0^\infty f_{X|W}(x|w)\ f_W(w; \lambda, \chi, \psi) dw \\
&= \int_0^\infty \frac{1}{\sqrt{2\pi w}}\ e^{-\frac{1}{2}\left(\frac{(x-(\mu+\beta w))^2}{w}\right)}\ \frac{\left(\frac{\psi}{\chi}\right)^{\frac{\lambda}{2}}}{2\,K_\lambda(\chi,\psi)} w^{\lambda-1} e^{-\frac{1}{2}\left(\chi w^{-1}+\psi w\right)}\ dw \qquad \text{(from (2.1.1))} \\
&= \int_0^\infty \frac{1}{\sqrt{2\pi w}}\ e^{-\frac{1}{2}\left(\frac{(x-(\mu+\beta w))^2}{w}\right)}\ \frac{1}{k_\lambda(\chi,\psi)}\ w^{\lambda-1}\ e^{-\frac{1}{2}(\chi w^{-1}+\psi w)}\ dw \qquad \text{(from (B.10))} \\
&= \frac{1}{\sqrt{2\pi}\ k_\lambda(\chi,\psi)} \int_0^\infty w^{(\lambda-\frac{1}{2})-1}\ e^{-\frac{1}{2}\left(w^{-1}((x-\mu)-\beta w))^2\right)}\ e^{-\frac{1}{2}(\chi w^{-1}+\psi w)}\ dw \\
&= \frac{1}{\sqrt{2\pi}\ k_\lambda(\chi,\psi)} \int_0^\infty w^{(\lambda-\frac{1}{2})-1}\ e^{-\frac{1}{2}\left(w^{-1}((x-\mu)^2-2\beta w(x-\mu)+\beta^2 w^2))\right)}\ e^{-\frac{1}{2}(\chi w^{-1}+\psi w)}\ dw \\
&= \frac{1}{\sqrt{2\pi}\ k_\lambda(\chi,\psi)} \int_0^\infty w^{(\lambda-\frac{1}{2})-1}\ e^{-\frac{1}{2}\left(w^{-1}((x-\mu)^2+\chi)+w(\beta^2+\psi)-2\beta(x-\mu)\right)}\ dw \\
&= \frac{1}{\sqrt{2\pi}\ k_\lambda(\chi,\psi)}\ e^{\beta(x-\mu)} \int_0^\infty w^{(\lambda-\frac{1}{2})-1}\ e^{-\frac{1}{2}\left(w^{-1}(\chi+(x-\mu)^2)+w(\psi+\beta^2)\right)}\ dw \\
&= \frac{1}{\sqrt{2\pi}\ k_\lambda(\chi,\psi)}\ e^{\beta(x-\mu)}\ k_{\lambda-\frac{1}{2}}(\chi+(x-\mu)^2, \psi+\beta^2) \qquad \text{(from (B.8))}. \qquad (2.1.4)
\end{aligned}
$$

Using result (B.10), (2.1.4) can be rewritten as

$$
\begin{aligned}
f_X(x; \lambda, \psi, \beta, \chi, \mu) &= \frac{e^{\beta(x-\mu)}}{\sqrt{2\pi}\ \left(\frac{\chi}{\psi}\right)^{\frac{\lambda}{2}} K_\lambda(\sqrt{\chi\psi})} \left(\frac{\chi+(x-\mu)^2}{\psi+\beta^2}\right)^{\frac{1}{2}(\lambda-\frac{1}{2})} \\
&\quad \times K_{\lambda-\frac{1}{2}}\left(\sqrt{(\chi+(x-\mu)^2)(\psi+\beta^2)}\right) \\
&= a(\lambda,\psi,\beta,\chi)\ e^{\beta(x-\mu)} \\
&\quad \times\ K_{\lambda-\frac{1}{2}}\left(\sqrt{(\chi+(x-\mu)^2)(\psi+\beta^2)}\right) \\
&\quad \times \left((\chi+(x-\mu)^2)(\psi+\beta^2)\right)^{\frac{1}{2}(\lambda-\frac{1}{2})}
\end{aligned}
$$

where $a(\lambda, \psi, \beta, \chi)$ is the norming constant as given in (2.1.3) □

The above parameterization is the natural parameterization that arises from the mixture representation of the GH distribution. The parameterization that follows was first proposed in Barndorff-Nielsen (1978), and has since become the dominant parameterization used when working with the GH distribution. This formulation is a simple transformation of (2.1.2), obtained by setting $\chi = \delta^2$

and $\psi = \alpha^2 - \beta^2$.

**Theorem 2.1.2.** Let $\chi = \delta^2$ and $\psi = \alpha^2 - \beta^2$, then $X$ is said to follow a generalized hyperbolic distribution with pdf given by

$$\begin{aligned} \tilde{f}_X(x; \lambda, \alpha, \beta, \delta, \mu) &= a(\lambda, \alpha, \beta, \delta)\, e^{\beta(x-\mu)} \\ &\quad \times K_{\lambda-\frac{1}{2}}\left(\alpha\sqrt{\delta^2 + (x-\mu)^2}\right) \\ &\quad \times \left(\delta^2 + (x-\mu)^2\right)^{\frac{1}{2}(\lambda-\frac{1}{2})} \qquad (x \in \mathbb{R}) \end{aligned} \tag{2.1.5}$$

where $\lambda, \alpha, \beta, \delta$ and $\mu$ are the parameters, the expression $K(.)$ denotes the modified Bessel function of the third kind as defined in (B.1), and $a_\lambda(\alpha, \beta, \delta)$ is a norming constant given by

$$a(\lambda, \alpha, \beta, \delta) = \frac{(\alpha^2 - \beta^2)^{\frac{\lambda}{2}}}{\sqrt{2\pi}\ \alpha^{\lambda-\frac{1}{2}}\ \delta^\lambda\ K_\lambda\left(\delta\sqrt{\alpha^2 - \beta^2}\right)}. \tag{2.1.6}$$

The domain of variation of the parameter space is given by

$$\begin{aligned} &\alpha > 0, \quad |\beta| < \alpha, \quad \delta \geq 0, \qquad \text{if } \lambda > 0, \\ &\alpha > 0, \quad |\beta| < \alpha, \quad \delta > 0, \qquad \text{if } \lambda = 0, \\ &\alpha \geq 0, \quad |\beta| \leq \alpha, \quad \delta > 0, \qquad \text{if } \lambda < 0 \end{aligned}$$

where $\mu \in \mathbb{R}$.

**Proof.** Let $X \in \mathbb{R}$ be a random variable such that (A.3) and (A.4) hold and assume $W \sim$

GIG$(\lambda, \delta^2, \alpha^2 - \beta^2)$. Then in a similar fashion as Theorem 2.1.1 the pdf of $X$ is given by

$$
\begin{aligned}
& f_X(x; \lambda, \alpha, \beta, \delta, \mu) \\
&= \int_0^\infty f_{X|W}(x|w)\, f_W(w; \lambda, \delta^2, \alpha^2 - \beta^2) dw \\
&= \int_0^\infty \frac{1}{\sqrt{2\pi w}}\, e^{-\frac{1}{2}\left(\frac{(x-(\mu+\beta w))^2}{w}\right)}\, \frac{\left(\frac{\alpha^2-\beta^2}{\delta^2}\right)^{\frac{\lambda}{2}}}{2\, K_\lambda(\delta^2, \alpha^2 - \beta^2)} w^{\lambda-1} e^{-\frac{1}{2}\left(\delta^2 w^{-1} + \alpha^2 - \beta^2 w\right)}\, dw \qquad \text{(from (2.1.1))} \\
&= \int_0^\infty \frac{1}{\sqrt{2\pi w}}\, e^{-\frac{1}{2}\left(\frac{(x-(\mu+\beta w))^2}{w}\right)}\, \frac{1}{k_\lambda(\delta^2, \alpha^2 - \beta^2)}\, w^{\lambda-1}\, e^{-\frac{1}{2}(\delta^2 w^{-1} + (\alpha^2 - \beta^2) w)}\, dw \qquad \text{(from (B.10))} \\
&= \frac{1}{\sqrt{2\pi}\, k_\lambda(\delta^2, \alpha^2 - \beta^2)} \int_0^\infty w^{(\lambda-\frac{1}{2})-1}\, e^{-\frac{1}{2}\left(w^{-1}((x-\mu)-\beta w))^2\right)}\, e^{-\frac{1}{2}(\delta^2 w^{-1} + (\alpha^2-\beta^2) w)}\, dw \\
&= \frac{1}{\sqrt{2\pi}\, k_\lambda(\delta^2, \alpha^2 - \beta^2)} \int_0^\infty w^{(\lambda-\frac{1}{2})-1}\, e^{-\frac{1}{2}\left(w^{-1}((x-\mu)^2 - 2\beta w(x-\mu) + \beta^2 w^2))\right)}\, e^{-\frac{1}{2}(\delta^2 w^{-1} + (\alpha^2-\beta^2) w)}\, dw \\
&= \frac{1}{\sqrt{2\pi}\, k_\lambda(\delta^2, \alpha^2 - \beta^2)} \int_0^\infty w^{(\lambda-\frac{1}{2})-1}\, e^{-\frac{1}{2}\left(w^{-1}((x-\mu)^2 + \delta^2) + w(\beta^2 + \alpha^2 - \beta^2) - 2\beta(x-\mu)\right)}\, dw \\
&= \frac{1}{\sqrt{2\pi}\, k_\lambda(\delta^2, \alpha^2 - \beta^2)}\, e^{\beta(x-\mu)} \int_0^\infty w^{(\lambda-\frac{1}{2})-1}\, e^{-\frac{1}{2}\left(w^{-1}(\delta^2 + (x-\mu)^2) + w(\alpha^2)\right)}\, dw \\
&= \frac{1}{\sqrt{2\pi}\, k_\lambda(\delta^2, \alpha^2 - \beta^2)}\, e^{\beta(x-\mu)}\, k_{\lambda-\frac{1}{2}}(\delta^2 + (x-\mu)^2, \alpha^2) \qquad \text{(from (B.8))}
\end{aligned}
\tag{2.1.7}
$$

Using result (B.10), (2.1.7) can be rewritten as

$$
\begin{aligned}
f_X(x; \lambda, \alpha, \beta, \delta, \mu) &= \frac{e^{\beta(x-\mu)}}{\sqrt{2\pi}\, 2\left(\frac{\delta^2}{\alpha^2-\beta^2}\right)^{\frac{\lambda}{2}} K_\lambda(\delta\sqrt{\alpha^2 - \beta^2})}\, 2\left(\frac{\delta^2 + (x-\mu)^2}{\alpha^2}\right)^{\frac{1}{2}(\lambda-\frac{1}{2})} \\
&\quad \times K_{\lambda-\frac{1}{2}}\left(\alpha\sqrt{\delta^2 + (x-\mu)^2}\right) \\
&= a(\lambda, \alpha, \beta, \delta)\, e^{\beta(x-\mu)} \\
&\quad \times K_{\lambda-\frac{1}{2}}\left(\alpha\sqrt{\delta^2 + (x-\mu)^2}\right) \\
&\quad \times \left(\delta^2 + (x-\mu)^2\right)^{\frac{1}{2}(\lambda-\frac{1}{2})}
\end{aligned}
\tag{2.1.8}
$$

where $a(\lambda, \alpha, \beta, \delta)$ is a norming constant as given in (2.1.6). □

### 2.1.1 Parameters

A more in depth description of the parameters is now given (see Paolella, 2007, pp.329-330).

$\lambda$: Commonly seen as the index parameter as it gives rise to many distinctions amongst the subfamilies of the generalized hyperbolic distribution. It also influences the shape of the pdf.

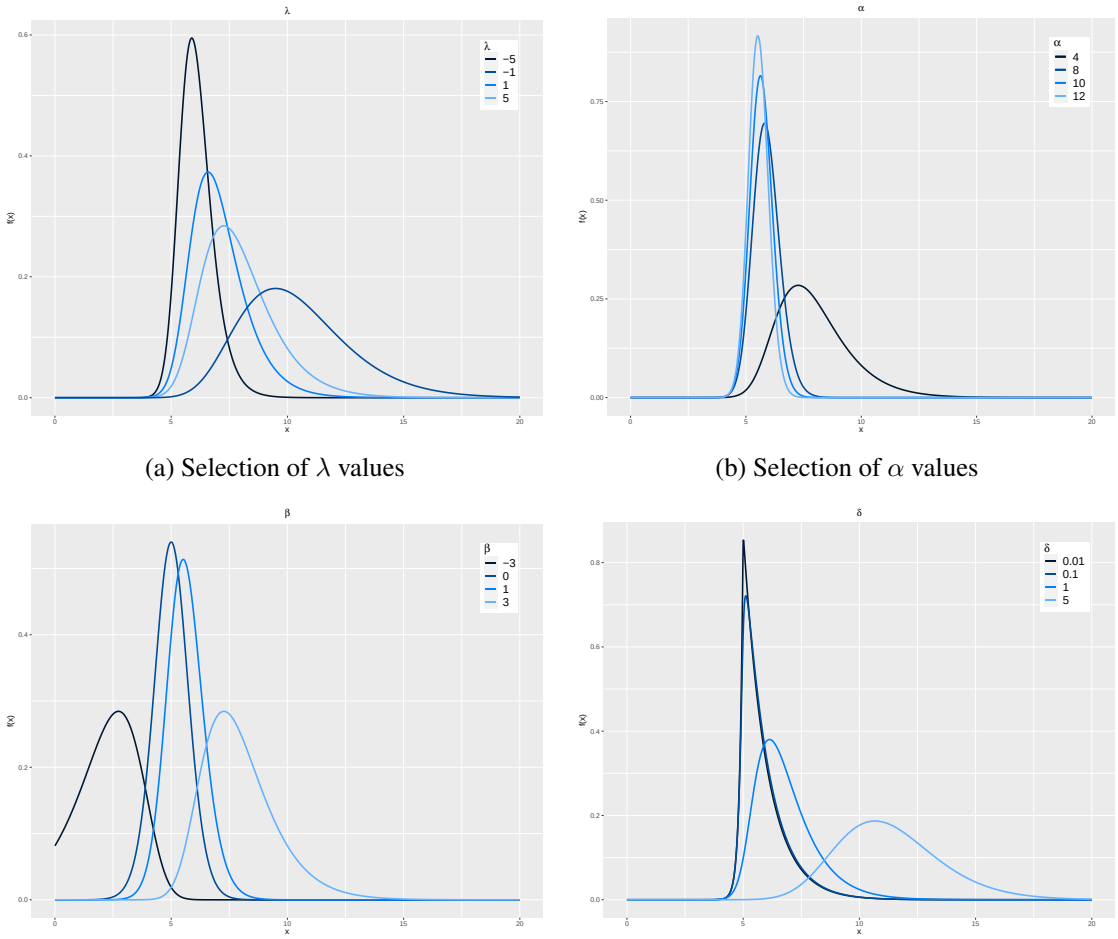
$\alpha$: The tail parameter that regulates the "fatness" of the tails. The larger the value of $\alpha$, the lighter the tails of the distribution.

$\beta$: The skewness parameter, with $|\beta| < \alpha$. An in increase in $\beta$ compared to $\alpha$ will result in an increase in the skewness. For $\beta = 0$ the distribution is symmetric.

$\delta$: Influences the shape of the pdf near its mode, and as such is often referred to as the "peakedness" parameter. Larger values of $\delta$ will result in an overall flatter peak of the pdf.

$\mu$: The location parameter. In the instance that $\beta = 0$, the distribution is symmetric and $\mu$ coincides with the mean (if the first moment exists).

In Figure 2.2 we can observe the behaviour of the pdf as individual parameters are varied, as well as the influence that each individual parameter has on the overall shape of the pdf. If we look at Figure 2.2(c), we can see the influence of the skewness parameter $\beta$ on the overall shape of the pdf, especially when $\beta = 0$, which corresponds to a symmetric case of the GH distribution. One can also observe the impact of the tail parameter $\alpha$ in Figure 2.2(b), in particular how the tails are lighter for increasing values of $\alpha$, which is to be expected.




(a) Selection of $\lambda$ values

(b) Selection of $\alpha$ values

(c) Selection of $\beta$ values

(d) Selection of $\delta$ values

Figure 2.2: Model representation of GH (see 2.1.5) for different values of $\lambda$, $\alpha$, $\beta$,and $\delta$.

## 2.2 Subfamilies

One of the major appeals of the GH distribution lies in its flexibility. This does, however, come at the cost of complexity. The various important subfamilies of the GH distribution will now be discussed. For an overview of each subfamily as well as the relationship between the mixing weight $W$ and the resulting distribution, please see Tables 2.1 and 2.2.

### 2.2.1 The variance-gamma distribution

If the mixing weight $W$ (see Definition A.1) follows the gamma distribution, the resulting pdf is that of the variance-gamma distribution (VG). This is equivalent to setting $\lambda > 0$, $\alpha > 0$,

$\beta \in (-\alpha, \alpha)$, and $\delta = 0$ in (2.1.5). The GH distribution and the VG distribution are related as follows

$$\text{VG}(\lambda, \alpha, \beta, \mu) = \text{GH}(\lambda, \alpha, \beta, 0, \mu).$$

Due to the constraint $\delta = 0$, result (B.15) needs to be used when deriving the pdf by means of (2.1.5). The resulting pdf is given by

$$f_X(x; \lambda, \alpha, \beta, \mu) = \frac{2\left(\frac{\alpha^2-\beta^2}{2}\right)^{\lambda}}{\sqrt{2\pi}\Gamma(\lambda)} \left(\frac{|x-\mu|}{\alpha}\right)^{\lambda-\frac{1}{2}} K_{\lambda-\frac{1}{2}}(\alpha|x-\mu|)\, e^{\beta(x-\mu)} \qquad (x \in \mathbb{R}). \tag{2.2.1}$$

In Figures 2.3 and 2.4, we observe how the shape of the variance-gamma pdf changes for different values of the tail parameter ($\alpha$) and the skewness parameter ($\beta$) respectively. The VG distribution was popularized by Madan & Seneta (1990) in a study of share market returns. This distribution is desirable as a model for financial data due to its longer tails, decreasing at a slower rate compared to the normal distribution.

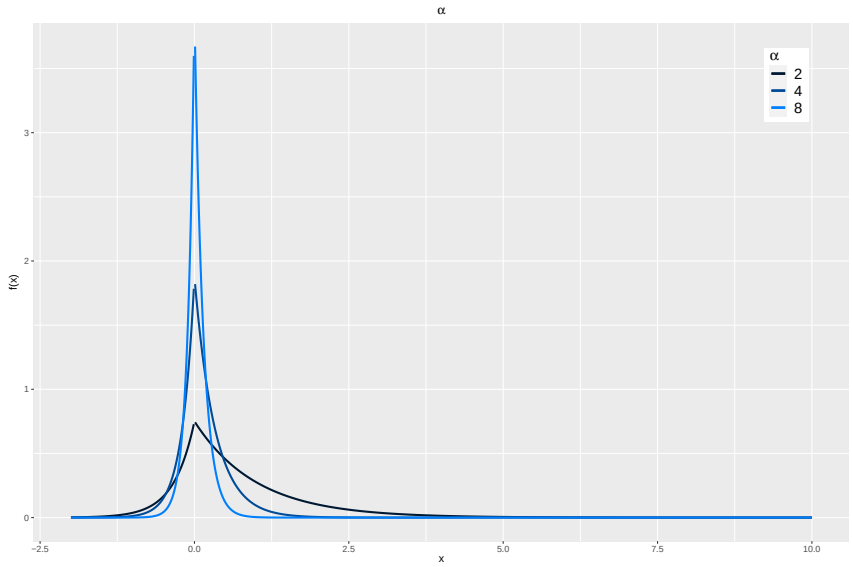



Figure 2.3: The plots of the variance-gamma pdf (2.2.1) for $\alpha = 2, 4$, and $8$.

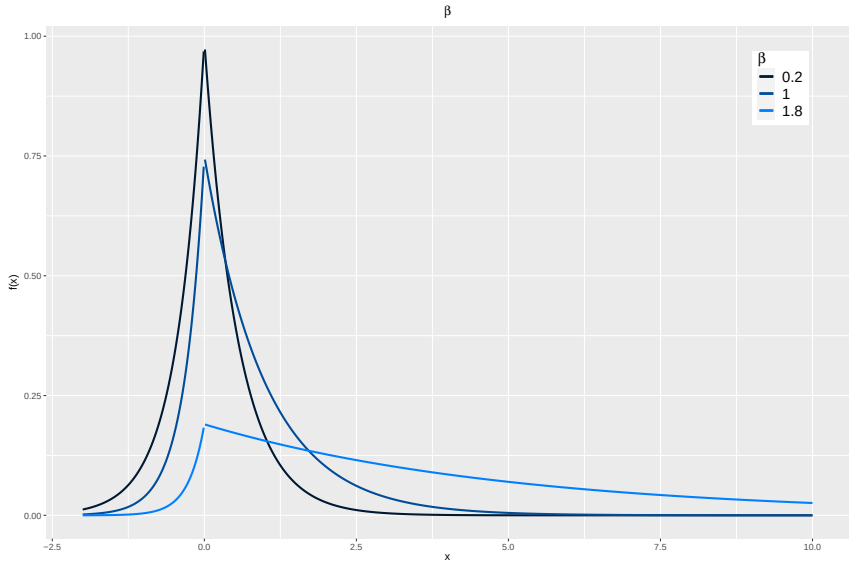



Figure 2.4: The plots of the variance-gamma pdf (2.2.1) for $\beta = 0.2, 1$, and $1.8$.

### 2.2.2 The asymmetric Laplace distribution

If the mixing weight $W$ (see Definition A.1) follows an exponential distribution, the resulting pdf is that of the asymmetric Laplace distribution (ALap). This is equivalent to setting $\lambda = 1$, $\alpha > 0$, $\beta \in (-\alpha, \alpha)$, and $\delta = 0$ in (2.1.5). The GH distribution and the ALap distribution are related as follows:

$$\text{ALap}(\alpha, \beta, \mu) = \text{GH}(1, \alpha, \beta, 0, \mu).$$

The asymmetric Laplace distribution is a special case of the variance-gamma distribution with $\lambda = 1$. This can also be deduced from the fact that we are using an exponential mixing weight, and the exponential distribution is a special case of the gamma distribution. The pdf of the Asymmetric Laplace distribution is now given:

$$f_X(x; \alpha, \beta, \mu) = \left(\frac{\alpha^2 - \beta^2}{2\alpha}\right) e^{-\alpha|x-\mu|+\beta(x-\mu)} \qquad (x \in \mathbb{R}). \tag{2.2.2}$$

For $\beta = 0$, the distribution reduces to that of the Laplace distribution (Lap) with the following pdf:

$$f_X(x; \alpha, \mu) = \left(\frac{\alpha}{2}\right) \; e^{-\alpha|x-\mu|}. \tag{2.2.3}$$

In other words we have the relation $\text{GH}(1, \alpha, 0, 0, \mu) = \text{Lap}(\mu, \alpha^{-1})$. In Figures 2.5 and 2.6 we observe the behaviours of the asymmetric Laplace pdf for varying values of the tail parameter ($\alpha$) and the skewness parameter ($\beta$) respectively.

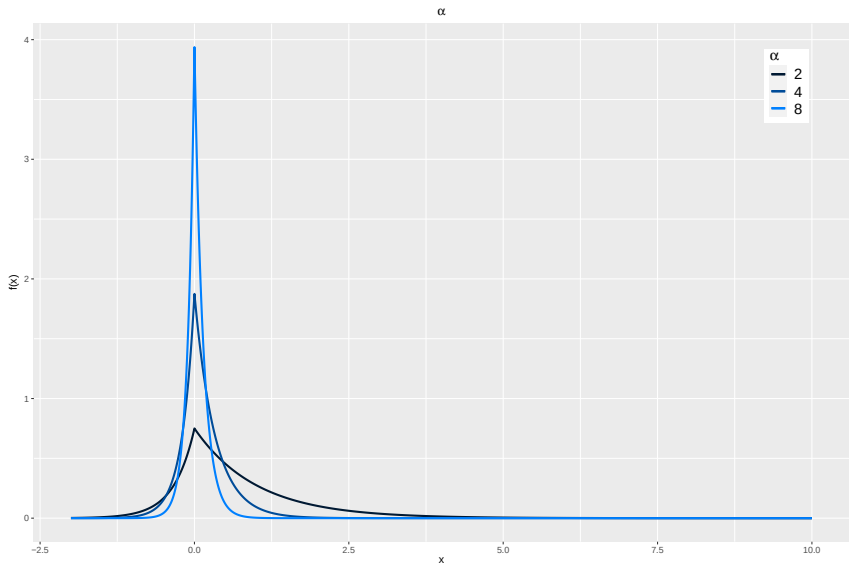



Figure 2.5: The plots of the asymmetric Laplace pdf (2.2.2) for $\alpha = 2, 4$, and $8$.

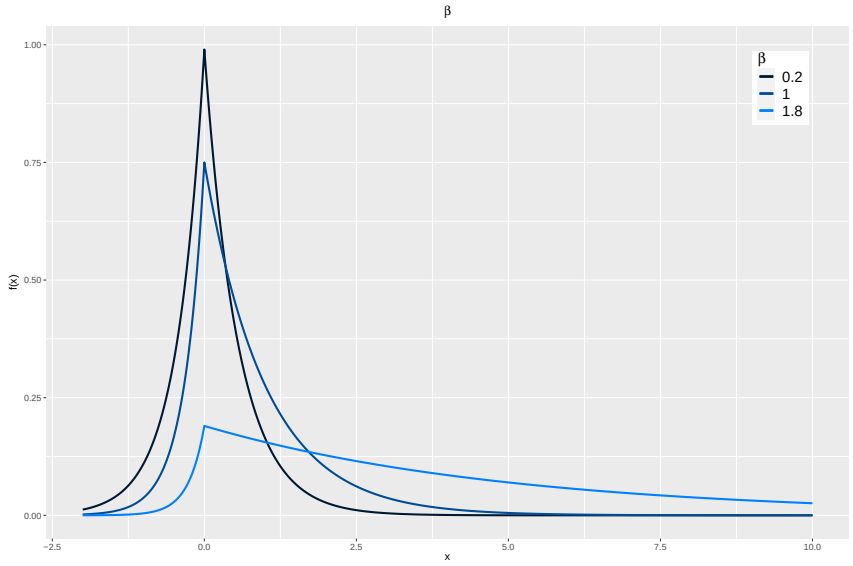



Figure 2.6: The plots of the asymmetric Laplace pdf (2.2.2) for $\beta = 0.2, 1$, and $1.8$.

### 2.2.3 The hyperbolic distribution

The hyperbolic distribution (Hyp) is a special case of the generalized hyperbolic distribution with $\lambda = 1$. The GH distribution and the Hyp distribution are related as follows

$$\text{Hyp}(\alpha, \beta, \delta, \mu) = \text{GH}(1, \alpha, \beta, \delta, \mu).$$

The pdf of the hyperbolic distribution is now given

$$f_X(x; \alpha, \beta, \delta, \mu) = \frac{\sqrt{\alpha^2 - \beta^2}}{2\alpha\delta K_1(\delta\sqrt{\alpha^2 - \beta^2})} \exp\left(-\alpha\sqrt{\delta^2 + (x-\mu)^2} + \beta(x-\mu)\right) \qquad (x \in \mathbb{R}). \tag{2.2.4}$$

If we set $\delta = 0$ then the distribution reduces to that of the asymmetric Laplace distribution as in section 2.2.2. In Figures 2.7 and 2.8 we observe the shape of the pdf as we vary the tail parameter ($\alpha$) and the skewness parameter ($\beta$) respectively.

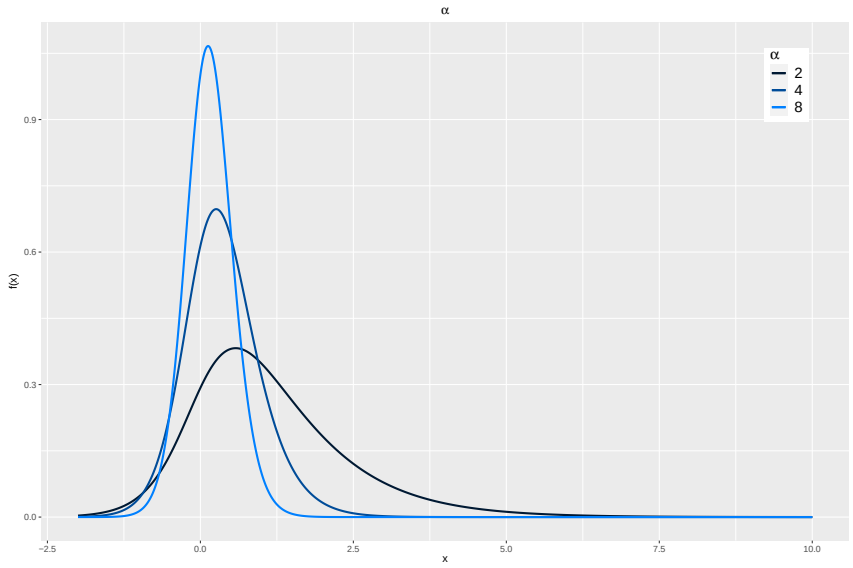



Figure 2.7: The plots of the hyperbolic pdf (2.2.4) for $\alpha = 2, 4$, and $8$.

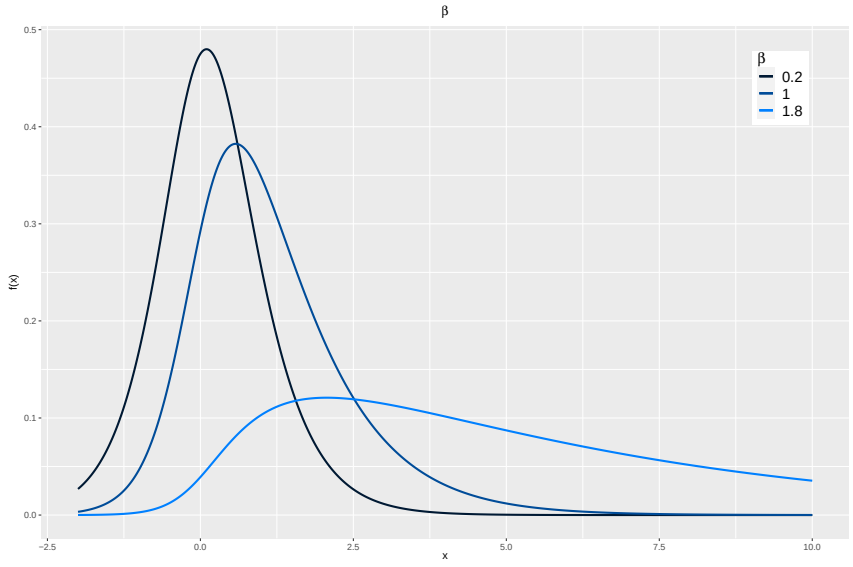



Figure 2.8: The plots of the hyperbolic pdf (2.2.4) for $\beta = 0.2, 1$, and $1.8$.

### 2.2.4 The hyperbolic asymmetric Student's *t* distribution

If the mixing weight $W$ (see Definition A.1) follows an inverse-gamma distribution, the resulting pdf is that of the hyperbolic asymmetric $t$ distribution (HA$t$). This is equivalent to setting the following constraints: $\lambda < 0$, $\alpha = |\beta|$, $\beta \in \mathbb{R}$, and $\delta > 0$ in (2.1.5). The GH distribution and the

HA$t$ distribution are related as follows:

$$\text{HA}t\,(\lambda,\beta,\delta,\mu) = \text{GH}(\lambda,|\beta|,\beta,\delta,\mu).$$

From the above constraints on the parameters, there arise two instances worthy of further exploration, namely $\alpha = |\beta| > 0$, and $\alpha = |\beta| = 0$. For $\alpha = |\beta| > 0$ the pdf is as follows

$$\begin{aligned} f_X(x;\lambda,|\beta|,\beta,\delta,\mu) = {} & \frac{2(\frac{\delta^2}{2})^{-\lambda}}{\sqrt{2\pi}\Gamma(-\lambda)}\left(\frac{\sqrt{\delta^2+(x-\mu)^2}}{|\beta|}\right)^{\lambda-\frac{1}{2}} \\ & \times K_{\lambda-\frac{1}{2}}\left(|\beta|\sqrt{\delta^2+(x-\mu)^2}\right)e^{\beta(x-\mu)} \qquad (x\in\mathbb{R}) \end{aligned} \tag{2.2.5}$$

For $\alpha = |\beta| = 0$ the pdf is derived using result (B.16)

$$\begin{aligned} f_X(x;\lambda,0,0,\delta,\mu) &= \int_0^\infty f_{X|W}(x|w)\; f_W(w;\lambda,\delta^2,\alpha^2-\beta^2)\; dw \\ &= \int_0^\infty \text{N}(x;\mu,w)\; f_W(w;\lambda,\chi,0)\; dw \\ &= \frac{k_{\lambda-\frac{1}{2}}((x-\mu)^2+\delta^2,0)}{\sqrt{2\pi}k_\lambda(\delta^2,0)}\,e^{0\cdot(x-\mu)} && \text{(from 2.1.5)} \\ &= \frac{((x-\mu)^2+\delta^2)^{\lambda-\frac{1}{2}}\Gamma(-\lambda+\frac{1}{2})}{\sqrt{2\pi}(\delta^2)^\lambda\Gamma(-\lambda)} && \text{(from B.16)} \\ &= \frac{\Gamma(\frac{-2\lambda+1}{2})}{\Gamma(\frac{-2\lambda}{2})}\frac{1}{\sqrt{\delta^2\pi}}\left(1+\frac{(x-\mu)^2}{\delta^2}\right)^{-\frac{-2\lambda+1}{2}} \end{aligned} \tag{2.2.6}$$

where $\text{N}(.;\mu,w)$ denotes a normal pdf with mean $\mu$ and variance $w$. For the case $\beta = 0$, the distribution is symmetric about $\mu$, and if $\delta^2 = -2\lambda$, (2.2.6) sumplifies to

$$f_X(x;\mu,\delta) = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})}\frac{1}{\sqrt{\delta^2\pi}}\left(1+\frac{x-\mu}{\delta}\right)\Bigg)^{-\frac{n+1}{2}}, \tag{2.2.7}$$

which is the Student's $t$ distribution with $n$ degrees of freedom.

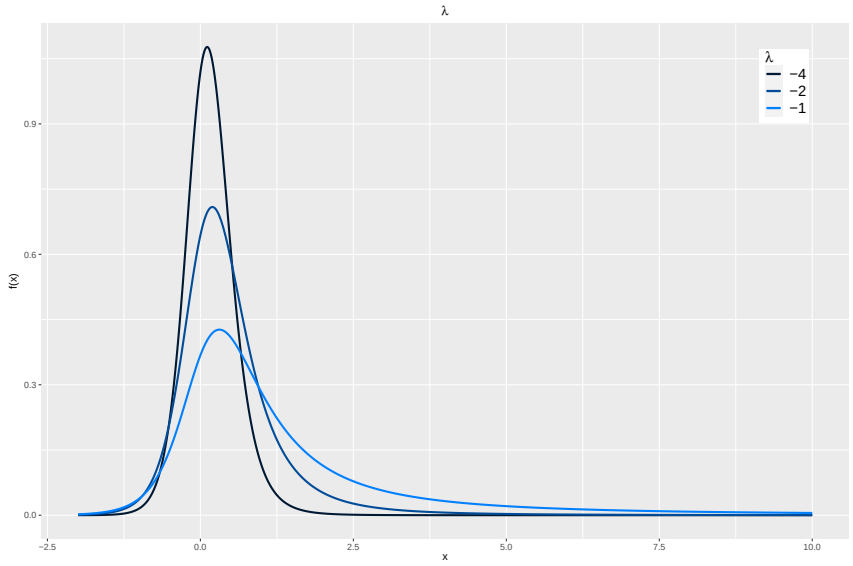



Figure 2.9: The plots of the hyperbolic asymmetric $t$ pdf (2.2.5) for $\lambda = -4, -2$, and $-1$.

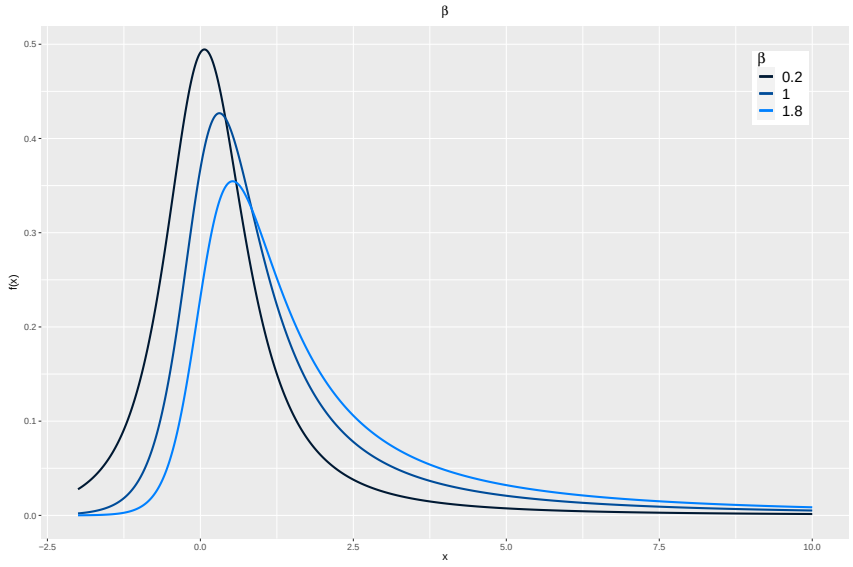



Figure 2.10: The plots of the hyperbolic asymmetric $t$ pdf (2.2.5) for $\beta = 0.2, 1$, and $1.8$.

### 2.2.5 The asymmetric Cauchy distribution

If the mixing weight $W$ (see Definition A.1) follows a Lévy distribution, the resulting pdf is that of the asymmetric Cauchy distribution (AC). This is equivalent to setting $\lambda = -0.5$, $\alpha = |\beta|$,

$\beta \in \mathbb{R}$, and $\delta > 0$ in (2.1.5). The GH distribution and the AC distribution are related as follows

$$\text{AC}(\beta, \delta, \mu) = \text{GH}(\lambda, |\beta|, \beta, \delta, \mu).$$

The pdf of the asymmetric Cauchy subclass is given by

$$f_X(x; \beta, \delta, \mu) = \frac{2(\frac{\delta^2}{2})^{\frac{1}{2}}}{\sqrt{2\pi}\Gamma(\frac{1}{2})}(\frac{\sqrt{\delta^2 + (x-\mu)^2}}{|\beta|})^{-1} K_{-1}(|\beta|\sqrt{\delta^2 + (x-\mu)^2})\, e^{\beta(x-\mu)}. \quad (2.2.8)$$

Setting $\beta = 0$ will yield the symmetric case, i.e. the Cauchy distribution with pdf:

$$f_X(x; \delta, \mu) = \frac{\delta}{\pi(\delta^2 + (x-\mu)^2)}. \quad (2.2.9)$$

In Figures 2.11 and 2.12, we observe how the behaviour of the asymmetric Cauchy pdf for different values of the skewness parameter ($\beta$) and the peakedness parameter ($\delta$) respectively.

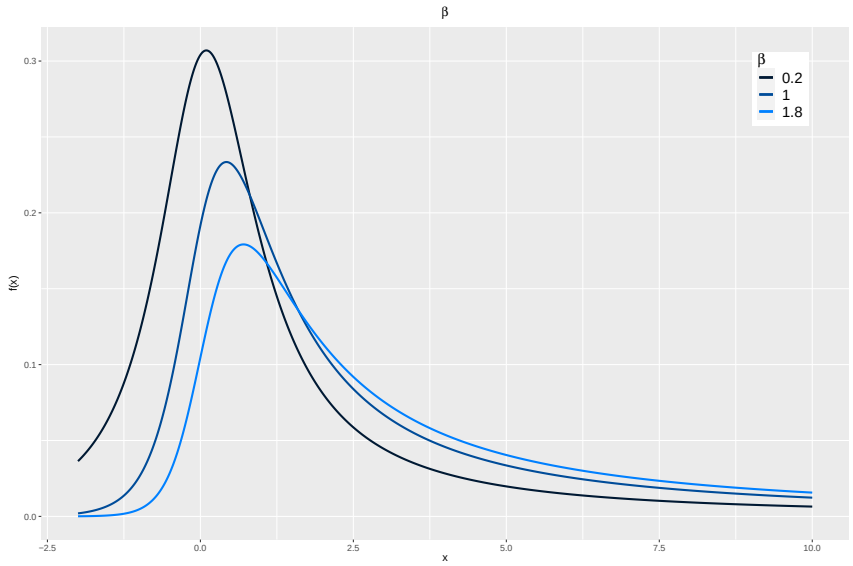



Figure 2.11: The plots of the asymmetric Cauchy pdf (2.2.8) for $\beta = 0.2, 1$, and $1.8$.

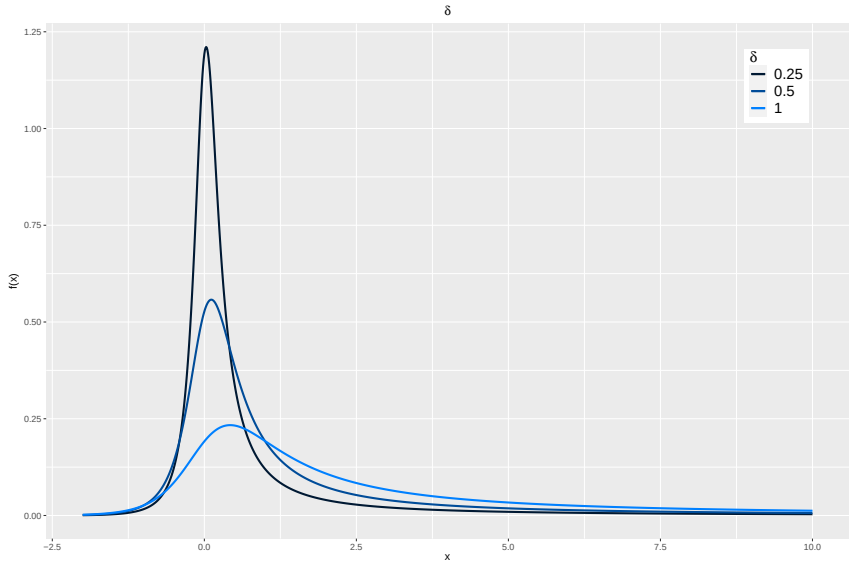



Figure 2.12: The plots of the asymmetric Cauchy pdf (2.2.8) for $\delta = 0.25, 0.5$, and 1.

### 2.2.6 The normal inverse Gaussian distribution

The normal inverse Gaussian (NIG) distribution results when the mixing weight $W$ (see Definition A.1) follows an inverse Gaussian distribution. This is equivalent to the constraints $\lambda = -\frac{1}{2}$, $\alpha > 0$, $\beta \in (-\alpha, \alpha)$, and $\delta > 0$ in (2.1.5). The GH distribution and the NIG distribution are related as follows

$$\text{NIG}(\alpha, \beta, \delta, \mu) = \text{GH}(-0.5, \alpha, \beta, \delta, \mu).$$

The pdf of the normal inverse Gaussian (NIG) subclass is now given

$$f_X(x; -\frac{1}{2}, \alpha, \beta, \delta, \mu) = e^{\delta\sqrt{\alpha^2-\beta^2}} \frac{\alpha\delta}{\pi\sqrt{\delta^2+(x-\mu)^2}} K_1\left(\alpha\sqrt{\delta^2+(x-\mu)^2}\right) e^{\beta(x-\mu)}. \tag{2.2.10}$$

One can see that for $\alpha = |\beta|$, this pdf reduces to that of the asymmetric Cauchy distribution in section 2.2.5. The AC distribution is thus a limiting case of the NIG distribution. A useful property of the NIG distribution lies in the simplified forms for the mean, variance and skewness compared to that of the generalized hyperbolic distribution. The mean, variance and skewness are given by:

- $E(X) = \mu + \beta\nu$,

- $V(X) = \nu + \beta^2 \frac{\nu^2}{\omega}$,
- $\mu_3(X) = 3\beta\frac{\nu^2}{\omega} + 3\beta^3\frac{\nu^3}{\omega^2}$

where $\nu = \frac{\delta}{\sqrt{\alpha^2-\beta^2}}$ and $\omega = \delta\sqrt{\alpha^2-\beta^2}$. The moment generating function is given by:

$$M_X(t) = e^{\mu t} e^{\delta\left(\sqrt{\alpha^2-\beta^2}-\sqrt{\alpha^2-(\beta+t)^2}\right)}. \tag{2.2.11}$$

In Figures 2.13 and 2.14, we observe how the shape of the normal inverse Gaussian pdf changes for different values of the tail parameter ($\alpha$) and the skewness parameter ($\beta$) respectively.

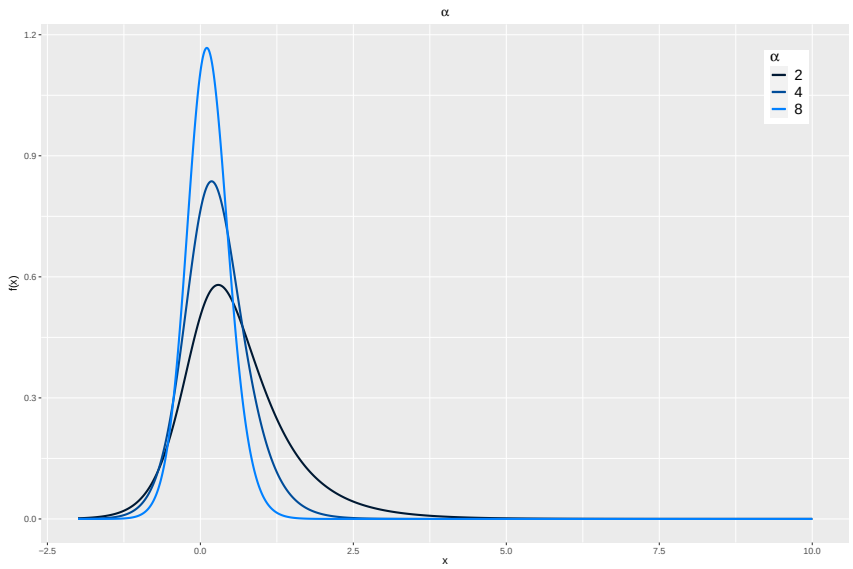



Figure 2.13: The plots of the normal inverse Gaussian pdf (2.2.10) for $\alpha = 2, 4$, and $8$.

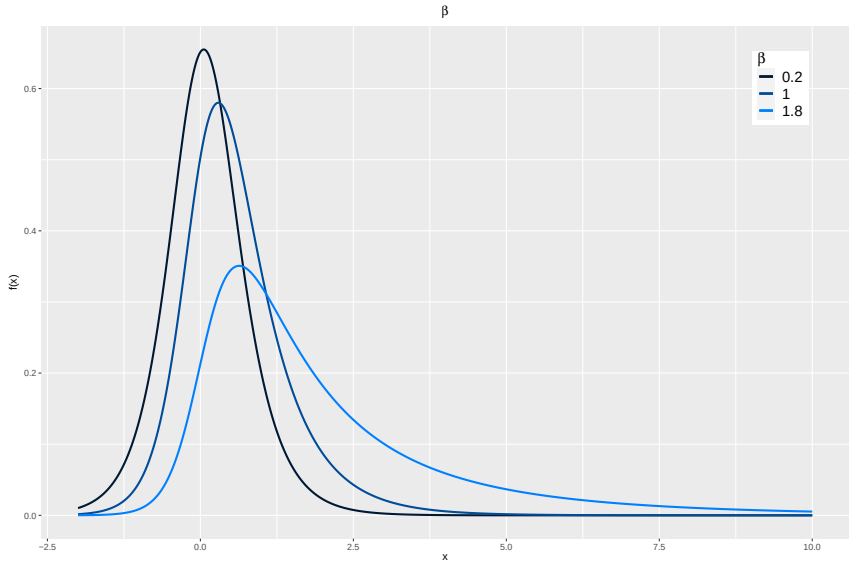



Figure 2.14: The plots of the normal inverse Gaussian pdf (2.2.10) for $\beta = 0.2, 1$, and $1.8$.

Table 2.1: Subfamilies of the GH distribution and corresponding parameter spaces, with location parameter $\mu \in \mathbb{R}$ in each case.

| **Distribution** | **Abbrev.** | **Parameter Space** | | | |
|---|---|---|---|---|---|
| variance-gamma | VG | $\lambda > 0$ | $\alpha > 0$ | $\lvert\beta\rvert < \alpha$ | $\delta = 0$ |
| asymmetric Laplace | ALap | $\lambda = 1$ | $\alpha > 0$ | $\lvert\beta\rvert < \alpha$ | $\delta = 0$ |
| Laplace | Lap | $\lambda = 1$ | $\alpha > 0$ | $\beta = 0$ | $\delta = 0$ |
| hyperbolic | Hyp | $\lambda = 1$ | $\alpha > 0$ | $\lvert\beta\rvert < \alpha$ | $\delta > 0$ |
| hyperbolic asymmetric *t* | HA*t* | $\lambda < 0$ | $\alpha = \lvert\beta\rvert$ | $\beta \geq 0$ | $\delta > 0$ |
| Student's *t* | *t* | $\lambda < 0$ | $\alpha = 0$ | $\beta = 0$ | $\delta > 0$ |
| asymmetric Cauchy | AC | $\lambda = -\frac{1}{2}$ | $\alpha = \lvert\beta\rvert$ | $\beta \in \mathbb{R}$ | $\delta > 0$ |
| Cauchy | Ca | $\lambda = -\frac{1}{2}$ | $\alpha = 0$ | $\beta = 0$ | $\delta > 0$ |
| normal inverse Gaussian | NIG | $\lambda = -\frac{1}{2}$ | $\alpha > 0$ | $\lvert\beta\rvert < \alpha$ | $\delta > 0$ |

Table 2.2: The mixing weights as well as the corresponding distribution outcome for each of the GH distribution subfamilies

| **Mixing weight** | **Resulting Distribution** |
|---|---|
| gamma | variance-gamma |
| exponential | asymmetric Laplace |
| exponential (with $\beta = 0$) | Laplace |
| inverse gamma | hyperbolic asymmetric *t* |
| inverse gamma (with $\beta = 0$) | Student's *t* |
| Lévy | asymmetric Cauchy |
| Lévy (with $\beta = 0$) | Cauchy |
| inverse Gaussian | normal inverse Gaussian |

## 2.3 Alternative parametrizations

The following four parameterizations are only defined for the generalized hyperbolic distribution, but with the added restriction that $\chi, \psi > 0$. The parameterizations were taken from Paolella (2007).

1. The $(\lambda, \omega, \beta, \eta, \mu)$ parameterization with $\omega = \sqrt{\chi\psi} = \delta\sqrt{\alpha^2 - \beta^2}$ and $\eta = \sqrt{\frac{\chi}{\psi}} = \frac{\delta}{\sqrt{\alpha^2-\beta^2}}$

2. The $(\lambda, \bar{\alpha}, \bar{\beta}, \delta, \mu)$ parameterization with $\bar{\alpha} = \alpha\delta$ and $\bar{\beta} = \beta\delta$. In this parameterization $\mu$ and $\delta$ are location and scale parameters and $\lambda, \bar{\alpha}, \bar{\beta}$ are both location- and scale- invariant.

3. The $(\lambda, \zeta, \rho, \delta, \mu)$ parameterization with $\zeta = \delta\sqrt{\alpha^2 - \beta^2} = \omega$ and $\rho = \frac{\alpha}{\beta}$. In this parameterization $\mu$ and $\delta$ are location and scale parameters and $\lambda, \zeta, \rho$ are both location- and scale-invariant as both $\zeta$ and $\rho$ can be expressed in terms of $\bar{\alpha}$ and $\bar{\beta}$.

4. The $(\lambda, \xi, q, \delta, \mu)$ parameterization with $\xi = \frac{1}{\sqrt{1+\zeta}}$ and $q = \rho\xi = \frac{\rho}{\sqrt{1+\zeta}}$. Since both $\xi$ and $q$ are defined in terms of $\zeta$ and $\rho$, and since $\zeta$ and $\rho$ are both location- and scale- invariant, it follows that $\xi$ and $q$ are also location- and scale- invariant

## 2.4 Properties

### 2.4.1 Moment generating function

The moment generating function of the generalized hyperbolic distribution is computed in Prause (1999) as follows

$$M(u) = e^{u\mu}\left(\frac{\alpha^2-\beta^2}{\alpha^2-(\beta+u)^2}\right)^{\frac{\lambda}{2}} \frac{K_\lambda\left(\delta\sqrt{(\alpha^2-(\beta+u)^2)}\right)}{K_\lambda\left(\delta\sqrt{\alpha^2-\beta^2}\right)}. \tag{2.4.1}$$

**Proof.** First, without loss of generality, assume $\mu = 0$. Then from (2.1.5), for $|\beta+u| < \alpha$ we have

$$\begin{aligned}
M(u) &= \int e^{ux}\text{GH}(x;\lambda,\alpha,\beta,\delta,0)\,dx \\
&= \int e^{ux} a(\lambda,\alpha,\beta,\delta)\left(\delta^2+x^2\right)^{\frac{1}{2}(\lambda-\frac{1}{2})} K_{\lambda-\frac{1}{2}}\left(\alpha\sqrt{\delta^2+x^2}\right)e^{\beta x}\,dx \\
&= a(\lambda,\alpha,\beta,\delta)\int e^{ux}\left(\delta^2+x^2\right)^{\frac{1}{2}(\lambda-\frac{1}{2})} K_{\lambda-\frac{1}{2}}\left(\alpha\sqrt{\delta^2+x^2}\right)e^{\beta x}\,dx \\
&= \frac{a(\lambda,\alpha,\beta,\delta)}{a(\lambda,\alpha,\beta+u,\delta)} \\
&= \frac{(\alpha^2-\beta^2)^{\frac{1}{2}}}{\sqrt{2\pi}\delta^\lambda\alpha^{\lambda-\frac{1}{2}}K_\lambda\left(\delta\sqrt{\alpha^2-\beta^2}\right)} \frac{\sqrt{2\pi}\delta^\lambda\alpha^{\lambda-\frac{1}{2}}K_\lambda\left(\delta\sqrt{\alpha^2-(\beta+u)^2}\right)}{(\alpha^2-(\beta+U)^2)^{\frac{1}{2}}} \\
&= \left(\frac{\alpha^2-\beta^2}{\alpha^2-(\beta+u)^2}\right)^{\frac{\lambda}{2}} \frac{K_\lambda\left(\delta\sqrt{\alpha^2-(\beta+u)^2}\right)}{K_\lambda\left(\delta\sqrt{\alpha^2-\beta^2}\right)}.
\end{aligned} \tag{2.4.2}$$

□

Finally, for location parameter $\mu$ the moment generating function is given by

$$M(u) = e^{u\mu}\left(\frac{\alpha^2-\beta^2}{\alpha^2-(\beta+u)^2}\right)^{\frac{\lambda}{2}}\frac{K_\lambda\left(\delta\sqrt{\alpha^2-(\beta+u)^2}\right)}{K_\lambda\left(\delta\sqrt{\alpha^2-\beta^2}\right)}. \tag{2.4.3}$$

Note that the restriction $|\beta+u|<\alpha$ in (2.4.1) follows from the domain of variation of the parameters of the GH distribution defined in Theorem 2.1.2. We are now able to calculate the mean and variance of the GH distribution:

$$\mathbb{E}(X) = \mu + \frac{\beta\delta}{\sqrt{\alpha^2-\beta^2}}\frac{K_{\lambda+1}(\xi)}{K_\lambda(\xi)}, \tag{2.4.4}$$

$$\text{Var}(X) = \delta^2\left(\frac{K_{\lambda+1}(\xi)}{\xi K_\lambda(\xi)} + \frac{\beta^2}{\alpha^2-\beta^2}\left[\frac{K_{\lambda+2}(\xi)}{K_\lambda(\xi)} - \left(\frac{K_{\lambda+1}(\xi)}{K_\lambda(\xi)}\right)^2\right]\right). \tag{2.4.5}$$

### 2.4.2 Moments of the generalized hyperbolic distribution

The moments of the generalized hyperbolic distribution are given by (see Scott et al., 2011)

$$\begin{aligned}
M_1 &= (\frac{\delta^2}{\zeta})\beta\frac{K_{\lambda+1}(\zeta)}{K_\lambda(\zeta)},\\
M_2 &= \frac{(\frac{\delta^2}{\zeta})K_{\lambda+1}(\zeta)+(\frac{\delta^2}{\zeta})^2\beta^2K_{\lambda+2}(\zeta)}{K_\lambda(\zeta)},\\
M_3 &= \frac{(3\frac{\delta^2}{\zeta}^2)\beta K_{\lambda+2}(\zeta)+(\frac{\delta^2}{\zeta})^3\beta^3K_{\lambda+3}(\zeta)}{K_\lambda(\zeta)},\\
M_4 &= \frac{(3\frac{\delta^2}{\zeta}^2)K_{\lambda+2}(\zeta)+6(\frac{\delta^2}{\zeta})^3\beta^2K_{\lambda+3}(\zeta)+(\frac{\delta^2}{\zeta})^4\beta^4K_{\lambda+4}(\zeta)}{K_\lambda(\zeta)},
\end{aligned} \tag{2.4.6}$$

where $\zeta = \delta\sqrt{\alpha^2-\beta^2}$ as in (2.3).

# Chapter 3

# Estimation and other inferential aspects

This section will contain a brief outline of maximum likelihood estimation, as this will be the primary method of estimating the parameters of the GH distribution and its subclasses. An outline on the concept of profile likelihood is also given, as it will be of relevance in the sections that follow. The information that follows was extracted from Bain and Engelhardt (1987), and Silvey (1970).

## 3.1 Maximum likelihood estimation

The use of least squares regression is justified by the fact that we require no knowledge of the distribution of the error vector, only its mean and variance matrices, and the method can be applied without access to said knowledge. The method of maximum likelihood, on the other hand, is used mainly in situations where we have information about the distribution of the sample space.

Maximum likelihood estimation is usually applied when the possible distributions on the sample space can be labelled/represented by a finite parameter vector $\boldsymbol{\theta}$[1]. The application of maximum likelihood is also restricted to the case where the distributions possess a pdf that can be represented as some measure on the sample space such a a counting measure (discrete), or a Lebesgue measure (continuous).

**Definition 3.1.1.** The joint pdf of $n$ random variables $X_1, X_2, \ldots, X_n$ evaluated at $x_1, x_2, \ldots, x_n$

[1]While the process can also be in terms of a scalar parameter $\theta$, we use the vector representation as it is consistent with the nature of the parameters of the generalized hyperbolic distribution

is referred to as the likelihood function. For fixed $x_1, x_2, \ldots, x_n$ the likelihood function is a function of $\boldsymbol{\theta}$ and is often denoted $L(\boldsymbol{\theta})$.

**Definition 3.1.2.** Let $L(\boldsymbol{\theta}) = f(\mathbf{x}; \boldsymbol{\theta})$, with $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\boldsymbol{\theta} \in \boldsymbol{\Omega}$. For a given set of observations $x_1, x_2, \ldots, x_n$, the value of $\hat{\boldsymbol{\theta}} \in \boldsymbol{\Omega}$ at which $L(\boldsymbol{\theta})$ is a maximum is referred to as the *maximum likelihood estimate* (MLE) of $\boldsymbol{\theta}$. In other words, $\hat{\boldsymbol{\theta}}$ is the value of $\boldsymbol{\theta}$ that satisfies

$$f(\mathbf{x}; \hat{\boldsymbol{\theta}}) = \max_{\boldsymbol{\theta} \in \boldsymbol{\Omega}} f(\mathbf{x}; \boldsymbol{\theta}). \tag{3.1.1}$$

### 3.1.1 The profile likelihood

When using the method of maximum likelihood for low-dimensional parameter vectors, we can easily visualize this with a graph. Provided the likelihood is smooth, it will resemble the shape of an upside down parabola (at least locally), with the peak representing the ML estimate. When working with higher-dimensional parameter vectors we can no longer use the likelihood function in this way, as it is no longer possible to graphically visualize the problem.

One way of overcoming this problem is by using the profile likelihood rather than the full likelihood (see Barndorff-Nielsen and Cox 2017 and Murphy and Van der Vaart 2000). We begin by partitioning the set of parameters $\boldsymbol{\theta}$ as follows:

- A set of low-dimensional parameters of interest $\boldsymbol{\xi}$.
- A set of high- or low-dimensional nuisance parameters $\boldsymbol{\eta}$.

If the full likelihood is defined as $L(\boldsymbol{\xi}, \boldsymbol{\eta})$, then the profile likelihood is defined as follows

$$L_p(\boldsymbol{\xi}) = \sup_{\boldsymbol{\eta}} L(\boldsymbol{\xi}, \boldsymbol{\eta}). \tag{3.1.2}$$

It is customary to use the curvature of the profile likelihood function as an estimate of the variability of $\hat{\boldsymbol{\xi}}$. For a Euclidean parameter this was justified by Patefield (1977), who showed that in parametric models, the inverse of the observed profile information is equal to the $\boldsymbol{\xi}$ aspect of the full observed inverse information. Further discussion in the parametric context is given by Barndorff-Nielsen and Cox 2017, p. 1.

Thus it seems the profile likelihood can be used and visualised in the same way as a full parametric likelihood. This should be an obvious enough reason to recommend the use of the profile likelihood. A definition and derivation of the profile likelihood will now be given.

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with pdf $f(x;\boldsymbol{\theta})$, where the objective is to estimate $\boldsymbol{\theta} = (\boldsymbol{\xi}, \boldsymbol{\eta})$. The log-likelihood function is then defined as

$$\ell(\boldsymbol{\xi}, \boldsymbol{\eta}) = \sum_{i=1}^{n} \log f(X_i; \boldsymbol{\xi}, \boldsymbol{\eta}), \tag{3.1.3}$$

where $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ represent the parameter/s of interest and nuisance parameter/s respectively. The profile log-likelihood function is thus defined similarly to (3.1.2) as

$$\ell_p(\boldsymbol{\xi}) = \sup_{\boldsymbol{\eta}} \ell(\boldsymbol{\xi}, \boldsymbol{\eta}). \tag{3.1.4}$$

Instead of simultaneously solving for the entire parameter vector $\boldsymbol{\theta}$, we partition the set of parameters into $\boldsymbol{\theta} = (\boldsymbol{\xi}, \boldsymbol{\eta})$ and estimate the parameter/s of interest $\boldsymbol{\xi}$ while eliminating the nuisance parameter $\boldsymbol{\eta}$ from the profile likelihood by the maximization operation in (3.1.4). We are thus only estimating the MLE's of $\boldsymbol{\xi}$ for some fixed choice of $\boldsymbol{\eta}$. The question that should arise now is how does one select an appropriate choice of nuisance parameter/s $\boldsymbol{\eta}$ to get an accurate estimate for our parameters of interest.

The solution is to repeat the process of estimating $\hat{\boldsymbol{\xi}}$ for an array of choices of the nuisance parameters. It is important that this array adequately represents a range of values that the nuisance parameters can take on as to ensure that the true maximum is not overlooked and we do not get caught up in a local maximum. The final estimate $\hat{\boldsymbol{\xi}}$ is then the one corresponding to the choice of $\boldsymbol{\eta}$ that has the highest profile log-likelihood value.

It is already clear that the use of the profile likelihood allows us to both visualize and intuitively interpret the log-likelihood function as well as analyse its behaviour. It will be shown later that the method of profiling also has a positive impact on the likelihood of the procedure converging to the global maximum, as the number of parameters that are simultaneously estimated seem to have an impact on the MLE estimation process.

# Chapter 4

# Numerical algorithms

This section contains an overview of the estimation methods that will be used to estimate the GH distribution parameters. The following numerical methods are considered:

- Nelder-Mead simplex method.
- EM algorithm.
- Profile likelihood based alternating algorithm.

## 4.1 The Nelder-Mead simplex method

The Nelder-Mead simplex method is one of the most widely used and popular methods of multi-dimensional, unconstrained optimization (Nelder and Mead, 1965). An advantage of this method is that it minimizes the objective function using only function values. In other words, it does not make use of derivatives (explicit or implicit). As one would expect this is quite useful as many situations arise when the derivative of the function to be maximized is not available, especially if the function is defined from a complex or convoluted computational structure (Han, 2006).

For a given function of $n$ variables, the method minimises this function by comparing the function values at $(n+1)$ vertices of a general simplex[1], after which the vertex corresponding to the highest value is replaced by another point. What is useful about this method, is the fact that

[1] A simplex is the generalization of the notion of a triangle to an arbitrary number of dimensions. For example, a simplex in two dimensions is a triangle, in three dimensions a tetrahedron etc.

the simplex is adaptive, as it changes depending on the landscape of the function being evaluated, and eventually contracts to the local or global minimum.

At each iteration of the process, there is a working simplex defined by $n+1$ vertices $x_1, x_2, \ldots, x_{n+1}$, each a point in $\mathbb{R}^n$ with corresponding function values $f(x_1)$, $f(x_2)$, $\ldots$, $f(x_{n+1})$. Each iteration begins with the ordering and labelling of the current set of vertices $x_1^{[k]}, x_2^{[k]}, \ldots, x_{n+1}^{[k]}$ such that

$$f(x_1^{[k]}) \leq f(x_2^{[k]}) \leq f(x_{n+1}^{[k]}). \tag{4.1.1}$$

Since the objective is to minimize our function, $f(x_1^{[k]})$ would naturally be the "best" point as it corresponds the smallest function value, and it logically follows that $x_{n+1}^{(k)}$ is the worst point. In order for the algorithm to perform properly and be well defined, consistent tie-breaking rules are required (Lagarias et al., 1998). After the calculation of one or more trial points and evaluating the function value $f$ at each of these points, the $k^{th}$ iteration generates a set of $n+1$ vertices that define a different simplex for the next iteration.

There are four possible operations: **reflection**, **expansion**, **contraction**, and **shrinkage**, each with an associated scalar parameters $\rho$, $\eta$, $\gamma$, and $\nu$ respectively. These coefficients should satisfy the following constraints:

$$\rho > 0, \quad \eta > 1, \quad 0 < \gamma < 1, \quad \text{and} \quad 0 < \nu < 1. \tag{4.1.2}$$

Common initial choices for the parameters in (4.1.2) are

$$\rho = 1, \quad \eta = 2, \quad \gamma = \frac{1}{2}, \quad \text{and} \quad \nu = \frac{1}{2}. \tag{4.1.3}$$

The generic algorithm has two possible outcomes. The first is a single new vertex (the accepted point), which will replace the current worst point $x_{n+1}$. In the second a shrink step is performed (see Figure 4.1), and a set of $n$ new points together with the best point $x_1$ will form the simplex at the next iteration. A kind of "search direction" is defined by $x_{n+1}$ and $\bar{x}$, the centroid of all vertices except $x_{n+1}$. The Nelder-Mead algorithm at each iteration $k$ is now outlined.

**Step 1: Order**

First we order the $n+1$ vertices such that $f(x_1) \leq f(x_2) \leq ... \leq f(x_{n+1})$, using a consistent tie-breaking rule as shown.

**Step 2: Reflect**

Next we compute the reflection point $x_r$ from the following

$$x_r = \bar{x} + \rho(\bar{x} - x_{n+1}), \tag{4.1.4}$$

where $\bar{x}$ is the centroid of the $n$ best vertices (excluding $x_{n+1}$). In other words $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Next evaluate $f_r = f(x_r)$, and if $f_1 \leq f_r < f_n$, we accept the reflected point $x_r$ and terminate the iteration.

**Step 3: Expand**

If however $f_r < f_1$, we calculate the expansion point $x_e$ from

$$x_e = \bar{x} + \eta(x_r - \bar{x}) \tag{4.1.5}$$

and evaluate $f_e = f(x_e)$. If $f_e < f_r$, accept the expansion point $x_e$ and terminate the iteration.

**Step 4: Contract**

Conversely, if $f_r \geq f_n$, we perform a contraction operation between $\bar{x}$ and the better point between $x_{n+1}$ and $x_r$. There are two possible contraction operations depending on the relation between $f_r$ and $f_{n+1}$.

**1. Outside contraction:**

If $f_n \leq f_r < f_{n+1}$ (i.e. $x_r$ is strictly better than $x_{n+1}$), we then perform what is known as an outside contraction

$$x_c = \bar{x} + \gamma(x_r - \bar{x}) \tag{4.1.6}$$

and evaluate $f_c = f(x_c)$. If $f_c \leq f_r$, we accept $x_c$ and terminate the iteration; otherwise we proceed to step 5.

**2. Inside contraction:**

If $f_r \geq f_{n+1}$ (i.e. $x_{n+1}$ is better than $x_r$), we then perform what is known as an inside

contraction

$$x_c' = \bar{x} - \gamma(\bar{x} - x_{n+1}), \tag{4.1.7}$$

and evaluate $f_c' = f(x_c)'$. If $f_c' \leq f_{n+1}$, we accept $x_c'$ and terminate the iteration, otherwise we proceed to step 5.

**Step 5: Perform a shrink step**

Here we define $n$ new vertices from

$$v_i = x_1 + \nu(x_i - x_1) \qquad i = 2, ..., n+1 \tag{4.1.8}$$

and evaluate $f$ at these points. The vertices of the simplex at the next iteration will then be $x_1, v_2, ..., v_{n+1}$.

Something that is not explicitly stated in Nelder and Mead (1965) is how the points should be ordered in the case of equal function values, otherwise known as a tie-breaking criterion. The following tie-break rules are defined for a step when a shrink occurs, and for when a non-shrink step occurs:

1. **Non-shrink tie-break rule**

   When a non-shrink step occurs, the worst point $x_{n+1}^{[k]}$ is discarded. The point created during the $k^{th}$ iteration, which we denote $v^{[k]}$, becomes a new vertex and takes the $j + 1^{th}$ position among the vertices where

   $$j = \max_{0 \leq \ell \leq n} \{\ell | f(v^{[k]}) < f(x_{\ell+1}^{[k]})\}. \tag{4.1.9}$$

   In this step all other vertices retain their relative ordering.

2. **Shrink tie-break rule**

   In the instance of a shrink step, the only point that is carried over is the best point $x_1^{[k]}$. There is thus only one tie-breaking rule for the case where two or more points are tied for the best point. If

   $$\min\{f(v_2^{[k]}), \ldots, f(v_{n+1}^{[k]}) = f(x_1^{[k]})\}, \tag{4.1.10}$$

then set $x_1^{k+1} = x_1^{[k]}$.

A simple graphical outline of how each step works is given in Figure 4.1.

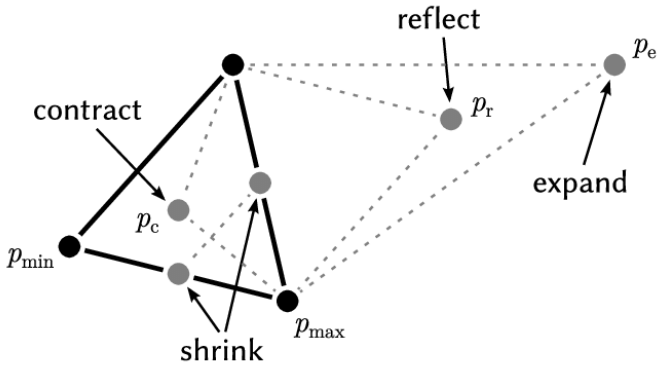



Figure 4.1: An illustration of the steps of the Nelder-Mead simplex method. Obtained from Cheng and Mailund (2015).

## 4.2 The EM algorithm

The Expectation-maximization (EM) algorithm is an iterative method developed by Dempster, Laird, and Rubin (1977) with the aim of computing maximum likelihood estimates in the presence of incomplete (or missing) data. It is also possible to make use of this method by reframing the problem as if there were missing data. The name of the method is derived from the fact that the process consists of an expectation step (E-step) followed by a maximization step (M-step).

The following will serve as a general introduction to the EM algorithm (Moon, 1996). Let $\mathcal{Y}$ be the sample space of observations, with $\mathbf{y} \in \mathbb{R}^m$ an observation from $\mathcal{Y}$, and let $\mathcal{X}$ be the underlying sample space, where $\mathbf{x} \in \mathbb{R}^n$ and $m < n$. We refer to $\mathbf{x}$ as the complete data, which is not observed directly, but only through $\mathbf{y}$ such that $\mathbf{y} = \mathbf{y}(\mathbf{x})$ where $\mathbf{y}(\mathbf{x})$ is a many-to-one mapping. Let $f_X(\mathbf{x}|\boldsymbol{\theta})$ denote the pdf of the complete data with $\boldsymbol{\theta} \in \boldsymbol{\Omega}$ the set of parameters of $f$. The pdf of the incomplete data is given by

$$g_Y(\mathbf{y}|\boldsymbol{\theta}) = \int_{\mathcal{X}(y)} f_X(\mathbf{x}|\boldsymbol{\theta})\, d\mathbf{x} \tag{4.2.1}$$

which also defines the relation between the complete and incomplete data specification. A useful

aspect of the EM algorithm is that even though the problem may not be one of incomplete data, by formulating it as such we simplify the computation of the maximum likelihood estimates. Let $L(\boldsymbol{\theta}) = g_Y(y|\boldsymbol{\theta})$ denote the incomplete-data likelihood function and let

$$\log L(\boldsymbol{\theta}) = \log g_Y(y|\boldsymbol{\theta}) \tag{4.2.2}$$

denote the corresponding log-likelihood function. The objective of the EM algorithm is to maximise the complete-data log-likelihood function $\log f_X(\mathbf{x}|\boldsymbol{\theta})$. The issue, however, is that we do not possess the data $\mathbf{x}$ to compute this log-likehood. What the EM algorithm does is circumvent this problem by maximising the expectation of $\log f_X(\mathbf{x}|\boldsymbol{\theta})$ given the observed data $\mathbf{y}$ and the current estimate for $\boldsymbol{\theta}$. In other words it indirectly solves the incomplete-data log-likelihood function in (4.2.2) by iteratively solving in terms of $\log f_X(\mathbf{x}|\boldsymbol{\theta})$. But, since we cannot observe $\mathbf{x}$, we instead maximize the conditional expectation of $\log f_X(\mathbf{x}|\boldsymbol{\theta})$ given $\mathbf{y}$. Thus, for the $(k+1)^{th}$ iteration of the EM algorithm we compute the following (in the form of the aforementioned E- and M-steps):

**E-Step:** Compute $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{[k]})$ where

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) &= \int_{\mathcal{X}(y)} \log f(\mathbf{x}|\boldsymbol{\theta}) f(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}^{[k]})\, d\mathbf{x} \\ &= E[\log f(\mathbf{x}|\boldsymbol{\theta})|\mathbf{y}, \boldsymbol{\theta}^{[k]}]. \end{aligned} \tag{4.2.3}$$

**M-Step:** Choose $\boldsymbol{\theta}^{[k+1]}$ such that

$$Q(\boldsymbol{\theta}^{[k+1]}, \boldsymbol{\theta}^{[k]}) \geq Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{[k]}) \qquad \text{for all} \quad \boldsymbol{\theta} \in \boldsymbol{\Omega}. \tag{4.2.4}$$

The process alternates the E- and M- steps above until some form of convergence criterion is met. Some common criteria include the use of a suitable norm $||\cdot||_p$ such that

$$||\boldsymbol{\theta}^{[k+1]} - \boldsymbol{\theta}^{[k]}||_p < \epsilon \tag{4.2.5}$$

for some choice of $\epsilon > 0$. Common choices of norms are the $L_1$ norm ($p = 1$), and the $L_2$, or Euclidean norm ($p = 2$). Another common stop criterion simply involves assessing the change in

log-likelihood function value at each iteration for the current set of parameters:

$$\ell(\boldsymbol{\theta}^{[k+1]}) - \ell(\boldsymbol{\theta}^{[k]}) \tag{4.2.6}$$

and to stop when the change is deemed insignificant. Although the EM algorithm can be summed up with the expectation and maximization steps outlined above, it may also be useful to expand upon this somewhat. The following steps provide some more detail as to the inner workings of the process:

**Step 1:** For $k = 0$, where $k$ is the current iteration, take a sensible initial estimate $\boldsymbol{\theta}^{[k]}$ for the set of parameters $\boldsymbol{\theta}$.

**Step 2:** Using this current estimate for $\boldsymbol{\theta}$ as well as the observed data $\mathbf{y}$, we calculate the conditional pdf $f(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}^{[k]})$ for the complete data $\mathbf{x}$.

**Step 3:** With this conditional pdf calculated in Step 2, we can formulate the conditional expected log-likelihood as in the E-step:

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) &= \int_{\mathcal{X}(y)} \log f(x|\boldsymbol{\theta}) f(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}^{[k]})\, d\mathbf{x} \\ &= E[\log f(\mathbf{x}|\boldsymbol{\theta})|\mathbf{y}, \boldsymbol{\theta}^{[k]}]. \end{aligned} \tag{4.2.7}$$

**Step 4:** Find the value of $\boldsymbol{\theta}$ that maximizes (4.2.7). Set the resulting estimate as the new estimate $\boldsymbol{\theta}^{[k+1]}$.

**Step 5:** If the chosen convergence criteria are met, then terminate the process. Otherwise increment $m$ to $m = m + 1$ and return to Step 2.

It can be proved that, when performing the iterations of the EM algorithm, the resulting estimate cannot get worse, which is not to say that it will improve with each iteration. The method will often find a peak at the top of the likelihood function, but, in the event that there are multiple potential maxima present (multiple peaks), the EM will not necessarily converge to the true global maximum. As such it is often necessary to perform the EM estimation process for an array of initial values as to discern which point may be the global maximum. Other issues such as the flatness

of the log-likelihood function can also play a part in the ability of the algorithm to converge to a global, or even local maximum point (see Prause 1999 and Barndorff-Nielsen and Blaesild 1981).

### 4.2.1 Estimation of the parameters of the generalized hyperbolic distribution using the EM algorithm

One would not typically think that the EM algorithm can be applied to the generalized hyperbolic distribution. However, the mean-variance representation (see (2.1.2)) of the GH distribution comes in handy here and is well suited for EM estimation. The structure of the EM algorithm that follows is taken from McNeil et al. (2015, pp. 81-83) and Hu (2005, pp. 27-35).

Assume we have a dataset $x_1, x_2, \ldots, x_n$ to which we want to fit a univariate generelized hyperbolic distribution or one of its subclasses. Let $\boldsymbol{\theta} = (\lambda, \chi, \psi, \beta, \mu)$ be the set of parameters we wish to estimate. We then maximize

$$\ell(\boldsymbol{\theta}; \boldsymbol{x}) = \sum_{i=1}^{n} \log f_X(x_i; \boldsymbol{\theta}) \tag{4.2.8}$$

where $f_X(x_i, \boldsymbol{\theta})$ is the pdf of the generalized hyperbolic distribution in (2.1.2). It should already be clear that the estimation is no easy task due to the large number of parameters. However, if we were able to observe the latent mixing variable $W$ in (A.3), this would make the problem much easier. Thus the problem is now to solve the following augmented log-likelihood function:

$$\ell(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{w}) = \sum_{i=1}^{n} \log f_{X_i, W_i}(x_i, w_i; \boldsymbol{\theta}). \tag{4.2.9}$$

Using the normal mean-variance mixture representation of the generalized hyperbolic distribution (see (2.1.7)), we are able to rewrite the log-likelihood function as

$$\begin{aligned} \tilde{\ell}(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{w}) &= \sum_{i=1}^{n} \log f_{X_i|W_i}(x_i|w_i; \mu, \beta) + \sum_{i=1}^{n} \log f_{W_i}(w_i; \lambda, \chi, \psi) \\ &= L_1(\mu, \beta; \boldsymbol{x}|\boldsymbol{w}) + L_2(\lambda, \chi, \psi; \boldsymbol{w}) \end{aligned} \tag{4.2.10}$$

where $X|(W = w) \sim \mathrm{N}(\mu + \beta w, w)$ and $f_{X|W}(x|w)$ is the pdf of the conditional normal distribution, and $f_W(w)$ is the pdf of the GIG mixing variable such that $W \sim \mathrm{GIG}(\lambda, \psi, \chi)$. What makes the use of this representation so powerful is it allows us to maximise $L_1$ and $L_2$ separately, meaning we can maximize $\mu, \beta$ and $\lambda, \psi, \chi$ separately. Since $X|(W = w) \sim \mathrm{N}(\mu + \beta w, w)$, we can write the pdf as

$$f_{X|W}(x|w) = \frac{1}{\sqrt{2\pi w}} e^{-\frac{1}{2}\left(\frac{(x-(\mu+\beta w))^2}{w}\right)} \tag{4.2.11}$$

and we can get the explicit form for the log-likelihood function $L_1$ as

$$\begin{aligned} L_1(\mu, \beta; \boldsymbol{x}|\boldsymbol{w}) = &-\frac{n}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{n}\log w_i \\ &- \frac{1}{2}\sum_{i=1}^{n}\frac{1}{w_i}\left(x_i - (\mu + \beta w_i)\right)^2. \end{aligned} \tag{4.2.12}$$

From (2.1.1), we get the explicit form of the log-likelihood function $L_2$ as

$$\begin{aligned} L_2(\lambda, \chi, \psi; \boldsymbol{w}) = &\frac{n\lambda}{2}\log\psi - \frac{n\lambda}{2}\log(\chi) - 2\log\left(2K_\lambda(\sqrt{\chi\psi})\right) \\ &+ (\lambda - 1)\sum_{i=1}^{n}\log w_i - \frac{1}{2}\sum_{i=1}^{n}(\chi w_i^{-1} + \psi w_i). \end{aligned} \tag{4.2.13}$$

The estimates for $\mu$ and $\beta$ are obtained by maximizing $L_1$. If we suppose that the latent mixing variable $W$ is observable, then we take the partial derivatives of $L_1$ with respect to the parameters $\mu$ and $\beta$ and set them equal to zero as follows

$$\begin{aligned} \frac{\partial L_1}{\partial \mu} &= 0, \\ \frac{\partial L_1}{\partial \beta} &= 0. \end{aligned} \tag{4.2.14}$$

These equations are commonly referred to as the likelihood equations and are typically solved in a simultaneous fashion to get the corresponding maximum likelihood estimates. Solving the above set of equations gives us the following MLEs for $\mu$ and $\beta$

$$\hat{\mu} = \frac{\sum_{i=1}^{n} w_i^{-1}x_i - n\beta}{\sum_{i=1}^{n} w_i^{-1}} \tag{4.2.15}$$

and

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i - n\mu}{\sum_{i=1}^{n} w_i}. \tag{4.2.16}$$

The estimates for $\lambda$, $\chi$, and $\psi$ are similarly obtained through the maximization of $L_2$. For the time being $\lambda$ will be fixed, as it does not seem possible to calibrate $\lambda$ as well. As with $L_1$, we take the partial derivatives of $L_2$ w.r.t $\chi$ and $\psi$ and set them equal to zero as follows

$$\begin{aligned} \frac{\partial L_2}{\partial \chi} &= 0, \\ \frac{\partial L_2}{\partial \psi} &= 0. \end{aligned} \tag{4.2.17}$$

Solving for the above set of likelihood equations is not as straightforward as with $L_1$. We first have to solve for $\zeta = \sqrt{\chi\psi}$ from the following equation

$$n^{-2} \sum_{i=1}^{n} w_i \sum_{j=1}^{n} w_j^{-1} K_\lambda^2(\zeta)\zeta + 2_\lambda(\zeta) - \zeta K_\lambda^2(\zeta) = 0. \tag{4.2.18}$$

After solving for $\zeta$, we are able to get the following expressions for $\chi$ and $\psi$

$$\hat{\chi} = \frac{n^{-1}\zeta \sum_{i=1}^{n} w_i K_\lambda(\zeta)}{K_{\lambda+1}(\zeta)} \tag{4.2.19}$$

and

$$\hat{\psi} = \frac{\zeta^2}{\chi}. \tag{4.2.20}$$

However, contrary to previous assumptions, the latent mixing variables $W_1, W_2, \ldots, W_n$ are not observable. We thus need an iterative setup consisting of an E-step and M-step as with the EM-algorithm. In the E-step, we calculate the conditional expectation of the augmented log-likelihood function given the current parameter estimates and the data. Suppose were are at step $k$ in the iterative procedure. The goal then is to calculate the following conditional expectation and get a new function to maximize:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{[k]}) = E\left(\log \tilde{L}(\boldsymbol{\theta}; \boldsymbol{x}, W_1, \ldots, W_n) | \boldsymbol{x}; \boldsymbol{\theta}^{[k]}\right). \tag{4.2.21}$$

In the M-step, we maximize the function in (4.2.21) and get a new set of estimates $parm^{[k+1]}$. We can observe from (4.2.12) and (4.2.13), that this is equivalent to replacing the $w_i$, $w_i^{-1}$, and $\log w_i$ terms in the augmented log-likelihood function by their conditional estimates $E(W_i|x_i;\boldsymbol{\theta}^{[k]})$, $E(W_i^{-1}|x_i;\boldsymbol{\theta}^{[k]})$, and $E(\log(W_i)|x_i;\boldsymbol{\theta}^{[k]})$. The function $Q(\boldsymbol{\theta},\boldsymbol{\theta}^{[k]})$ is thus re-expressed by observations and known conditional expectations such that it can be maximized. The following expressions follow if $W \sim \text{GIG}(\lambda,\chi,\psi)$, and will be needed later:

$$E(W^\alpha) = \left(\frac{\chi}{\psi}\right)^{\frac{\alpha}{2}} \frac{K_{\lambda+\alpha}(\sqrt{\chi\psi})}{K_\lambda(\sqrt{\chi\psi})}, \tag{4.2.22}$$

and

$$E(\log W) = \frac{dE(W^\alpha)}{d\alpha}|_{\alpha=0}. \tag{4.2.23}$$

While we can use (4.2.22) directly, especially for $\alpha = -1$ and $\alpha = 1$ as required, (4.2.23) will require numerical methods to solve. Fortunately, if one looks at (4.2.13), the $E(\log(W_i)|x_i;\boldsymbol{\theta}^{[k]})$ term is only needed if we solve for $\lambda$, and since we fix $\lambda$ this term is not needed. Following the notation as in McNeil et al. (2015, pp. 81-83), let

$$\eta_i^{[k]} = E\left(W_i|x_i;\boldsymbol{\theta}^{[k]}\right), \tag{4.2.24}$$

$$\delta_i^{[k]} = E\left(W_i^{-1}|x_i;\boldsymbol{\theta}^{[k]}\right), \tag{4.2.25}$$

$$\xi_i^{[k]} = E\left(\log\left(W_i\right)|x_i;\boldsymbol{\theta}^{[k]}\right). \tag{4.2.26}$$

Then, by using (4.2.22) and (4.2.23) we get the following

$$\eta_i^{[k]} = \sqrt{\frac{\chi}{\psi}}\frac{K_{\lambda+1}(\sqrt{\chi\psi})}{K_\lambda(\sqrt{\chi\psi})}, \tag{4.2.27}$$

and

$$\delta_i^{[k]} = \sqrt{\frac{\psi}{\chi}}\frac{K_{\lambda-1}(\sqrt{\chi\psi})}{K_\lambda(\sqrt{\chi\psi})}. \tag{4.2.28}$$

Now we just replace the latent variables in the M-step with the corresponding conditional expec-

tations, giving us the following estimates at the $k^{th}$ iteration:

$$\mu^{[k+1]} = \frac{\sum_{i=1}^{n} \delta_i^{[k]} x_i - n\beta^{[k]}}{\sum_{i=1}^{n} \delta_i^{[k]}}, \tag{4.2.29}$$

$$\beta^{[k+1]} = \frac{\sum_{i=1}^{n} x_i - n\mu^{[k+1]}}{\sum_{i=1}^{n} \eta_i^{[k]}}, \tag{4.2.30}$$

$$\chi^{[k+1]} = \frac{n^{-1}\zeta \sum_{i=1}^{n} \eta_i^{[k]} K_\lambda(\zeta)}{K_{\lambda+1}(\zeta)}, \tag{4.2.31}$$

$$\psi^{[k+1]} = \frac{\zeta^2}{\chi^{[k+1]}}. \tag{4.2.32}$$

All that follows now is to repeat the process and update the estimates at each iteration until the selected convergence criteria are met.

## 4.3 A new method: profile likelihood based alternating algorithm

One of the issues of using the full log-likelihood function is that the overall region is quite flat. This leads to optimization methods either stopping/getting stuck before reaching the global maximum, or progressing in the wrong search direction altogether. This issue is not limited to the numerical methods used here, but is a consequence of numerical methods as a whole when a flat function region is present. This stems from the fact that whether the method makes use of the function gradient or simply the function values, the steepness of the functions region, especially around a global or local maximum will have an impact on the methods ability to correctly converge.

In much the same way we use the profile log-likelihood to allow for interpretable visual results, we can use it to split the estimation process up into two parts instead of simultaneously estimating all the parameters. By analysing the marginal (profile) log-likelihood functions and splitting the parameter set in two, this creates a desirable situation whereby the functions to be maximised are better behaved than the full log-likelihood. This is due to the lower dimensional space in which the function is minimized which in turn allows for better performance of the numerical methods.

The formulation that follows is for the subclasses of the GH distribution that have fixed $\lambda$.

In a similar fashion to the profile likelihood, the parameter vector $\boldsymbol{\theta} = (\alpha, \beta, \delta, \mu)$ is split into $\boldsymbol{\xi} = (\alpha, \beta)$, and $\boldsymbol{\eta} = (\delta, \mu)$. This choice of split was decided based on trial and error, and it was found that this pairing led to convergence of the algorithm. It is of course possible that other pairings could very well lead to the same outcome, but this pairing seemed to have the most consistent convergence. The first log-likelihood function will maximized on $\boldsymbol{\xi} = (\alpha, \beta)$, with $\boldsymbol{\eta} = (\delta, \mu)$ fixed, and the second will be maximized on $\boldsymbol{\eta} = (\delta, \mu)$, with $\boldsymbol{\xi} = (\alpha, \beta)$ fixed. The equations to be maximized are thus given by

$$L_{p_1}(\boldsymbol{\xi}) = \sup_{\boldsymbol{\eta}} L(\boldsymbol{\xi}, \boldsymbol{\eta}) \tag{4.3.1}$$

and

$$L_{p_2}(\boldsymbol{\eta}) = \sup_{\boldsymbol{\xi}} L(\boldsymbol{\xi}, \boldsymbol{\eta}). \tag{4.3.2}$$

This algorithm is implemented as follows:

**Step 1:** Set $k = 0$. Select a set of initial values $\boldsymbol{\theta_k} = (\alpha_k, \beta_k, \delta_k, \mu_k)$.

**Step 2:** Maximize $L_{p_2}(\boldsymbol{\eta})$ with respect to $\boldsymbol{\eta} = (\delta, \mu)$ giving us parameters $\mu_{k+1}$ and $\delta_{k+1}$.

**Step 3:** Set $\mu_k = \mu_{k+1}$ and $\delta_k = \delta_{k+1}$.

**Step 4:** Maximize $L_{p_1}(\boldsymbol{\xi})$ with respect to $\boldsymbol{\xi} = (\alpha, \beta)$ giving us parameters $\alpha_{k+1}$ and $\beta_{k+1}$.

**Step 5:** Set $\alpha_k = \alpha_{k+1}$ and $\beta_k = \beta_{k+1}$.

**Step 6:** If $|\boldsymbol{\theta_{k+1}} - \boldsymbol{\theta_k}| > \epsilon$, set $k = k + 1$ and return to step 2.

**Step 7:** If the maximum $|\boldsymbol{\theta_{k+1}} - \boldsymbol{\theta_k}| < \epsilon$ we terminate the process, where $\epsilon$ is the smallest acceptable variation for the process to repeat (typically chosen to be between $10^{-3}$ and $10^{-6}$).

There was a rather interesting discovery, in that the addition of $\lambda$ into the estimation process is actually possible. In fact the resulting estimates corresponded with those of the other methods used. For clarity, $\lambda$ was introduced into the estimation process by setting $\boldsymbol{\xi} = (\lambda, \alpha, \beta)$, $\boldsymbol{\eta} = (\delta, \mu)$, and simply proceeding as before. While including $\lambda$ in this fashion yielded positive results, it was found when testing other parameter combinations that success is not guaranteed.

It is therefore advised that the reader defines $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ as sensibly as possible, with the alternative being simple trial and error. In practice it was found that setting $\boldsymbol{\eta} = (\delta, \mu)$, and setting $\boldsymbol{\xi}$ to the remaining parameters was quite reliable. A step-by-step outline of the algorithm is given in Algorithm 1.

**Algorithm 1** Profile likelihood based alternating algorithm

**Step 1:** Set iteration number $k = 0$.

**Step 2:** Choose initial values $\boldsymbol{\theta_k} = (\alpha_k, \beta_k, \delta_k, \mu_k)$.

**Step 3:** Maximize $L_{p_2}(\boldsymbol{\eta})$ w.r.t. $\boldsymbol{\eta} = (\delta, \mu)$, yielding parameters $\mu_{k+1}$ and $\delta_{k+1}$.

**Step 4:** Set $\mu_k = \mu_{k+1}$ and $\delta_k = \delta_{k+1}$.

**Step 5:** Maximize $L_{p_1}(\boldsymbol{\xi})$ w.r.t. $\boldsymbol{\xi} = (\alpha, \beta)$, yielding parameters $\alpha_{k+1}$ and $\beta_{k+1}$.

**Step 6:** Set $\alpha_k = \alpha_{k+1}$ and $\beta_k = \beta_{k+1}$.

**Step 7:** If $|\boldsymbol{\theta_{k+1}} - \boldsymbol{\theta_k}| > \epsilon$, set $k = k + 1$ and return to step 2.

**Step 8:** If $|\boldsymbol{\theta_{k+1}} - \boldsymbol{\theta_k}| < \epsilon$, where $\epsilon$ is smallest acceptable variation for the process to repeat (typically chosen to be between $10^{-3}$ and $10^{-6}$).

# Chapter 5

# Application

This chapter contains a full discussion of the findings when fitting the GH distribution to data. A full exploratory analysis for a selection of popular datasets is performed, whereby we look at the behaviour of the log-likelihood function, and the resulting impact on the estimation process. We also look at the importance of initial value selection, as well as the importance of selecting an appropriate subclass of the GH distribution. The estimation results are analysed and compared using some popular goodness-of-fit statistics, and finally, a simulation study is conducted to shed more light on some of the findings when fitting the GH distribution to the real world data sets.

The exploration aspect is carried out in the R Software environment, with some deviation to other platforms that provide similar function optimization routines. This is, however, purely for comparative reasons and as such the focus of the discussion that follows will be on the process in R. Since we are working with the generalized hyperbolic distribution, it is natural to make use of any packages that accommodate this distribution and make the process easier. There are three notable packages that deal with the generalized hyperbolic distribution, namely:

1. The ghyp package .
2. The GeneralizedHyperbolic package.
3. The HyperbolicDist package.

The above packages offer very similar functions and toolkits, and as such there is some overlap in the core functions offered. Much of the variation comes from the auxiliary functions and options

and as such these packages will be used in tandem for the most part. It is worth noting that there is some dependency between the above packages, and although some function calls appear the same, they are in fact uniquely defined in each package and as a result some care is needed when working between the packages. An apparent example can be seen with the function `pghyp()`. Below is an illustration of how the functions are defined in each of the respective packages :

1. **ghyp package**

```
pghyp(q, object = ghyp(), n.sim = , subdivisions = ,
rel.tol = ,abs.tol = , lower.tail = )
```

2. **GeneralizedHyperbolic package**

```
pghyp(q, mu = , delta = , alpha = , beta = , lambda = ,
param = c(mu, delta, alpha, beta, lambda),
lower.tail = , subdivisions = ,ntTol = ,
valueOnly = , ...)
```

3. **HyperbolicDist package**

```
pghyp(q, Theta, small = , tiny = ,deriv = ,
subdivisions = ,accuracy = , ...)
```

At the time of writing, the `HyperbolicDist` package provided the more tractable function options and, as a result, was the package primarily used.

## 5.1 Data sets

### 5.1.1 NYSE composite index

The NYSE Composite Index measures the performance of all common stocks listed on the New York Stock Exchange, including American Depositary Receipts issued by foreign companies, Real Estate Investment Trusts and tracking stocks. The weights of the index constituents are calculated on the basis of their free-float market capitalization. The index itself is calculated on the basis of price return and total return, which includes dividends. The breadth of the NYSE Composite

Index makes it a far better indicator of market performance than narrow indexes that have far fewer components.

This data set is chosen as it is used in Prause (1999), which is widely considered a useful resource on the topic of the GH distribution and related subfamilies. This data is readily available on most stock exchange websites as this is a relatively popular index. The NYSE composite index dataset covers the daily high from January 2, 1990 to November 29, 1996. We fit the GH distribution and subclasses to the log-returns of the data. The log-return of a price $(S_t)_{t\geq 0}$ for time interval $\vec{\Delta}t$ (in this instance one day) is defined as

$$X_t = \log S_t - \log S_{t-\vec{\Delta}t}. \tag{5.1.1}$$

where $X_t$ is the stock price at time $t$. The return during $n$ periods is thus the sum of the single period returns.

### 5.1.2 S&P 500 index

The following data set provides the year end prices of Standard and Poorś most notable stock market price index, the S&P 500. It contains the year end price of the index from 1800 through to 2001 and contains 201 observations. The data is taken from Brown et al. (2002), and can also be found in the GeneralizedHyperbolic package in the R software environment. For this study we will be looking at the proportional changes of the stock price:

$$Y_t = \frac{X_{t-1}}{X_t}, \tag{5.1.2}$$

where $Y_t$ is the ratio of the stock price at time $t-1$ to the ratio at time $t$.

## 5.2 Fitting GH distributions to the data

Due to the large number of parameters in the generalized hyperbolic distribution, as well as the form of the pdf, numerical methods will be needed in order to estimate the parameters. Fortunately, R has an abundance of resources to deal with this. Commonly used functions for optimization in-

clude nlminb(), which performs unconstrained as well as box constrained optimizations using PORT routines (Quasi-Newton), and the optim() function, which performs general-purpose optimization base on Nelder-Mead, quasi-Newton and conjugate-gradient algorithms, and the constrOptim() function, which provides the same functionality as the optim() function, while also allowing for the inclusion of linear inequality constraints between the parameters. There are naturally many other functions that perform similar optimization routines but with some slight variation.

There are a number of aspects of the Generalized Hyperbolic distribution that have an impact on the estimation process, and the consequent restrictions this imposes on available methods.The biggest issue in the context of exploring the log-likelihood is the lack of a closed-form derivative for the pdf (see equation (2.1.5)). This is primarily due to the presence of Bessel functions in the pdf (see equations (2.1.5) and (B.1)). It is thus not possible to make use of gradient based methods when performing numerical optimization, and in the few cases that the procedure functions normally, the relevant optimization routine attempts to calculate and work with an approximation of the gradient of the log-likelihood. This is, however, extremely unreliable, as in the majority of the practical findings this does not work.

Another aspect of the GH distribution that needs to be considered is the parameters themselves. Not only are there a large number of parameters to estimate, but there are a staggering number of parameterizations that have been proposed in the literature over the years. The two most commonly used parametrizations are the $(\lambda, \psi, \beta, \chi, \mu)$ parametrization that follows naturally from the derivation of the GH distribution (see (2.1.2)), and the $(\lambda, \alpha, \beta, \delta, \mu)$ parametrization proposed by Barndorff-Nielsen (1978). For the remaining parametrizations please refer to section (2.3).

The most widely used parametrization seems to be the $(\lambda, \alpha, \beta, \delta, \mu)$ parametrization proposed by Barndorff-Nielsen (1978), and as such this will be the parametrization used to fit the GH distribution and related subtypes to our real world datasets. An important aspect that comes with using this parametrization is not only the general parameter constraints inherent to the GH distribution, but the bounded relationship between the $\alpha$ and $\beta$ parameters as well. The constraint $|\beta| < |\alpha|$ creates a boundary that is rather problematic for general purpose optimization routines, especially those that do not explicitly account for this constraint. This is generally only a problem when

the parameters are simultaneously estimated with methods such as the Nelder-Mead algorithm, or related simultaneous routines.

There is, fortunately, a function in R that can account for not only this constraint, but any constraints that the function to be maximized may have. The constrOptim() function performs optimization subject to a linear inequality using an adaptive barrier algorithm. The function essentially allows for the explicit specification of constraints on the parameter space beforehand, and takes this into account when performing the optimization, thus mitigating the aforementioned issue. Preliminary findings indicate a substantial improvement when using constrOptim() over the existing methods that do not account for the constraints.

A last note, but certainly not the least important, is that of the shape of the log-likelihood function of the GH distribution. Prause (1999), Protassov (2004), Aas and Haff (2006), Barndorff-Nielsen and Blaesild (1981), and Snoussi and Idier (2006) all report on the GH distribution having a flat log-likelihood function. There are many consequences to this, the first being that it is rather difficult to accurately estimate $\lambda$ (the parameter responsible for the subclasses). This is especially true when smaller sample sizes are used, and in fact it will later be shown that even in larger sample instances, the estimation of this parameter is quite unstable. To mitigate this issue, many papers resort to fixing $\lambda$ to a value corresponding to a known subclass.

While this may be a decent circumvention to the problem, it is by no means a clean cut solution. In Prause (1997) it is indicated that the NIG subclass is a rather good fit for financial asset data. It is, however, also shown in Barndorff-Nielsen (1995) that different subclasses can have near identical densities, where a hyperbolic ($\lambda = 1$), and a normal inverse Gaussian ($\lambda = -0.5$) distribution are shown to be almost the same. This is an unfortunate caveat of having a flat log-likelihood function, as not only is it more difficult for numerical methods to find the global maximum of the log-likelihood, but there may be multiple estimate permutations leading to seemingly identical densities.

For the process of fitting the GH distribution to the NYSE composite index data, as well as the S&P 500 data, we will make use of the following methods:

1. Nelder-Mead simplex method (see section 4.1).

2. Expectation Maximization (EM) algorithm(see section 4.2).

3. Alternating profile likelihood based algorithm(see section 4.3).

The Nelder-Mead simplex method is a widely used numerical method and is especially useful in that it does not require the calculation or use of derivatives. This is very useful when working with the GH distribution as there does not exist a closed form for the derivative of the pdf. It is also one of the method options within the optim() function, a commonly used function for numerical optimization in R. The use of this method will also serve as a good baseline for assessing the performance of the ML estimation of the GH parameters in the methods that follow.

An extension of the optim() function, namely the constrOptim() function, will also be considered. This function allows for specification of linear inequality constraints while using the Nelder-Mead simplex method to minimize the objective function. This allows us to consider $\alpha > 0$, $\delta > 0$, and especially the bounded relationship $|\beta| < \alpha$.

The EM algorithm is one of the most commonly used methods in the literature for the estimation of the GH parameters. Prause (1999), McNeil et al. (2015), Aas and Haff (2006), Hu (2005), Karlis (2002), Hellmich and Kassberger (2011), and Panahi (2018) all make use of the EM algorithm to fit the GH model or one of its subclasses to data. See section 4.2.1 for a breakdown of the EM algorithm in the context of the GH distribution. A limitation of the EM algorithm as given in 4.2.1 is that it is not possible to report is inspired. The current solution is to proceed by fixing the value of $\lambda$ to a value corresponding to one of the GH distribution subclasses, as is done in majority in the literature.

Before diving into the specifics, it is useful to begin with an illustration of the log-likelihood function to get a better sense of what we are working with. Since the log-likelihood function is a function of five variables (four if we count $\lambda$ as fixed), we need to make use of profiling the log-likelihood function in order to allow a visual representation in lower-dimensional space. As outlined in section 3.1.1, we reduce the dimensionality of the log-likelihood function by fixing a set of nuisance parameters. The majority of what follows will focus on the $\alpha$ and $\beta$ parameters of the GH distribution, as the bounded relationship between these parameters, specifically that $|\beta| < \alpha$, and the ensuing feasible region it creates is of particular interest in the context of exploration. As such we fix the $\mu$ and $\delta$ parameters giving us the following log-likelihood function as in section

3.1.1:

$$\ell_p(\boldsymbol{\xi}) = \sup_{\boldsymbol{\eta}} \ell(\boldsymbol{\xi}, \boldsymbol{\eta}) \tag{5.2.1}$$

Figure 5.1 provides a graphical representation of the profile log-likelihood in terms of $\alpha$ and $\beta$. The profile log-likelihood is transformed into the likelihood ratio test statistic (also called deviance) as follows

$$D(\alpha, \beta) = 2\left\{\ell_p(\hat{\alpha}, \hat{\beta}) - \ell_p(\alpha, \beta)\right\}, \tag{5.2.2}$$

where $\ell_p(\hat{\alpha}, \hat{\beta})$ is the overall maximum of $\ell_p(\alpha, \beta)$.

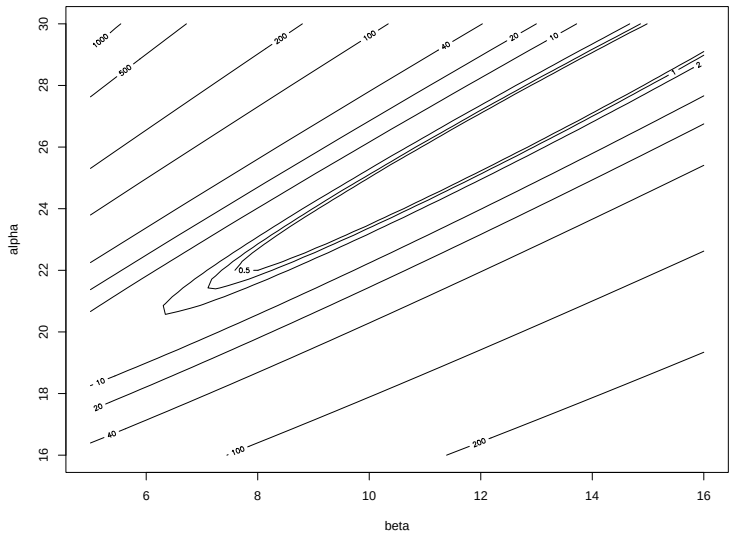


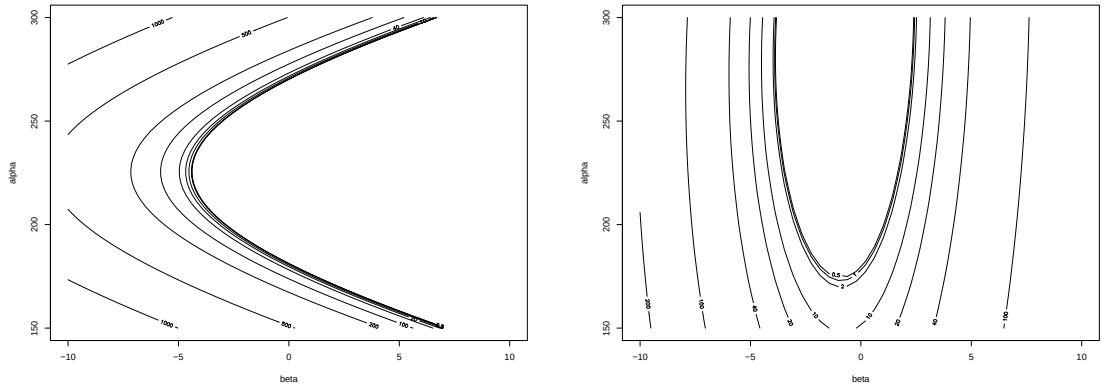

Figure 5.1: Contour level plot of the deviance function. Data simulated from GH(4,20,10,1,0) with sample size $n = 50$

One of the issues with displaying the profile likelihood in this fashion, i.e. by fixing the nuisance parameters, is that one does not always capture the behaviour of the log-likelihood function in full. This is due to the fact that by fixing $\mu$ and $\delta$ in this way, we are effectively "snapshotting" the log-likelihood function at these values. As a consequence, we are no longer able to get an idea of the behaviour of this function in its entirety. To expand on this, consider the instance that we standardize $\mu$ and $\delta$, i.e. set $(\mu, \delta) = (0, 1)$. This may create and entirely different region for $D(\alpha, \beta)$ compared to, say, setting $(\mu, \delta) = (0, 0.001)$.

This can also be seen in Figures 5.2a and 5.2b, where different selections for $\mu$ and $\delta$ result

in significantly different behaviours of the log-likelihood function. Raue et al. (2009) discusses two different kinds of parameter non-identifiability that can be seen here. The first is structural non-identifiability, which relates to the model itself and does not depend on the underlying data. The second is practical non-identifiability, which takes into account the underlying data.

Structural non-identifiability is a likely indicator of redundant parameters, whereas practical non-identifiability indicates that the sample size may be to small, or that the data itself may not be suited to the model being considered. Figure 5.2a displays similar visual properties to the structural non-identifiability described in Raue et al. (2009), whereas Figure 5.2b indicates practically non-identifiable behaviour. For a more detailed breakdown and description please see Raue et al. (2009).




(a) deviance contour for $(\mu, \delta) = (0, 1)$.

(b) Deviance contour for $(\mu, \delta) = (0, 0.001)$.

Figure 5.2: Contour plots in terms of the deviance (as defined in (5.2.2)) for the $\alpha$ and $\beta$ parameters.

To circumvent the issues of displaying the profile likelihood in this fashion, the following is proposed. Instead of analysing the behaviour of the profile log-likelihood function as before, we initialize a grid of starting values $\boldsymbol{\theta_0} = (\alpha_0, \beta_0, \delta_0, \mu_0)$, and proceed to fit a GH model to the both datasets using the Nelder-Mead simplex method. The resulting final estimates $(\hat{\alpha}, \hat{\beta})$ for each starting value choice are plotted against the corresponding log-likelihood value at each point. The purpose of this is to provide a means of illustrating the behaviour of the log-likelihood function without needing to fix the values of the nuisance parameters $\mu$ and $\delta$. The expectation is that this perceived behaviour of the log-likelihood function will be better illustrated.

First, for each starting value in the grid, unconstrained optimization is performed using the Nelder-Mead simplex method. This is done by means of the optim() function in R. The resulting estimates and corresponding log-likelihood values are recorded for each value in the grid. This is done to allow us to visualise the behaviour of the log-likelihood function in terms of $\alpha$ and $\beta$, but without having to fix $\delta$ and $\mu$, and lose out on the impact these parameters may have on the shape of the log-likelihood function. This process is done for the generalized hyperbolic model, as well as the hyperbolic, and normal inverse Gaussian subclasses.

The resulting estimates for $\alpha$ and $\beta$ are plotted against one another. Each $(\alpha, \beta)$ point is grouped according to its associated log-likelihood value as this allows us to artificially construct a contour-like region, and thus graphically depict the perceived behaviour of the log-likelihood function. In Figures 5.3-5.5 we observe the results of fitting the generalized hyperbolic, hyperbolic, and normal inverse Gaussian models to the NYSE composite index data using optim().


Figure 5.3: Scatterplot of $(\hat{\alpha}, \hat{\beta})$ estimates: NYSE Composite Index GH fit.

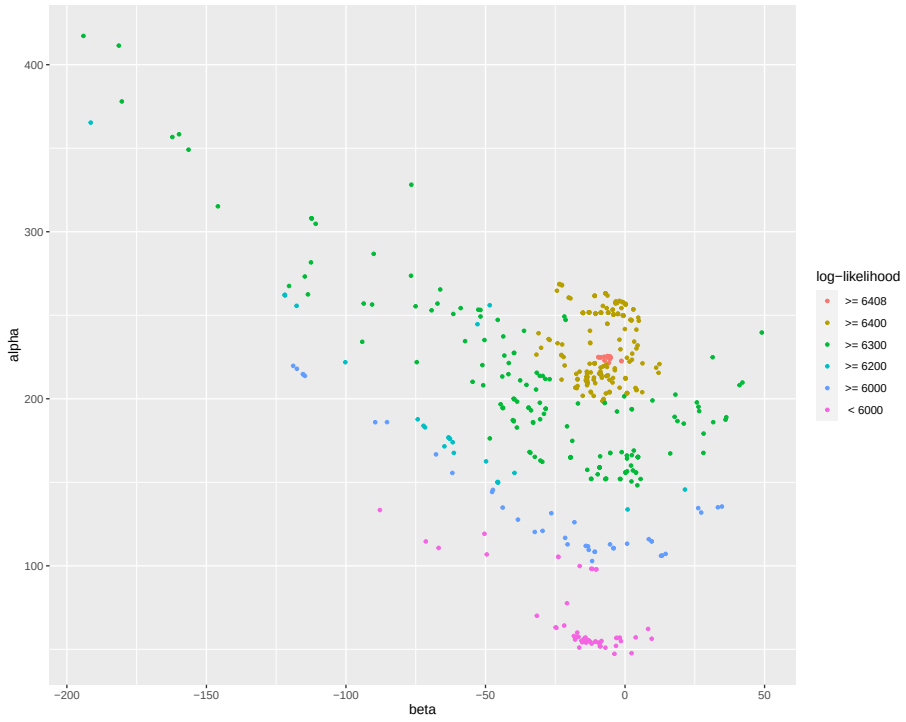



Figure 5.4: Scatterplot of $(\hat{\alpha}, \hat{\beta})$ estimates: NYSE Composite Index hyperbolic fit.


Figure 5.5: Scatterplot of $(\hat{\alpha}, \hat{\beta})$ estimates: NYSE Composite Index NIG fit.

Looking at Figures 5.3 and 5.4, there is clear indication of this perceived flatness of the log-likelihood function. This is especially apparent in Figure 5.3, where we have this consistent region spanning from $\alpha = 0$ all the way to $\alpha = 200$ with a negligible difference in log-likelihood values. This corresponds to statements made in the literature on the impact that $\lambda$ has on the flatness of the log-likelihood function. We also see that even when $\lambda$ is fixed in the hyperbolic case, we still have somewhat of a flat linear region, although now not as pronounced. Another interesting aspect of the hyperbolic fit in Figure 5.4 is that there seems to be some linearity along the lower bound of the $|\beta| < \alpha$ constraint, and to a lesser extent the upper bound. This linearity could indicate possible identifiability issues stemming from the $\alpha$ and $\beta$ parameters. This linearity can also be seen in Figure 5.3, although not as much.

When looking at Figure 5.5, we see that the points are rather nicely spread in a circular fashion around a single peak. This peak does in fact corresponded to the global maximum, and although the log-likelihood values are still quite close in magnitude, this function seems to be much better behaved than those corresponding to the hyperbolic and generalized hyperbolic fits. It has been stated in the literature (see Prause, 1999) that the normal inverse Gaussian subclass is well suited to model financial returns and financial data, and this is confirmed by the outcome in Figure 5.5.

In order to account for potential point overlap, Tables 5.1 - 5.3 have been included to compliment Figures 5.3 - 5.5. This is done simply to ensure that the figures are not misleading, and to give us an idea of how many points lie in each of the pre-defined log-likelihood brackets.

This process will now be repeated using the constrOptim() function. Recall the constrOptim() function has the same functionality as optim(), but allows for the specification of linear constraints. This is useful as three out of the five parameters of the GH distribution are bound by such constraints, and as will now be shown, the consideration of these constraints are quite important when estimating the GH distribution or one of its subtypes. In Figures 5.6 - 5.8 we observe the results of fitting the generalized hyperbolic, hyperbolic, normal inverse Gaussian, and hyberbolic asymmetric *t* models to the NYSE composite index data using constrOptim().

Comparing the value spread in Figure 5.6 with that of Figure 5.3, it is clear that the consideration of constraints has a positive impact on the outcome of the estimation process. Although there now seems to be a form indicating a local as well as a global maximum, this is much better than

Table 5.1: The number of estimates falling in each log-likelihood bracket corresponding to the GH distribution as in Figure 5.3.

| **Log-likelihood Value** | **Number of Estimates** |
|---|---|
| $\geq 6406$ | 240 |
| $\geq 6400$ | 1611 |
| $\geq 6300$ | 835 |
| $\geq 6200$ | 34 |
| $\geq 6100$ | 7 |
| $\geq 6000$ | 9 |
| $< 6000$ | 258 |

Table 5.2: The number of estimates falling in each log-likelihood bracket corresponding to the hyperbolic subclass as in Figure 5.4.

| **Log-likelihood Value** | **Number of Estimates** |
|---|---|
| $\geq 6408$ | 117 |
| $\geq 6400$ | 202 |
| $\geq 6300$ | 172 |
| $\geq 6200$ | 35 |
| $\geq 6000$ | 44 |
| $< 6000$ | 55 |

the large region spanning $\alpha = 0$ to $\alpha = 200$ that was found in Figure 5.3. The same improvements can be observed in Figures 5.7 and 5.8, where the points are much more concentrated about the maximum. The only problem that seems to persist is that of the perceived collinearity between $\alpha$ and $\beta$. Even with the inclusion of the parameter constraints in the estimation process we still observe this relation, in fact it is made even clearer in this instance. It is again observed that the

Table 5.3: The number of estimates falling in each log-likelihood bracket corresponding to the NIG subclass as in Figure 5.5.

| **Log-likelihood Value** | **Number of Estimates** |
|---|---|
| $\geq 6406$ | 149 |
| $\geq 6400$ | 195 |
| $\geq 6300$ | 281 |

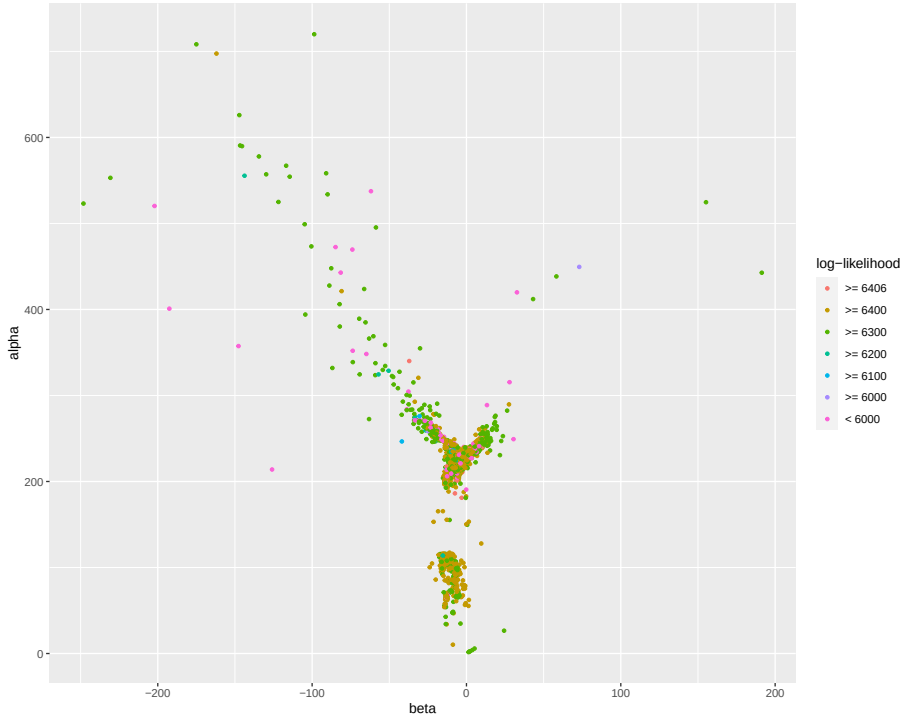



Figure 5.6: Scatterplot of $(\hat{\alpha}, \hat{\beta})$ estimates for GH fit: NYSE composite index.

difference in log-likelihood values is quite small even in the instances where $\lambda$ is fixed. As before, Tables 5.4 - 5.6 have been included to compliment Figures 5.6 - 5.8.

The chosen grid spans 625 points when $\lambda$ is fixed at the outset, and 3125 when we estimate $\lambda$ (GH model). To give a crude idea of how the final estimates compare with the initial grid of points, Figure 5.9 provides the same scatterplot of points as Figure 5.3, but with the inclusion of a visual indicator of the initial value grid for $\alpha$ and $\beta$ (in RED). It is clear that many of initial value permutations resulted in divergent behaviour rather then staying in the realm of the global

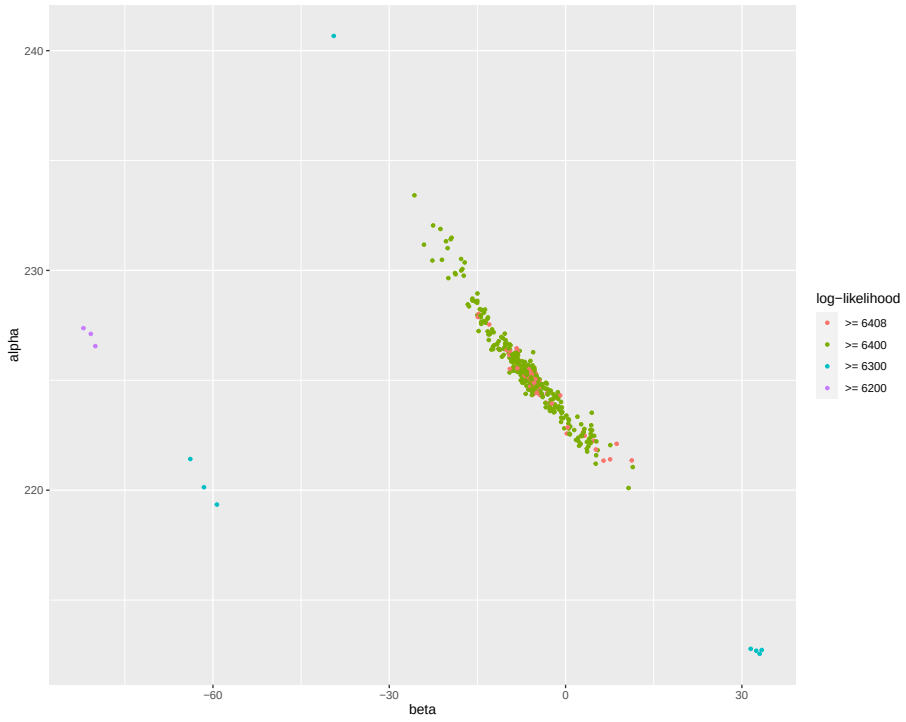



Figure 5.7: Scatterplot of $(\hat{\alpha}, \hat{\beta})$ estimates for hyperbolic fit: NYSE composite index.

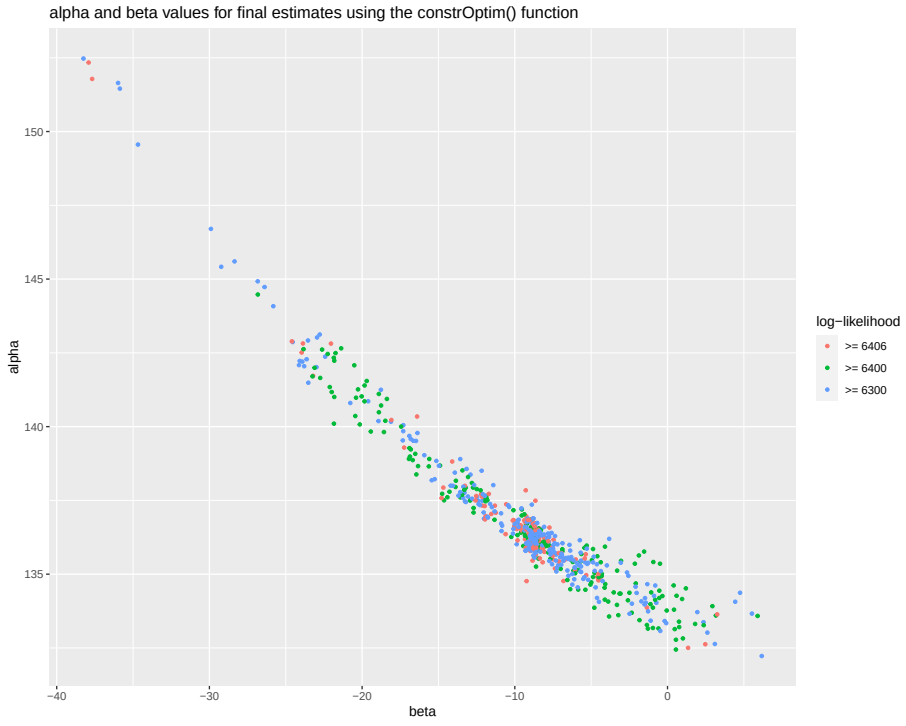



Figure 5.8: Scatterplot of $(\hat{\alpha}, \hat{\beta})$ estimates for NIG fit: NYSE composite index.

Table 5.4: The number of estimates falling in each log-likelihood bracket corresponding to the GH distribution as in Figure 5.6.

| **Log-likelihood Value** | **Number of Estimates** |
|---|---|
| $\geq 6406$ | 241 |
| $\geq 6400$ | 1611 |
| $\geq 6300$ | 897 |
| $\geq 6200$ | 55 |
| $\geq 6100$ | 13 |
| $\geq 6000$ | 17 |
| $< 6000$ | 291 |

Table 5.5: The number of estimates falling in each log-likelihood bracket corresponding to the hyperbolic subclass as in Figure 5.7.

| **Log-likelihood Value** | **Number of Estimates** |
|---|---|
| $\geq 6408$ | 117 |
| $\geq 6400$ | 202 |
| $\geq 6300$ | 172 |
| $\geq 6200$ | 35 |
| $\geq 6000$ | 44 |
| $< 6000$ | 55 |

maximum. This may be due the the associated $\mu$ and $\delta$ values, but upon observing the resulting set of points and comparing them with the initial values, this cannot be said as the values conflict with this notion.

Table 5.6: The number of estimates falling in each log-likelihood bracket corresponding to the NIG subclass as in Figure 5.8.

| **Log-likelihood Value** | **Number of Estimates** |
|---|---|
| $\geq 6406$ | 149 |
| $\geq 6400$ | 195 |
| $\geq 6300$ | 281 |

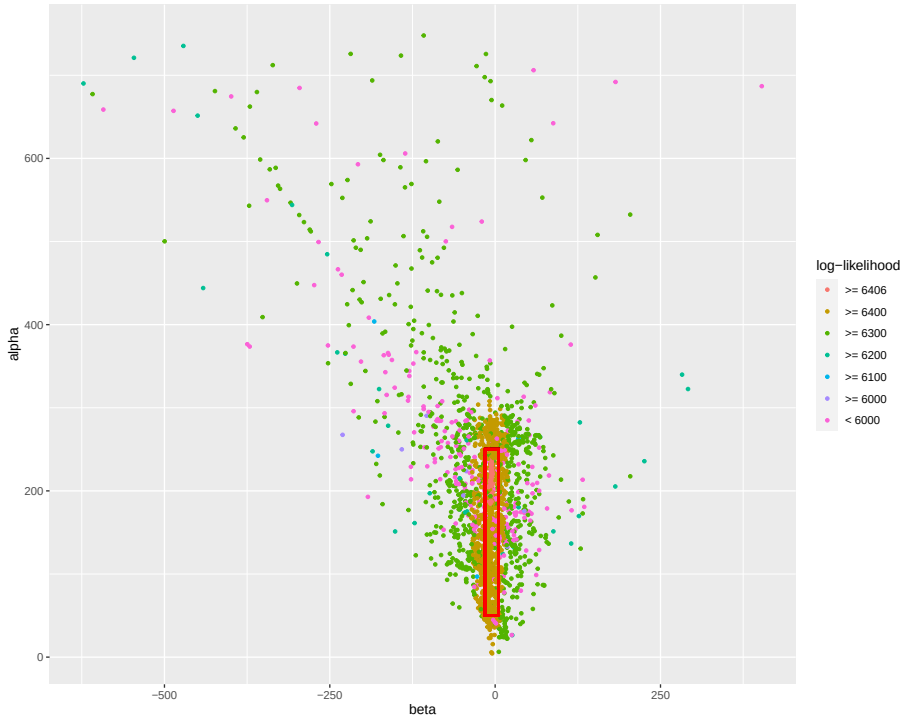



Figure 5.9: Scatterplot of $(\hat{\alpha}, \hat{\beta})$ estimates for GH fit with initial value overlay: NYSE composite index.

We observe similar behaviours when fitting GH models to the S&P 500 Index data. As before, in Figures 5.10-5.12 we observe the results of fitting the generalized hyperbolic, hyperbolic, and normal inverse Gaussian models to the data. What stands out here, is while there were indicators of dependency between the $\alpha$ and $\beta$ parameters when fitting the models to the NYSE data, Figures 5.11 and 5.12 provide a strong indication of dependency between $\alpha$ and $\beta$. There is a strong linear region for the $\alpha$ and $\beta$ parameters, all corresponding to approximately the same log-likelihood

value. This is a strong indicator of a potential structural non-identifiability (see Raue et al., 2009).


Figure 5.10: Scatterplot of $(\hat{\alpha}, \hat{\beta})$ estimates: S&P 500 Index GH fit.

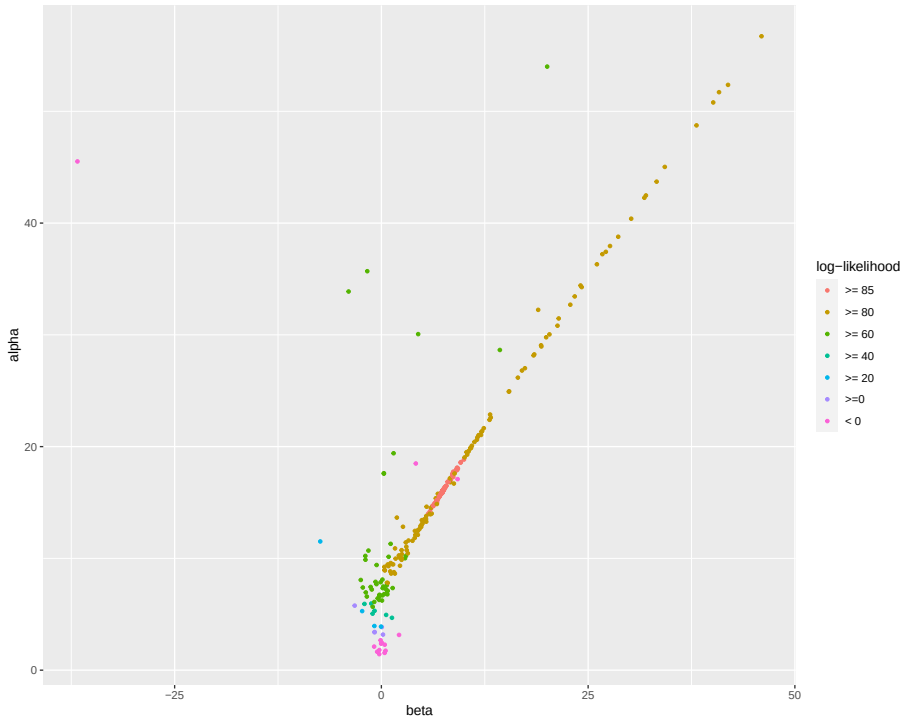



Figure 5.11: Scatterplot of $(\hat{\alpha}, \hat{\beta})$ estimates: S&P 500 Index hyperbolic fit.

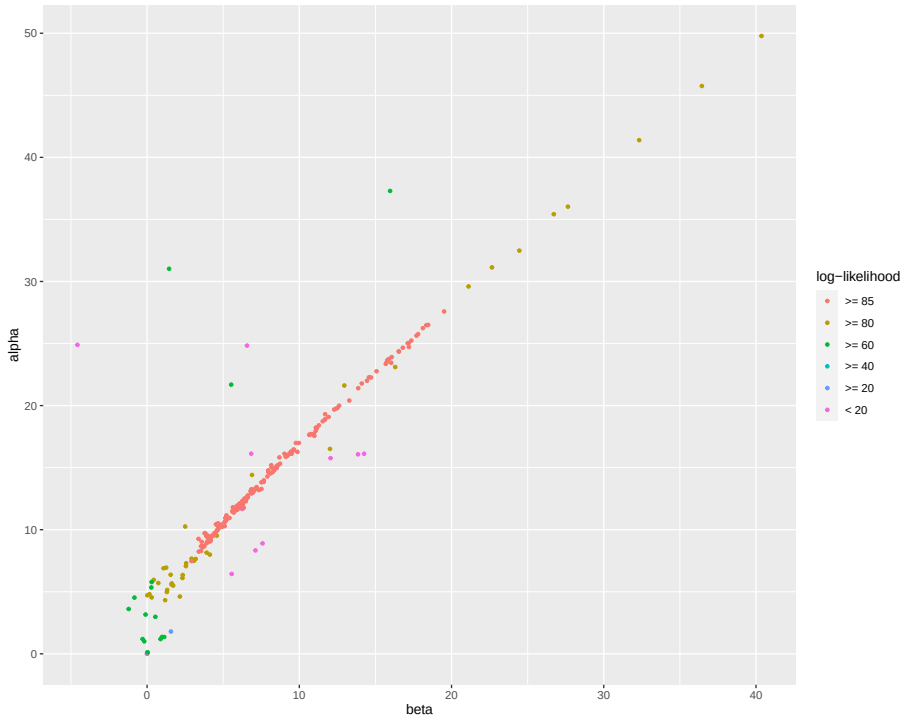



Figure 5.12: Scatterplot of $(\hat{\alpha}, \hat{\beta})$ estimates: S&P 500 Index NIG fit.

This analysis provides clear emphasis on the importance of initial value selection, as not only does this influence the likelihood of converging to the global maximum, but clearly can result in divergent behaviour by moving away from the maximum point and not towards. The problem is exacerbated when variation in $\lambda$ is introduced. Despite this clear dependence on the initial value when using the Nelder-Mead algorithm, it will be shown later that in most instances the resulting estimate is quite good when an appropriate initial value is chosen. It is, however, worth mentioning that there is no guarantee that an sensibly chosen initial value will converge, and as such it is advised to consider more commonly used methods such as the EM algorithm, or the proposed method that is outlined in section 4.3.

The next method that was considered is the EM algorithm, most notably for its aforementioned prevalence in literature for fitting GH models to data. Sections 4.2 and 4.2.1 provide an overview of the EM algorithm, as well as the EM algorithm in the context of the GH distribution and how to implement it. The discussion that follows will be focused on the performance of the EM algorithm as well as the limitations that were found when implementing it.

While this analysis proved useful with a simultaneous optimization routine such as the Nelder-

Mead algorithm, the same visual representation is not really possible with the EM algorithm. In the case of the EM algorithm one of two things happened. Either the choice of initial value resulted in the process converging to the global maximum value, or the algorithm got caught in a local maximum. This is, however, a known behaviour of the EM algorithm and as such did not raise a particular concern.

Of course, it is not always feasible to do an entire analysis across a grid of what can very easily be a substantial number of initial values. This can be an extremely costly process, especially when the underlying data contains a great many points, and if the estimation method being used has a significant time to completion. It also needs to be factored, as with any iterative method, that convergence time is not consistent across initial value choices. A solution to this is of course to put a upper limit on the number of iterations before terminating the procedure. It has, however, been found in some instances that while the process may slow down and exceed the chosen iteration limit of $10^3$, if left to run it would eventually converge to the global maximum.

When this analysis across the same grid of points is applied to the method proposed in section 4.3, the results where quite promising. Where the other methods had a significant subset of iterations either failing to reach the global maximum, or getting caught in a local maximum, this method had a 100% convergence rate to the global maximum point. This means that, for every permutation of initial value in the chosen grid, the algorithm ended up converging to the global maximum point of the given log-likelihood function for this specific case.

This method is clearly more resistant to the impact of the initial value choice. This increased robustness can likely be attributed to the fundamental principle of the algorithm. By splitting the likelihood function into parts, we reduce the dimensionality of the subsequent functions to maximize. This has the effect of improving the overall shape and behaviour of the function to be maximized. Another convenient aspect of this method lies in its construction. The optimization routine used for each profiled likelihood function is decided by the user, and as such can be adapted and even applied in numerous ways and to other distributions as well.

## 5.3 Initial value selection

This section will contain a discussion on the findings related to the choice of initial value, as well as the impact this had on each estimation method. When making use of numerical methods to find the maximum a function, it is necessary to provide an initial value to start off the process. The importance of this selection should not be understated, as it can often be the deciding factor in the successful convergence of the algorithm. This can clearly be seen by the results in section 5.2, where for even seemingly sensible initial value choices the algorithm either got caught in a local maximum, or failed to converge. The process of selecting initial values can also provide some valuable insights regarding the fitting process. These values can provide a good indicator of the parameter ranges, allowing us to better gauge the permissible range of the underlying parameters.

In terms of initial value selection for fitting a GH model to data, there is a recurring method in the literature. Aas and Haff (2006), Panahi (2018), Karlis (2002), and Rathgeber et al. (2017) all make use of the moment estimates as a starting point for the estimation of the GH parameters. It is important to clarify that the method of moments estimates are only viable when  is fixed to a corresponding subclass such as the hyperbolic or normal inverse Gaussian models. An appealing aspect of the NIG subclass lies in the rather simple and straightforward moments (see section 2.2.6) that allow for easy moment estimation. In the other subclass instances we are fortunate to have packages at our disposal allowing for straightforward calculations of the moment estimates for both the hyperbolic and hyperbolic asymmetric $t$ subclasses. Please refer to the GeneralizedHyperbolic and SkewHyperbolic packages for the required functions to get the moment estimates.

When it comes to initial value selection for the full GH model, the process is not so simple. While the moment estimates of the NIG, HYP, and HA$t$ subclasses are tractable, and as a result more easily obtainable, the same cannot be said for the full GH model. The moments in their standard form, as in section 2.4.2, cannot be readily solved to obtain moment estimates. There does exist a rather advanced methodology in Rathgeber et al. (2017), but that is beyond the scope of this study, and as will be shown shortly, it is not a necessary measure to obtain initial values for the GH model.

Prause (1999) propose a useful starting point, whereby they set $\beta = 0$, resulting in a symmetric model, as well choosing a reasonable kurtosis value (e.g. $\xi \approx 0.7$). It was also found that the

sample mean and sample standard deviation are good starting choices for $\mu_0$ and $\delta_0$ respectively. This allows us to make use of the fourth alternate parameterization in section (2.3) to solve for $\alpha_0$, namely

$$\xi = \frac{1}{\sqrt{1+\zeta}}, \tag{5.3.1}$$

and using $\zeta = \delta\sqrt{\alpha^2 - \beta^2}$ we get

$$\xi = \frac{1}{\sqrt{1+\delta\sqrt{\alpha^2-\beta^2}}}, \tag{5.3.2}$$

and since we have initial estimates for all but $\alpha$, this allows us to easily solve for $\alpha$ using (5.3.2). All that then remains is to find an initial estimate for $\lambda$. Since we have estimates for the other four parameters, a reasonable solution is simply to find the maximum likelihood estimate for $\lambda$ given these initial estimates, much in the same fashion as profile likelihood estimation.

This process (adapted from Prause (1999)) gives us the following initial values:

$$(\lambda_0, \alpha_0, \beta_0, \delta_0, \mu_0) = (0.48, 158.34, 0, 0.0066, 0.00040).$$

Comparing this vector with the resulting global maximum (see Table 5.7 and Prause (1999, p. 34))

$$(\hat{\lambda}, \hat{\alpha}, \hat{\beta}, \hat{\delta}, \hat{\mu}) = (0.81, 212.56, -5.93, 0.0022, 0.00066),$$

this method seems to generate reasonable starting values for the full GH model, and for all instances, save the unconstrained Nelder-Mead algorithm, performed quite well in converging to the global maximum (see Table 5.7).

## 5.4 Estimation results

Tables 5.7 and 5.8 contain the resulting estimates for the GH, NIG, hyberbolic and hyperbolic asymmetric *t* distributions fitted to the NYSE Composite Index and S&P 500 index datasets respectively. The tables contain the resulting estimates from the following methods:

1. Nelder-Mead simplex method.

2. Nelder-Mead simplex method with linear constraints.

3. EM algorithm.

4. Profile likelihood based alternating estimation.

In Table 5.7 we observe the result of fitting the models to the NYSE Composite Index data. For both the hyperbolic and normal inverse Gaussian subclasses all the estimation methods performed rather well, in that they were all able to converge to their respective global maximum points. In the full GH estimation it is clear that the Nelder-Mead simplex method did not adequately converge to the global maximum point. This is clearly a consequence of the flatness of the log-likelihood function, further amplified by the inclusion of $\lambda$ in the estimation process, and referring back to Figure 5.3 this is not surprising. The hyberbolic asymmetric $t$ fit had some stability issues when various initial values were tested, but when the moment estimates were used as initial values the process seems to converge to the required maximum.

Something that was found when fitting the GH model or one of the relevant subclasses, is that it is not always correct to assume a subclass will be a good fit. For some of the fitted subclasses, the likelihood of stability issues and inadequate convergence was higher. There were also instances of non-convergence, and sometimes even divergence. Another aspect that must be carefully considered is the proximity of the initial values for $\alpha$ and $\beta$ to the boundary of the constraint $|\beta| < \alpha$. It is found that when $\alpha_0$ $\beta_0$ are close in magnitude, the estimation algorithm can get "stuck" on the line that governs the constraint between these parameters, and consequently fail to converge to the global maximum point.

Another scenario that can occur is the global maximum point itself being close to this boundary. This is often an indication that the current distribution being fitted is not the most suitable and another subclass should be fitted. A direct example can be found in Table 5.8, where the full GH fit is rather close to the hyperbolic asymmetric $t$ fit for the S&P 500 index dataset. In this instance it would be better suited to assume a hyperbolic asymmetric $t$ model from the outset.

Looking at Table 5.8, namely the estimation results for the S&P 500 dataset, there are somewhat more instances of non-convergence, and potentially even divergence of the estimation algorithms. In the hyperbolic subclass fit, the EM algorithm fails to converge to the global maximum point, seemingly exhibiting the behaviour more commonly seen with the Nelder-Mead simplex

method and other general purpose optimization strategies. When the NIG model is considered, both iterations of the Nelder-Mead simplex method fail to converge to the global maximum point, instead getting caught in a local maximum. The EM algorithm shows a similar outcome, but with a slightly more desirable estimate. The new alternating method actually managed to converge to a point that could resemble a global maximum point, if one were to compare log-likelihood values across the subclass fits. An important note, however, is that the alternating method took much longer to converge when the NIG subclass is fit. Most iterative methods would have some built-in stop criterion that would prevent this point from ever being reached. This is most likely an indication that the NIG fit is not ideal for this data.

If we compare the full GH fit with the hyperbolic asymmetric $t$ fit, we see that they are really quite similar. This is a strong indicator that the hyperbolic asymmetric $t$ fit is ideal for the S&P 500 dataset. When fitting the full GH model in this fashion and the parameters are leaning towards one of the subclasses, it is advised to instead fit the relevant subclass, as it has already been shown what impact the variation of $\lambda$ has on the estimation process.

## 5.5 Model fit assessment

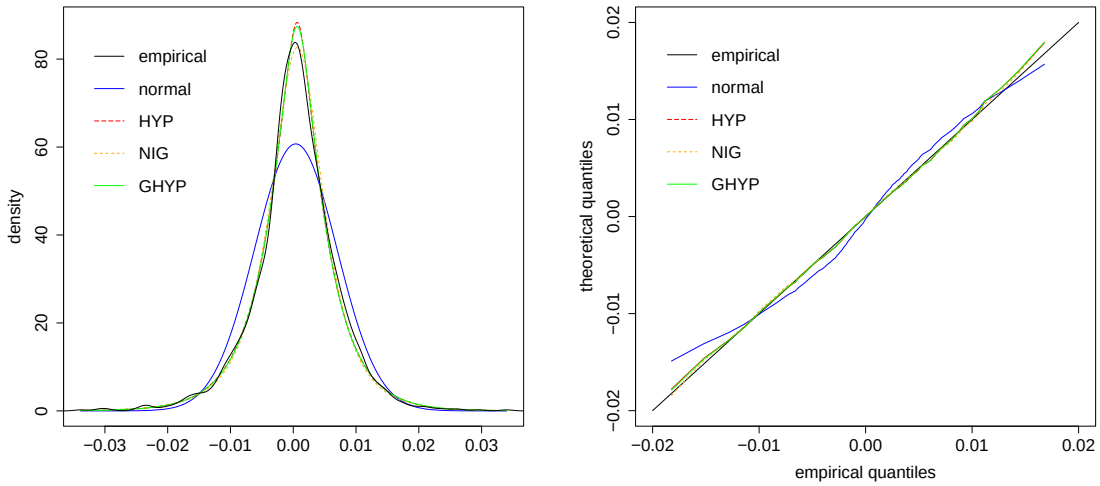
In this section we will be analysing and comparing each of the model fits for both the NYSE data and the S&P 500 data. Figure 5.13 provides a graphical overlay of the densities of the various fitted distributions, as well as a normal fit, with the empirical pdf. In Figure 5.13b we have multiple QQ-plots overlayed in a similar pattern. It is clear here that the normal fit is not appropriate to this data. This is to be expected given the nature of the chosen datasets, and the typical tail properties present in financial data. Looking at Figures 5.13a and 5.13b, it seems as if the full GH distribution, as well the selected subclasses all provide an adequate fit to the data.

Table 5.7: Estimates for the GH distribution and relevant subclasses: NYSE composite index.

| | $\lambda$ | $\alpha$ | $\beta$ | $\delta$ | $\mu$ | Log-L |
|---|---|---|---|---|---|---|
| *Hyperbolic* | | | | | | |
| Nelder-Mead | 1 | 225.04 | -5.68 | 0.0016 | 0.00064 | 6408.27 |
| Nelder-Mead (constrained) | 1 | 224.91 | -6.12 | 0.0016 | 0.00066 | 6408.27 |
| EM-algorithm | 1 | 226.33 | -5.91 | 0.0016 | 0.00065 | 6408.26 |
| New method | 1 | 225.03 | -5.84 | 0.0016 | 0.00064 | 6408.27 |
| *Normal inverse Gaussian* | | | | | | |
| Nelder-Mead | -0.5 | 136.36 | -8.89 | 0.0059 | 0.00079 | 6406.74 |
| Nelder-Mead (constrained) | -0.5 | 136.50 | -8.85 | 0.0059 | 0.00079 | 6406.74 |
| EM-algorithm | -0.5 | 139.13 | -9.19 | 0.0060 | 0.00080 | 6406.72 |
| New method | -0.5 | 135.84 | -8.80 | 0.0059 | 0.00078 | 6406.74 |
| *Generalized hyperbolic* | | | | | | |
| Nelder-Mead | 0.67 | 203.49 | -7.09 | 0.0027 | 0.00070 | 6408.26 |
| Nelder-Mead (constrained) | 0.84 | 215.08 | -6.72 | 0.0022 | 0.00068 | 6408.30 |
| New method | 0.81 | 212.56 | -5.93 | 0.0022 | 0.00066 | 6408.31 |
| *Hyperbolic asymmetric t* | | | | | | |
| Nelder-Mead | -1.94 | 9.71 | -9.71 | 0.0093 | 0.00084 | 6402.09 |
| Nelder-Mead (constrained) | -1.94 | 9.66 | -9.66 | 0.0093 | 0.00084 | 6402.09 |
| New method | -1.96 | 9.91 | -9.91 | 0.0094 | 0.00084 | 6402.09 |

Table 5.8: Estimates for the GH distribution and relevant subclasses: S&P 500 index.

| | $\lambda$ | $\alpha$ | $\beta$ | $\delta$ | $\mu$ | Log-L |
|---|---|---|---|---|---|---|
| *Hyperbolic* | | | | | | |
| Nelder-Mead | 1 | 15.96 | 7.35 | 0.1734 | 0.8352 | 85.06 |
| Nelder-Mead (constrained) | 1 | 15.98 | 7.37 | 0.1735 | 0.8349 | 85.06 |
| EM-algorithm | 1 | 8.37 | 1.43 | 4.14e-14 | 0.9423 | 80.80 |
| New method | 1 | 15.09 | 6.58 | 0.1650 | 0.8464 | 85.05 |
| *Normal inverse Gaussian* | | | | | | |
| Nelder-Mead | -0.5 | 414.44 | -392.76 | 0.4381 | 2.2882 | 38.39 |
| Nelder-Mead (constrained) | -0.5 | 415.99 | -394.85 | 0.4298 | 2.839 | 38.38 |
| EM-algorithm | -0.5 | 151.51 | -95.82 | 1.9947 | 2.6133 | 60.88 |
| New method | -0.5 | 11.94 | 6.05 | 0.2183 | 0.8557 | 85.58 |
| *Generalized hyperbolic* | | | | | | |
| Nelder-Mead | -4.37 | 7.8088 | 7.8087 | 0.3648 | 0.8299 | 86.35 |
| Nelder-Mead (constrained) | -4.36 | 7.8119 | 7.8117 | 0.3644 | 0.8298 | 86.35 |
| New method | -4.44 | 8.1109 | 8.1106 | 0.3666 | 0.8256 | 86.35 |
| *Hyperbolic asymmetric t* | | | | | | |
| Nelder-Mead | -4.49 | 8.20 | 8.20 | 0.3684 | 0.8244 | 86.35 |
| Nelder-Mead (constrained) | -4.52 | 8.24 | 8.24 | 0.3699 | 0.8236 | 86.35 |
| New method | -4.47 | 8.19 | 8.19 | 0.3676 | 0.8247 | 86.35 |




(a) Densities: NYSE composite index. (b) QQ-plot: NYSE Composite Index.

Figure 5.13: NYSE Composite Index.

The first measure that will be used to assess the various fitted models will be the Kolmogorov-Smirnov (KS) distance (see Massey Jr, 1951). The KS distance is used to test the equality of two one-dimensional distributions by comparing the empirical CDF with the fitted CDF. We define the KS distance as the supremum of the difference between the empirical CDF and the fitted CDF as follows

$$KS = \sup_x |F_n(x) - F(x)|, \tag{5.5.1}$$

where $F_n(x)$ refers to the empirical CDF and is defined as

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{X_i \leq x} \tag{5.5.2}$$

where $1_{X_i \leq x}$ is an indicator function and is defined as

$$I_{X_i \leq x} = \begin{cases} 1 & \text{if } X_i \leq x \\ 0 & \text{otherwise} \end{cases} \tag{5.5.3}$$

We will also be using the Anderson-Darling (AD) statistic as a measure of the goodness-of-fit

(see Anderson and Darling, 1954). The AD statistic is defined as

$$AD = \sup_x \frac{|F_n(x) - F(x)|}{\sqrt{F_n(x)(1 - F_n(x))}}. \tag{5.5.4}$$

While both measures are similarly constructed, the AD test is more sensitive to the tail structure of the distribution, whereas the KS test is more focused on the central structure of the distribution. Considering that the tail behaviour is an appealing aspect of the generalized hyperbolic model, and thus an important aspect to consider, we will likely favour this metric over the KS distance.

As a final addition we will also consider Akaike information criterion (AIC) (see Akaike, 1998). For a given model with $k$ parameters to estimate, the AIC is defined as:

$$AIC = 2k - 2\,\ell(\hat{\boldsymbol{\theta}}), \tag{5.5.5}$$

where $\ell(\hat{\boldsymbol{\theta}})$ is the log-likelihood value corresponding the the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ as before. The resulting measures can be found in Table 5.9. It is important to note that the values in Table 5.9 are only test statistics, not significance values, and they cannot be interpreted in an inferential sense.

For the NYSE Composite Index dataset, we see that for both the KS test and the AD test the full GH distribution and the NIG subclass are favoured, with the NIG subclass being preferred in the latter case. It is, however, clear that while there is a favourite in each case, the differences in the test statistics are rather small. This points back to the flatness of the log-likelihood function, as well as the potential overfitting problem faced by this class of distributions, especially when $\lambda$ is arbitrary. Interestingly enough, the AIC metric actually favours the NIG fit the least, with the hyperbolic fit being the most favourable according to this test. It should again be noted that the differences in AIC values are really quite small, again pointing to the undesirable shape of the log-likelihood function of the GH distribution and its subclasses.

For the S&P 500 index we observed a similar trend. In this instance we see that the full GH model and the hyperbolic asymmetric $t$ subclass are favoured in terms of the KS and AD statistics. These values are very close in magnitude, but this is to be expected considering the striking similarity of the full GH and HA$t$ estimates in Table 5.8. Here we see that the AIC

statistic also favours the hyperbolic asymmetric *t* distribution over the others. We also observe the same trend as before, specifically the small differences in magnitudes of the measures in Table 5.9 between the different model fits. This seems to be a recurring problem of the GH distribution, especially considering that the S&P 500 index contains 200 observations, compared to the 1749 observations in the NYSE Composite Index dataset.

Table 5.9: Goodness-of-fit metrics for the GH, NIG, Hyperbolic and hyperbolic asymmetric *t* distributions fitted to the NYSE Composite Index and S&P 500 Index data.

| | Kolmogorov-Smirnov Statistic | Anderson-Darling Statistic | AIC |
|---|---|---|---|
| *NYSE Composite Index* | | | |
| GH | 0.0160 | 0.0497 | -12806.61 |
| HYP | 0.0185 | 0.0548 | -12808.53 |
| NIG | 0.0176 | 0.0404 | -12805.48 |
| HAt | 0.0223 | 0.0508 | -12796.18 |
| *S&P 500 Index* | | | |
| GH | 0.0397 | 0.1120 | -162.699 |
| HYP | 0.0411 | 0.2246 | -162.102 |
| NIG | 0.0406 | 0.1555 | -163.169 |
| HAt | 0.0393 | 0.1122 | -164.699 |

## 5.6 Simulation study

In order to get a better understanding of the stability of fitting the generalized hyperbolic distribution to data, a simulation study will be performed. In this study we simulate datasets of size 2000 from the full GH distribution, as well as the NIG and hyperbolic subclasses. The parameter values in each case are chosen to be the maximum likelihood estimates for each of the subclasses fitted to the NYSE Composite Index data.

This will allow for a more consistent comparison, as these estimates come from fitting notable subclasses of the GH distribution, as well as the full GH distribution to the same dataset. Simulating in this fashion also ensures that a meaningful set of parameters is chosen to simulate from in each case, and helps paint a picture as to the stability of the fitting the GH model and its subclasses to real world datasets.

A description of the process to generate the ensuing tables will now be given. For varying sample sizes, we sample with replacement, much in the fashion of a bootstrap, after which we fit the relevant GH model or subclass to this sampled dataset. This process is repeated 100 times for each sample size, thus allowing us to compute the mean and standard errors of the parameters in each case.

In Tables 5.10 and 5.11 provide the means and standard errors of the parameters corresponding to the normal inverse Gaussian subclass fit. While the $\delta$ and $\mu$ parameters are relatively stable throughout, it is clear that this is not the case with $\alpha$ and $\beta$. It is only from sample size 1000 onwards the we begin to observe reasonable stability in the $\alpha$ and $\beta$ parameters in terms of the mean and standard error.

In Tables 5.12 and 5.13 we observe that while the values of $\delta$ and $\mu$ are again stable throughout, The $\alpha$ and $\beta$ parameters are again showing some instability. While it can be argued that the $\alpha$ and $\beta$ parameters have a more consistent convergence pattern to the MLE estimates as we increase the sample size $n$, what is alarming is that even at a sample of size 2000 these values are still significantly off from the MLE. From Table 5.7, for the hyperbolic fit we see that the estimates for $\alpha$ and $\beta$ are 225.03 and -5.84 respectively. While this $\alpha$ is not so far from average $\alpha$ value of 230.12 we observe when $n = 2000$, the average $\beta$ value of -16.30 is quite far from the corresponding MLE value.

What is most alarming, however, is what can be seen in Tables 5.14 and 5.15, that is, the means and standard errors when the full GH model is fitted. While the $\delta$ and $\mu$ parameters are exhibiting the same behaviours as before, we again observe some stability issues in terms of $\alpha$ and $\beta$. What is especially interesting, is what can be observed when we look at the stability of the $\lambda$ parameter. For smaller sample sizes the average estimate for $\lambda$ is significantly different from the estimated value of 0.81 in Table 5.7. Not only this, but at the observed rate of change, it seems entirely possible, if not likely, that an average value close to the estimated value of 0.81 will only happen for sample sizes of 10000 or more.

This occurance serves as a reinforcement of the same pattern of behaviours stemming from the estimation of the $\lambda$ parameter, namely a negative impact not only on the shape and behaviour of the log-likelihood function of the GH distribuution, but on the estimation process itself and the quality of the resulting parameter estimates. This surely calls into question whether $\lambda$ should be included in the estimation process, as this analysis has shown that it seems a much more sensible choice to fix $\lambda$ and fit the relevant subclass/es to the data.

Table 5.10: Mean of the estimates of the NIG subclass (see (2.2.10)) at different sample sizes: NYSE Composite Index; 100 iterations

| Sample Size | Mean $\alpha$ | Mean $\beta$ | Mean $\delta$ | Mean $\mu$ |
|---|---|---|---|---|
| 50 | 183.16 | -5.13 | 0.0073 | 0.00060 |
| 100 | 178.63 | -2.48 | 0.0074 | 0.00076 |
| 150 | 169.89 | -7.57 | 0.0075 | 0.00084 |
| 200 | 169.27 | -7.36 | 0.0072 | 0.00082 |
| 250 | 158.46 | -10.20 | 0.0070 | 0.00088 |
| 500 | 147.41 | -8.82 | 0.0066 | 0.00095 |
| 750 | 145.85 | -9.90 | 0.0066 | 0.00095 |
| 1000 | 140.59 | -9.05 | 0.0065 | 0.00091 |
| 1500 | 139.83 | -7.68 | 0.0064 | 0.00089 |
| 2000 | 139.23 | -7.70 | 0.0064 | 0.00090 |

Table 5.11: Standard error of the estimates of the NIG subclass (see (2.2.10)) at different sample sizes: NYSE Composite Index; 100 iterations

| Sample Size | Std error $\alpha$ | Std error $\beta$ | Std error $\delta$ | Std error $\mu$ |
|---|---|---|---|---|
| 50 | 9.158 | 5.601 | 0.00030 | 0.00018 |
| 100 | 6.866 | 3.572 | 0.00022 | 0.00013 |
| 150 | 6.360 | 2.621 | 0.00023 | 0.00011 |
| 200 | 6.362 | 2.466 | 0.00020 | 0.000097 |
| 250 | 4.822 | 2.115 | 0.00017 | 0.000079 |
| 500 | 3.257 | 1.460 | 0.000094 | 0.000056 |
| 750 | 2.067 | 1.314 | 0.000082 | 0.000048 |
| 1000 | 1.764 | 1.011 | 0.000061 | 0.000044 |
| 1500 | 1.568 | 0.9977 | 0.000048 | 0.000041 |
| 2000 | 1.278 | 0.8514 | 0.000046 | 0.000040 |

Table 5.12: Mean of the estimates of the hyperbolic subclass (see (2.2.4)) at different sample sizes: NYSE Composite Index; 100 iterations

| Sample Size | Mean $\alpha$ | Mean $\beta$ | Mean $\delta$ | Mean $\mu$ |
|---|---|---|---|---|
| 50 | 259.12 | -22.29 | 0.0016 | 0.00097 |
| 100 | 261.75 | -22.68 | 0.0020 | 0.00084 |
| 150 | 242.28 | -17.00 | 0.0015 | 0.00082 |
| 200 | 244.28 | -20.63 | 0.0017 | 0.00087 |
| 250 | 244.94 | -18.86 | 0.0019 | 0.00090 |
| 500 | 234.88 | -17.33 | 0.0015 | 0.00083 |
| 750 | 229.72 | -18.76 | 0.0013 | 0.00090 |
| 1000 | 230.39 | -17.48 | 0.0013 | 0.00085 |
| 1500 | 230.29 | -15.73 | 0.0013 | 0.00080 |
| 2000 | 230.12 | -16.30 | 0.0013 | 0.00079 |

Table 5.13: Standard error of the estimates of the hyperbolic subclass (see (2.2.4)) at different sample sizes: NYSE Composite Index; 100 iterations

| Sample Size | Std error $\alpha$ | Std error $\beta$ | Std error $\delta$ | Std error $\mu$ |
|---|---|---|---|---|
| 50 | 5.589 | 4.772 | 0.00028 | 0.00016 |
| 100 | 5.625 | 4.898 | 0.00029 | 0.00014 |
| 150 | 3.889 | 2.959 | 0.00021 | 0.00011 |
| 200 | 3.924 | 2.251 | 0.00019 | 0.000071 |
| 250 | 3.359 | 1.929 | 0.00022 | 0.000071 |
| 500 | 2.226 | 1.966 | 0.00014 | 0.000062 |
| 750 | 1.611 | 2.136 | 0.00012 | 0.000080 |
| 1000 | 1.414 | 1.798 | 0.00010 | 0.000062 |
| 1500 | 1.073 | 0.8536 | 0.000077 | 0.000029 |
| 2000 | 1.012 | 0.9682 | 0.000080 | 0.000034 |

Table 5.14: Mean of the estimates of the GH distribution (see (2.1.2)) at different sample sizes: NYSE Composite Index; 100 iterations

| Sample Size | Mean $\lambda$ | Mean $\alpha$ | Mean $\beta$ | Mean $\delta$ | Mean $\mu$ |
|---|---|---|---|---|---|
| 50 | -3.594 | 235.10 | -44.27 | 0.0075 | 0.00140 |
| 100 | -2.430 | 177.70 | -12.47 | 0.0064 | 0.00072 |
| 150 | -0.1369 | 168.94 | -11.57 | 0.0042 | 0.00069 |
| 200 | 0.1706 | 175.22 | -13.64 | 0.0034 | 0.00078 |
| 250 | 0.2892 | 175.34 | -12.16 | 0.0031 | 0.00071 |
| 500 | 0.4392 | 184.19 | -13.93 | 0.0028 | 0.00080 |
| 750 | 0.4548 | 180.05 | -12.25 | 0.0027 | 0.00069 |
| 1000 | 0.4677 | 182.68 | -12.77 | 0.0028 | 0.00072 |
| 1500 | 0.4689 | 180.83 | -12.47 | 0.0028 | 0.00075 |
| 2000 | 0.5424 | 186.97 | -11.43 | 0.0026 | 0.00071 |

Table 5.15: Standard error of the estimates of the GH distribution (see (2.1.2)) at different sample sizes: NYSE Composite Index; 100 iterations.

| Sample Size | Std error $\lambda$ | Std error $\alpha$ | Std error $\beta$ | Std error $\delta$ | Std error $\mu$ |
|---|---|---|---|---|---|
| 50 | 1.468 | 17.82 | 16.22 | 0.0015 | 0.00033 |
| 100 | 1.365 | 9.39 | 7.202 | 0.0013 | 0.00019 |
| 150 | 0.2160 | 7.51 | 2.790 | 0.00053 | 0.000090 |
| 200 | 0.1192 | 6.02 | 2.230 | 0.00039 | 0.000079 |
| 250 | 0.0971 | 6.02 | 1.858 | 0.00031 | 0.000064 |
| 500 | 0.0680 | 4.30 | 1.157 | 0.00025 | 0.000041 |
| 750 | 0.0593 | 4.01 | 0.9486 | 0.00020 | 0.000034 |
| 1000 | 0.0552 | 3.75 | 0.8244 | 0.00019 | 0.000029 |
| 1500 | 0.0468 | 2.91 | 0.6241 | 0.00017 | 0.000023 |
| 2000 | 0.0447 | 3.09 | 0.6404 | 0.00016 | 0.000023 |

# Chapter 6

# Synthesis

This final chapter serves as a conclusion to the study, with a summary of the findings, a discussion on the study's significance, and lastly a brief outline of potential avenues of departure for future work.

## 6.1 Findings

As it can clearly be seen in the exploratory analysis undertaken in chapter 5, there exist clear problems that need to be navigated when fitting the GH distribution to data by means of the log-likelihood function. This stems not only from the formulation of the GH distribution, but from the behaviour of the ensuing log-likelihood function. There is clear evidence that the dimensionality of the parameter space causes behavioural problems in terms of the log-likelihood. This seems to stem not only from a potential over-parameterization, but from the numerical limitations imposed by the funcional form of the GH distribution.

This is not to say that fitting the GH model to data is a fruitless pursuit. This simply means that a certain degree of care needs to be exercised when fitting the model. It was, for example, found that the inclusion of the $\lambda$ parameter had a significant impact on not only the likelihood of convergence, but on the likelihood of converging to an adequate estimate. This is in large part due the the flatness of the log-likelihood function, especially that contributed by the $\lambda$ parameter. This can be seen when comparing the regions in Figures 5.4, 5.5, and 5.3.

There also seems to be some linearity in the relationship between the $\alpha$ and $\beta$ parameters, as can be seen in Figures 5.3, 5.4, 5.7, and 5.8. This is a possible indication of a redundant parameterization, but can also be an indication of a practical identifiability issue stemming from the data (see Raue et al., 2009). The simulation study conducted also serves to support this notion, as it was shown in Tables 5.10, 5.12, and 5.14 that we only begin to see convergence to the true parameter values for sample sizes of 2000 and above. In the GH instance specifically (see Table 5.14), there is still a discrepancy between the average estimate values and true parameter values at the sample of size 2000.

## 6.2 Significance of study

It is easy to overlook issues stemming from the behaviour of the log-likelihood function or the parameters themselves, especially when the role they play is not so obvious. In the context of the GH distribution, it was shown that different subclass can have largely similar shapes, and in some instances can have virtually identical fits to the data. It is also shown in Figures 5.3 and 5.4 that vastly different parameter estimates have near identical log-likelihood values. This makes it somewhat dangerous to simply accept the resulting estimate without further investigation. There is also the issue of getting caught in a local maximum, which, in tandem with the the log-likelihood function behaviour, further reduces the validity of the resulting estimate. The inability of visualising the behaviour of the log-likelihood function due to its dimension is also a contributing factor. It is therefore of critical importance to be aware of these potential pitfalls. In this study it is shown that when certain measures are taken, such as fixing certain problematic parameters or breaking up the log-likelihood function to allow for better behaviour, and an understanding of the underlying distribution exists, then there is a much better likelihood of successfully fitting the model to the data.

## 6.3 Future prospects

This study serves to create a platform as a byproduct of an exploratory analysis into maximum likelihood estimation in the context of the generalized hyperbolic model. What has been found,

however, is that while some of the findings are likely unique to the underlying distribution, there are certainly instances where these findings can also apply to similar distributions. The GH distribution is considered a flexible distribution class, whereby the pdf is regulated by four parameters, thus allowing for greater variation in terms of measures of skewness and of kurtosis. There are an array of flexible distributions (see Ley, 2015), and it is therefore expected that these issues of log-likelihood behaviour and potential over-parameterization, as well as the issues they bring, may also be present within distributions conforming the the flexible distribution structure.

# Appendix A

# Continuous normal mixture distributions

The following information is extracted from McNeil, Frey, and Embrechts (2005, pg. 73-78) and will serve as a brief overview of the mechanics of normal mean-variance mixtures, and is essential in the derivation of the GH distribution and its various subclasses, as well as the estimation of the GH parameters by means of the EM algorithm. The process of normal mean-variance mixtures involves the introduction of randomness into the variance component, as well as the mean component of the normal distribution by means of a positive mixing variable which will be denoted by $W$ throughout the text that follows.

### Normal variance mixtures

**Definition A.1.** The random variable $X \in \mathbb{R}$ is said to have a normal variance mixture distribution if

$$X \stackrel{d}{=} \mu + \sqrt{W}Z \tag{A.1}$$

where

1. $Z \sim N(0,1)$,
2. $W \geq 0$ is any non-negative, scalar-valued random variable independent of $Z$, and
3. $\mu$ is a parameter in $\mathbb{R}$.

We refer to these distributions as variance mixtures, since by conditioning on the mixing variable $W$ we observe that

$$X|(W = w) \sim N(\mu, W). \tag{A.2}$$

The distribution of the random variable $X$ is thought of as a composite distribution, constructed by taking a set of univariate normal distributions with equal means and equal variances up to a multiplicative constant $w$. The mixture distribution is then constructed by randomly drawing from the set of composite normal distributions according to the weighting determining by the mixing variable $W$. It needs to be noted that this mixture, namely the distribution of $X$ is not itself a normal distribution.

### Normal mean-variance mixtures

The resulting mixture distributions from normal variance mixtures have what is called elliptical symmetry. This does not align with the inherent structure of the typical dataset to which we fit the GH distribution. As previously stated the GH distribution is a popular choice when modelling financial data such as stock returns, which tend to have heavier tales for negative returns than for positive returns. In contrast to normal variance mixtures where the resulting mixture distributions have elliptical symmetry, normal mean-variance mixtures add asymmetry to the process by mixing normal distributions that have different means as well as different variances.

**Definition A.2.** The random variable $X \in \mathbb{R}$ is said to have a normal mean-variance mixture distribution if

$$X \overset{d}{=} \mu + \beta W + \sqrt{W} Z \tag{A.3}$$

where

1. $Z \sim N(0, 1)$,

2. $W \geq 0$ is any non-negative, scalar-valued random variable independent of $Z$, and

3. $\mu$ and $\beta$ are parameters in $\mathbb{R}$.

From this definition we have that

$$X|(W = w) \sim N(\mu + \beta W, W). \tag{A.4}$$

# Appendix B

# Modified Bessel functions

In this appendix a few key results and properties of modified Bessel functions are discussed which are useful in deriving and working with the generalized hyperbolic distributions. The following information is taken from Paolella (2007) as well as Bibby & Sørensen (2003). The modified Bessel function of the third kind with index $\lambda$ is defined by the following integral expression

$$K_\lambda(x) = \frac{1}{2}\int_0^\infty t^{\lambda-1} e^{-\frac{1}{2}x(t+\frac{1}{t})}\, dt, \quad x > 0. \tag{B.1}$$

What makes this expression interesting is its similarity to the gamma function, and it will later be shown how these two functions are related.

The modified Bessel function of the third kind has the following key properties:

$$K_{-\lambda}(x) = K_\lambda(x) \tag{B.2}$$

$$K_{\lambda+1}(x) = \frac{2\lambda}{x}K_\lambda(x) + K_{\lambda-1}(x) \tag{B.3}$$

$$K'_\lambda(x) = \frac{-\lambda}{x}K_\lambda(x) - K_{\lambda-1}(x) \tag{B.4}$$

For $\lambda = n + \frac{1}{2}$, with $n = 0, 1, 2, \ldots$, we have

$$K_{n+\frac{1}{2}}(x) = \sqrt{\frac{\pi}{2x}}e^{-x}\left(1 + \sum_{i=1}^{n}\frac{(n+i)!}{(n-i)!\,i!}(2x)^{-i}\right) \tag{B.5}$$

For $x = 0$, the modified Bessel function has a singular point (or singularity), and for small values of $x$ we have

$$K_\lambda(x) \sim -\ln(x) \quad \text{for } x \to 0,\ \lambda = 0. \tag{B.6}$$

$$K_\lambda(x) \sim \Gamma(\lambda) 2^{|\lambda|-1} x^{-|\lambda|} \quad \text{for } x \to 0,\ \lambda \neq 0. \tag{B.7}$$

The following integral expression is closely related to the Bessel function, and is of importance as it is used in the derivation of the generalized hyperbolic distributions as well as in many of the subfamily pdf function expressions:

$$k_\lambda(\chi, \psi) = \int_0^\infty x^{\lambda-1} e^{-\frac{1}{2}(\chi x^{-1} + \psi x)}\, dx. \tag{B.8}$$

This integral converges for arbitrary $\lambda \in \mathbb{R}$ and $\chi, \psi > 0$. By setting $\eta = \sqrt{\frac{\chi}{\psi}}$, $\omega = \sqrt{\chi\psi}$, and using the substitution $x = \eta y$ we get

$$\begin{aligned}
k_\lambda(\chi, \psi) &= \int_0^\infty x^{\lambda-1} e^{-\frac{1}{2}(\chi x^{-1} + \psi x)}\, dx. \\
&= \int_0^\infty x^{\lambda-1} e^{-\frac{1}{2}\omega((\frac{x}{\eta})^{-1} + \frac{x}{\eta})}\, dx. \\
&= \int_0^\infty (\eta y)^{\lambda-1} e^{-\frac{1}{2}\omega(y^{-1} + y)} \eta\, dy. \\
&= 2\eta^\lambda \frac{1}{2} \int_0^\infty y^{\lambda-1} e^{-\frac{1}{2}\omega(y^{-1} + y)}\, dy. \\
&= 2\eta^\lambda K_\lambda(\omega).
\end{aligned} \tag{B.9}$$

This gives us the following

$$k_\lambda(\chi, \psi) = 2\eta^\lambda K_\lambda(\omega) = 2\left(\frac{\chi}{\psi}\right)^{\frac{\lambda}{2}} K_\lambda\left(\sqrt{\chi\psi}\right). \tag{B.10}$$

The expression in (B.8) also has two boundary cases to which it converges. The first case occurs if $\chi = 0$ and $\psi > 0$, and convergence occurs if and only if $\lambda > 0$. For this boundary case, we use

of the substitution $y = \frac{\psi}{2}x$ which gives us the following:

$$\begin{aligned} k_\lambda(0,\psi) &= \int_0^\infty x^{\lambda-1}e^{-\frac{1}{2}\psi x}\,dx \\ &= \left(\frac{\psi}{2}\right)^{-\lambda}\int_0^\infty y^{\lambda-1}e^{-y}\,dy \\ &= \left(\frac{\psi}{2}\right)^{-\lambda}\Gamma(\lambda). \end{aligned} \tag{B.11}$$

(B.12)

The second boundary case occurs if $\chi > 0$ and $\psi = 0$, and convergence occurs if and only if $\lambda < 0$. For this case we make use of the substitution $y = \frac{\psi}{2x}$, which gives us

$$\begin{aligned} k_\lambda(\chi,0) &= \int_0^\infty x^{\lambda-1}e^{-\frac{1}{2}\chi x^{-1}}\,dx \\ &= \left(\frac{\chi}{2}\right)^{\lambda}\int_0^\infty y^{-\lambda-1}e^{-y}\,dy \\ &= \left(\frac{\chi}{2}\right)^{\lambda}\Gamma(-\lambda). \end{aligned} \tag{B.13}$$

(B.14)

This gives us the expressions

$$k_\lambda(0,\psi) = \left(\frac{\psi}{2}\right)^{-\lambda}\Gamma(\lambda) \tag{B.15}$$

and

$$k_\lambda(\chi,0) = \left(\frac{\chi}{2}\right)^{\lambda}\Gamma(-\lambda). \tag{B.16}$$

The function in (B.8) possesses the following useful properties

$$k_\lambda(\chi,\psi) = k_{-\lambda}(\psi,\chi), \tag{B.17}$$

and

$$k_\lambda(\chi,\psi) = r^\lambda k_\lambda(r^{-1}\chi, r\psi) \quad \forall r > 0. \tag{B.18}$$

# Bibliography

Aas, K. and Haff, I. H. The generalized hyperbolic skew studentâst-distribution. *Journal of financial econometrics*, 4(2):275–309, 2006.

Akaike, H. Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pages 199–213. Springer, 1998.

Anderson, T. W. and Darling, D. A. A test of goodness of fit. *Journal of the American statistical association*, 49(268):765–769, 1954.

Bagnold, R. A. *The physics of blown sand and desert dunes*. Methuen Co., Ltd., London, 1941.

Bain, L. J. and Engelhardt, M. *Introduction to probability and mathematical statistics*. Duxbury Press, Boston, 1987.

Barndorff-Nielsen, O. E. Exponentially decreasing distributions for the logarithm of particle size. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 353(1674):401–419, 1977.

Barndorff-Nielsen, O. E. Hyperbolic distributions and distributions on hyperbolae. *Scandinavian Journal of statistics*, pages 151–157, 1978.

Barndorff-Nielsen, O. E. *Normal inverse Gaussian processes and the modelling of stock returns*. Aarhus Universitet. Department of Theoretical Statistics, 1995.

Barndorff-Nielsen, O. E. and Blaesild, P. Hyperbolic distributions and ramifications: Contributions to theory and application. In *Statistical distributions in scientific work* (ed. C. Taillie, G. Patil and B. Baldessari), pages 19–44. Springer, 1981.

Barndorff-Nielsen, O. E. and Cox, D. R. *Inference and asymptotics*. Routledge, 2017.

Behr, A. and Pötter, U. Alternatives to the normal model of stock returns: Gaussian mixture, generalised logf and generalised hyperbolic models. *Annals of Finance*, 5(1):49–68, 2009.

Bibby, B. M. and Sørensen, M. Hyperbolic processes in finance. In *Handbook of heavy tailed distributions in finance* (ed. S. T. Rachev), pages 211–248. Elsevier, 2003.

Brown, B. W.; Spears, F. M., and Levy, L. B. The log f: a distribution for all seasons. *Computational Statistics*, 17(1):47–58, 2002.

Cheng, J. Y. and Mailund, T. Ancestral population genomics using coalescence hidden markov models and heuristic optimisation algorithms. *Computational biology and chemistry*, 57:80–92, 2015.

Dempster, A. P.; Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1): 1–22, 1977.

Eberlein, E.; Keller, U., and others, . Hyperbolic distributions in finance. *Bernoulli*, 1(3):281–299, 1995.

L.and NeumannHan, M. Effect of dimensionality on the Nelder-Mead simplex method. *Optimization Methods and Software*, 21(1):1–16, 2006.

Hellmich, M. and Kassberger, S. Efficient and robust portfolio optimization in the multivariate generalized hyperbolic framework. *Quantitative Finance*, 11(10):1503–1516, 2011.

Hu, W. *Calibration of multivariate generalized hyperbolic distributions using the EM algorithm, with applications in risk management, portfolio optimization and portfolio credit risk*. PhD thesis, Florida State University, 2005.

Jørgensen, B. *Statistical properties of the generalized inverse Gaussian distribution*, volume 9. Springer Science & Business Media, 1982.

Karlis, D. An em type algorithm for maximum likelihood estimation of the normal–inverse gaussian distribution. *Statistics & probability letters*, 57(1):43–52, 2002.

Küchler, U.; Neumann, K.; Sørensen, M., and Streller, A. Stock returns and hyperbolic distributions. *Mathematical and Computer Modelling*, 29(10-12):1–15, 1999.

Lagarias, J. C.; Reeds, J. A.; Wright, M. H., and Wright, P. E. Convergence properties of the nelder–mead simplex method in low dimensions. *SIAM Journal on optimization*, 9(1):112–147, 1998.

Ley, Christophe. Flexible modelling in statistics: past, present and future. *Journal de la Société Française de Statistique*, 156(1):76–96, 2015.

Madan, D. B. and Seneta, E. The variance gamma model for share market returns. *Journal of business*, pages 511–524, 1990.

Massey Jr, F. J. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

McNeil, A. J.; Frey, R., and Embrechts, P. *Quantitative risk management: concepts, techniques and tools-revised edition*. Princeton university press, 2015.

Moon, T. K. The expectation-maximization algorithm. *IEEE Signal processing magazine*, 13(6): 47–60, 1996.

Murphy, S. A. and Van der Vaart, A. W. On profile likelihood. *Journal of the American Statistical Association*, 95(450):449–465, 2000.

Nelder, J. A. and Mead, R. A simplex method for function minimization. *The computer journal*, 7(4):308–313, 1965.

Panahi, H. Discriminating between the normal inverse gaussian and generalized hyperbolic skew-t distributions with a follow-up the stock exchange data. *Yugoslav Journal of Operations Research*, 28(2):185–199, 2018.

Paolella, M. S. *Intermediate probability: A computational approach*. John Wiley & Sons, 2007.

Patefield, W. M. On the maximized likelihood function. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 92–96, 1977.

Prause, K. Modelling financial data using generalized hyperbolic distributions. *Freiburg Center for Data Analysis and Modelling Preprint*, 48, 1997.

Prause, K. *The generalized hyperbolic model: Estimation, financial derivatives and risk measures*. PhD thesis, Albert Ludwig University, 1999.

Protassov, R. S. Em-based maximum likelihood parameter estimation for multivariate generalized hyperbolic distributions with fixed $\lambda$. *Statistics and Computing*, 14(1):67–77, 2004.

Puig, P. and Stephens, M. A. Goodness-of-fit tests for the hyperbolic distribution. *Canadian Journal of Statistics*, 29(2):309–320, 2001.

Rathgeber, A. W.; Stadler, J., and Stöckl, S. Fitting generalized hyperbolic processes-new insights for generating initial values. *Communications in Statistics-Simulation and Computation*, 46(7): 5752–5762, 2017.

Raue, A.; Kreutz, C.; Maiwald, T.; Bachmann, J.; Schilling, M.; Klingmüller, U., and Timmer, J. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–1929, 2009.

Rdocumentation.org, . constroptim: Linearly constrained optimization, a. https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/constrOptim.

Rdocumentation.org, . nlminb: Optimization using port routines, b. https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/nlminb.

Rdocumentation.org, . optim: General-purpose optimization, c. https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/optim.

Scott, D. *HyperbolicDist: The hyperbolic distribution*, 2009. URL https://CRAN.R-project.org/package=HyperbolicDist. R package version 0.6-2.

Scott, D. *GeneralizedHyperbolic: The Generalized Hyperbolic Distribution*, 2018. URL https://CRAN.R-project.org/package=GeneralizedHyperbolic. R package version 0.8-4.

Scott, D. and Grimson, F. *SkewHyperbolic: The Skew Hyperbolic Student t-Distribution*, 2018. URL https://CRAN.R-project.org/package=SkewHyperbolic. R package version 0.4-0.

Scott, D. J.; Würtz, D.; Dong, C., and Tran, T. T. Moments of the generalized hyperbolic distribution. *Computational statistics*, 26(3):459–476, 2011.

Silvey, S. D. *Statistical Inference*. Harmondsworth: Penguin Press, 1970.

Snoussi, H. and Idier, J. Bayesian blind separation of generalized hyperbolic processes in noisy and underdeterminate mixtures. *IEEE Transactions on Signal Processing*, 54(9):3257–3269, 2006.

Weibel, M.; Luethi, D., and Breymann, W. *ghyp: Generalized Hyperbolic Distribution and Its Special Cases*, 2020. URL https://CRAN.R-project.org/package=ghyp. R package version 1.6.1.