5-2022

Finding a Representative Distribution for the Tail Index Alpha, α, for Stock Return Data from the New York Stock Exchange

Jett Burns
East Tennessee State University

Finding a Representative Distribution for the Tail Index Alpha, α, for Stock Return Data from the New York Stock Exchange

---

A thesis

presented to

the faculty of the Department of Mathematics & Statistics

East Tennessee State University

In partial fulfillment

of the requirements for the degree

Master of Science in Mathematical Sciences

---

by

Jett Burns

May 2022

JeanMarie Hendrickson, Ph.D., Chair

Michele Joyner, Ph.D.

Nicole Lewis, Ph.D.

Robert Price, Ph.D.

Gary Shelley, Ph.D.

Keywords: statistics, risk, parameter estimation, maximum likelihood estimation, tail index

ABSTRACT

Finding a Representative Distribution for the Tail Index Alpha, α, for Stock Return Data from the New York Stock Exchange

by

Jett Burns

Statistical inference is a tool for creating models that can accurately display real-world events. Special importance is given to the financial methods that model risk and large price movements. A parameter that describes tail heaviness, and risk overall, is α. This research finds a representative distribution that models α. The absolute value of standardized stock returns from the Center for Research on Security Prices are used in this research. The inference is performed using R. Approximations for α are found using the *ptsuite* package. The *GAMLSS* package employs maximum likelihood estimation to estimate distribution parameters using the CRSP data. The distributions are selected by using AIC and worm plots. The Skew t family is found to be representative for the parameter α based on subsets of the CRSP data. The Skew t type 2 distribution is robust for multiple subsets of $\hat{\alpha}$ values calculated from the CRSP stock return data.

DEDICATION

This research is dedicated to my wonderful husband, Jacob.

ACKNOWLEDGMENTS

I want to thank William Trainor, Ph.D. for providing access to data that made this research possible.

TABLE OF CONTENTS

ABSTRACT 2
DEDICATION 4
ACKNOWLEDGMENTS 5
LIST OF TABLES 7
LIST OF FIGURES 8
1 INTRODUCTION 9
1.1 Alpha-Stable Distributions 11
1.2 The Lévy Distribution and the Applications of Stable Distributions in Quantitative Finance 12
1.3 The Tail Index Alpha 14
1.4 Review of Software Based Approaches 15
1.5 Use of the ptsuite Package 15
1.6 Use of the GAMLSS Package 17
2 THE DATA AND DATA CLEANING PROCESS 21
2.1 Example of the Process using Altria Group 22
2.2 Use of Different Sample Sizes 23
2.3 Creating the $\hat{\alpha}$ Vectors 24
3 METHODOLOGY AND R IMPLEMENTATION 27
3.1 The m_0 Model with Varying s_o Values 27
3.2 The m_1 Model with $s_o = 2500$ 32
3.3 The m_2 Model with $s_0 = 5000$ 37
3.4 The m_3 Model with $s_0 = 10000$ 42
3.5 Discovering a Final Model Overall 46
4 OVERVIEW OF THE SELECTION RESULTS AND FITTED MODELS 50
4.1 The m_0 Fitted Model 50
4.2 The m_1 Fitted Model 55
4.3 The m_2 Fitted Model 57
4.4 The m_3 Fitted Model 60
4.5 Comparing Different Fitted m_0 and m_1 Models 62
5 INTERPRETATIONS AND CONCLUSIONS 66
5.1 Discussion of the $\hat{\alpha}$ Vectors 66
5.2 Finding the Representative Distribution for α 67
5.3 Relevant Further Research 69
BIBLIOGRAPHY 71
APPENDIX: R Code 74
VITA 87

LIST OF TABLES

1 Overview of Model s_n and AIC Values 50
2 Link Functions for the ST3 Distribution 52
3 Summary Output of the Fitted ST3 Distribution 53
4 Link Functions for the ST1 Distribution 56
5 Summary Output of the Fitted ST1 Distribution 56
6 Link Functions for the ST2 Distribution 59
7 Summary Output of the Fitted ST2 Distribution 59
8 Summary Output of the Fitted ST2 Distribution 61
9 Summary Output of the $\hat{\alpha}_0$ Fitted ST2 Distribution 64
10 Summary Output of the $\hat{\alpha}_1$ Fitted ST2 Distribution 64
11 Summary of Measures of Center for $\hat{\alpha}$ Vectors 67

LIST OF FIGURES

1 Dow Jones Index vs Gaussian Simulation, found in [19] 10
2 The 3 Special Cases of Alpha-Stable Distributions, from [15] 12
3 Histogram of the Absolute Value Standardized Return for MO 23
4 Histogram of the $\hat{\alpha}$ values for the varying sample size data set 28
5 Worm Plot for the m_0 SEP2 model 29
6 Worm Plot for the m_0 ST3 model 30
7 Worm Plot for the m_0 SST model 31
8 Histogram of the $\hat{\alpha}_1$ values for $s_o = 2500$ subset 32
9 Worm Plot for the m_1 ST1 model 34
10 Worm Plot for the m_1 ST5 model 35
11 Worm Plot for the m_1 JSUo model 36
12 Histogram of the $\hat{\alpha}_2$ values for $s_o = 5000$ subset 37
13 Worm Plot for the m_2 ST2 model 39
14 Worm Plot for the m_2 ST5 model 40
15 Worm Plot for the m_2 ST3 model 41
16 Histogram of the $\hat{\alpha}_3$ values for $s_o = 10000$ subset 42
17 Worm Plot for the m_3 ST2 model 43
18 Worm Plot for the m_3 ST3 model 44
19 Worm Plot for the m_3 SST model 45
20 Worm plot for the $\hat{\alpha}_1$ fitted with the ST2 distribution 47
21 Worm plot for the $\hat{\alpha}_2$ fitted with the ST2 distribution 48
22 Histogram of $\hat{\alpha}_0$ values with the fitted ST3 distribution 51
23 Histogram of $\hat{\alpha}_1$ values with the fitted ST1 distribution 55
24 Histogram of $\hat{\alpha}_2$ values with the fitted ST2 distribution 58
25 Histogram of $\hat{\alpha}_3$ values with the Fitted ST2 distribution 60
26 Histogram of $\hat{\alpha}_0$ values with the Fitted ST2 distribution 63
27 Histogram of $\hat{\alpha}_1$ values with the Fitted ST2 distribution 63

1 INTRODUCTION

Statistical inference, especially in the late 20th century, has relied heavily on the work of Carl Friedrich Gauss and Louis Bachelier. Around 1900, Bachelier developed a random walk model based on probability theory, which laid the foundation for Brownian motion and what became known as the 'normal distribution' [19]. He used this to model bets on coin tosses, and then option prices in the financial market; his thesis went relatively unnoticed until 1964, when it was translated to English and became a cornerstone of 20th century economics and finance [19]. The late 1900s, especially the 1970s, stands out as a distinct period for the creation of financial models. For instance, the Markowitz portfolio theory, the Bachelier market model, the Sharpe asset-pricing, the Capital Asset Pricing Model and its modifications, and the Black-Scholes formula, among others, remain valuable tools in finance [18] [23] [27]. Since the tools' inception, they have been continually modified using stable distributions, computational mathematics, and software-based modeling [23].

For example, a Gaussian distribution allows for an event outside of three standard deviations, '3σ', to happen approximately 0.13% of the time [19]. Historical data for the Standard and Poor's 500 index indicate that '3σ' events happened at an almost 1% frequency, 8 times the Gaussian prediction [19][31]. Eugene Fama, a student of Mandlebrot, investigated the 30 stocks in the Dow Jones Industrial index one-by-one, finding that price changes of five standard deviations or more occurred two thousand more times than expected, or, to quote Mandlebrot, "you should have encountered such drama only once every seven thousand years; in fact, ... it happened once every three or four years" [9] [19]. In the currency market, there is research by Citigroup that found many price changes outside of the '3σ' boundary, the largest being a

'10.7σ' event, which under Gaussian odds, to quote Mandlebrot again, "would not have happened once if Citigroup had been trading dollars and yen every day since the Big Bang 15 billion years ago" [9] [19].

A graph of daily price changes of the Dow Jones Industrial Average index is found in Figure 1. The graph contains the size of changes in standard deviations on the horizontal axis, and the frequency of these changes on the vertical axis [19]. The black bars are the real-world data, and the grey are the Gaussian simulation. The data is right-skewed, and the Gaussian simulation fails to account for the heavy tails. Evidently, as the preceding examples explain, there is a fundamental issue with the 'normal distribution', especially when applied in the complex setting of financial modeling [19] [26]. This research focuses on compiling data on stock returns, estimating the heaviness of the tail index, and then finally, finding a representative distribution for the tail index. In order to do this, there must be a discussion of the family of stable distributions.

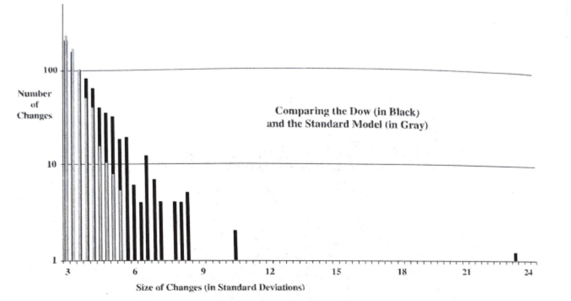

Figure 1: Dow Jones Index vs Gaussian Simulation, found in [19]

1.1 Alpha-Stable Distributions

As will be discussed later, Figure 1 is integral to this research. Overall, the Gaussian problem, or 'normal distribution' problem, has been identified by financial modelers and academics for some time [26]. The Gaussian distribution is one of three special cases of a 'stable distribution', the others being the Cauchy distribution and the Lévy Distribution [26]. A stable distribution is defined as the following: If X_0, X_1, and X_2 are independent, identically distributed (iid) random variables, then the distribution of these random variables is 'stable' if for every pair of positive real numbers, a and b, there exists a positive c and real d so that $cX_0 + d$ is distributed the same as $aX_1 + bX_2$ [8].

Alpha-Stable distributions are specified by four parameters [8] [13]. We can write a stable distribution as follows: $X_s \sim (\alpha_s, \beta, \mu, \sigma)$, using the notation of Taleb [26]. We will define each parameter as follows: α_s is the stability index for $\alpha_s \in (0, 2]$, also known as the tail index, which determines tail behavior; β is the skewness parameter for $\beta \in [-1, 1]$, which indicates a right skew for positive β, a left skew for negative β, and symmetry at 0; μ is the shift parameter, which measures the shift of the mode or rather the 'peak' of the distribution; σ is the scale parameter, which is a positive number determining the width and dispersion [14]. Besides the three special cases of Gaussian, Cauchy, and Lévy, the probability density function (PDF) and cumulative distribution function (CDF) of stable distributions do not have a closed form [15][26]. The Gaussian distribution has parameters $\alpha_s = 2$ and $\beta = 0$, with μ and σ. [14] The Cauchy distribution has parameters $\alpha_s = 1$ and $\beta = 0$, with μ and σ [15]. The Lévy distribution has parameters $\alpha_s = 1/2$ and $\beta = 1$, with μ and σ [14].

Figure 2 shows the three cases where $\beta = 0$ for the Gaussian and Cauchy and

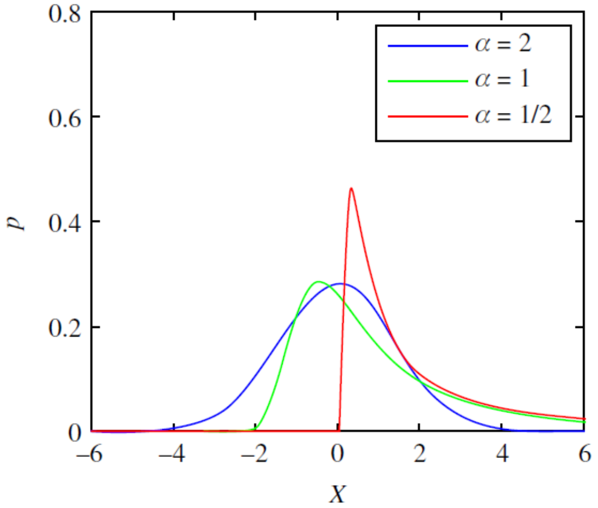

Figure 2: The 3 Special Cases of Alpha-Stable Distributions, from [15]

$\beta = 1$ for the Levy, $\mu = 0$ and $\sigma = 1$, with Gaussian being blue, Cauchy being green, and Lévy in red [15]. Stable distributions have heavy tails, except for the Gaussian, making stable distributions more appropriate for modeling systems that behave randomly or involve risk [26]. The red line in Figure 2 bears a resemblance to the behavior of the black bars in Figure 1. Of particular interest is the Lévy distribution and similar generalized distributions, as they have become widely applied in quantitative finance as a solution to the Gaussian problem described above [14] [15] [26] [19].

1.2 The Lévy Distribution and the Applications of Stable Distributions in Quantitative Finance

The Lévy distribution is named after mathematician Paul Lévy, who began much of the work on stable distributions [19] [26]. Mandlebrot worked extensively with these distributions; he termed stable distributions with $1 < \alpha_s < 2$, and $\beta = 1$ as 'Pareto- Lévy distributions' and found these better for financial markets than the

Gaussian distribution [19]. The PDF of the Lévy distribution with $\alpha_s = \dfrac{1}{2}$ and $\beta = 1$ is as follows:

$$f_x(x|\mu, \sigma, \frac{1}{2}, 1) = \left(\frac{\sigma}{2\pi}\right)^{1/2} \frac{1}{(x-\mu)^{3/2}} e^{-\frac{\sigma}{2(x-\mu)}},$$

where $\mu < x < \infty$ [2] [14] [18].

As discussed, financial engineers have attempted to fix the 'Gaussian problem' with the family of stable distributions with $\alpha_s < 2$ [19] [26] [32]. Specifically, Lévy and Pareto-Lévy distributions and processes, such as Brownian motion with drift, Poisson and compound Poisson jump processes, and pure jump Lévy processes have all been applied in quantitative finance, with examples such as Wu [31], Todorov and Tauchen [27], Figueroa- López [11], [18], Choi and Yoon [6], Hirsa and Neftci [12], Choudhry [7], and Xiong [32].

Wu states that a Lévy process that generates an infinite number of jumps is more suitable to capture daily changes in financial securities [31]. In addition, Wu concludes that infinite-activity jumps perform better than finite-activity jumps, and that $1 \leq \alpha_s < 2$ generates sample paths with infinite variation, which provides smooth transitions from large jumps to small jumps to continuous movement [31]. Wu closes by describing Lévy processes as integral to modern finance, and that different Lévy-based models can be used to model both continuous and discontinuous movement [31].

Most of the other technical research is driven by modeling price behavior using Lévy processes, like Todorov and Tauchen, who introduce a bivariate mixture of Gamma models for driving the Lévy process [27].

The main interest in describing the Lévy distribution is to focus on the parameter α_s, the tail index. The tail index is the parameter of interest in this research because of the significance α holds with modern financial models, portfolio risk analysis, and Lévy processes [20] [17]. Therefore, a focus on modeling α is the beginning of developing and influencing models for financial risk as a whole [20].

1.3 The Tail Index Alpha

The tail index α measures the heaviness of the tail of the distribution, and thus α helps measure the risk of a large price movement for a financial asset [20] [26] [22]. According to Mittinik, it may be more beneficial to approximate the tail index alpha for a financial asset directly, rather than model the entire distribution [20].

Therefore, α is generally regarded as a tool to model the risk of large movements in the price of a financial asset and its value plays an integral role in the modeling of financial risk [17] [20]. The consequences of an unexpected downside event in financial markets can have drastic effects in the real-world economy [30]. Naturally, individuals, private businesses, and governments have a vested interest to prepare for any potential downside events [30].

This research intends to fortify the complex models described above by finding a representative distribution for the tail index α overall by using a software-based approach to stock returns obtained from the New York Stock Exchange. Ultimately, the α parameter inside of the models described above or new models can be viewed as a random variable based on the representative distribution found by using real data.

Thus, there are two research questions proposed here:

1. Using numerical estimation methods, can estimates for α be obtained, and are these α's approximately .50, akin to the Lévy distribution?
2. Can a representative distribution, and the estimated parameters of that distribution, be found that can accurately model α?

In order to answer both of these questions, there must be a discussion on the numerical approach to the data using R statistical software.

1.4 Review of Software Based Approaches

Using the statistical software R, numerical approximation was used to estimate the tail index, denoted $\hat{\alpha}$, and numerical searches were used to develop representative models for $\hat{\alpha}$. R is an open source programming language used predominantly in academia because of its ease of use and versatility. The foundation of R is built with the development of specific packages and functions for specific problems, especially when there are no present functions available.

R was introduced in 1996, built upon the software S, and became publicly available in 2000 with version 1.0.0 [16]. This research is done solely in R version 4.1.2. R packages can be built by individuals or institutions and can be accessed on the R repository for public use. There are two main packages that are used in this research. For a full list of R packages used here, and for all the code used to find the results, see appendix.

1.5 Use of the ptsuite Package

The package *ptsuite* was built by Ranjiva Munasinghe, Pathum Kossinna, Dovini Jayasinghe, and Dilanka Wijeratne with the specific intention of finding estimates for the tail index $\hat{\alpha}$ of heavy-tailed distributions [21]. The team behind *ptsuite* developed unique functions that would estimate α using multiple different approximation methods [21]. In this research, the numerical approximation of α was performed using the geometric mean modified percentile method of Bhatti that was built inside of the package *ptsuite* [5]. The geometric mean modified percentile method (GMMP) did not perform as well in Bhatti's Monte Carlo simulations compared to other percentile methods and maximum likelihood methods. [5] However, when the GMMP method was performed and compared to the other methods for the data in this research, the GMMP method was the best performer. Performance was judged based on miscalculations or outlying observations by the approximation method. Additionally, the geometric mean has long been used in quantitative finance for calculations involving data that may not be independent or time-series. [28]. Thus, it was determined that α estimation for this research would be done using the geometric mean modified percentile estimator. The equation 1 is the approximation method for $\hat{\alpha}$ [21]. For clarity, α represents the unknown true value of the parameter, and $\hat{\alpha}$ represents the estimated tail index specific to the data. The equation can found on page 4 of [21].

$$\hat{\alpha} = \frac{1 - \ln(4)}{\frac{1}{N}\sum_{i=1}^{N} \ln x_i - \ln(P_{75}^{*})} \tag{1}$$

We have that P_q^* is the q^{th} percentile of the data [21]. Percentiles are the recommended approach for parameter approximation; the idea behind percentiles begins with equating two values of the cumulative distribution function with the respective percentiles and then solving for the unknown parameters [5]. See both [5] and [21] for more details.

1.6 Use of the GAMLSS Package

The Generalized Additive Models for Location, Scale, and Shape (GAMLSS) package and model procedure were developed by Robert Rigby, Mikis Stasino- poulos, Calliope Akantziliotou, and others in 2005 to overcome limitations associated with the Generalized Linear and the Generalized Additive Models at the time [29].

Specifically, GAMLSS utilizes Maximum Likelihood Estimation procedures and numerical search methods to find estimates for parameters and fit distributions over quantitative data sets [24]. Below is an example of the Maximum Likelihood Estimation procedure under GAMLSS for the gamma distribution. Theoretically, the data in this example will be fit to a two parameter gamma distribution, defined as $GA(\mu, \sigma)$ [24].

Let $GA(\mu, \sigma)$ be defined with the following probability density function:

$$\frac{y^{1/\sigma^2-1}e^{-y/(\sigma^2\mu)}}{(\sigma^2\mu)^{1/\sigma^2}\Gamma(\sigma^{-2})}$$

The mean, $E(Y) = \mu$, and the variance, $Var(Y) = \sigma^2\mu^2$ are also defined [24].

Now, the likelihood function can be developed in the following way:

$$L(\mu, \sigma) = \Pi_{i=1}^{n} \frac{1}{(\sigma^2\mu)^{1/\sigma^2}} \frac{y_i^{1/\sigma^2-1}e^{-y_i/(\sigma^2\mu)}}{\Gamma(1/\sigma^2}$$

Likewise, the log-likelihood is defined as:

$$\ell(\mu,\sigma) = \sum_{i=1}^{n}[-\frac{1}{\sigma^2}(\log\sigma^2 + \log\mu) + (\sigma^{-2} - 1)\log(y_i) - \frac{y_i}{\sigma^2\mu} - \log\Gamma(\sigma^{-2})]$$

$$= -\frac{n}{\sigma^2}(\log\sigma^2 + \log\mu) - n\log\Gamma(\sigma^{-2}) + (\sigma^{-2} - 1)\sum_{i=1}^{n}\log y_i - \frac{\sum_{i=1}^{n} y_i}{\sigma^2\mu}$$

In Maximum Likelihood Estimation, the derivations are typically easier to handle with the log-likelihood [24]. The next step in the MLE process is to take derivatives of the log-likelihood with respect to each of the parameters [24]. In this case, the derivative is taken with respect to μ and σ.

$$\frac{\partial\ell(\mu,\sigma)}{\partial\mu} = \frac{\sum_{i=1}^{n} y_i - n\mu}{\sigma^2\mu^2}$$

$$\frac{\partial\ell(\mu,\sigma)}{\partial\sigma} = \frac{2}{\sigma^3}[(\frac{\sum_{i=1}^{n} y_i}{\mu}) - (\sum_{i=1}^{n}\log y_i) + n\log(\mu) + n\log(\sigma^2) - n + n\Psi(\frac{1}{\sigma^2})],$$

where $\Psi(x) = \frac{d}{dx}\log\Gamma(x)$ [24]. In order to find the maximum likelihood estimate (MLe) for each of the parameters, the derivatives for each parameter are set equal to zero and solved [24]. Analytically, $\hat{\mu} = \bar{y}$ can be found [24]. However, $\hat{\sigma}$ cannot be found using a closed form, and thus must be found using numerical search methods [24]. The numerical search methods involves searching along a vector of values and finding the value which minimizes the equation in terms of absolute value, therefore finding the value for which the equation equals zero [29]. The package GAMLSS, and the included functions, are built specifically to streamline the process for Maximum Likelihood Estimation of parameters for many different distributions [29]. For example, using the function $gamlssML()$, the likelihood functions, derivatives, and numerical searches are performed internally, especially since many of the distributions' log-likelihood derivations do not have closed forms, like the gamma distribution example above [29] [24]. This numerical search procedure will be employed in order to find the representative distribution of $\hat{\alpha}$, and the parameters for the distribution. For more information, please refer to [24] [29].

The Akaike information criterion is used for comparing models to one another. There are many model selection techniques, including the Schwarz Bayesian criterion (SBC); however, the AIC is chosen for this research since it is generous in model selection [24]. While this can create complicated models, the AIC's generosity is favorable for situations where the distribution and parameters are completely unknown [24]. The AIC value has no meaning on its own, but has meaning when compared to other AIC values for the same data, and the typical rule is the smallest value indicates the best model fit [29]. In addition, checking the normalized quantile residuals using worm plots, which are detrended Q-Q plots and found using the function $wp()$, provides guidance as to the model fit [24]. Worm plots provide approximate 95% bands, represented as dotted lines, and plot the observations' residuals based on deviance from 0 [24]. The larger each observations deviance from zero then the less a fitted model is representing the data [24].

The Generalized Likelihood Ratio (GLR) test is used to describe the significance of parameters in models that are nested [29]. That is, one model is a complex extension of another; the GLR test determines if the parameters of the complex model are useful [24]. For instance, let M_1 and M_2 be two models, where M_1 is nested within M_2 [24]. Therefore, M_1 is the simpler model and M_2 is the more complex model [24]. The hypothesis test becomes [24]:

1. $H_0 : M_1$

2. $H_1 : M_2$

Both of these models have fitted likelihood functions, $\hat{L}_1$ and $\hat{L}_2$, respectively [24]. Thus, the likelihood ratio is $LR = \dfrac{\hat{L}_1}{\hat{L}_2}$ [24]. The GLR test statistic becomes:

$$\Lambda = -2\log(LR)$$

There is an asymptotic distribution $\Lambda \sim \chi^2_d$ [24]. Since this data is based explicitly on data with no explanatory variable, the null hypothesis will be rejected if $\Lambda \geq 3.84$, which corresponds to χ^2_1 [24].

2 THE DATA AND DATA CLEANING PROCESS

The research questions proposed in Section 1.3 require financial data that is both significant in the amount of observations and significant in the relevance that the parameter α would provide. Thus, it was determined that data on stocks, specifically the standardized returns on stocks, would be appropriate for analysis; the parameter α would provide an estimate on the tail risk for large movements in the stock, as measured by the standardized return [17] [28]. Standardized returns represent the financial return on a certain stock, standardized over time for analysis of financial performance [1].

All calculations of $\hat{\alpha}$ were calculated based on stock return data from Center for Research in Security Prices (CRSP) at the University of Chicago Booth School of Business [1]. The data begins at the date 01/04/1926 and ends with the date 12/31/2020 for a sample of 1642 stocks in the New York Stock Exchange (NYSE), resulting in 25,044 entries in the data set. However, not all the stocks in the data set have been actively traded since 1926, resulting in incomplete columns of data for a large proportion of the stocks. In the data set, there is a -77, -88, -99 or similar value, in the place of the missing data. Because of these values, unique functions were written to go column by column and remove these values in order to have a column of data that was just the stock returns themselves. See appendix for these functions.

Additionally, the focus on the α parameter implies that the interest is on strictly price swings; whether the swings are in the positive or negative direction are not of particular interest. Likewise, returns that are approximately zero cannot be used since this implies no price movement, and the zeros cannot be used in the estimation of α using (1) [21]. Thus, the final data column for each stock is the absolute value of the standardized return (AVSR) for each trading day since the stock's introduction on the NYSE, excluding those returns that are approximately zero. This process will be referred to as the data cleaning process.

2.1 Example of the Process using Altria Group

For example, the histogram of the AVSR for Altria Group, ticker symbol MO, is found in Figure 3. The histogram in Figure 3 is extremely right skewed, with a minimum value of 0.000062, a median value of 0.009075, and a maximum value of 0.28, with a total of $22,690$ observations. Thus, we can define this as a distribution with a heavy tail [26]. It is the tail of this distribution in Figure 3 for which an estimate for α, designated as $\hat{\alpha}$ once it is assigned a value, will be obtained in order to attempt to describe the heaviness of the tail with a singular parameter estimate [17]. For instance, using (1), Altria group has an $\hat{\alpha} = 0.5815435$. While this value has meaning on its own, it is important to analyze this value compared to other α values in the data set. This process of data cleaning and $\hat{\alpha}$ estimation will be repeated for each stock in the data set, allowing for the distribution of the $\hat{\alpha}$'s to be analyzed.

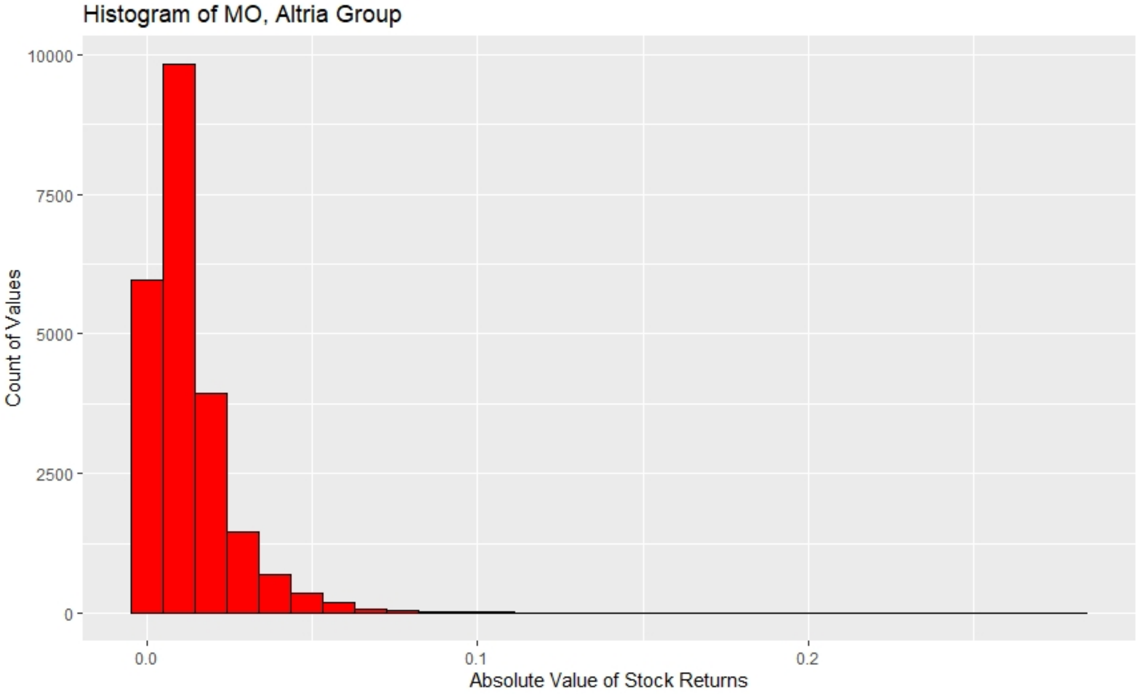

Figure 3: Histogram of the Absolute Value Standardized Return for MO

2.2 Use of Different Sample Sizes

The data cleaning process implies that each $\hat{\alpha}$ value will be calculated with different sample sizes for each stock, depending on when that stock was introduced on the NYSE. In order to account for this, a unique function called $lengthfunc()$ was written to record the number of observed values for each stock after the cleaning process has taken place. Therefore, this function can be used to find stocks that share a certain amount of observations and enable the partitioning of the data set into subsets where each subset has the same amount of observations per stock. However, this process will sacrifice overall sample size in order to equalize the observations. For clarity, following notation will be used in reference to the data and subsets of the data:

1. s_n = the number of stocks in the data set, or subset of the data

2. $s_o =$ the number of observations for each stock in a subset of the data

For example, for the overall data set, $s_n = 1642$, but s_o varies for each stock. Using the function $lengthfunc()$ and the indexing process in R, 3 subsets of the data were created. See appendix for details. The subsets are as follows:

1. A subset containing $s_o = 2500$, and $s_n = 1284$.

2. A subset containing $s_o = 5000$, and $s_n = 905$.

3. A subset containing $s_o = 10000$ and $s_n = 270$

Increasing the s_o threshold limits s_n, the number of stocks that can be analyzed. There were subsets past the s_o value of $10,000$ created, but this decreased the s_n value so much that statistical inference did not seem appropriate. The interest and purpose for creating the subsets is three-fold: first, ensuring a model was created based on an equal number of observations for each $\hat{\alpha}$, second, comparing these models together and to the model with different sample sizes, and finally, to see if similar results appear for each of the four models.

2.3 Creating the $\hat{\alpha}$ Vectors

Each of these subsets and the original data were analyzed to create the vector of $\hat{\alpha}$ values that correspond to each stock in the data set, such as Figure 3. The $\hat{\alpha}$ vectors were used inside the $gamlssML()$ function to create a $GAMLSS$ object [29]. This object is used for the numerical search to find the most approrpiate distribution and the MLe's for the parameters of that distribution [24] [29].

Each model was named in R, using the following format: m_0 represents the model created using the entire data set with varying s_o values, m_1 represents the model created using the $s_o = 2500$ subset, m_2 represents the model created using the $s_o = 5000$, and finally m_3 represents the model created using the $s_o = 10000$ subset.

There are two unique function written to find the $\hat{\alpha}$ for each stock, called $alpha\_estimate()$ and $alpha\_estimate2()$, each with the same basic application. The input would be a column of data, and the output would be the $\hat{\alpha}$ value using (1). This function was applied in conjunction with the $lapply()$ function, which allowed for each column (stock) of the data sets to be recognized, $\hat{\alpha}$ to be calculated, and the output be a vector containing s_n number of $\hat{\alpha}$ values. See appendix for details. This process was repeated to create an $\hat{\alpha}$ vector corresponding to the data set overall and to each subset. These sets of $\hat{\alpha}$ vectors will be labeled with a subscript corresponding to the subscript of m_i for each subset. Thus, $\hat{\alpha}_0$ will correspond to the $\hat{\alpha}$ vector used to create the m_0 model, the model with varying s_o values. Likewise, $\hat{\alpha}_1$ will correspond to the $\hat{\alpha}$ vector used to create the m_1 model, the model with $s_o = 2500$.

During this process of creating these vectors, there were some data irregularites that made it past the cleaning process that was intended when using the $lengthfunc()$ and indexing procedures. For example, when creating the $\hat{\alpha}_1$ vector, there were issues with some of the $\hat{\alpha}$ estimates. The $\hat{\alpha}_1$ vector contained three NaN values for columns 48, 63, and 147 in the $s_o = 2500$ data subset. Additionally, columns 23 and 1157 returned $\hat{\alpha}$ estimates of -1.968654 and 6.865167, which are well outside the assumed standard for stable distributions and their tails [14]. Upon further investigation, these estimates were the result of stocks that made it through the cleaning process that did not have 2500 observations, with -77 as a value for missing observations, which skewed the $\hat{\alpha}$ estimate for these two columns. These two columns were removed, along with the columns that produced NaN values. This same process of amending the $\hat{\alpha}$ vector after it was calculated was repeated for each subset, although the other subsets had far less issues with $\hat{\alpha}$ estimation and NaN values.

The data cleaning process and the creation of the $\hat{\alpha}$ vectors produced four sets of $\hat{\alpha}$ values, one for the original data set with differing s_o values, and three corresponding to each of the subsets partitioned by the s_o threshold. The $\hat{\alpha}$ vectors are noted as $\hat{\alpha}_0$ for the original data set, and then $\hat{\alpha}_1$, $\hat{\alpha}_2$, and $\hat{\alpha}_3$ for each of the subsets, respectively. These four sets are used to estimate representative distributions for α overall.

3 METHODOLOGY AND R IMPLEMENTATION

3.1 The m_0 Model with Varying s_o Values

Beginning with m_0, the model with varying sample sizes, the $\hat{\alpha}_0$ vector was used as the input into *gamlssML*() function to create a *GAMLSS* object. Next, the *chooseDist*() function employed the Maximum Likelihood Estimation process discussed in Section 1.6. This process fit 29 distributions in total, and the AIC values for each are found using the *getOrder*() function [29]. Importantly, the selection procedure will be performed with the following guidelines: comparing each model's AIC value, creating worm plots, and if necessary, examining the model's parameter significance. Parameter significance will be discussed in detail in Section 4, including parameter estimates for the final selected models. The guidelines were decided due to the close proximity of many of the AIC values in the following results, and the analysis of worm plots played a significant role in model selection. Going forward, the three models with the lowest AIC values will be analyzed, including the worm plots for each. See appendix for more details and code used for model selection. The final models selected in this section are discussed in detail in Section 4.

The histogram of the $\hat{\alpha}_0$ values can be found in Figure 4. Overall, there are 1642 $\hat{\alpha}_0$ values.

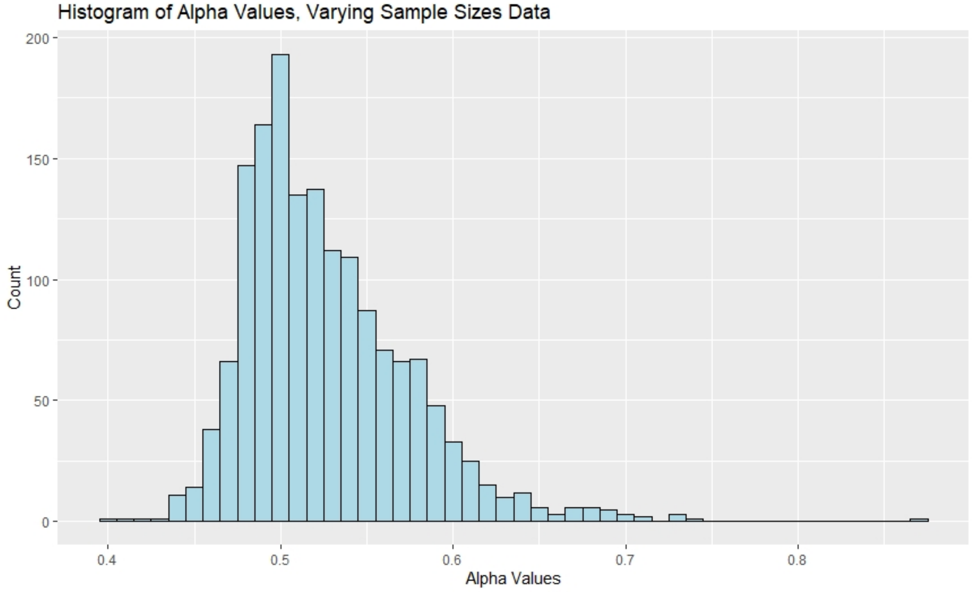

Figure 4: Histogram of the $\hat{\alpha}$ values for the varying sample size data set

According to AIC value, the top two distributions to fit this data were the Skew Exponential Power type 2 (SEP2) and the Skew t type 3 (ST3), with AIC values -5513.77 and -5512.18, respectively. The Skew Student t (SST) distribution is third with an AIC value of -5512.18. The SST distribution is a reparameterization of the ST3 distribution, thus the fit for both are the same, although the parameters may differ [24]. The final model for the $\hat{\alpha}_0$ is discussed in detail in Section 4.1.

The worm plots for the SEP2 model, the ST3 model, and the SST model are found in Figure 5, Figure 6, and Figure 7 respectively.

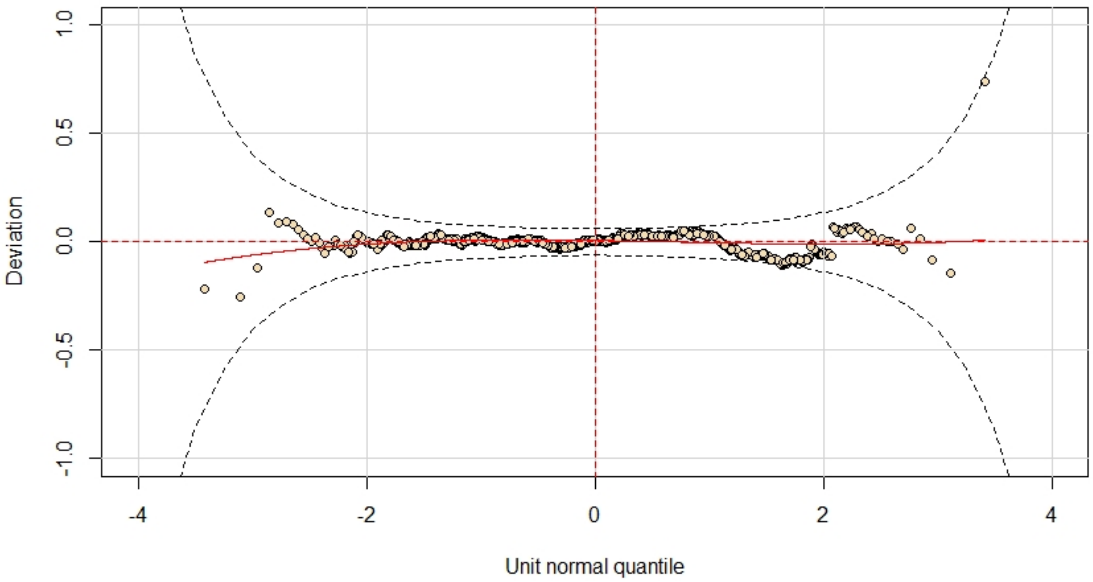

Figure 5: Worm Plot for the m_0 SEP2 model

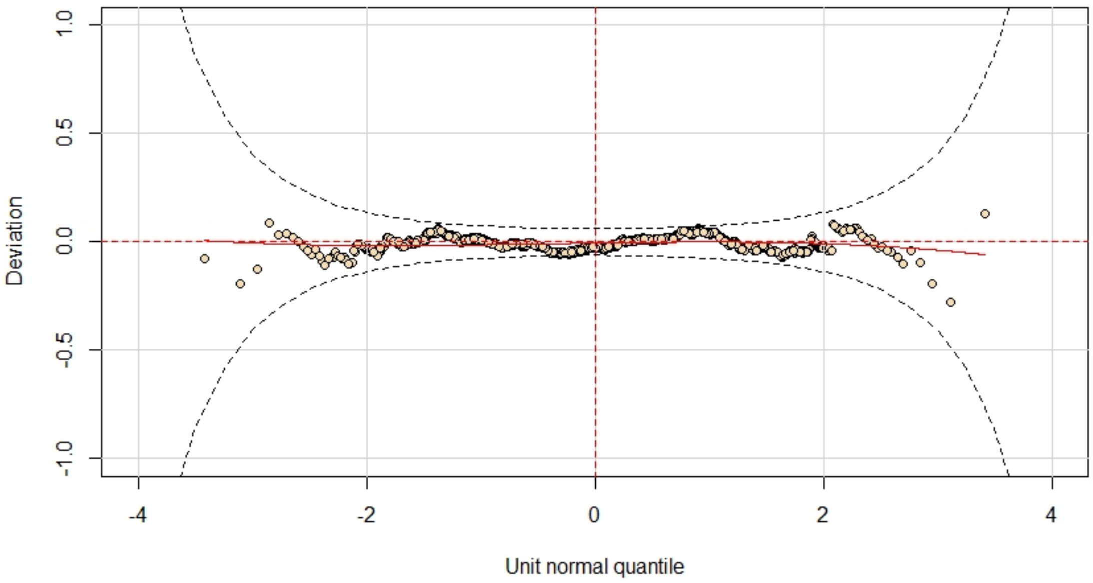

Figure 6: Worm Plot for the m_0 ST3 model

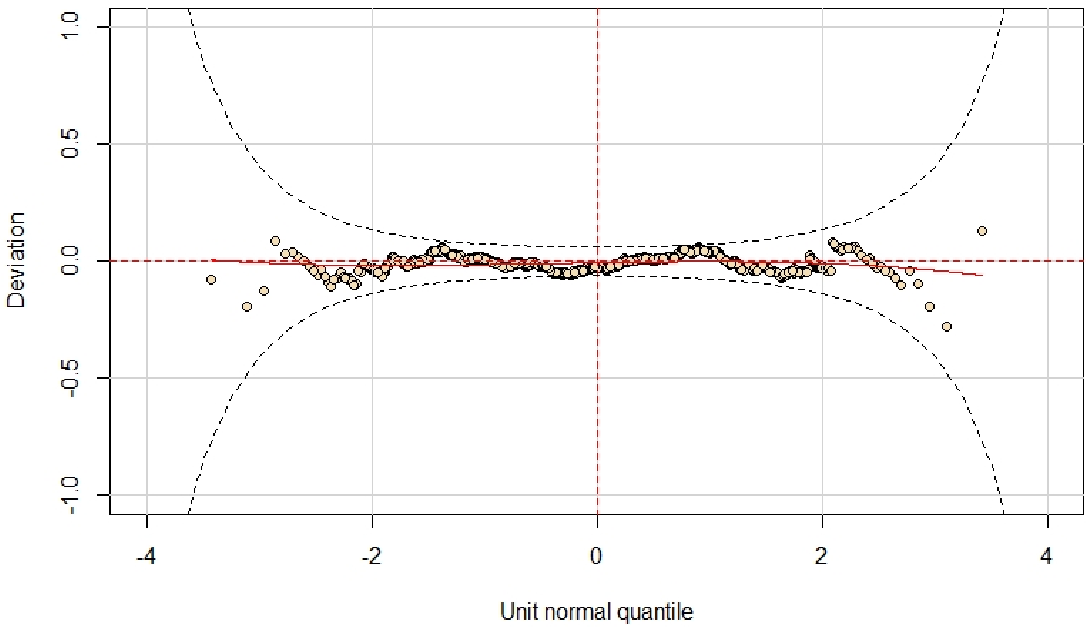

Figure 7: Worm Plot for the m_0 SST model

In the comparison of these three plots, which each lend credence to the individual models' overall fit to the $\hat{\alpha}_0$ values, there is more deviance from 0 in the SEP2 model than in the ST3 model. The dotted lines represent approximate 95% level confidence bands [24]. All of the observation points fall within these bands for the ST3 model, whereas in the SEP3 model, there is a strong dip which seems to violate the confidence band, in addition to a strongly outlying observation on the right hand side of the graph. Both of these considered and in comparison with the ST3 model plot, the ST3 model seems to actually have a more robust fit on the data. The SST worm plot is nearly identical to the ST3 model. Thus, the ST3 model is chosen to represent the $\hat{\alpha}_0$ data.

3.2 The m_1 Model with $s_o = 2500$

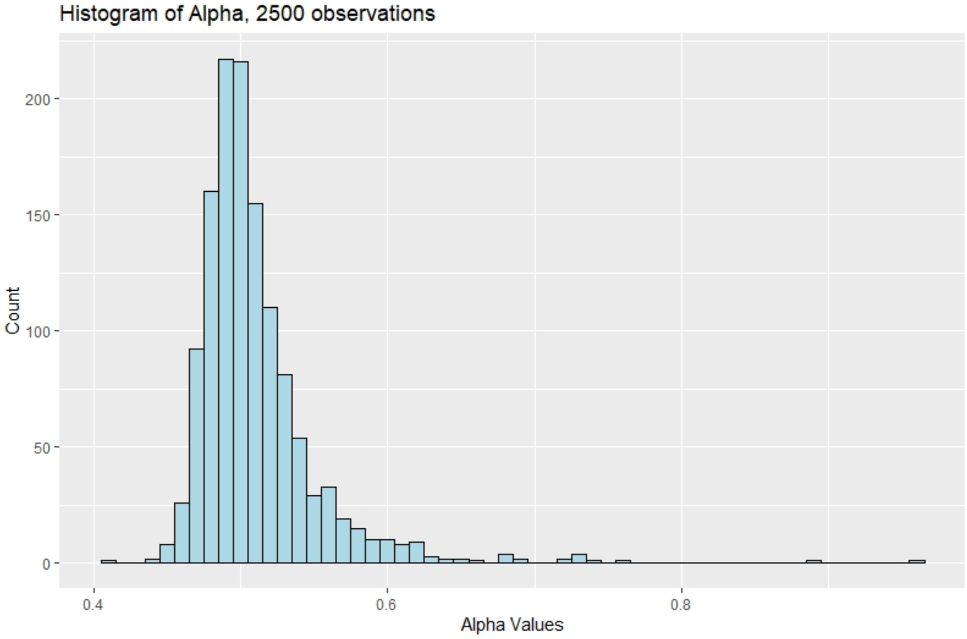

Figure 8: Histogram of the $\hat{\alpha}_1$ values for $s_o = 2500$ subset

The m_1 model is fit using the partition of the $\hat{\alpha}_1$ vector of values with $s_o = 2500$, thus making $s_n = 1284$. After creating the m_1 $GAMLSS$ object and fitting it using $gamlssML()$, the $getOrder()$ function is used to order the distributions by AIC value. The histogram of the $\hat{\alpha}_1$ values can be found in Figure 8. Overall, there are 1284 $\hat{\alpha}_1$ values. The top three distributions are as follows: the Skew t Azzalini type 1 distribution (ST1) has AIC value -5320.814 and the worm plot is shown in Figure 9, the Skew t type 5 distribution (ST5) has AIC value -5320.810 and the worm plot is shown in Figure 10, and the Johnson SU original distribution has AIC value -5320.435 and the worm plot is shown in Figure 11.

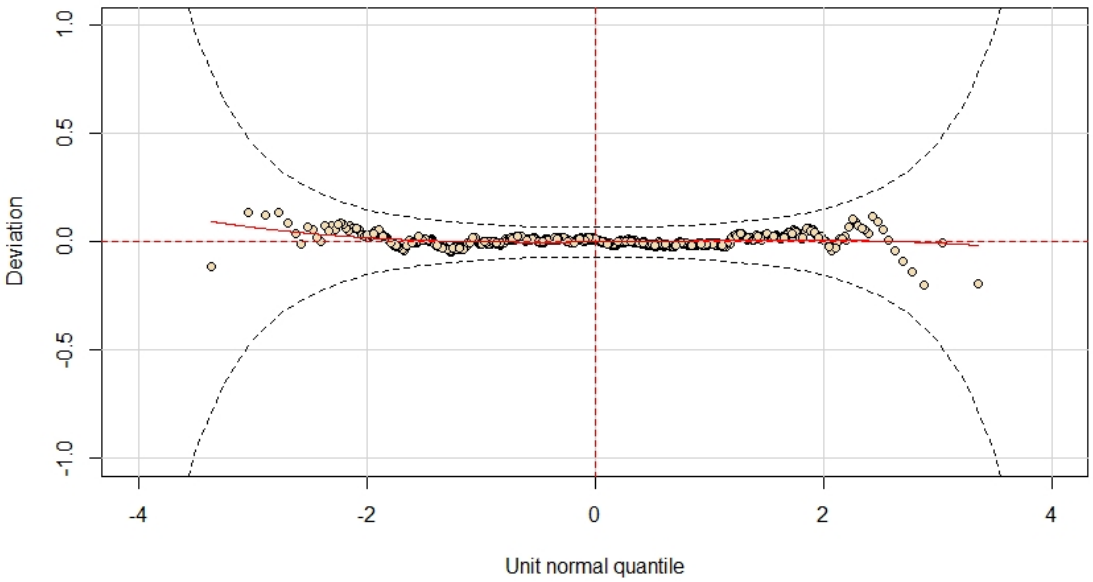

Figure 9: Worm Plot for the m_1 ST1 model

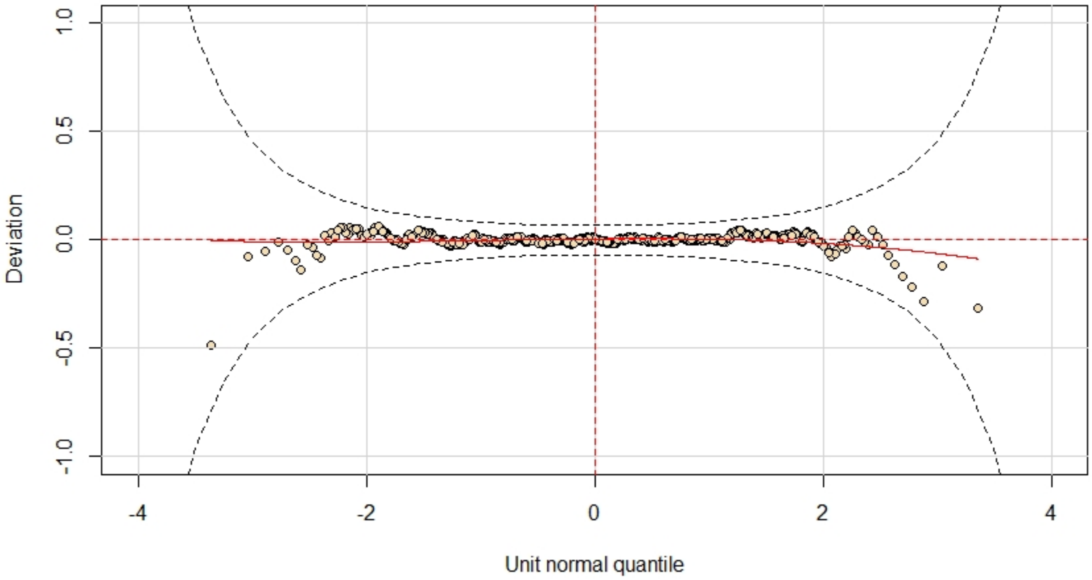

Figure 10: Worm Plot for the m_1 ST5 model

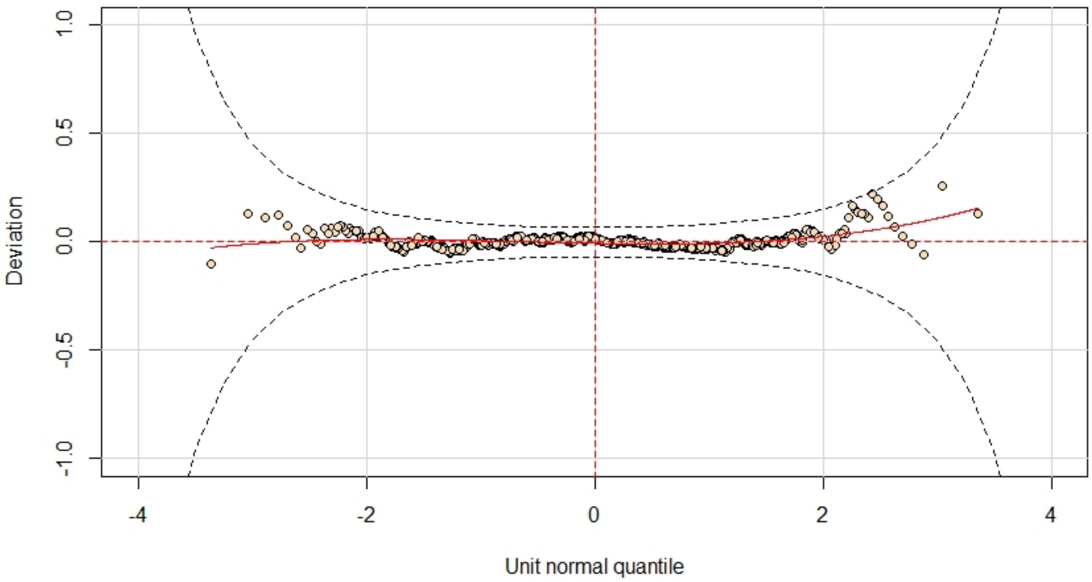

Figure 11: Worm Plot for the m_1 JSUo model

After examining all three worm plots and comparing the AIC values for each of the distributions, the ST1 distribution is chosen for the $\hat{\alpha}_1$ values. The ST1 distribution shows the least deviance in the observations in the worm plot, which makes it a suitable distribution along with the the model having the best overall AIC value [24]. The fitted model is discussed in Section 4.2.

3.3 The m_2 Model with $s_0 = 5000$

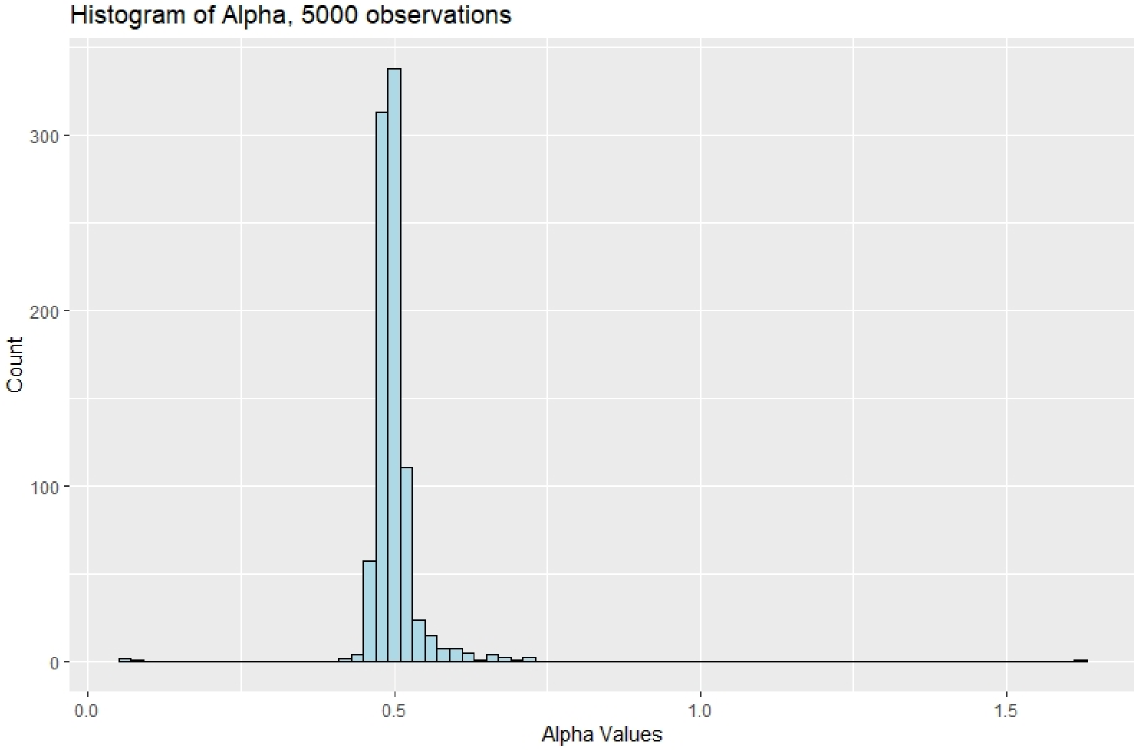

Figure 12: Histogram of the $\hat{\alpha}_2$ values for $s_o = 5000$ subset

The m_2 model is fit using the $\hat{\alpha}_2$ vector with $s_o = 5000$, thus making $s_n = 905$. The histogram of the $\hat{\alpha}_2$ values can be found in Figure 12. Overall, there are 905 $\hat{\alpha}$. The $getOrder()$ function returned the Skew t Azzalini type 2 (ST2) distribution as the distribution with the lowest AIC value, -4169.4904. The next five distributions fit for this distribution are all members of the Skew t family of distributions, including ST5, ST3, SST, and ST1. The worm plot for the m_2 model fit with the ST2 distribution is found in Figure 13. The AIC values for the ST5 distribution and ST3 distribution are -4167.6077 and -4164.1521, and the worm plots are found in Figure 14 and Figure 15, respectively.

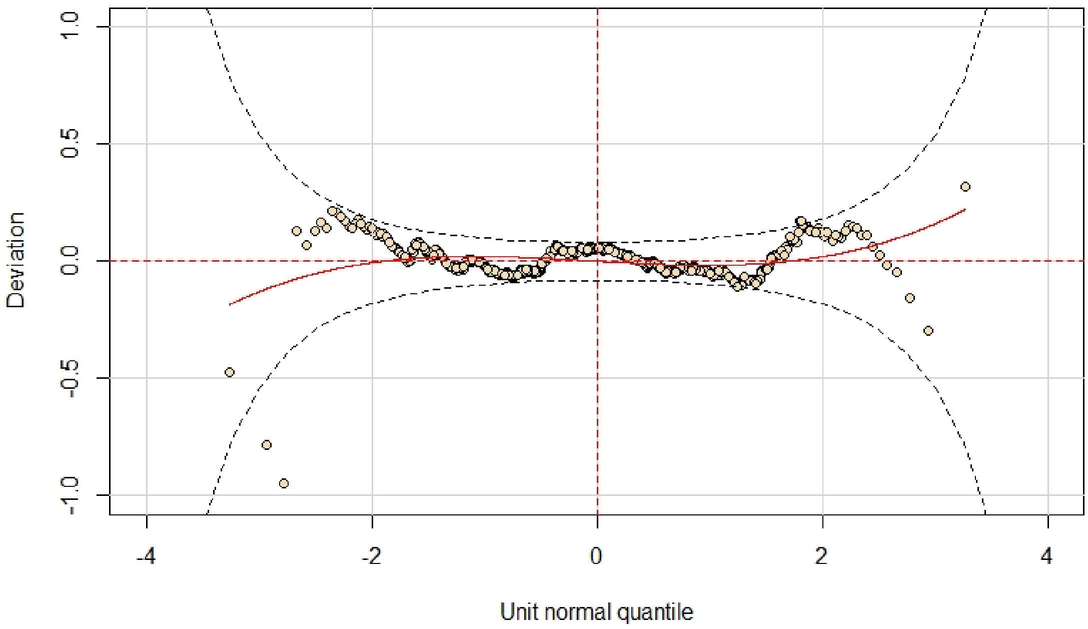

Figure 13: Worm Plot for the m_2 ST2 model

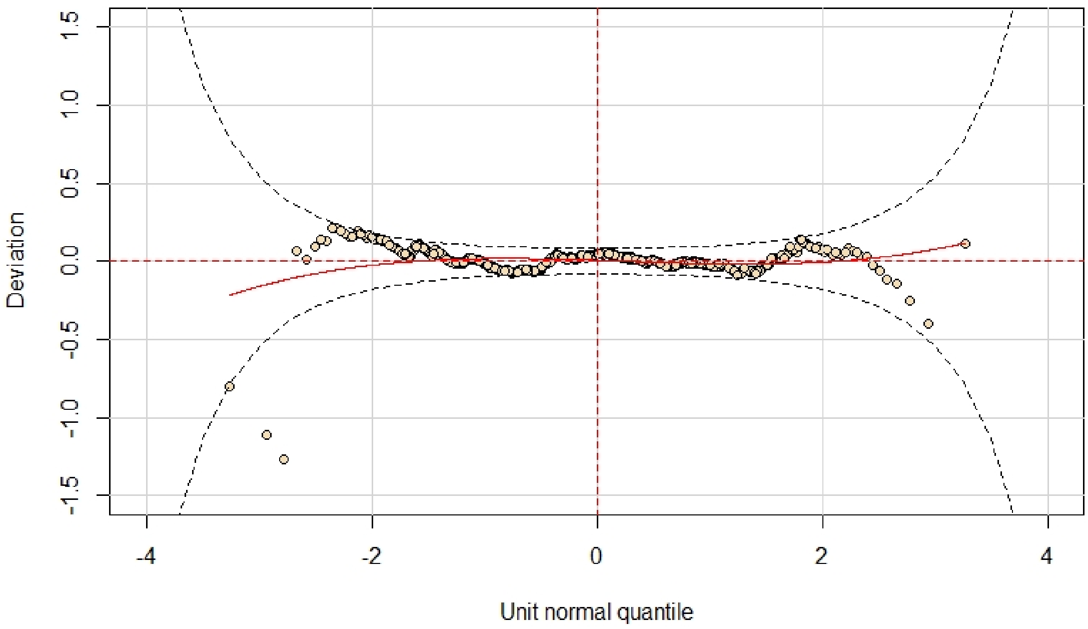

Figure 14: Worm Plot for the m_2 ST5 model

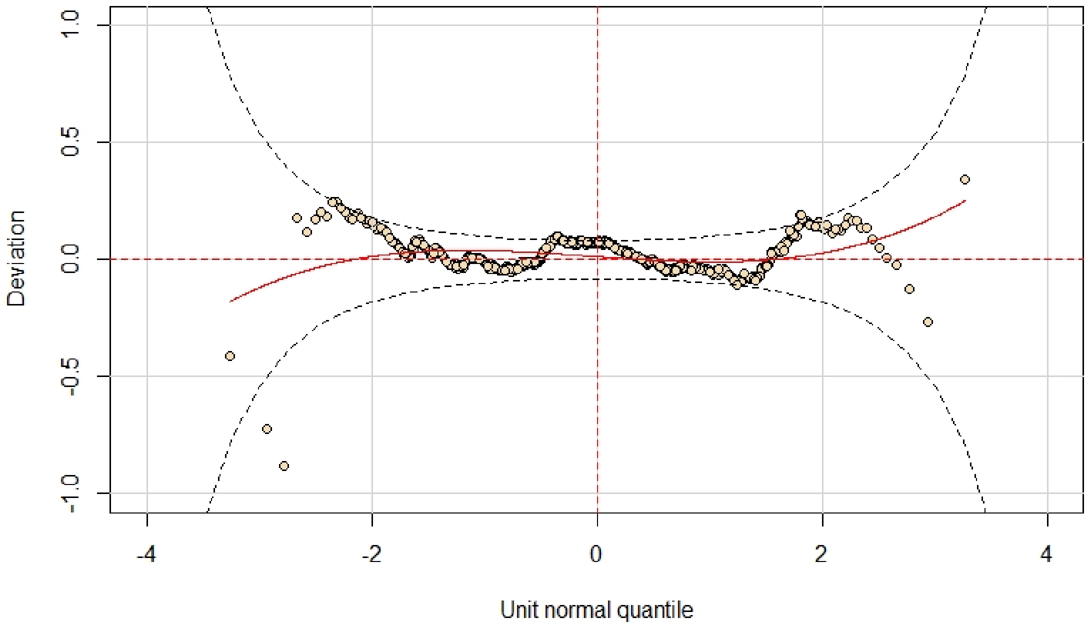

Figure 15: Worm Plot for the m_2 ST3 model

The worm plots for ST5 and ST3 show more deviance of the residuals for the respective models. The y-axis in Figure 14 had to be extended in order to capture the deviating observations. Thus, The ST2 distribution is chosen to represent the $\hat{\alpha}_2$ set of values based on the AIC value and the worm plot for the model.

3.4 The m_3 Model with $s_0 = 10000$

The m_3 model is fit using the partition of the data set with $s_o = 10000$, thus making $s_n = 270$. The histogram of the $\hat{\alpha}_3$ values can be found in Figure 16. Overall, there are 270 $\hat{\alpha}_3$ values. The $getOrder()$ function returned the Skew t Azzalini type 2 as the distribution with the lowest AIC value, -1020.9639, followed by the Skew t type 3 with an AIC value -1020.84, and by Skew Student t with an AIC value -1020.8436. The worm plots for the the ST2, the ST3, and the SST fits are given in Figure 17, Figure 18, and Figure 19, respectively.

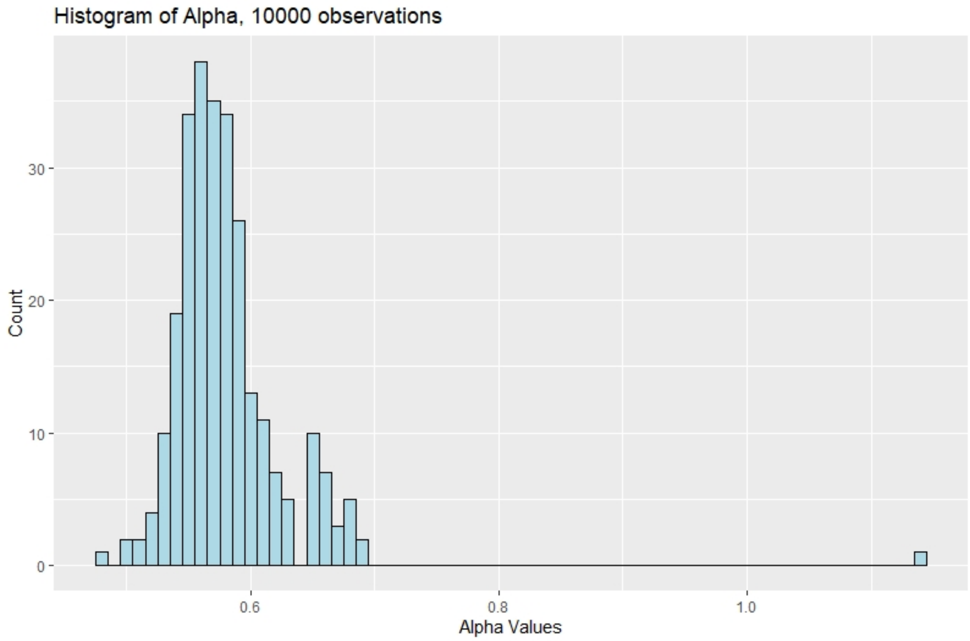

Figure 16: Histogram of the $\hat{\alpha}_3$ values for $s_o = 10000$ subset

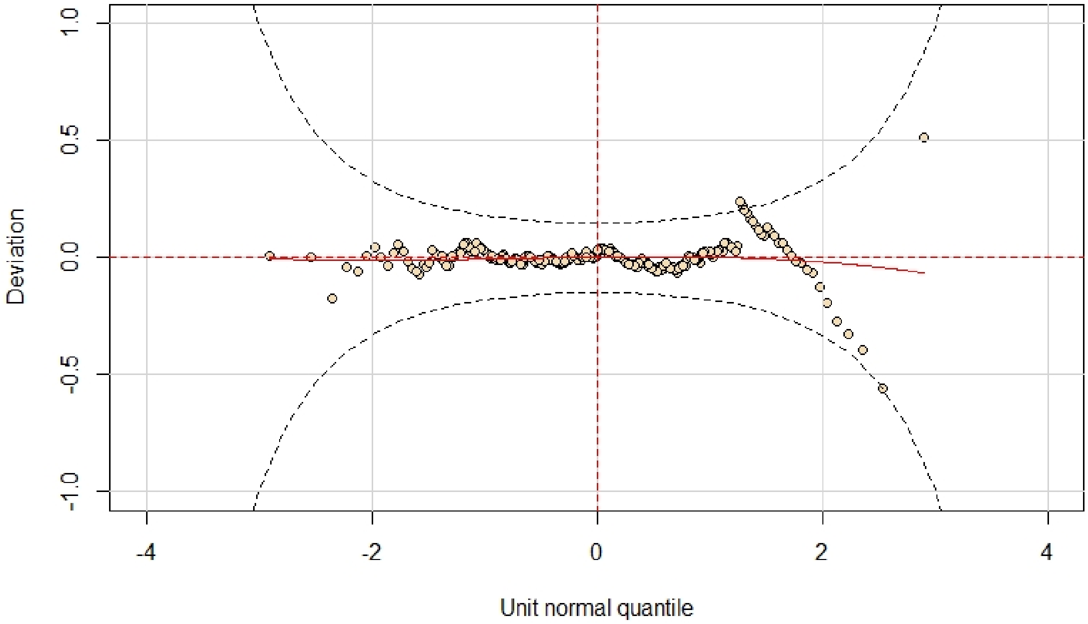

Figure 17: Worm Plot for the m_3 ST2 model

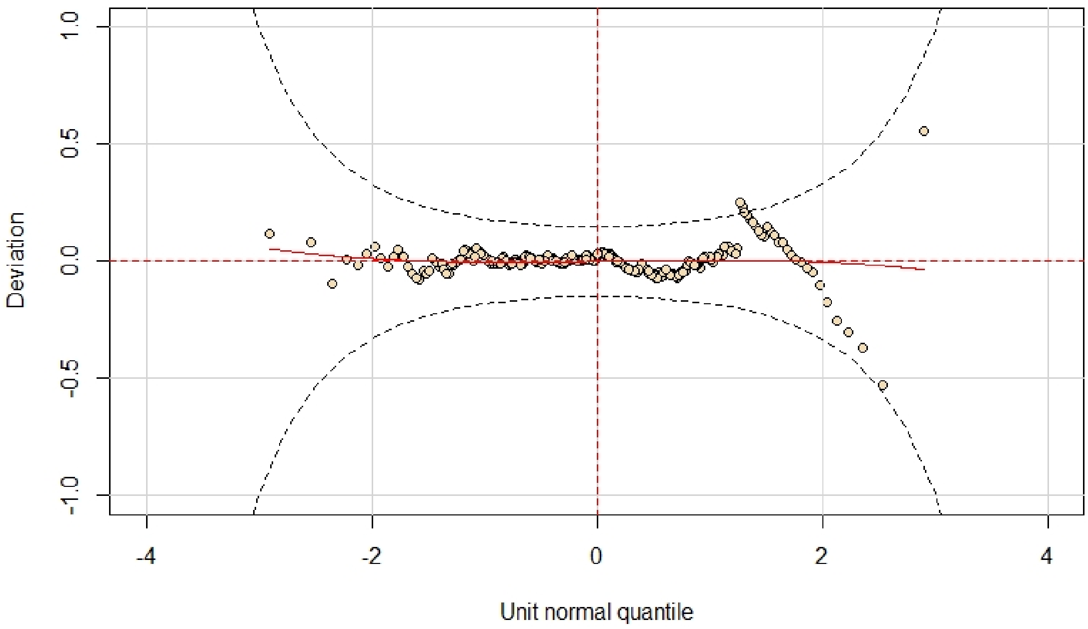

Figure 18: Worm Plot for the m_3 ST3 model

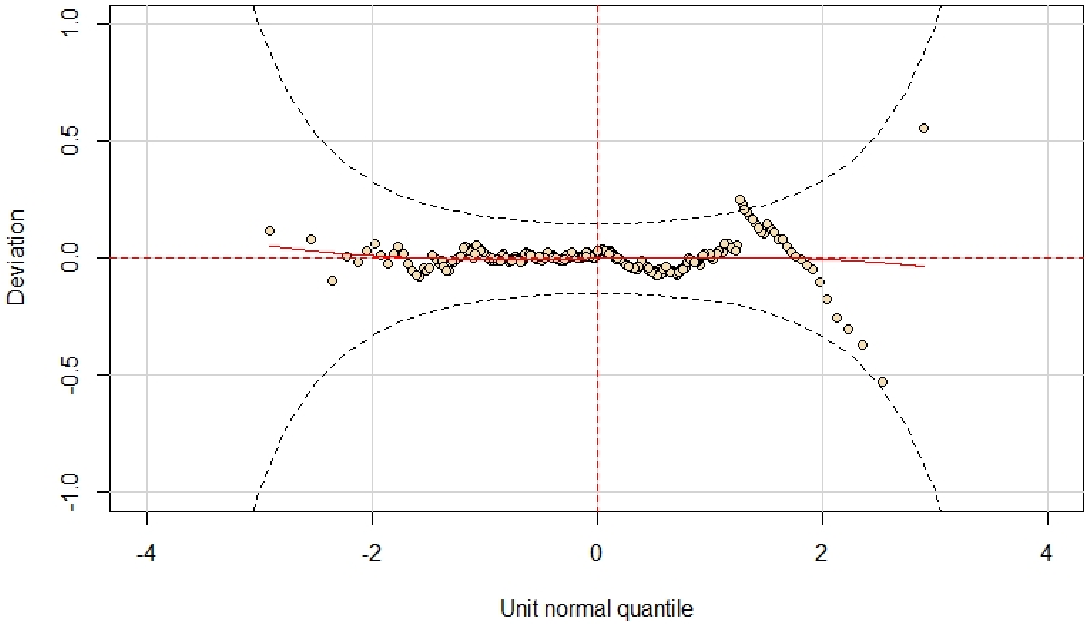

Figure 19: Worm Plot for the m_3 SST model

While the differences are subtle, there is more deviation in the ST3 worm plot than in the ST2 worm plot, and likewise for SST. Again, the SST and ST3 distributions are closely related, and their fits are virtually identical [24] [10]. Analyzing these plots, and the respective AIC values leads the ST2 distribution to be chosen for the $\hat{\alpha}_3$ values.

3.5 Discovering a Final Model Overall

Although there is a decrease in the sample size, s_n, for the $\hat{\alpha}_2$ and $\hat{\alpha}_3$ subsets, the fitting procedures still lead to a distribution within the Skew t family, the ST2 distribution. The significance of this pertains to the robustness of the ST2 distribution and its application to larger sample sizes. It is possible that the ST2 distribution could be a proper distribution to model α for each set of data. In order to investigate this, the *getOrder*() function was used on the m_0 and m_1 *GAMLSS* model objects; the AIC value for the ST2 distribution is within 8.462 units of the top distribution for the m_0 model and within 1.409 units of the top distribution for the m_1 model.

If the discussion divulges into choosing a final model for α overall, regardless of sample size, then the ST2 distribution seems a suitable candidate. Thus, worm plots for both the $\hat{\alpha}_0$ and $\hat{\alpha}_1$ fit with the ST2 distribution for the purpose of establishing a definitive final distribution. These are found in Figure 20 and Figure 21.

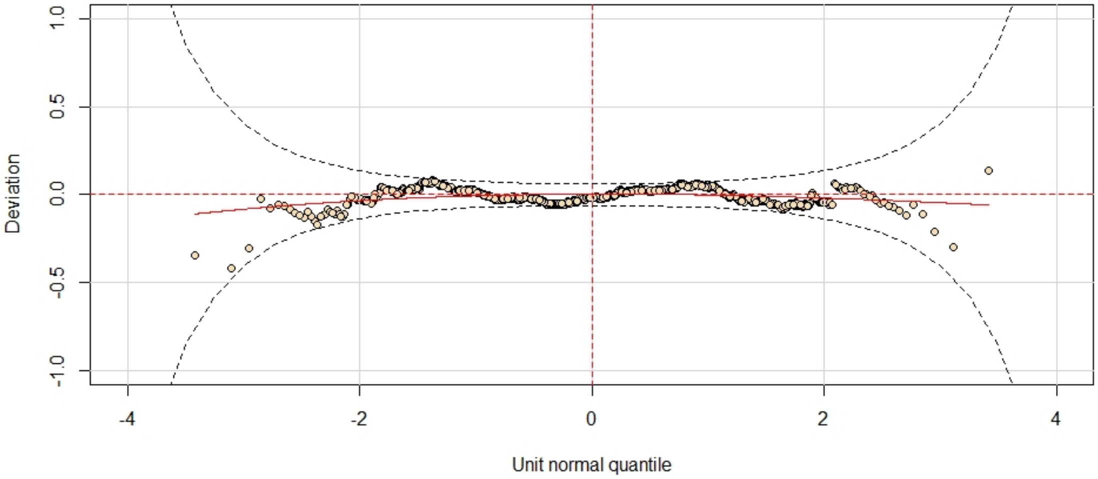

Figure 20: Worm plot for the $\hat{\alpha}_1$ fitted with the ST2 distribution

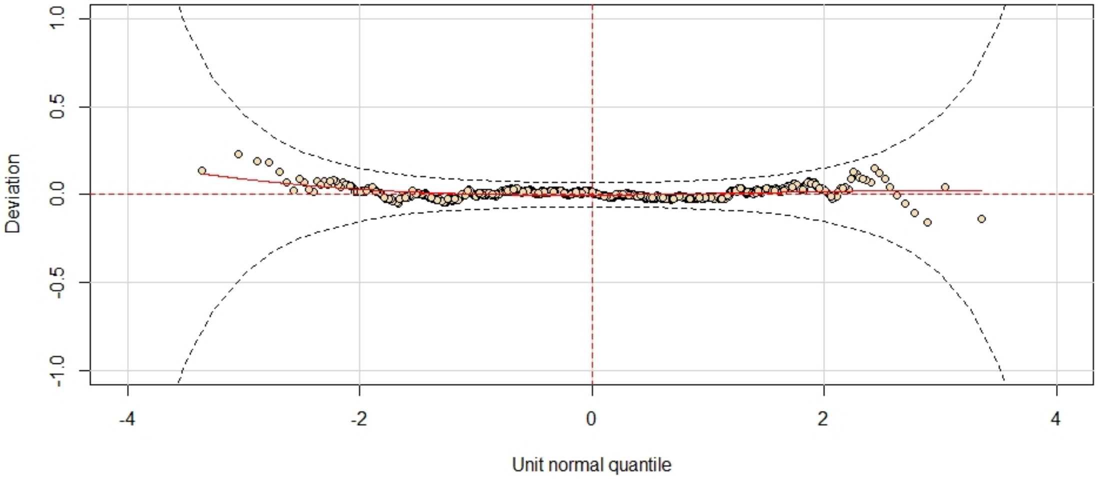

Figure 21: Worm plot for the $\hat{\alpha}_2$ fitted with the ST2 distribution

The model selction procedures in Sections 3.1 and 3.2 found other distributions for m_0 and m_1, namely the ST3 and ST1, that are better suited for the data strictly based on AIC values and worm plots. However, the worm plots hold up well compared to the plots presented in each of the models' selection procedures. If all four subsets are considered to be representative of the parameter α, then the ST2 distribution fit over a data set with a representative s_n value does seem to be a robust candidate for modeling the tail index.

4 OVERVIEW OF THE SELECTION RESULTS AND FITTED MODELS

The model selection process produced four models, one for each set of data. All four of the distributions selected, the Skew t type 3 (ST3) for the m_0 model, the Skew t Azzalini type 1 (ST1) for the m_1 model, Skew t Azzalini type 2 (ST2) for the m_2 model, and the Skew t Azzalini type 2 (ST2) for m_3 model, are from the same family of distribution, the Skew t family of distributions [24]. The model selection process also singled out the ST2 distribution as a candidate robust enough to model α overall. The Skew t Azzalini type 2 was developed by Azzanlini and Capitanio [4]. The Skew t type 3 is based on the work of Fernandez and Steel [10]. The Skew t Azzalini type 1 was developed by Azzalini [3]. The table below, Table 1, summarizes the model selection results. What follows is a discussion of each fitted model, along with summaries of the parameters and probability density functions.

Table 1: Overview of Model s_n and AIC Values

Model	$\mathbf{s_n}$ **value**	**Distribution**	**AIC**
m_0	1642	ST3	-5512.18
m_1	1284	ST1	-5320.814
m_2	905	ST2	-4169.4904
m_3	270	ST2	-1020.8436

4.1 The m_0 Fitted Model

First, the m_0 model which is based on the varying s_0 values for each stock. The model selection procedures settled on the Skew t type 3. The ST3 distribution was chosen over the Skew Exponential type 2 due to the worm plot, since the AIC values of the two distributions were very close. The fitted distribution is below in Figure 22.

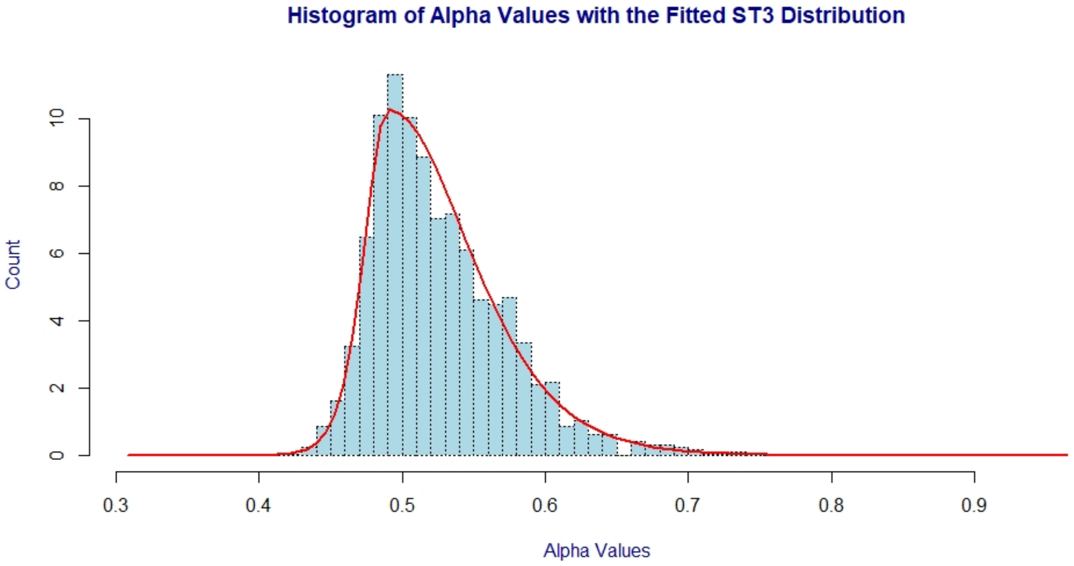

Figure 22: Histogram of $\hat{\alpha}_0$ values with the fitted ST3 distribution

The ST3 distribution is fit to the histogram using parameters found using maximum likelihood estimation within the *GAMLSS* package [25]. The Skew t type 3 is described as a 'scale-spliced' distribution [10]. Splicing is used to create skewness inside of a symmetric distribution [24]. The ST3 distribution is defined by four parameters:

1. μ, the location shift parameter, which can take any value $-\infty$ to ∞.

2. σ, the scaling parameter, which can take any value 0 to ∞.

3. ν, the skewness parameter, which can take any value 0 to ∞.

4. τ, the kurtosis parameter, which can take any value 0 to ∞.

The probability density function that defines the ST3 distribution is as follows:

$$f_Y(y|\mu,\sigma,\nu,\tau) = \left\{ \begin{array}{l} \frac{c}{\sigma}(1+\frac{\nu^2 z^2}{\tau})^{-(\tau+1)/2} \text{ if } y < \mu \\ \frac{c}{\sigma}(1+\frac{z^2}{\nu^2 z^2})^{-(\tau+1)/2} \text{ if } y \geq \mu \end{array} \right\}, \tag{2}$$

where y represents the observations in the data. Also, $z = (y-\mu)/\sigma$, and $c = 2\nu[(1+\nu^2)B(1/2,\tau/2)\tau^{1/2}]^{-1}$ [10]. Here, B represents the use of the gamma distribution inside the PDF of ST3 [10] [24]. Thus, numerical approximation, as opposed to the use of analytical methods, is the standard approach to estimate parameters in the PDF of the ST3 distribution, just as the MLE example in Section 1.6 [24]. When the *GAMLSS* procedure estimates the model parameters, the fitting method uses link functions for each of the parameters to ensure that the parameters remain in the respective intervals [24] [25]. For instance, the log-link function is used for σ in the ST3 distribution to ensure that the σ parameter is greater than 0 [24]. The link functions for the ST3 distribution are given in Table 2. Importantly, ν and τ are used for modeling aspects of the kurtosis of the distribution in question [24].

Table 2: Link Functions for the ST3 Distribution

Parameter	Estimate
μ	Identity
σ	Log Link
ν	Log Link
τ	Log Link

The summary output for the fitted ST3 distribution, the m_0 model, is found in Table 3.

The summary output is devised into the following columns from left to right: the parameter in question, the estimate according to the link function used for that parameter, the standard error for calculating that estimate, the t test statistic for the hypothesis test, and the p-value for the hypothesis test [24]. The parameters

Table 3: Summary Output of the Fitted ST3 Distribution

Parameter	Estimate	Standard Error	t-value	p-value
η_μ	0.49013525	0.00180533	271.4934	0
η_σ	-3.46108206	.03809905	-90.8443	0
η_ν	0.61461599	0.04022275	15.2803	0
η_τ	2.00311813	0.17698011	11.3183	0

are represented as $\hat{\eta}_\mu$, $\hat{\eta}_\sigma$, $\hat{\eta}_\nu$, and $\hat{\eta}_\tau$, and to clarify these are parameter estimates determined by the respective link function [29]. The hypothesis test for each output, the m_0 model and the others, is the asymptotic Wald test, which tests the following hypotheses:

1. $H_0 : \eta_\mu = \mu = 0$, $H_1 : \eta_\mu \neq 0$
2. $H_0 : \eta_\sigma = 0$, $H_1 : \eta_\sigma \neq 0$
3. $H_0 : \eta_\nu = 0$, $H_1 : \eta_\nu \neq 0$
4. $H_0 : \eta_\tau = 0$, $H_1 : \eta_\tau \neq 0$

According to the $GAMLSS$ manual and creators, the t tests above are dubious when fitting distributions to data, and instead, the Generalized Likelihood Ratio (GLR) test is a better indicator of parameter significance overall [24] [29]. The GLR test was performed using the symmetric generalized t (GT) distribution, since the Skew t distributions are extensions upon the symmetric generalized t distribution [29]. The null hypothesis for the GLR test is the GT distribution and the alternative hypothesis is the ST3 distribution [24]. Thus, for the m_0 model, the GLR test returned a test statistic of 187.841, which is greater than 3.84 as described in Section 1.6, confirming parameter significance and confirming that it is appropriate to use an extension of the GT distribution [24].

Thus, the fitted parameters, and 95% confidence intervals for each parameter, for the ST3 distribution are below. These parameters are transformed via the respective link functions, along with the confidence interval endpoints [24]. The confidence intervals are calculated using the standardized 95% confidence Z value, 1.96 and the form:

$$\text{estimate} \pm 1.96 \times SE_{estimate}$$

1. $\hat{\mu} = 0.49013525$, with a 95% confidence interval $[0.4865968, 0.4936737]$

2. $\hat{\sigma} = 0.03139577$, with a 95% confidence interval $[0.02913672, 0.03382998]$

3. $\hat{\nu} = 1.848946$, with a 95% confidence interval $[1.70878, 2.000611]$

4. $\hat{\tau} = 7.412132$ with a 95% confidence interval $[5.239557, 10.48546]$

Therefore, the fitted probability density function for the m_0 model can be written in the following way:

$$f_Y(y|\mu,\sigma,\nu,\tau) = \left\{ \begin{array}{ll} \dfrac{c}{0.03139577}(1+\dfrac{1.848946^2 z^2}{7.412132})^{-(7.412132+1)/2} & \text{if } y < 0.49013525 \\ \dfrac{c}{0.03139577}(1+\dfrac{z^2}{1.848946^2 z^2})^{-(7.412132+1)/2} & \text{if } y \geq 0.49013525 \end{array} \right\}, \tag{3}$$

where z and c are equal to the following values:

$$z = (y - 0.49013525)/0.03139577$$

$$c = 2 * 1.848946 * [(1 + 1.848946^2)B(1/2, 7.412132/2)7.412132^{1/2}]^{-1}$$

[10].

4.2 The m_1 Fitted Model

The m_1 model was determined to be fit with the Skew t Azzalini type 1 distribution. The fitted distribution is given in Figure 23. The m_1 model is based on 2500 observations for each stock with an $s_n = 1284$.

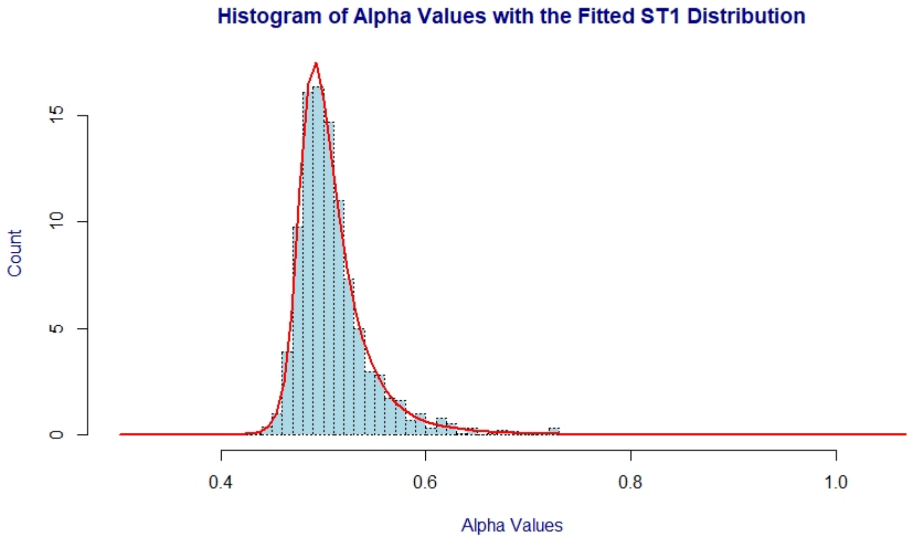

Figure 23: Histogram of $\hat{\alpha}_1$ values with the fitted ST1 distribution

The ST1 distribution is defined by four parameters [3]:

1. μ, the location shift parameter, which can take any value $-\infty$ to ∞.

2. σ, the scaling parameter, which can take any value greater than 0 and to ∞.

3. ν, the skewness parameter, which can take any value $-\infty$ to ∞.

4. τ, the kurtosis parameter, which can take any value greater than 0 and to ∞.

The probability density function that defines the ST1 distribution is as follows:

$$f_Y(y|\mu,\sigma,\nu,\tau) = \frac{2}{\sigma} f_{Z_1}(z) F_{Z_1}(\nu z), \tag{4}$$

where y is the observations of the data and $z = (y - \mu)\sigma$ [24]. In addition, f_{Z_1} and F_{Z_1} are the probability density function and the cumulative density function of $Z_1 \sim TF(0, 1, \tau) = t_\tau$, the t distribution with $\tau > 0$ degrees of freedom, and τ is a continuous parameter [24] [3].

The link functions for the ST1 distribution are given below in Table 4.

Table 4: Link Functions for the ST1 Distribution

Parameter	Estimate
μ	Identity
σ	Log Link
ν	Identity
τ	Log Link

The summary output for the fitted ST1 distribution, the m_1 model, is found in Table 5.

Table 5: Summary Output of the Fitted ST1 Distribution

Parameter	Estimate	Standard Error	t-value	p-value
η_μ	0.47743006	0.00149467	319.42231	0
η_σ	-3.46196753	0.06616602	-52.32244	0
η_ν	2.93949024	0.45788404	6.41973	0
η_τ	1.05352688	0.10474601	10.05792	0

The GLR test was performed with the null hypothesis as the GT distribution, which is fit on the data using the same *GAMLSS* procedure, and the alternative hypothesis as the ST1 distribution, and returned a test statistic of 274.958 which is greater than 3.84 [24]. The GLR test confirms the signficance of using the extension of the generalized t distribution, the Skew t type 1, found via *GAMLSS* [24].

Now, the fitted parameters, and 95% confidence intervals for each parameter in the ST1 distribution are below. Just as Section 4.1, the parameters are transformed via the respective link functions, along with the confidence interval endpoints [24].

1. $\hat{\mu} = 0.47743006$, with a 95% confidence interval $[0.4745005, 0.4803596]$
2. $\hat{\sigma} = 0.03136798$, with a 95% confidence interval $[0.02755275, 0.03571152]$
3. $\hat{\nu} = 2.93949024$, with a 95% confidence interval $[2.042038, 3.836943]$
4. $\hat{\tau} = 2.867748$ with a 95% confidence interval $[2.335497, 3.521296]$

Therefore, the fitted probability density function for the m_1 model is written in the following way:

$$f_Y(y|\mu, \sigma, \nu, \tau) = \frac{2}{0.03136798} f_{Z_1}(z) F_{Z_1}(2.93949024z), \tag{5}$$

where $z = (y - 0.47743006)/0.03136798$, and $Z_1 \sim TF(0, 1, 2.867748) = t_{2.867748}$

4.3 The m_2 Fitted Model

The m_2 model, with $s_0 = 5000$ and $s_n = 905$, was determined to be fit using the Skew t Azzalini type 2 distribution. The fitted histogram is found in Figure 24.

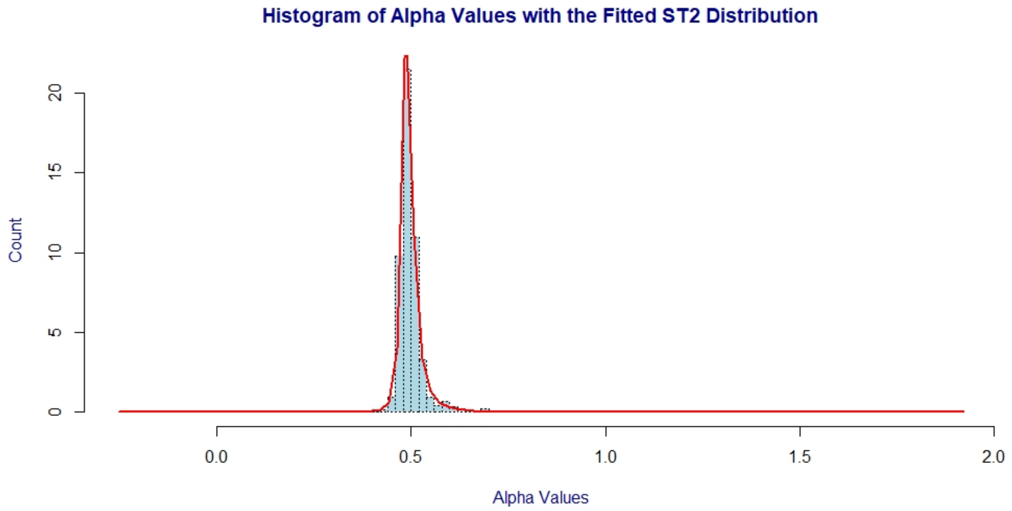

Figure 24: Histogram of $\hat{\alpha}_2$ values with the fitted ST2 distribution

The ST2 distribution is defined by four parameters [4].

1. μ, the location shift parameter, which can take any value $-\infty$ to ∞.

2. σ, the scaling parameter, which can take any value greater than 0 and to ∞.

3. ν, the skewness parameter, which can take any value $-\infty$ to ∞.

4. τ, the kurtosis parameter, which can take any value greater than 0 and to ∞.

The probability density function of the ST2 distribution is given in Equation 6 [4].

$$f_Y(y|\mu,\sigma,\nu,\tau) = \frac{2}{\sigma} f_{Z_1}(z) F_{Z_2}(\omega), \tag{6}$$

where y is the data and $z = (y-\mu)\sigma$, $\omega = \nu\lambda^{1/2}z$, and $\lambda = (\tau+1)/(\tau+z^2)$ [24]. Also, $f_{Z_1}(.)$ is the PDF of $Z_1 \sim TF(0,1,\tau) = t_\tau$, and $F_{Z_2}(.)$ is the CDF of $Z_2 \sim TF(0,1,\tau+1) = t_{\tau+1}$ [24] [4].

The link functions for fitting the ST2 distribution are given in Table 6.

Table 6: Link Functions for the ST2 Distribution

Parameter	Estimate
μ	Identity
σ	Log Link
ν	Identity
τ	Log Link

The summary output for the fitted ST2 distribution, the m_2 model, is found in Table 10.

Table 7: Summary Output of the Fitted ST2 Distribution

Parameter	Estimate	Standard Error	t-value	p-value
η_μ	0.4812267	0.0015055	319.64628	0
η_σ	-4.0522605	0.0682193	-59.40048	0
η_ν	1.1554209	0.2057751	5.61497	0
η_τ	0.7285679	0.0802370	9.08020	0

The GLR test was performed with the null hypothesis as the GT distribution and the alternative hypothesis as the ST2 distribution. The GLR test returned a test statistic of 41.04102 which is greater than 3.84, confirming the significance of the ST2 parameters [24]. The fitted parameters and the 95% SE-based confidence intervals for each parameter in the fitted ST2 distribution are below. These parameters are transformed via their respective link functions.

1. $\hat{\mu} = 0.4812267$, with a 95% confidence interval $[0.4782759, 0.4841775]$

2. $\hat{\sigma} = 0.01738304$, with a 95% confidence interval $[0.01520745, 0.01986987]$

3. $\hat{\nu} = 1.1554209$, with a 95% confidence interval $[0.7521017, 1.55874]$

4. $\hat{\tau} = 2.072111$ with a 95% confidence interval $[1.770573, 2.425002]$

Therefore, the fitted probability density function for the ST2 distribution is written as in equation 7.

$$f_Y(y|\mu,\sigma,\nu,\tau) = \frac{2}{0.01738304} f_{Z_1}(z) F_{Z_2}(\omega), \tag{7}$$

where y is the data and $z = (y - 0.4812267)/0.01738304$, $\omega = 1.1554209\lambda^{1/2}z$, and $\lambda = (3.072111)/(2.072111 + z^2)$ [24]. Also, $f_{Z_1}(.)$ is the PDF of $Z_1 \sim TF(0, 1, 2.072111) = t_{2.072111}$, and $F_{Z_2}(.)$ is the CDF of $Z_2 \sim TF(0, 1, 3.072111) = t_{3.072111}$ [24] [4].

4.4 The m_3 Fitted Model

The m_3 model, with $s_0 = 10000$ and $s_n = 270$ was also determined to be fit using the Skew t Azzalini type 2 distribution. The fitted histogram is found in Figure 25.

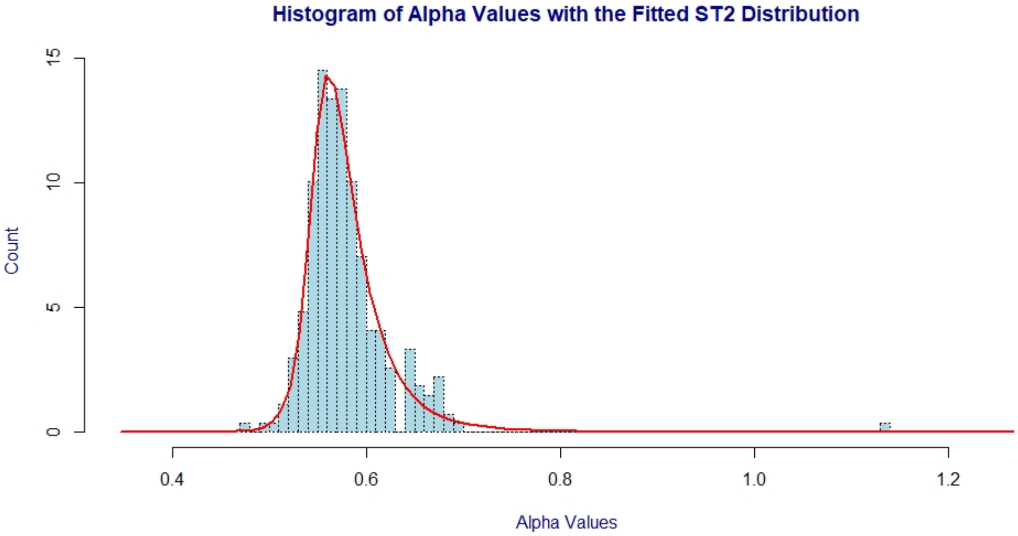

Figure 25: Histogram of $\hat{\alpha}_3$ values with the Fitted ST2 distribution

The parameters are the same as described in Section 4.3. The probability density function is the same as in equation 6. The link functions are also represented in

Section 4.3 Table 6. Thus, the summary output of the fitted ST2 distribution is given in Table 8.

Table 8: Summary Output of the Fitted ST2 Distribution

Parameter	Estimate	Standard Error	t-value	p-value
η_μ	0.54448659	0.00358428	151.90940	0
η_σ	-3.29159559	0.12303963	-26.75232	0
η_ν	2.08117642	0.53352565	3.90080	0
η_τ	1.19147471	0.23322566	5.10868	0

Again, the GLR test was performed with the null hypothesis as the GT distribution and the alternative hypothesis as the ST2 distribution. The GLR test returned a test statistic of 30.75573 which is greater than 3.84, confirming the significance of the ST2 parameters [24]. The fitted parameters and the 95% SE-based confidence intervals for each parameter in the fitted ST2 distribution are below. These parameters are transformed via their respective link functions.

1. $\hat{\mu} = 0.54448659$, with a 95% confidence interval $[0.5374614, 0.5515118]$

2. $\hat{\sigma} = 0.03719445$, with a 95% confidence interval $[0.02922435, 0.04733819]$

3. $\hat{\nu} = 2.08117642$, with a 95% confidence interval $[1.035466, 3.126887]$

4. $\hat{\tau} = 3.291932$ with a 95% confidence interval $[2.084074, 5.19968]$

Therefore, the fitted probability density function for the ST2 distribution in the m_3 model is written as in equation 8.

$$f_Y(y|\mu, \sigma, \nu, \tau) = \frac{2}{0.03719445} f_{Z_1}(z) F_{Z_2}(\omega), \tag{8}$$

where y is the data and $z = (y - 0.54448659)/0.03719445$, $\omega = 2.08117642\lambda^{1/2}z$, and $\lambda = (4.291932)/(3.291932{+}z^2)$ [24]. Also, $f_{Z_1}(.)$ is the PDF of $Z_1 \sim TF(0, 1, 3.291932) = t_{3.291932}$, and $F_{Z_2}(.)$ is the CDF of $Z_2 \sim TF(0, 1, 4.291932) = t_{4.291932}$ [24] [4].

4.5 Comparing Different Fitted m_0 and m_1 Models

In Section 3.5, the ST2 distribution is floated as a possible candidate for α overall due to handling smaller sample sizes and sharing a family with the larger s_n models. The m_0 and m_1 models are fitted with the ST3 and ST1 distributions, while both the m_2 and m_3 models are fit with the ST2 distribution. Thus, there is an interest from Section 3.5 in evaluating the fit of the ST2 distribution on both the $\hat{\alpha}_0$ and $\hat{\alpha}_1$ vectors. These new models were compared to the original m_0 and m_1 models using the Generalized Likelihood Ratio (GLR) test. There are two different GLR test results below, one for the $\hat{\alpha}_0$ vector and one for $\hat{\alpha}_1$ vector.

The $\hat{\alpha}_0$ vector of values returned a test statistic of 6.874987 when the ST2 model is the null hypothesis and the ST3 model is the alternative hypothesis. Since this test statistic is larger than the baseline $\chi^2 = 3.84$, the ST3 distribution still seems appropriate for the $\hat{\alpha}_0$ vector.

The $\hat{\alpha}_1$ vector of values returned a GLR test statistic of 1.408177 when the ST2 model is the null hypothesis and the ST3 model is the alternative hypothesis. The test statistic fails to cross the baseline $\chi^2 = 3.84$, determining that the parameters are not different from one another according to the GLR test [24]. Thus, the ST2 distribution may be appropriate for modeling the $\hat{\alpha}_1$ vector.

The histograms of both the $\hat{\alpha}_0$ and $\hat{\alpha}_1$ values fit with the ST2 distribution are given below in Figure 26 and Figure 27.

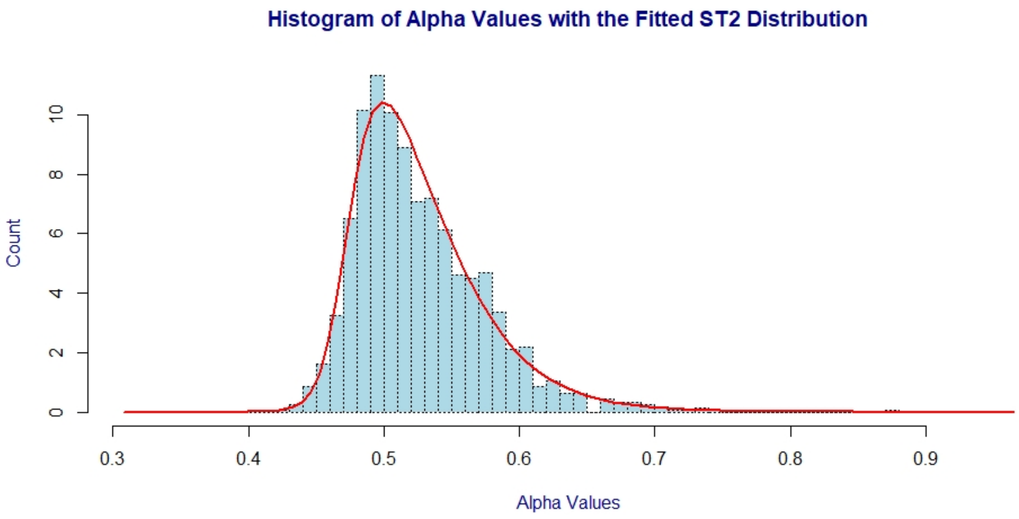

Figure 26: Histogram of $\hat{\alpha}_0$ values with the Fitted ST2 distribution

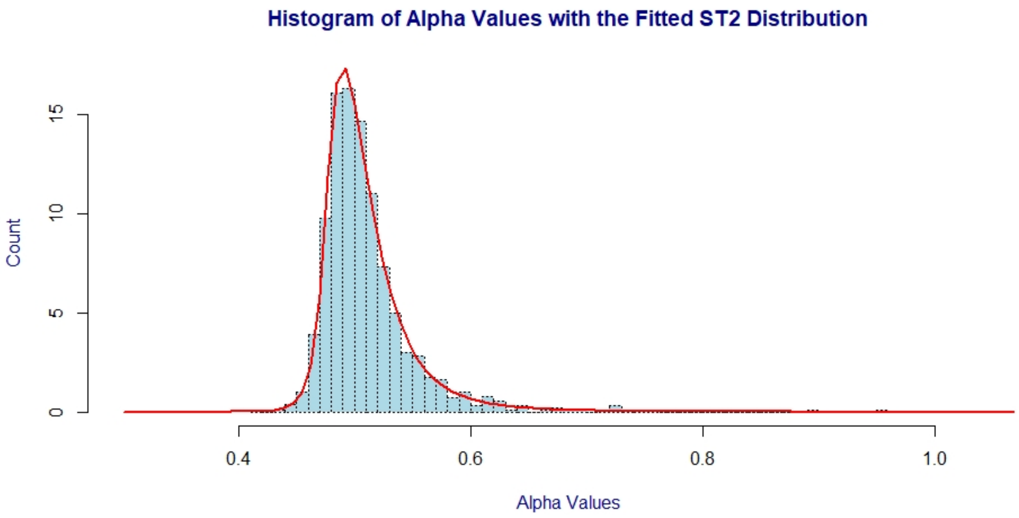

Figure 27: Histogram of $\hat{\alpha}_1$ values with the Fitted ST2 distribution

The summary output of each fitted model are given below.

Table 9: Summary Output of the $\hat{\alpha}_0$ Fitted ST2 Distribution

Parameter	Estimate	Standard Error	t-value	p-value
η_μ	0.47337390	0.00146635	322.8243	0
η_σ	-2.76784764	0.03976934	-69.5975	0
η_ν	3.69041529	0.33726875	10.9421	0
η_τ	2.12429935	0.20668927	10.2777	0

Table 10: Summary Output of the $\hat{\alpha}_1$ Fitted ST2 Distribution

Parameter	Estimate	Standard Error	t-value	p-value
η_μ	0.4766338	0.0012223	389.94717	0
η_σ	-3.4250096	0.0540141	-63.40955	0
η_ν	2.7502377	0.3264104	8.42571	0
η_τ	1.1065670	0.966680	11.44708	0

Using the link functions, the estimates and the 95% confidence intervals for the respective models are as follows:

For the $\hat{\alpha}_1$ ST2 model.

1. $\hat{\mu} = 0.47337390$, with a 95% confidence interval $[0.4704999, 0.4762479]$

2. $\hat{\sigma} = 0.06279702$, with a 95% confidence interval $[0.05808801, 0.06788773]$

3. $\hat{\nu} = 3.69041529$, with a 95% confidence interval $[3.029369, 4.351462]$

4. $\hat{\tau} = 2.12429935$ with a 95% confidence interval $[1.719188, 2.52941]$

For the $\hat{\alpha}_2$ ST2 model.

1. $\hat{\mu} = 0.4766338$, with a 95% confidence interval $[0.4742381, 0.4790295]$

2. $\hat{\sigma} = 0.03254897$, with a 95% confidence interval $[0.02927923, 0.03618386]$

3. $\hat{\nu} = 2.7502377$, with a 95% confidence interval $[2.110473, 3.390002]$

4. $\hat{\tau} = 3.023959$ with a 95% confidence interval $[2.502018, 3.65478]$

The fitted PDFs for each model, the $\hat{\alpha}_0$ ST2 model and the $\hat{\alpha}_1$ ST2 model are below in equation 9 and 10, respectively.

For the $\hat{\alpha}_0$ ST2 model.

$$f_Y(y|\mu, \sigma, \nu, \tau) = \frac{2}{0.06279702} f_{Z_1}(z) F_{Z_2}(\omega), \tag{9}$$

where y is the data and $z = (y - 0.47337390)/0.06279702$, $\omega = 3.69041529\lambda^{1/2}z$, and $\lambda = (3.12429935)/(2.12429935 + z^2)$ [24]. Also, $f_{Z_1}(.)$ is the PDF of $Z_1 \sim TF(0, 1, 2.12429935) = t_{2.12429935}$, and $F_{Z_2}(.)$ is the CDF of $Z_2 \sim TF(0, 1, 3.12429935) = t_{3.12429935}$ [4] [24].

For the $\hat{\alpha}_1$ ST2 model.

$$f_Y(y|\mu, \sigma, \nu, \tau) = \frac{2}{0.03254897} f_{Z_1}(z) F_{Z_2}(\omega), \tag{10}$$

where y is the data and $z = (y - 0.4766338)/0.03254897$, $\omega = 2.7502377\lambda^{1/2}z$, and $\lambda = (4.023959)/(3.023959 + z^2)$ [24]. Also, $f_{Z_1}(.)$ is the PDF of $Z_1 \sim TF(0, 1, 3.023959) = t_{3.023959}$, and $F_{Z_2}(.)$ is the CDF of $Z_2 \sim TF(0, 1, 4.023959) = t_{4.023959}$ [4] [24].

5 INTERPRETATIONS AND CONCLUSIONS

This research began by proposing two questions in Section 1.3:

1. Using numerical estimation methods, can estimates for α be obtained, and are these α's approximately .50, akin to the Lévy distribution?
2. Can a representative distribution, and the estimated parameters of that distribution, be found that can accurately model α?

In the search for the answers to both of these, many statistical methods were employed and over 18 R packages are loaded in the finalized R code. For a full list of packages used, see appendix. In order to provide a setting to explore α, standardized stock return data were obtained from the CRSP. A frequentist approach was used, with the approximation of distribution parameters coming directly from the CRSP data [24]. Ultimately, the *ptsuite* package provided the framework to estimate the tail index α using equation 1. The $GAMLSS$ package provided many useful tools for both parameter estimation and distribution fitting by using the maximum likelihood estimation methods explained in Section 1.6.

5.1 Discussion of the $\hat{\alpha}$ Vectors

In the answer to the first question, estimates for $\hat{\alpha}$ were found successfully using equation 1. Overall, for both the data with varying s_0 values and for each of the subsets with fixed s_0 values, the center of the $\hat{\alpha}$ vectors fell around the hypothesized $\alpha = 0.50$ value. The m_0 model used the $\hat{\alpha}_0$, m_1 model used the $\hat{\alpha}_1$ vector, the m_2 model used the $\hat{\alpha}_2$ vector, and the m_3 model used $\hat{\alpha}_3$ vector. The means and medians for each $\hat{\alpha}$ vector are summarized in Table 11. Typically, the five number summaries are used for skewed distributions, but the means and medians are close in proximity and provide an adequate description of the center of each vector.

Table 11: Summary of Measures of Center for $\hat{\alpha}$ Vectors

$\hat{\alpha}$ Vector	Mean	Median
$\hat{\alpha}_0$	0.5264747	0.5167716
$\hat{\alpha}_1$	0.5104384	0.5013612
$\hat{\alpha}_2$	0.4989941	0.4936408
$\hat{\alpha}_3$	0.5803399	0.5721385

These values, with an exception of $\hat{\alpha}_3$, are close to the hypothesized value of 0.50 for the tail index α. Again, $\hat{\alpha}_3$ is calculated with $s_n = 270$. Overall, the summary of these $\hat{\alpha}$ lend credence to the idea that overall, the tail index α for standardized stock returns can be described as having a tail index similar to that of a Lévy distribution.

5.2 Finding the Representative Distribution for α

In order to answer the second question, numerical methods had to be employed. These methods were found inside the *GAMLSS* package and described in detail in Section 1.6. Overall, for each of the four $\hat{\alpha}$ vectors, 29 distributions, 29 distributions, 32 distributions, and 32 distributions, respectively, were fit and AIC values returned for each. The top three distributions for each model, based on the AIC value, were singled out and worm plots created for each. This procedure helped determine which distribution was chosen for each model. Finally, through the model selection procedures detailed in Section 3, a distribution was chosen to model each of the $\hat{\alpha}$ vectors. The distributions and AIC values are detailed in Table 1.

The ST3, ST1, and ST2 distributions are all members of the Skew t family of distributions, with variations on each of the probability density functions. Each of the

PDFs contain the same four parameters, μ, σ, ν, and τ. The generalized PDFs for the ST3, ST1, and ST2 distributions can be found in equations 2, 4, and 6, respectively. The fitted PDFs for models m_0, m_1, m_2, and m_3 can be found in equations 3, 5, 7, and 8, respectively. Since each of these Skew t distribution variations are extensions of the Generalized t (GT) distribution, a generalized likelihood ratio (GLR) test was used to confirm the parameter significance of each model. Although it is reasonable enough to assume that since the $GAMLSS$ procedure did not return the GT distribution, these parameters are useful, it is still good practice to confirm that a more generalized model would be inappropriate [24]. Each of these GLR tests confirmed parameter significance.

Overall, the Skew t family appears to be a representative family of distributions to model α for the standardized stock return data. Specifically, it is the ST3 , ST1 , and ST2 distributions that are identified respective to the specific data, as in Table 1 [4] [3] [10]. The most favorable AIC values are for the data with varying s_o values and the data with $s_o = 2500$, which have $s_n = 1642$ and $s_n = 1284$, respectively. Notably, the s_o values vary greatly in the $\hat{\alpha}_0$ data, with more than 20 stocks over $s_o = 20000$, compared to the $\hat{\alpha}_1$ data where s_o is fixed considerably lower at 2500. The conclusion here is that since these AIC values are close, and the distributions for each are in the same family, that the s_n parameter plays a more important role than s_o. Thus, a larger sample size of stocks, s_n, seems to lead to a more accurate model, based on AIC value.

Each set of $\hat{\alpha}$ values has its own distribution that seems to model that specific data set the best. If the aim is to find a distribution that can model $\hat{\alpha}$ regardless of the s_n values, Sections 4.3, 4.4 and 4.5 describe the $\hat{\alpha}_2$, $\hat{\alpha}_3$, $\hat{\alpha}_0$, and $\hat{\alpha}_1$ values fitted with the ST2 distribution. The ST2 distribution when taking the AIC values collectively, seems to be the an accurate distribution to model α. While the Skew t family is appropriate, the specific Skew t type 2 is robust and accurate in modeling the tail index of standardized stock returns.

5.3 Relevant Further Research

The results presented above provide a foundation for many future research opportunities. The most prevalent future research ideas will be summarized below. First, there are different ways to estimate the tail index, $\hat{\alpha}$, such as general percentile methods and modified percentile methods that are different than the geometric percentile method employed in this paper [5]. The geometric mean is used for data that may not be independent and data that is specific to stock or economic returns, which is why the geometric mean percentile method was used here [17]. However, there are other methods that may change the ultimate estimate of $\hat{\alpha}$, and it would be of interest if the Skew t family remains the most appropriate model with different parameters.

The main model selection procedures employed here are AIC values and worm plots. These are powerful but generous model building methods [24]. It would be beneficial to test these models, or new models, using more rigorous and definitive model selection procedures. Further, only three subsets of the standardized stock return data are described here. There would be an interest in the trade-off between the s_n value and the AIC, or other model selection criteria, that would maximize model accuracy while maintaining generality for new data.

Finally, some modern quantitative finance models are built using Lévy distributions and Lévy processes that contain a tail index parameter. Examples of quantitative finance models can be found in [31] [27],[6], [12], and [7]. One of the interests, and ultimate goals, in modeling α is to discover if the models discussed in Section 1.2 and beyond can be improved by having a Skew t distribution, such as ST2, included in the financial model. The hope is that the results here can be used to fortify existing economic models or lead to the creation of new methods for approaching risk.

BIBLIOGRAPHY

[1] Center for research in security prices (crsp).

[2] Nevorov V. B. Ahsanullah, M. Some inferences on the lévy distribution. *Journal of Statistical Theory and Applications*, 13(3):205–211, 2014.

[3] A. Azzalini. Futher results on a class of distributions which includes the normal ones. *Statistica*, 46:199–208, 1986.

[4] Capitanio A. Azzalini, A. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65:367–389, 2003.

[5] Sajjad Haider Bhatti, Shahzad Hussain, Tanvir Ahmad, Muhammad Aslam, Muhammad Aftab, and Muhammad Ali Raza. Efficient estimation of pareto model: Some modified percentile estimators. *PLOS ONE*, 13(5):1–15, 05 2018.

[6] Sun-Yong Choi and Ji-Hun Yoon. Modeling and risk analysis using parametric distributions with an application in equity-linked securities. *Mathematical Problems in Engineering*, 2020:1–20, 03 2020.

[7] Choudhry M. Choudhry, M. Advanced fixed income analysis. *Elsevier Science*, pages 13–36, 2004.

[8] John Cook. Power laws and the generalized clt.

[9] Eugene Fama. The distribution of daily differences of stock prices: A test of mandelbrot's stable paretian hypothesis. *Ph. D Dissertation*, 1964.

[10] Steel M.F. Fernandez, C. On bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association*, 93:359–371, 1998.

[11] Jose Enrique Figueroa-Lopez. Estimation methods for lévy based models of asset prices. *Financial Mathematics Seminar*, 2006.

[12] Neftci S. N. Hirsa, Ali. The wiener process, lévy processes, and rare events in financial markets. *An Introduction to the Mathematics of Financial Derivatives*, pages 123–144, 2014.

[13] Robert Hogg. *Introduction to Mathematical Statistics.* 2019.

[14] Oliver C. Ibe. Levy processes. *Markov Processes for Stochastic Modeling*, pages 329–347, 2013.

[15] Oliver C. Ibe. Lévy walk. *Elements of Random Walk and Diffusion Processes*, pages 175–195, 2013.

[16] Gentleman R. Ihaka, R. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314, 1996.

[17] de Vries C. Jansen, D. On the frequency of large stock return: Putting booms and busts into perspective. *The Review of Economics and Statistics*, 73(1):18–24, 1991.

[18] S.G. Kou. Lévy processes in asset prices. *Encyclopedia of Quantitative Risk Analysis and Assessment*, 2008.

[19] Hudson R. L. Mandelbrot, B. B. *The (mis)behavior of markets: a fractal view of financial turbulence.* Basic Books, 2008.

[20] Paolella M. Rachev S. Mittnik, S. A tail estimator for the index of the stable paretian distribution. *Communications in Statistics - Theory and Methods*, 27(5):1239–1262, 1998.

[21] Ranjiva Munasinghe, Pathum Kossinna, Dovini Jayasinghe, and Dilanka Wijeratne. Tail index estimation for power law distributions in r.

[22] J. Neto. Power laws heavy tail distributions.

[23] Shige Peng. A new central limit theorem under sublinear expectations. *Institute of Mathematics*, 2008.

[24] et al. Rigby, R. *Distributions for Modeling Location, Scale, and Shape: Using GAMLSS in R.* 2020.

[25] Stasinopoulos D. M. Rigby, R. A. Generalized additive models for location, scale and shape,(with discussion). *Appl. Statist.*, 54:507–554, 2005.

[26] N. N. Taleb. *Statistical consequences of fat tails: real world preasymptotics, epistemology, and applications: papers and commentary.* STEM Academic Press, 2020.

[27] Viktor Todorov. Simulation methods for lévy-driven continuous-time autoregressive moving average. *Journal of Business Economic Statistics*, 34(4):455–469, 2006.

[28] Young W. Trent, R. Geometric mean approximations of individual security and portfolio performance. *The Journal of Financial and Quantitative Analysis*, 4(2):179–199, 1969.

[29] Young W. Trent, R. Instructions on how to use the gamlss package in r: Second edition. 2008.

[30] Cindy Sin-Yi Tsai. The real world is not normal. *Morningstar Alternative Investments Observer*, 2011.

[31] Liuren Wu. Modeling financial security returns using lévy processes. *Handbooks in Operations Research and Management Science*, 15, 2008.

[32] James X. Xiong. Using truncated lévy flight to estimate downside risk. *Journal of Risk Management in Financial Institutions*, 3(3):231–242, 2010.

APPENDIX: R Code

```
####################################################
#Packages Loaded
####################################################

library(Rcpp)
library(ptsuite)
library(readxl)
library(MASS)
library(TeachingDemos)
library(extraDistr)
library(fitdistrplus)
library(ggplot2)
library(FAdist)
library(logspline)
library(actuaryr)
library(ExtDist)
library(optimx)
library(actuar)
library(expint)
library(gamlss)
library(gamlss.dist)
library(gamlss.add)

####################################################
#Written Functions
####################################################

lengthfunc <- function(x){
  H <- x[-c(which(x == 0))]
  G <- H[-c(which(H == -88))]
  D <- G[-c(which(G < -30))]
  A <- length(D)
  return(A)
}

alpha_estimate <- function(x){
  H <- x[-c(which(x == 0))]
  G <- H[-c(which(H == -88))]
  B <- G[-c(which(G < -30))]
  A <- alpha_geometric_percentile(abs(B))$shape
  return(A)
}

alpha_estimate2 <- function(x){
  B <- abs(x)
  D <- B[-c(which(B == 0))]
  A <- alpha_geometric_percentile(D)$shape
  return(A)
}

##

####################################################
# Stock Ticker MO, Altria Group Example
####################################################

stockdata2_R <- as.data.frame(stockdata2_R)

example <- stockdata2_R$X0.019169

example <- example[-c(which(example < -30))]

example <- example[-c(which(example == 0 ))]

example <- abs(example)

alpha_geometric_percentile(example$example)

hist(example, main = "Histogram␣of␣MO,␣Altria␣Group")

example <- data.frame(example)

ggplot(data = example, aes(x = example)) +
  geom_histogram(color = 'black', fill = 'red') +
  labs(title = "Histogram␣of␣MO,␣Altria␣Group",
       x = "Absolute␣Value␣of␣Stock␣Returns",
       y = "Count␣of␣Values")

####################################################
# Data Creation and Cleaning
```

```
####################################################

lengthvec <- lapply(thesis_data, lengthfunc)

lengthvec <- cbind(lengthvec)

rownames(lengthvec) = c()

colnames(lengthvec) = c()

lengthfinal <- as.numeric(lengthvec)

length2500 <- which(lengthfinal < 2500)

length5000 <- which(lengthfinal < 5000)

length10000 <- which(lengthfinal < 10000)

dim(thesis_data)

thesis_data2500 <- thesis_data[,-c(length2500)]

data2500 <- thesis_data2500[22545:25044,]

dim(data2500)

thesis_data5000 <- thesis_data[, -c(length5000)]

data5000 <- thesis_data5000[20045:25044,]

dim(data5000)

thesis_data10000 <- thesis_data[, -c(length10000)]

data10000 <- thesis_data10000[14045:24044,]

dim(data10000)

####################################################
# Alpha Vector Creation and Cleaning
####################################################
```

```
alpha_vec1 <- lapply(thesis_data, alpha_estimate)
alpha_cols1 <- cbind(alpha_vec1)
rownames(alpha_cols1) = c()
colnames(alpha_cols1) = c()
alpha_cols1 <- as.numeric(alpha_cols1)
hist(alpha_cols1)
which(is.nan(alpha_cols1))
alpha_cols1 <- alpha_cols1[-c(which(is.nan(alpha_cols1)))]


alpha_vec2 <- lapply(data2500, alpha_estimate2)
alpha_cols2 <- cbind(alpha_vec2)
rownames(alpha_cols2) = c()
colnames(alpha_cols2) = c()
alpha_cols2 <- as.numeric(alpha_cols2)
hist(alpha_cols2)
which(is.nan(alpha_cols2))
alpha_cols2 <- alpha_cols2[-c(which(is.nan(alpha_cols2)))]

alpha_vec3 <- lapply(data5000, alpha_estimate2)
alpha_cols3 <- cbind(alpha_vec3)
rownames(alpha_cols3) = c()
```

```
colnames(alpha_cols3) = c()

alpha_cols3 <- as.numeric(alpha_cols3)


alpha_vec4 <- lapply(data10000, alpha_estimate2)

alpha_cols4 <- cbind(alpha_vec4)

rownames(alpha_cols4) = c()

colnames(alpha_cols4) = c()

alpha_cols4 <- as.numeric(alpha_cols4)


####################################################
# GAMLSS Model Creation and Relevant Plots
####################################################
# m0 Model and Plots
####################################################

m0 <- gamlssML(alpha_cols1)

model_m0 <- chooseDist(m0, type = "realline", try.gamlss =
    T, trace = F)

getOrder(model_m0)

m0_ST3 <- gamlssML(alpha_cols1, family = "ST3")

summary(m0_ST3)

resid <- resid(m0_ST3)

resid.df <- as.data.frame(resid)

ggplot(data = resid.df, aes(x = resid)) +
  geom_histogram(color = 'black', fill = 'forestgreen',
     binwidth = .2) +
  labs(title = "Histogram␣of␣Residuals,␣m0␣model",
       x = "Residuals",
       y = "Count")

wp(m0_ST3, ylim.all = T)

#####################################################

m0_SEP2 <- gamlssML(alpha_cols1, family = "SEP2")

summary(m0_SEP2)

hist(resid(m0_SEP2))

wp(m0_SEP2, ylim.all = T)

#####################################################

m0_SST <- gamlssML(alpha_cols1, family = "SST")

summary(m0_SST)

hist(resid(m0_SST))

wp(m0_SST, ylim.all = T)

#####################################################

m0_ST2 <- gamlssML(alpha_cols1, family = "ST2")

summary(m0_ST2)

wp(m0_ST2, ylim.all = T)

#####################################################


histDist(alpha_cols1, family = "ST3", line.col = "red",
   col.hist = "lightblue",
         nbins = 47, border.hist = "black",
         main = "Histogram␣of␣Alpha␣Values␣with␣the␣Fitted
```

```
           ␣ST3␣Distribution",
         xlab = "Alpha␣Values", ylab = "Count", col.axis =
             "black", fg.hist = "black")


histDist(alpha_cols1, family = "ST2", line.col = "red",
   col.hist = "lightblue",
         nbins = 47, border.hist = "black",
         main = "Histogram␣of␣Alpha␣Values␣with␣the␣Fitted
           ␣ST2␣Distribution",
         xlab = "Alpha␣Values", ylab = "Count", col.axis =
             "black", fg.hist = "black")


####################################################
# m1 Model and Plots
####################################################

m1 <- gamlssML(alpha_cols2)

model_m1 <- chooseDist(m1, type = "realline", try.gamlss =
    T, trace = F)

getOrder(model_m1)

m1_ST1 <- gamlssML(alpha_cols2, family = "ST1")

summary(m1_ST1)

hist(resid(m1_ST1))

resid <- resid(m1_ST1)

resid.df <- as.data.frame(resid)

ggplot(data = resid.df, aes(x = resid)) +
  geom_histogram(color = 'black', fill = 'forestgreen',
     binwidth = .2) +
  labs(title = "Histogram␣of␣Residuals,␣m1␣model",
       x = "Residuals",
       y = "Count")
```

```
wp(m1_ST1, ylim.all = T)

ST1()

####################################################

m1_ST5 <- gamlssML(alpha_cols2, family = "ST5")

summary(m1_ST5)

ST5()

wp(m1_ST5, ylim.all = T)

####################################################

m1_JSUo <- gamlssML(alpha_cols2, family = "JSUo")

summary(m1_JSUo)

wp(m1_JSUo, ylim.all = T)

####################################################

m1_ST2 <- gamlssML(alpha_cols2, family = "ST2")

summary(m1_ST2)

wp(m1_ST2, ylim.all = T)

####################################################

histDist(alpha_cols2, family = "ST1", line.col = "red",
   col.hist = "lightblue",
         nbins = 47, border.hist = "black",
         main = "Histogram␣of␣Alpha␣Values␣with␣the␣Fitted
            ␣ST1␣Distribution",
         xlab = "Alpha␣Values", ylab = "Count", col.axis =
             "black",
         fg.hist = "black", ylim = c(0,17))
```

```
####################################################
# m2 Model and Plots
####################################################

m2 <- gamlssML(alpha_cols3)

model_m2 <- chooseDist(m2, type = "realline", try.gamlss =
    T, trace = F)

getOrder(model_m2)

m2_ST2 <- gamlssML(alpha_cols3, family = "ST2")

summary(m2_ST2)

hist(resid(m2_ST2))

resid <- resid(m2_ST2)

resid.df <- as.data.frame(resid)

ggplot(data = resid.df, aes(x = resid)) +
  geom_histogram(color = 'black', fill = 'forestgreen',
     binwidth = .2) +
  labs(title = "Histogram␣of␣Residuals,␣m2␣model",
       x = "Residuals",
       y = "Count")

wp(m2_ST2, ylim.all = T)

####################################################

m2_ST5 <-gamlssML(alpha_cols3, family = "ST5")

summary(m2_ST5)

wp(m2_ST5, ylim.all = 1.5 * sqrt(1/length(resid)))

wp(m2_ST5, ylim.all = T)
```

```
####################################################

m2_ST3 <- gamlssML(alpha_cols3, family = "ST3")

summary(m2_ST3)

wp(m2_ST3, ylim.all = T)

####################################################

m2_SST <- gamlssML(alpha_cols3, family = "SST")

summary(m2_SST)

wp(m2_SST, ylim.all = T)

####################################################

histDist(alpha_cols3, family = "ST2", line.col = "red",
   col.hist = "lightblue",
         nbins = 85, border.hist = "black",
         main = "Histogram␣of␣Alpha␣Values␣with␣the␣Fitted
            ␣ST2␣Distribution",
         xlab = "Alpha␣Values", ylab = "Count", col.axis =
             "black",
         fg.hist = "black")

####################################################
# m3 Model and Plots
####################################################

m3 <- gamlssML(alpha_cols4)

model_m3 <- chooseDist(m3, type = "realline", try.gamlss =
    T, trace = F)

getOrder(model_m3)

m3_ST2 <- gamlssML(alpha_cols4, family = "ST2")
```

```
summary(m3_ST2)

hist(resid(m3_ST2))

resid <- resid(m3_ST2)

resid.df <- as.data.frame(resid)

ggplot(data = resid.df, aes(x = resid)) +
  geom_histogram(color = 'black', fill = 'forestgreen',
     binwidth = .2) +
  labs(title = "Histogram␣of␣Residuals,␣m3␣model",
       x = "Residuals",
       y = "Count")

wp(m3_ST2, ylim.all = T)

####################################################

m3_ST3 <- gamlssML(alpha_cols4, family = "ST3")

summary(m3_ST3)

wp(m3_ST3, ylim.all = T)

####################################################

m3_ST1 <- gamlssML(alpha_cols4, family = "ST1")

summary(m3_ST1)

wp(m3_ST1, ylim.all = T)

####################################################

m3_SST <- gamlssML(alpha_cols4, family = "SST")

summary(m3_SST)

wp(m3_SST, ylim.all = T)
```

```
####################################################

histDist(alpha_cols4, family = "ST2", line.col = "red",
   col.hist = "lightblue",
         nbins = 47, border.hist = "black",
         main = "Histogram␣of␣Alpha␣Values␣with␣the␣Fitted
            ␣ST2␣Distribution",
         xlab = "Alpha␣Values", ylab = "Count", col.axis =
             "black",
         fg.hist = "black")

####################################################
# GLR Tests
####################################################

m0_GT <- gamlssML(alpha_cols1, family = "GT")

LR.test(m0_GT, m0_ST3)

LR.test(m0_ST2, m0_ST3)

histDist(alpha_cols2, family = "ST1", line.col = "red",
   col.hist = "lightblue",
         nbins = 47, border.hist = "black",
         main = "Histogram␣of␣Alpha␣Values␣with␣the␣Fitted
            ␣ST1␣Distribution",
         xlab = "Alpha␣Values", ylab = "Count", col.axis =
             "black",
         fg.hist = "black", ylim = c(0,17))

m1_GT <- gamlssML(alpha_cols2, family = "GT")

LR.test(m1_GT, m1_ST1)

histDist(alpha_cols2, family = "ST2", line.col = "red",
   col.hist = "lightblue",
         nbins = 47, border.hist = "black",
         main = "Histogram␣of␣Alpha␣Values␣with␣the␣Fitted
            ␣ST2␣Distribution",
         xlab = "Alpha␣Values", ylab = "Count", col.axis =
             "black",
         fg.hist = "black", ylim = c(0,17))

LR.test(m1_ST2, m1_ST1)

m2_GT <- gamlssML(alpha_cols3, family = "GT")

LR.test(m2_GT, m2_ST2)

LR.test(m2_SST, m2_ST2)

m3_GT <- gamlssML(alpha_cols4, family = "GT")

LR.test(m3_GT, m3_ST2)


####################################################
# All GAMLSS Models fit
####################################################

getOrder(model_m0)

getOrder(model_m1)

getOrder(model_m2)

getOrder(model_m3)

####################################################
```

VITA

JETT BURNS

Education:	B.A. Economics, East Tennessee State University, Johnson City, Tennessee 2020 M.S. Mathematical Sciences, East Tennessee State University, Johnson City, Tennessee 2022
Professional Experience:	Controller, Timber! in Johnson City, Johnsnon City, Tennessee, 2020–present