STATISTICAL DISTRIBUTION
THEORY WITH APPLICATIONS TO FINANCE

A THESIS SUBMITTED TO THE UNIVERSITY OF MANCHESTER
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
IN THE FACULTY OF SCIENCE AND ENGINEERING

2018

By
Jeffrey Chu
School of Mathematics
Contents

Abstract 15

Declaration 17

Copyright 18

Acknowledgements 19

1 Introduction 20

1.1 Aims ......................................................... 20

1.2 Motivation and contributions ...................................... 20

1.3 Publications ...................................................... 23

1.4 Conferences and seminars ......................................... 23

1.5 Future work ....................................................... 24

Bibliography 26

2 A saddlepoint approximation to the distribution of the sum of independent non-identical beta random variables 27

2.1 Introduction ...................................................... 27

2.2 Methodology ....................................................... 29
Bibliography

A Appendix to Chapter 2 140

B Appendix to Chapter 6 147
List of Tables

3.1 Comparison of the performances of the five approximations under gamma $(\alpha, \theta)$ distributed claim amounts and Poisson $(\lambda)$ distributed claim numbers, shown by the number/percentage of combinations of parameter values tested where the approximation(s) $(f)$ gave the smallest absolute $(\Delta_f)$ or relative $(\Omega_f)$ error. ................................................................. 44

3.2 Comparison of the performances of the five approximations under inverse Gaussian $(\eta, \theta)$ distributed claim amounts and Poisson $(\lambda)$ distributed claim numbers, shown by the number/percentage of combinations of parameter values tested where the approximation(s) $(f)$ gave the smallest absolute $(\Delta_f)$ or relative $(\Omega_f)$ error. ................................................................. 48

3.3 Comparison of the performances of the five approximations under gamma $(\alpha, \theta)$ distributed claim amounts and negative binomial $(r, \beta)$ distributed claim numbers, shown by the number/percentage of combinations of parameter values tested where the approximation(s) $(f)$ gave the smallest absolute $(\Delta_f)$ or relative $(\Omega_f)$ error. ................................................................. 52

3.4 Comparison of the performances of the five approximations under inverse Gaussian $(\eta, \theta)$ distributed claim amounts and negative binomial $(r, \beta)$ distributed claim numbers, shown by the number/percentage of combinations of parameter values tested where the approximation(s) $(f)$ gave the smallest absolute $(\Delta_f)$ or relative $(\Omega_f)$ error. ................................................................. 56
4.1 Relative biases of $\hat{N}$, $\hat{N}^*$ and $\hat{N}^+$ in a simulation study using small sample sizes. ................................. 66

4.2 Relative biases of $\hat{N}$, $\hat{N}^*$ and $\hat{N}^+$ in a simulation study using large sample sizes. 68

4.3 Comparison of mean collisions, reciprocal of mean collisions, and relative biases of $\hat{N}$, $\hat{N}^*$ and $\hat{N}^+$ from estimating the size of the Twitter network using different sample sizes. ............................... 71

4.4 Comparison of mean collisions, reciprocal of mean collisions, and relative biases of $\hat{N}$, $\hat{N}^*$ and $\hat{N}^+$ from estimating the size of the LiveJournal network using small sample sizes. ................................. 73

4.5 Comparison of mean collisions, reciprocal of mean collisions, and relative biases of $\hat{N}$, $\hat{N}^*$ and $\hat{N}^+$ from estimating the size of the LiveJournal network using large sample sizes. ................................. 75

5.1 AIC and BIC values for the four models fitted to the degree density distribution of market graphs constructed using threshold values $\theta \in [0.1, 0.4]$, for the period of 1st January 2000 - 31st December 2001. .............. 91

5.2 KS test statistics for the four models fitted to the degree density distribution of market graphs constructed using threshold values $\theta \in [0.1, 0.4]$, for the period of 1st January 2000 - 31st December 2001. .............. 92

5.3 AIC and BIC values for the four models fitted to the degree density distribution of complements of market graphs constructed using threshold values $\theta \in [-0.125, -0.025]$, for the period of 1st January 2000 - 31st December 2001. .............. 94

5.4 KS test statistics for the four models fitted to the degree density distribution of complements of market graphs constructed using threshold values $\theta \in [-0.125, -0.025]$, for the period of 1st January 2000 - 31st December 2001. .............. 95

5.5 AIC and BIC values for the four models fitted to the degree density distribution of market graphs constructed using threshold values $\theta \in [0.1, 0.6]$, for the period of 1st January 2006 - 31st December 2007. .............. 97
5.6 KS test statistics for the four models fitted to the degree density distribution of market graphs constructed using threshold values $\theta \in [0.5, 0.6]$, for the period of 1st January 2006 - 31st December 2007.

5.7 AIC and BIC values for the four models fitted to the degree density distribution of complements of market graphs constructed using threshold values $\theta \in [-0.125, -0.025]$, for the period of 1st January 2006 - 31st December 2007.

5.8 KS test statistics for the four models fitted to the degree density distribution of complements of market graphs constructed using threshold values $\theta \in [-0.075, -0.025]$, for the period of 1st January 2006 - 31st December 2007.

5.9 AIC and BIC values for the four models fitted to the degree density distribution of market graphs constructed using threshold values $\theta \in [0.1, 0.6]$, for the period of 1st January 2012 - 31st December 2013.

5.10 KS test statistics for the four models fitted to the degree density distribution of market graphs constructed using threshold values $\theta \in [0.1, 0.6]$, for the period of 1st January 2012 - 31st December 2013.

5.11 AIC and BIC values for the four models fitted to the degree density distribution of complements of market graphs constructed using threshold values $\theta \in [-0.125, -0.025]$, for the period of 1st January 2012 - 31st December 2013.

5.12 KS test statistics for the four models fitted to the degree density distribution of complements of market graphs constructed using threshold values $\theta \in [-0.075, -0.025]$, for the period of 1st January 2012 - 31st December 2013.


6.3 Best fitting distributions to the daily log returns of the exchange rates of
Bitcoin, Dash, Dogecoin, Litecoin, MaidSafeCoin, Monero, and Ripple versus
the US Dollar from 23 June 2014 until 28 February 2017, and parameter
estimates, with standard errors given in brackets. ................. 131

B.1 Fitted distributions and information criteria results for daily log returns of
the exchange rates of Bitcoin from 23 June 2014 until 28 February 2017. . . 147

B.2 Fitted distributions and information criteria results for daily log returns of
the exchange rates of Dash from 23 June 2014 until 28 February 2017. . . . 148

B.3 Fitted distributions and information criteria results for daily log returns of
the exchange rates of Dogecoin from 23 June 2014 until 28 February 2017. . 148

B.4 Fitted distributions and information criteria results for daily log returns of
the exchange rates of Litecoin from 23 June 2014 until 28 February 2017. . 149

B.5 Fitted distributions and information criteria results for daily log returns of
the exchange rates of MaidSafeCoin from 23 June 2014 until 28 February 2017.149

B.6 Fitted distributions and information criteria results for daily log returns of
the exchange rates of Monero from 23 June 2014 until 28 February 2017. . . 150

B.7 Fitted distributions and information criteria results for daily log returns of
the exchange rates of Ripple from 23 June 2014 until 28 February 2017. . . 150
List of Figures

2.1 Comparison of the relative errors (vertical axis) of the saddlepoint ($A_S$) and normal ($A_N$) approximations, versus $E_P = F_{\text{sim}}(x)$ (horizontal axis), for the sixteen cases. Cases are arranged as $\{1, 2, 3, 4\}$ (first row), $\{5, 6, 7, 8\}$ (second row), $\{9, 10, 11, 12\}$ (third row), and $\{13, 14, 15, 16\}$ (fourth row). Plots are coded as black for $n = 6$, red for $n = 8$, blue for $n = 10$, brown for $n = 12$; broken lines for $A_N$ and solid lines for $A_S$. 33

3.1 The combinations of the values of parameters ($\alpha, \theta, \lambda$), indicated by dots, for which the approximations based on the gamma (top left), exponentiated exponential (top right), Weibull (middle left), inverse Gaussian (middle right) and lognormal (bottom left) distributions gave the smallest absolute error ($\Delta_f$). 45

3.2 The combinations of the values of parameters ($\alpha, \theta, \lambda$), indicated by dots, for which the approximations based on the gamma (top left), exponentiated exponential (top right), Weibull (bottom left) and lognormal (bottom right) distributions gave the smallest relative error ($\Omega_f$). 46

3.3 The combinations of the values of parameters ($\eta, \theta, \lambda$), indicated by dots, for which the approximations based on the gamma (top left), exponentiated exponential (top right), Weibull (middle left), inverse Gaussian (middle right) and lognormal (bottom left) distributions gave the smallest absolute error ($\Delta_f$). 49
3.4 The combinations of the values of parameters \((\eta, \theta, \lambda)\), indicated by dots, for which the approximations based on the gamma (top left), exponentiated exponential (top right) and Weibull (bottom left) distributions gave the smallest relative error \((\Omega_f)\). .................................................. 50

3.5 The combinations of the values of parameters \((\alpha, r, \beta)\) from \((\alpha, \theta, r, \beta)\), indicated by dots, for which the approximations based on the exponentiated exponential (top left), Weibull (top right), inverse Gaussian (bottom left) and lognormal (bottom right) distributions gave the smallest absolute error \((\Delta_f)\). 53

3.6 The combinations of the values of parameters \((\alpha, r, \beta)\) from \((\alpha, \theta, r, \beta)\), indicated by dots, for which the approximations based on the gamma (top left), exponentiated exponential (top right), inverse Gaussian (bottom left) and lognormal (bottom right) distributions gave the smallest relative error \((\Omega_f)\). 54

3.7 The combinations of the values of parameters \((\eta, r, \beta)\) from \((\eta, \theta, r, \beta)\), indicated by dots, for which the approximations based on the exponentiated exponential (top left), Weibull (top right), inverse Gaussian (bottom left) and lognormal (bottom right) distributions gave the smallest absolute error \((\Delta_f)\). 57

3.8 The combinations of the values of parameters \((\eta, r, \beta)\) from \((\eta, \theta, r, \beta)\), indicated by dots, for which the approximations based on the gamma (top left), exponentiated exponential (top right), inverse Gaussian (bottom left) and lognormal (bottom right) distributions gave the smallest relative error \((\Omega_f)\). 58

4.1 Plot of relative biases of \(\hat{N}\) (dotted line above zero), \(\hat{N}^*\) (dashed line below zero) and \(\hat{N}^+\) (solid black line) in the simulation study versus logarithm of sample size \(n/10\). ................................................................. 69

4.2 Mean estimates of the true size of the Twitter network (solid horizontal line) generated by \(\hat{N}\), \(\hat{N}^*\) and \(\hat{N}^+\) at different sample sizes. ................... 72

4.3 Plot of relative biases of \(\hat{N}^*\) and \(\hat{N}^+\) from the estimation of the LiveJournal network size using small sample sizes. ................................. 74
4.4 Plot of relative biases of $\hat{N}$, $\hat{N}^*$ and $\hat{N}^+$ from the estimation of the LiveJournal network size using large sample sizes. 76

5.1 Comparison of the distributions of correlation coefficients, between pairs of stocks, in the three time periods: i) 1st January 2000 - 31st December 2001 (solid black line); ii) 1st January 2006 - 31st December 2007 (dashed red line); iii) 1st January 2012 - 31st December 2013 (dotted blue line). 88

5.2 Comparison of the four models fitted to the degree density distribution of market graphs constructed using threshold values: $\theta = 0.1$ (top left); $\theta = 0.2$ (top right); $\theta = 0.3$ (bottom left); $\theta = 0.4$ (bottom right), for the period of 1st January 2000 - 31st December 2001. Fitted models are coded as blue for Pareto I, green for Fréchet, yellow for lognormal, and purple for GPD. 90

5.3 Comparison of the four models fitted to the degree density distribution of complements of market graphs constructed using threshold values: $\theta = -0.025$ (top left); $\theta = -0.05$ (top right); $\theta = -0.075$ (bottom left); $\theta = -0.1$ (bottom right), for the period of 1st January 2000 - 31st December 2001. Fitted models are coded as blue for Pareto I, green for Fréchet, yellow for lognormal, and purple for GPD. 93

5.4 Comparison of the four models fitted to the degree density distribution of market graphs constructed using threshold values: $\theta = 0.1$ (top left); $\theta = 0.2$ (top right); $\theta = 0.3$ (middle left); $\theta = 0.4$ (middle right); $\theta = 0.5$ (bottom left); $\theta = 0.6$ (bottom right), for the period of 1st January 2006 - 31st December 2007. Fitted models are coded as blue for Pareto I, green for Fréchet, yellow for lognormal, and purple for GPD. 96
5.5 Comparison of the four models fitted to the degree density distribution of complements of market graphs constructed using threshold values: \( \theta = -0.025 \) (top left); \( \theta = -0.05 \) (top right); \( \theta = -0.075 \) (bottom left); \( \theta = -0.1 \) (bottom right), for the period of 1st January 2006 - 31st December 2007. Fitted models are coded as blue for Pareto I, green for Fréchet, yellow for lognormal, and purple for GPD.

5.6 Comparison of the four models fitted to the degree density distribution of market graphs constructed using threshold values: \( \theta = 0.1 \) (top left); \( \theta = 0.2 \) (top right); \( \theta = 0.3 \) (middle left); \( \theta = 0.4 \) (middle right); \( \theta = 0.5 \) (bottom left); \( \theta = 0.6 \) (bottom right), for the period of 1st January 2012 - 31st December 2013. Fitted models are coded as blue for Pareto I, green for Fréchet, yellow for lognormal, and purple for GPD.

5.7 Comparison of the four models fitted to the degree density distribution of complements of market graphs constructed using threshold values: \( \theta = -0.025 \) (top left); \( \theta = -0.05 \) (top right); \( \theta = -0.075 \) (bottom left); \( \theta = -0.1 \) (bottom right), for the period of 1st January 2012 - 31st December 2013. Fitted models are coded as blue for Pareto I, green for Fréchet, yellow for lognormal, and purple for GPD.


B.1 The Q-Q plots of the best fitting distributions for daily log returns of the exchange rates of Bitcoin (first row, left), Dash (first row, right), Dogecoin (second row, left), Litecoin (second row, right), MaidSafeCoin (third row, left), Monero (third row, right) and Ripple (last row) from 23 June 2014 until 28 February 2017.
B.2 The P-P plots of the best fitting distributions for daily log returns of the exchange rates of Bitcoin (first row, left), Dash (first row, right), Dogecoin (second row, left), Litecoin (second row, right), MaidSafeCoin (third row, left), Monero (third row, right) and Ripple (last row) from 23 June 2014 until 28 February 2017.

B.3 Value at risk for the best fitting distributions for daily log returns of the exchange rates of Bitcoin (first row, left), Dash (first row, right), Dogecoin (second row, left), Litecoin (second row, right), MaidSafeCoin (third row, left), Monero (third row, right) and Ripple (last row) from 23 June 2014 until 28 February 2017. Also shown are the value at risk for daily log returns of the exchange rates of the Euro, computed using the same best fitting distributions.

B.4 Expected shortfall for the best fitting distributions for daily log returns of the exchange rates of Bitcoin (first row, left), Dash (first row, right), Dogecoin (second row, left), Litecoin (second row, right), MaidSafeCoin (third row, left), Monero (third row, right) and Ripple (last row) from 23 June 2014 until 28 February 2017. Also shown are the expected shortfall values for daily log returns of the exchange rates of the Euro, computed using the same best fitting distributions.

B.5 Kupiec’s $p$-values for the best fitting distributions for daily log returns of the exchange rates of Bitcoin (first row, left), Dash (first row, right), Dogecoin (second row, left), Litecoin (second row, right), MaidSafeCoin (third row, left), Monero (third row, right) and Ripple (last row) from 23 June 2014 until 28 February 2017.

B.6 Histograms of standard deviations of daily log returns of the exchange rates of the seven cryptocurrencies over windows of width 20 days. Also shown is the histogram of standard deviations of daily log returns of the exchange rates of the Euro over windows of width 20 days.
Abstract

The whole thesis comprises six chapters, where the running theme focuses on the development of statistical methods and distribution theory, with applications to finance. It begins with Chapter 1, which provides the introduction and background to my thesis. This is then followed by Chapters 2 through to 6, which provide the main contributions.

The exact distribution of the sum of more than two independent beta random variables is not a known result. Even in terms of approximations, only the normal approximation is known for the sum. Motivated by Murakami (2014), Chapter 2 derives a saddlepoint approximation for the distribution of the sum. An extensive simulation study shows that it always gives better performance than the normal approximation.

Jin et al. (2016) proposed a novel moments based approximation based on the gamma distribution for the compound sum of independent and identical random variables, and illustrated their approximation through the use of six examples. Chapter 3 revisits four of their examples, and it is shown that moments based approximations based on simpler distributions can be good competitors. The moments based approximations are also shown to be more accurate than the truncated versions of the exact distribution of the compound sum. Results regarding the performances of the approximations are provided, which could be useful in determining which approximation should be used given a real data set.

The estimation of the size of large populations can often be a significant problem. Chapter 4 proposes a new population size estimator and provides a comparison of its performance with two recent estimators known in the literature. The comparison is based on a simulation study and applications to two real big data sets from the Twitter and LiveJournal social networks. The proposed estimator is shown to outperform the known estimators, at least for small sample sizes.

In recent years, with a growing interest in big or large datasets, there has been a rise in the application of large graphs and networks to financial big data. Much of this research has focused on
the construction and analysis of the network structure of stock markets, based on the relationships between stock prices. Motivated by Boginski et al. (2005), who studied the characteristics of a network structure of the US stock market, Chapter 5 constructs network graphs of the UK stock market using the same method. Four distributions are fitted to the degree density of the vertices from these graphs: the Pareto I, Fréchet, lognormal, and generalised Pareto distributions, and their goodness of fits are assessed. Results show that the degree density of the complements of the market graphs, constructed using a negative threshold value close to zero, can be fitted well with the Fréchet and lognormal distributions.

Chapter 6 analyses statistical properties of the largest cryptocurrencies (determined by market capitalisation), of which Bitcoin is the most prominent example. The analysis characterises their exchange rates versus the US Dollar by fitting parametric distributions to them. It is shown that cryptocurrency returns are clearly non-normal, however, no single distribution fits well jointly to all of the cryptocurrencies analysed. We find that for the most popular cryptocurrencies, such as Bitcoin and Litecoin, the generalised hyperbolic distribution gives the best fit, whilst for the smaller cryptocurrencies the normal inverse Gaussian distribution, generalised $t$ distribution, and Laplace distribution give good fits. The results are important for investment and risk management purposes.
Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.
Copyright

1. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

2. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

3. The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

4. Further information on the conditions under which disclosure, publication and commercialization of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://www.campus.manchester.ac.uk/medialibrary/policies/intellectualproperty.pdf), in any relevant Thesis restriction declarations deposited in the University Library, The University Library’s regulations (see http://www.manchester.ac.uk/library/aboutus/regulations) and in The University’s policy on presentation of Theses.
Acknowledgements

I am extremely grateful to my supervisor Saralees Nadarajah, for all his help, support and guidance during the course of my studies. I am also thankful for his patience, his influence on myself to become a better researcher, and for believing in me even when others did not.

I am grateful for the love and support from my mother, father, sister, and brother. Although it was not a simple decision, and it has not been an easy few years, I am thankful every day for this opportunity you have given me.

Finally, to friends near and far, old and new, and everyone I have met on this long and winding road, thank you.
Chapter 1

Introduction

1.1 Aims

The aim of this thesis is to present a selection of my research work and contributions made, over the course of my PhD studies, to the area of statistical distribution theory. This thesis is presented in the alternative format, where each of the main chapters is a research paper which has either been published or is under review for publication in a peer-reviewed journal. Each chapter loosely follows a similar format where the problem is introduced and defined, followed by the methodology and results, and is concluded with a discussion and summary of the results. The main focus of the work presented relates to the area of statistical distribution theory with applications to finance, and provides a balance between theory and application.

The remainder of this introduction lists the motivation and contributions of each of the main chapters, details of the chapters which have been published, conferences and seminars that I have attended and presented at during my PhD, and suggestions for the direction of future work.

1.2 Motivation and contributions

- Chapter 2: A saddlepoint approximation to the distribution of the sum of independent non-identical beta random variables – The sums of beta random variables appear in a wide variety of applied areas, such as cellular networks, software reliability, and even project management. However, to date, no one has been able to derive the exact distribution of the sum of more than two independent but non-identical beta random variables.
Taking inspiration from an existing saddlepoint approximation for the sum of independent and non-identical uniform random variables (Murakami, 2014), an approximation for the sum of independent and non-identical beta random variables is derived. An improvement over the standard normal approximation is shown, and it is hoped that this approximation method could be beneficial in situations such as project management and software reliability, where the accuracy of models has a direct influence on costs or profits.

- **Chapter 3: On moment based density approximations for aggregate losses** — Following on from the previous chapter, the sums of independent and identical random variables, and Pareto distributions are commonly used in the area of insurance. However, the sum of independent and identical Pareto random variables does not have a closed form distribution function, which means that it is not straightforward to use in insurance claim modelling. This chapter examines the problem of modelling insurance claims, more specifically the total claim over a period of time which is a sum of independent and identical random variables. These represent the individual claim amounts, which in turn are independent of another random variable representing the number of claims. However, closed form expressions for the distribution of this sum are often hard to find. Revisiting a novel moments based approximation method based on the gamma distribution (Jin et al., 2016), approximations based on other distributions are derived and shown to be good competitors. Suggestions regarding the most suitable approximation methods, in relation to the characteristics of real insurance data, are also provided.

- **Chapter 4: Bias reduction in the size estimation of large populations using small data sets** — Many data sets are commonly modelled as graphs or networks, however, in the analysis of large populations, a significant problem of the sheer size of the data means that many traditional methods and measures used to obtain representative samples for these real graphs can become problematic. Furthermore, determining the size of a graph or network can often be difficult due to limited access to the data, and even obtaining large samples can be tricky. Knowing the total population size of a large graph or network can be advantageous in many scenarios: determining the worth or overall performance of a social network; helping to choose the best (social) platform to maximise the impact of advertising; estimating the market size of particular financial markets; optimising networks and network traffic. The challenge then becomes how to use the smallest sample of data to give the best representative estimate of the true size of a real graph. This chapter proposes a new population size estimator, which overcomes some of the problems of existing methods, and is shown to give improved
performance (and lower bias) compared with recent estimators in the literature, especially when only small samples of data are available. Whilst this could bring improved estimation accuracy in the situations mentioned above, it is also hoped that this method could be particularly beneficial in the context of time-critical decisions or actions. For example, in the military, where such estimates may need to be computed quickly and efficiently, or where acquiring additional data may take too much time, be too expensive in terms of monetary value or may put lives at risk.

- **Chapter 5: A statistical analysis of UK financial networks** – As mentioned above, large data sets are often modelled as large graphs or networks, and they have useful applications in many different areas not limited to just science and engineering. In recent years, the applications in finance have increased significantly, in particular in the modelling of the network of financial markets. A popular analysis is to model the network structure of stock markets based on the relationship between stock prices or returns, and research already exists on the US, South American, and Chinese stock markets amongst others. In relation to the first chapter, it has often been found that statistical properties of these networks, such as the influence-strength relationship and degree distribution, can be modelled by a Pareto or power law distribution. Motivated by Boginski et al. (2015), this chapter contributes to the literature by providing the first statistical analysis of UK financial networks, over different time periods including before and after the most recent financial crisis. The analysis focuses on modelling the degree density distribution, using heavy tailed distributions including the Pareto and generalised Pareto distributions, of the UK stock market network constructed from the correlation between stock prices and returns. It is hoped that this initial analysis will encourage a deeper understanding of the complexities of the networks of UK financial markets.

- **Chapter 6: A statistical analysis of cryptocurrencies** – Over the past few months, there has (again) been a significant increase in the interest in Bitcoin and other cryptocurrencies. For example, the UK government is considering paying out research grants in Bitcoin; an increasing number of IT companies are stockpiling Bitcoin to defend against ransomware; growing numbers in China are buying into Bitcoin and see it as an investment opportunity. With this recent surge in interest, it was felt that now is the time to start studying Bitcoin and other cryptocurrencies as key pieces of financial technology, and not just as a novelty. In addition, an interesting property of cryptocurrencies is that it seems like almost anything and everything can affect the price, thus the volatility of digital currencies is much greater than traditional financial instruments. Whilst the majority of the existing literature focuses on
CHAPTER 1. INTRODUCTION

Bitcoin, this chapter acts as an indirect extension to Chu et al. (2015), and provides the first study investigating the statistical properties of cryptocurrencies going beyond just Bitcoin. Results relating to the best fitting distributions for modelling the returns of cryptocurrencies are given, as well as results for financial risk. The findings have implications in the area of risk management, where estimates for the Value at Risk and Expected Shortfall for the cryptocurrencies could be used to assist in decisions about the inclusion of cryptocurrencies in investment portfolios for diversification purposes.

1.3 Publications

Four out of the five main chapters in this thesis have been published in refereed journals:


1.4 Conferences and seminars

During the course of my PhD, I have represented the School of Mathematics at the University of Manchester at the following conferences and seminars:

- The 9th International Conference on Extreme Value Analysis (EVA), The University of Michigan, Ann Arbor, USA, June 15th – 19th 2015. Poster presentation based on the published paper “On moment based density approximations for aggregate losses”.
• The 4th Institute of Mathematical Statistics Asia Pacific Rim Meeting, The Chinese University of Hong Kong, Hong Kong, June 27th – 30th 2016. Contributed talk on the published paper “On moment based density approximations for aggregate losses”.


• The 1st International Conference on Econometrics and Statistics (EcoSta 2017), The Hong Kong University of Science and Technology, Hong Kong, June 15th – 17th 2017. Contributed talk on research entitled “Bias reduction in the population size estimation of big data sets”.

• IBM UK, STFC Daresbury Laboratory, Daresbury, Cheshire, UK, August 22nd 2017. Seminar entitled “Statistical methods and distribution theory with applications to finance and large data sets”.


1.5 Future work

• The moments based density approximations discussed in Chapter 3 consider only the univariate case of a sum of a single set of losses or insurance claims. However, this could be extended further to model the compound sums of independent and identical bivariate or multivariate random vectors of insurance claims. An insurer will likely deal with multiple sources or types of insurance losses, thus modelling the multivariate cases would allow one to analyse the interdependency of different sources of risk. For example, for vehicle insurers claims may come from individuals who are not the insured, from damages to the insured, or from property damage of those who are not the insured etc.

• The population size estimator proposed in Chapter 4 is only one of many possible estimators which could be derived. Further work could be to develop an even more accurate estimation method which could be generalised for use with both small and large samples from big data, and also for use across many different data sets or types originating from various sources.

• Chapter 5 provides a preliminary analysis of the network structure of UK stock markets, and future work could be to extend this analysis further to get a deeper understanding of
the underlying network characteristics. For example, analysing if (and how) the statistical properties and network structure of the UK stock market changes over time, and answering questions such as: which are the financial instruments that have the largest influence in the market? do the largest influencers change over time? how do instruments in one sector influence those in the same or different sectors? how do these properties and structures compare with those from other developed or developing countries?

- A direct extension to the analysis in Chapter 6 could be to look at alternative ways of modelling the prices and returns of cryptocurrencies, for example GARCH type models where the volatility can depend on covariates such as time. Indeed, cryptocurrencies appear to be much more volatile than, say, traditional fiat currencies, and arguably using monthly, weekly, and even daily data may not be enough. It is not uncommon for cryptocurrencies to exhibit sharp changes in the price over the course a 24 hour period, however, this may be masked by simply looking at daily data which may show little or no change at all over one day. Therefore, future work could also be to analyse higher frequency price and returns data for cryptocurrencies, such as 1-hour, 2-hour, 5-hour etc. returns, to see whether the same results hold as for traditional financial instruments.
Bibliography


Chapter 2

A saddlepoint approximation to the distribution of the sum of independent non-identical beta random variables

This chapter has been published as Nadarajah, S., Jiang, X. and Chu, J., 2015. A saddlepoint approximation to the distribution of the sum of independent non-identically beta random variables. Statistica Neerlandica, 69, pp. 102 – 114.

2.1 Introduction

Let $X_i, i = 1, 2, \ldots, n$ be independent beta random variables with shape parameters $(a_i, b_i)$, i.e., the probability density function of $X_i$ is given by

$$f_{X_i}(x) = \frac{x^{a_i-1}(1-x)^{b_i-1}}{B(a_i, b_i)}$$

for $0 < x < 1, a_i > 0$ and $b_i > 0$, where $B(a, b)$ denotes the beta function defined by

$$B(a, b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt.$$
and let \( S = X_1 + X_2 + \cdots + X_n \) denote the sum.

The sums of random variables crop up frequently in financial risk management, common examples include the use of linear combinations of random variables (with fixed weights) to represent the value of portfolios of individual financial assets in terms of their individual returns (Gulisashvili, 2016) i.e. the minimum variance portfolio, and modelling of products or items to be bought. The linear combinations of these random variables are then used as objective functions, to be minimised or maximised, in portfolio analysis (Lee et al., 2010).

The beta distribution also plays an important role in risk, for example in the modelling of recovery rates in credit risk, or in credit events where the losses given default can vary between 0 and 100% (Schroeck, 2002; Jiménez and Mencia, 2010). The beta distribution is very flexible and suitable for modelling these types of random variables, where they take values between zero and one. The distributions of linear combinations of, for example, correlated lognormal, Pareto, and gamma random variables have been studied extensively, and with regards to the approximation of the distribution of their sums. However, there appears to have been little work on the study of the distribution of the sums of beta random variables in relation to finance.

Outside of finance, beta random variables arise in many applied areas: analysis of PERT networks (Sculli and Wong, 1985); software reliability estimation (Adams, 1996); project management processes (Klein, 2000); maximal ratio combining in cellular MIMO-CDMA downlink systems (Ma et al., 2007); the impact of quantized channel feedback in guaranteeing secrecy with artificial noise (Liang et al., 2009); to mention but a few.

To date, no one has been able to derive the exact distribution of \( S \) for \( n > 2 \). Even in terms of approximations, only the normal approximation has been used, i.e.,

\[
Pr (S < x) \approx \Phi \left( \frac{x - \mu}{\sigma} \right) = A_N(x)
\]

(2.1)
say, where \( \Phi(\cdot) \) denotes the cumulative distribution function of a standard normal random variable,

\[
\mu = \sum_{i=1}^{n} \frac{a_i}{a_i + b_i}
\]

is the mean of \( S \), and

\[
\sigma^2 = \sum_{i=1}^{n} \frac{a_i b_i}{(a_i + b_i)^2 (a_i + b_i + 1)}
\]
is the variance of \( S \). For example, Klein (2000) states that “the project duration as a sum of beta-distributed variables is usually assumed to be approximately normally distributed”.

The aim of this chapter is to propose an approximation for the distribution of \( S \) that is better than the normal approximation. The approximation is based on saddlepoint methods and is motivated by Murakami (2014). Murakami (2014) proposed a saddlepoint approximation for the sum of independent and non-identical uniform random variables.

The contents of this chapter are organised as follows. The saddlepoint approximation for the distribution of \( S \) is derived in Section 2.2. Its relative performance versus the normal approximation is assessed via simulations in Section 2.3, and some conclusions are noted in Section 2.4.

### 2.2 Methodology

The moment generating function of \( X_i \) is (Johnson et al., 1994)

\[
M_{X_i}(s) = \frac{1}{1} F_1 \left( a_i; a_i + b_i; s \right),
\]

where \( \frac{1}{1} F_1 (a; b; x) \) denotes the confluent hypergeometric function defined by

\[
\frac{1}{1} F_1 (a; b; x) = \sum_{k=0}^{\infty} \frac{(a)_k x^k}{(b)_k k!},
\]

where \( (f)_k = f(f+1) \cdots (f+k-1) \) denotes the ascending factorial. Thus, the moment and cumulant generating functions of \( S \) are

\[
M_S(s) = \prod_{i=1}^{n} \frac{1}{1} F_1 \left( a_i; a_i + b_i; s \right)
\]

and

\[
\kappa_S(s) = \sum_{i=1}^{n} \log \left[ \frac{1}{1} F_1 \left( a_i; a_i + b_i; s \right) \right],
\]

respectively. The first four derivatives of \( \kappa_S(s) \) with respect to \( s \) are

\[
\kappa'_S(s) = \sum_{i=1}^{n} \frac{a_i}{a_i + b_i} \frac{1}{1} F_1 \left( a_i + 1; a_i + b_i + 1; s \right) \frac{1}{1} F_1 \left( a_i; a_i + b_i; s \right),
\]
CHAPTER 2. SADDLEPOINT APPROXIMATION FOR SUMS OF BETA RVS

\[ \kappa_S''(s) = \sum_{i=1}^{n} \frac{a_i (a_i + 1)}{(a_i + b_i) (a_i + b_i + 1)} \binom{F_1}{a_i + 2; a_i + b_i + 2; s} - \sum_{i=1}^{n} \frac{a_i^2}{(a_i + b_i)^2} \left[ \frac{F_1 (a_i + 1; a_i + b_i + 1; s)}{F_1 (a_i; a_i + b_i; s)} \right]^2, \]

\[ \kappa_S'''(s) = \sum_{i=1}^{n} \frac{a_i (a_i + 1) (a_i + 2)}{(a_i + b_i) (a_i + b_i + 1) (a_i + b_i + 2)} \binom{F_1}{a_i + 3; a_i + b_i + 3; s} - 3 \sum_{i=1}^{n} \frac{a_i^2 (a_i + 1)}{(a_i + b_i)^2 (a_i + b_i + 1)} \frac{F_1 (a_i + 1; a_i + b_i + 1; s) F_1 (a_i + 2; a_i + b_i + 2; s)}{[F_1 (a_i; a_i + b_i; s)]^2} + 2 \sum_{i=1}^{n} \frac{a_i^3 (a_i + 1)}{(a_i + b_i)^3 (a_i + b_i + 1)} \left[ \frac{F_1 (a_i + 1; a_i + b_i + 1; s)}{F_1 (a_i; a_i + b_i; s)} \right]^3, \]

and

\[ \kappa_S''''(s) = \sum_{i=1}^{n} \frac{a_i (a_i + 1) (a_i + 2) (a_i + 3)}{(a_i + b_i) (a_i + b_i + 1) (a_i + b_i + 2) (a_i + b_i + 3)} \binom{F_1}{a_i + 4; a_i + b_i + 4; s} - 4 \sum_{i=1}^{n} \frac{a_i^2 (a_i + 1) (a_i + 2)}{(a_i + b_i)^2 (a_i + b_i + 1) (a_i + b_i + 2)} \frac{F_1 (a_i + 3; a_i + b_i + 3; s) F_1 (a_i + 1; a_i + b_i + 1; s)}{[F_1 (a_i; a_i + b_i; s)]^2} + 12 \sum_{i=1}^{n} \frac{a_i^3 (a_i + 1)}{(a_i + b_i)^3 (a_i + b_i + 1)} \left[ \frac{F_1 (a_i + 1; a_i + b_i + 1; s)}{[F_1 (a_i; a_i + b_i; s)]^2} \right]^2 - 3 \sum_{i=1}^{n} \frac{a_i^4 (a_i + 1)^2}{(a_i + b_i)^4 (a_i + b_i + 1)^2} \left[ \frac{F_1 (a_i + 2; a_i + b_i + 2; s)}{F_1 (a_i; a_i + b_i; s)} \right]^2 - 6 \sum_{i=1}^{n} \frac{a_i^4 (a_i + 1)}{(a_i + b_i)^4} \left[ \frac{F_1 (a_i + 1; a_i + b_i + 1; s)}{F_1 (a_i; a_i + b_i; s)} \right]^4. \]

According to Daniels (1954, 1987), the first order approximation for the probability density function of \( S \) is

\[ f_S(x) = \left[ 2 \pi \kappa_S'' (\tilde{s}) \right]^{-1/2} \exp \left[ \kappa_S (\tilde{s}) - \tilde{s}x \right] \left\{ 1 + O \left( n^{-1} \right) \right\}, \]

where \( \tilde{s} \) is the root of \( \kappa_S'(s) = x \). Also according to Daniels (1954, 1987), the second order approximation for the probability density function of \( S \) is

\[ f_S(x) = \left[ 2 \pi \kappa_S'' (\tilde{s}) \right]^{-1/2} \exp \left[ \kappa_S (\tilde{s}) - \tilde{s}x \right] \left\{ 1 + \frac{1}{8} \frac{\kappa_S''' (\tilde{s})}{\kappa_S'' (\tilde{s})^2} \left[ \kappa_S (\tilde{s}) \right]^2 + O \left( n^{-2} \right) \right\}. \]
Therefore, a second order saddlepoint approximation for $\Pr(S < x)$ is

$$\Pr(S < x) \approx 1 - \int_{x}^{\infty} \left[ \frac{2\pi \kappa_S''(\hat{s})}{\kappa_S''(\hat{s})} \right]^{-1/2} \exp \left[ \frac{\kappa_S''(\hat{s})}{8 \kappa_S''(\hat{s})} \left( \frac{5}{24} \frac{\kappa_S'''(\hat{s})}{\kappa_S''(\hat{s})^3} \right) \right] dy = A_S(x) \quad (2.3)$$

say, where $\hat{s}$ is the root of $\kappa_S'(s) = y$. The integration in (2.3) can be performed numerically. In-built routines for the confluent hypergeometric function are widely available, and we used the function \texttt{genhypergeo} in the R contributed package \texttt{hypergeo} (Hankin, 2013; R Development Core Team, 2014).

The saddlepoint approach used in Murakami (2014) was different, as Murakami (2014) did not numerically integrate (2.2). Instead, an approximation of the right hand side of (2.2) was used to derive an expression for $\Pr(S < x)$. This may explain why the saddlepoint approximation in Murakami (2014) did not always perform better than the normal approximation. We feel that our approximation in (2.3) is much improved because it utilises the full power of (2.2).

### 2.3 Numerical study

Here, we compare the performances of the normal approximation in (2.1) and the saddlepoint approximation in (2.3) by simulation. We take sixty four different values for $a = (a_1, a_2, \ldots, a_n)$ and $b = (b_1, b_2, \ldots, b_n)$, and these values were grouped into sixteen cases: see Case 1 to Case 16. Each case gives four different values for $a$ and $b$ with $n = 6, 8, 10, 12$, as in Murakami (2014). The values for $a$ and $b$ were simulated from the same distributions considered in Murakami (2014): the uniform $[\alpha, \beta]$ distribution given by the probability density function

$$f(x) = \frac{1}{\beta - \alpha}$$

for $0 \leq \alpha \leq x \leq \beta$; the Weibull $(\alpha, \beta)$ distribution given by the probability density function

$$f(x) = \frac{\alpha x^{\alpha-1} \exp \left[ \left( \frac{x}{\beta} \right)^{\alpha} \right]}{\beta^\alpha}$$

for $x > 0, \alpha > 0$ and $\beta > 0$; the chi-squared $(\alpha)$ distribution given by the probability density function

$$f(x) = \frac{x^{\alpha/2-1} \exp(-x/2)}{2^{\alpha/2} \Gamma(\alpha/2)}$$
for \( x > 0 \) and \( \alpha > 0 \); the lognormal \((\alpha, \beta)\) distribution given by the probability density function

\[
f(x) = \frac{1}{\sqrt{2\pi}\beta x} \exp \left[ -\frac{(\log x - \alpha)^2}{2\beta^2} \right]
\]

for \( x > 0 \), \( \alpha > 0 \) and \( \beta > 0 \).

The sixteen cases can be summarised as follows:

- **Cases 1, 2, 3, 4:** \( a_i \) simulated from uniform \([0, 3]\), and \( b_i \) simulated from uniform \([0, 3]\), Weibull \((3, 5)\), chisquared \((3)\), and lognormal \((1, 3)\), respectively.
- **Cases 5, 6, 7, 8:** \( a_i \) simulated from Weibull \((3, 5)\), and \( b_i \) simulated from uniform \([0, 3]\), Weibull \((3, 5)\), chisquared \((3)\), and lognormal \((1, 3)\), respectively.
- **Cases 9, 10, 11, 12:** \( a_i \) simulated from chisquared \((3)\), and \( b_i \) simulated from uniform \([0, 3]\), Weibull \((3, 5)\), chisquared \((3)\), and lognormal \((1, 3)\), respectively.
- **Cases 13, 14, 15, 16:** \( a_i \) simulated from lognormal \((1, 3)\), and \( b_i \) simulated from uniform \([0, 3]\), Weibull \((3, 5)\), chisquared \((3)\), and lognormal \((1, 3)\), respectively.

Exact values of \( a_i \) and \( b_i \) used in each case can be found in Appendix A.
As mentioned in Section 2.1, the exact distribution is not known and so we approximate the exact distribution by simulating a sample of size one million from each $X_i$, generating a sample of size one million on $S$, and letting $F_{\text{sim}}$ denote the corresponding empirical cumulative distribution function.

Figure 2.1: Comparison of the relative errors (vertical axis) of the saddlepoint ($A_S$) and normal ($A_N$) approximations, versus $E_P = F_{\text{sim}}(x)$ (horizontal axis), for the sixteen cases. Cases are arranged as $\{1, 2, 3, 4\}$ (first row), $\{5, 6, 7, 8\}$ (second row), $\{9, 10, 11, 12\}$ (third row), and $\{13, 14, 15, 16\}$ (fourth row). Plots are coded as black for $n = 6$, red for $n = 8$, blue for $n = 10$, brown for $n = 12$; broken lines for $A_N$ and solid lines for $A_S$.

We need to know the exact distribution for $S$ to compare the performances of (2.1) and (2.3). As mentioned in Section 2.1, the exact distribution is not known and so we approximate the exact distribution by simulating a sample of size one million from each $X_i$, generating a sample of size one million on $S$, and letting $F_{\text{sim}}$ denote the corresponding empirical cumulative distribution function.

The performances of (2.1) and (2.3) were compared as follows: i) compute the relative error of
CHAPTER 2. SADDLEPOINT APPROXIMATION FOR SUMS OF BETA RVS

Page 34

\[ A_N \text{ as } |E_p - A_N(x)| / E_p \text{ at } x = F_{\text{sim}}^{-1}(E_p) \text{ for } E_p = 0.6, 0.7, 0.8, 0.9, 0.95, 0.975, 0.99; \]

ii) compute the relative error of \( A_S \) as \( |E_p - A_S(x)| / E_p \text{ at } x = F_{\text{sim}}^{-1}(E_p) \) for \( E_p = 0.6, 0.7, 0.8, 0.9, 0.95, 0.975, 0.99; \)

iii) plot the two relative errors versus \( E_p \) in log scale.

The plots for the sixteen cases are shown in Figure 2.1, and we can observe the following from the figure: the relative error for (2.3) is always less than that for (2.1); all of the relative errors are between \( 10^{-5} \) and \( 10^{-1} \); the relative errors generally decrease with increasing \( E_p \); the relative errors do not appear to show any pattern with respect to \( n \).

2.4 Conclusions

We have derived a second order saddlepoint approximation for the distribution of the sum of independent but non-identical beta random variables. We have argued that our approach for saddlepoint approximation improves upon the one presented in Murakami (2014). We have compared our approximation with the normal approximation by simulation, showing that the former always performs better. The simulation has presented sixty four comparisons of the saddlepoint and normal approximations, each corresponding to certain parameter values. Other parameter configurations not presented here showed the same result, i.e., the saddlepoint approximation performs better than the normal approximation.
Bibliography


Chapter 3

On moment based density approximations for aggregate losses

This chapter has been published as Nadarajah, S., Chu, J. and Jiang, X., 2016. On moment based density approximations for aggregate losses. *Journal of Computational and Applied Mathematics*, 298, pp. 152 – 166.

3.1 Introduction

Suppose $X_1, X_2, \ldots, X_N$ are insurance claims made over a fixed period, then the total claim over the period is

$$S = \sum_{i=1}^{N} X_i.$$ 

Closed form expressions for the distribution of $S$, the aggregate loss, are hard to find but are extremely important in insurance and actuarial science.

Aggregate losses have been widely examined in the insurance literature and determining the distribution of these aggregate losses is crucial to an insurer, for example in the pricing of insurance coverage (Heckman and Meyers, 1983; Klugman et al., 2008). Insurers offer coverage to many individuals, and all of these risks can conveniently be combined into an aggregate risk to be modelled. As the number of companies offering insurance coverage to consumers continues to increase, it is even more important to be able to compute accurate estimates of aggregate losses which result from
insurance contracts (Heckman and Meyers, 1983), in order to price insurance contracts competitively and attract consumers whilst also remaining profitable.

In particular, in rate making and reserving, the distribution of aggregate losses is needed to make predictions about future expected claims from groups of individuals in different classes of risk (Papush et al., 2001). Insurers also need to estimate the reserves required for their business, and these can be computed by using the percentiles of the distributions of aggregate losses. Furthermore, the financial risk of loss can be calculated from the aggregate loss distribution in terms of the Value at Risk and Expected Shortfall (Jin et al., 2016).

Jin et al. (2016) developed quite a novel moments based approximation based on the gamma distribution. By means of six examples, they showed that their approximations perform well. The six examples were: i) $X_1, X_2, \ldots, X_N$ are independent and identical gamma random variables and $N$ is a Poisson random variable independent of $X_1, X_2, \ldots, X_N$; ii) $X_1, X_2, \ldots, X_N$ are independent and identical inverse Gaussian random variables and $N$ is a Poisson random variable independent of $X_1, X_2, \ldots, X_N$; iii) $X_1, X_2, \ldots, X_N$ are independent and identical gamma random variables and $N$ is a negative binomial random variable independent of $X_1, X_2, \ldots, X_N$; iv) $X_1, X_2, \ldots, X_N$ are independent and identical inverse Gaussian random variables and $N$ is a negative binomial random variable independent of $X_1, X_2, \ldots, X_N$; v) $X_1, X_2, \ldots, X_N$ are independent and identical Pareto random variables and $N$ is a Poisson random variable independent of $X_1, X_2, \ldots, X_N$; vi) $X_1, X_2, \ldots, X_N$ are independent and identical Pareto random variables and $N$ is a negative binomial random variable independent of $X_1, X_2, \ldots, X_N$.

While showing that their approximations perform well, Jin et al. (2016) considered only one set of parameter values for each of the six examples. It is not clear if their approximations perform well for all possible values of the parameters. It is also not clear if other moments based approximations can perform better when parameter values are different. To investigate this, we revisit the first four of their examples and see how the gamma approximation compares to four other moments based approximations for a wide range of parameter values. The four other approximations are based on the exponentiated exponential (Gupta and Kundu, 1999), Weibull (Weibull, 1951), inverse Gaussian (Schrödinger, 1915) and lognormal distributions. We have not considered the last two of Jin et al. (2016)’s examples because the sum of independent and identical Pareto random variables does not have a closed form distribution function. The sum of independent and identical gamma random variables and the sum of independent and identical inverse Gaussian random variables do have closed form distribution functions.
The gamma distribution has an exponential tail while the inverse Gaussian distribution has a heavy tail. Both distributions are popular models in actuarial science, in particular for claims data. See Sections 5.2 and 5.4 of Kleiber and Kotz (2003) for details and references of applications.

The contents of this chapter are organised as follows. The gamma approximation due to Jin et al. (2016) and the four other moments based approximations are discussed in Section 3.2. The performances of these approximations for the first four examples in Jin et al. (2016) are compared in Section 3.3. Some concluding remarks are given in Section 3.4.

3.2 Approximations

The gamma approximation for the probability density function of $S$ proposed in Jin et al. (2016, Section 2) is

$$f_{\text{Approx, gamma}, t}(s) = \frac{s^{\alpha-1} \exp(-s/\theta)}{\theta^\alpha \Gamma(\alpha)} \sum_{r=0}^{t} c_r s^r,$$

where

$$\alpha = \frac{\mu^2(1)}{\mu(2) - \mu^2(1)}, \quad \theta = \frac{\mu(2) - \mu^2(1)}{\mu(1)},$$

and

$$
\begin{pmatrix}
c_0 \\
c_1 \\
\vdots \\
c_t
\end{pmatrix} = \begin{pmatrix} m_0 & \cdots & m_t \\ m_1 & \cdots & m_{t+1} \\ \vdots & \ddots & \vdots \\ m_t & \cdots & m_{2t} \end{pmatrix}^{-1} \begin{pmatrix} \mu(0) \\ \mu(1) \\ \vdots \\ \mu(t) \end{pmatrix},
$$

(3.2)

where

$$\mu(\alpha) = \left[ 1 - \Pr(N = 0) \right]^{-1} \frac{d^\alpha}{dt^\alpha} E \left[ \left\{ E(tX) \right\}^N \right] \bigg|_{t=0}$$

(3.3)

and $m_i = \theta^i \Gamma(\alpha + i) / \Gamma(\alpha)$. The inner expectation in (3.3) is with respect to $X$ and the outer expectation in (3.3) is with respect to $N$. As described in Jin et al. (2016, Section 2), $t$ is a positive integer determining the degree of accuracy of (3.1).
Following the method in Jin et al. (2016, Section 2), other approximations for the distributions of $S$ can be derived by replacing the two-parameter gamma probability density function, $s^{\alpha-1} \exp(-s/\theta)/[\theta^\alpha \Gamma(\alpha)]$, by some other two-parameter probability density function. We take that to be one of the two-parameter exponentiated exponential distribution, two-parameter Weibull distribution, two-parameter inverse Gaussian distribution or the two-parameter lognormal distribution. These and the gamma distribution are among the most popular two-parameter distributions defined on the positive half-line.

The resulting exponentiated exponential approximation is

$$f_{\text{Approx}, \text{ee}, t}(s) = \alpha \theta \left[ 1 - \exp(-\theta s) \right]^{\alpha-1} \exp(-\theta s) \sum_{r=0}^{t} c_r s^r, \quad (3.4)$$

where $c_r$ are given by (3.2) with

$$m_i = \frac{(-1)^i \alpha \Gamma(\alpha)}{\theta^i} \left. \frac{\partial^i}{\partial p^i} \left[ \frac{\Gamma(p+1-\alpha)}{\Gamma(p+1)} \right] \right|_{p=\alpha}$$

and $\mu(\alpha)$ are given by (3.3). In particular,

$$m_1 = \theta^{-1} [\psi(\alpha+1) - \psi(1)],$$

$$m_2 = \theta^{-2} \left\{ \psi'(1) - \psi'(\alpha+1) + [\psi(\alpha+1) - \psi(1)]^2 \right\},$$

where $\psi(x) = d \log \Gamma(x)/dx$ and $\psi'(x) = d\psi(x)/dx$. Solving $m_1 = \mu(1)$ and $m_2 = \mu(2)$, we obtain $\alpha$ as the root of

$$\sqrt{\frac{\psi'(1) - \psi'(x+1)}{\psi(x+1) - \psi(1)}} = \frac{\sqrt{\mu(2) - \mu^2(1)}}{\mu(1)},$$

and $\theta = [\psi(\alpha+1) - \psi(1)]/\mu(1)$.

The resulting Weibull approximation is

$$f_{\text{Approx}, \text{Weibull}, t}(s) = \alpha \theta^{-\alpha} s^{\alpha-1} \exp\left[-(s/\theta)^\alpha\right] \sum_{r=0}^{t} c_r s^r, \quad (3.5)$$

where $c_r$ are given by (3.2) with

$$m_i = \theta^i \Gamma(1 + i/\alpha)$$
and $\mu(\alpha)$ are given by (3.3). In particular,
\[ m_1 = \theta \Gamma \left( 1 + \frac{1}{\alpha} \right), \]
\[ m_2 = \theta^2 \Gamma \left( 1 + \frac{2}{\alpha} \right). \]

Solving $m_1 = \mu(1)$ and $m_2 = \mu(2)$, we obtain $\alpha$ as the root of
\[ \frac{\Gamma(1 + 2/x)}{[\Gamma(1 + 1/x)]^2} = \frac{\mu(2)}{\mu^2(1)} \]
and $\theta = \mu(1)/\Gamma(1 + 1/\alpha)$.

The resulting inverse Gaussian approximation is
\[ f_{\text{Approx,IG,t}}(s) = \sqrt{\lambda/2\pi s^3} \exp \left( -\lambda(s - u)^2 / 2su^2 \right) \sum_{r=0}^{t} c_r s^r, \tag{3.6} \]
where $c_r$ are given by (3.2) with
\[ m_i = \exp \left( \frac{\lambda}{u} \right) \sqrt{\frac{2\lambda}{u}} u^{-1/2} K_{i-1/2} \left( \frac{\lambda}{u} \right) \]
and $\mu(\alpha)$ are given by (3.3), where $K_{\nu}()$ denotes the modified Bessel function of the third kind of order $\nu$. In particular,
\[ m_1 = u, \]
\[ m_2 = \frac{u^3}{\lambda} + u^2. \]

Solving $m_1 = \mu(1)$ and $m_2 = \mu(2)$, we obtain $u = \mu(1)$ and $\lambda = \mu^3(1) / [\mu(2) - \mu^2(1)]$.

The resulting lognormal approximation is
\[ f_{\text{Approx,LN,t}}(s) = \frac{1}{\sqrt{2\pi} \sigma s} \exp \left( -\frac{(\log s - u)^2}{2\sigma^2} \right) \sum_{r=0}^{t} c_r s^r, \tag{3.7} \]
where $c_r$ are given by (3.2) with
\[ m_i = \exp (iu + i^2 \sigma^2 / 2) \]
and $\mu(a)$ are given by (3.3). In particular,

\[
\begin{align*}
    m_1 &= \exp \left( u + \frac{\sigma^2}{2} \right), \\
    m_2 &= \exp \left( 2u + 2\sigma^2 \right).
\end{align*}
\]

Solving $m_1 = \mu(1)$ and $m_2 = \mu(2)$, we obtain $\sigma = \sqrt{\log\mu(2) - 2\log\mu(1)}$ and $u = 2\log\mu(1) - 2^{-1}\log\mu(2)$.

## 3.3 The four examples

Here, we compare the performances of the five approximations given by (3.1), (3.4), (3.5), (3.6), (3.7) with respect to the first four examples in Jin et al. (2016). We also compare the approximations versus the truncated version of the exact distribution of $S$. A wide range of parameter values is considered for each of the examples. For examples 1 and 2, we let each parameter take the values \{1, 2, ..., 20\}, so a total of $20^3 = 8000$ sets of parameter values were considered. For examples 3 and 4, we let each parameter take the values \{1, 2, ..., 20\}, so a total of $20^4 = 160000$ sets of parameter values were considered. The reported results were similar for other parameter values not considered here. All computations were performed in Maple.

### 3.3.1 Example 1

In this example, $X_i$ are independent and identical gamma ($\alpha, \theta$) random variables and $N$ is a Poisson ($\lambda$) random variable independent of the $X_i$. So, the exact probability density function of $S$ is

\[
    f_{\text{Exact}}(s) = \sum_{n=0}^{\infty} \frac{\exp(-\lambda) \lambda^n s^{n\alpha-1} \exp(-s/\theta)}{n! \Gamma(n\alpha) \theta^{n\alpha}}.
\]

The approximations given by (3.1), (3.4), (3.5), (3.6), (3.7) can be computed by knowing

\[
    \mu(a) = \left. \frac{1}{1 - \exp(-\lambda)} \frac{d^{\alpha}}{dt^{\alpha}} \left[ \exp \left\{ \lambda \left[ (1 - \theta t)^{-\alpha} - 1 \right] \right\} \right] \right|_{t=0}.
\]

We computed the discrepancy measures

\[
    \Delta_f(\alpha, \theta, \lambda) = \min_{1 \leq t \leq 100} \int_{\min} \left| f_{\text{Exact}}(s) - f_{\text{Approx},f,t}(s) \right| ds,
\]
CHAPTER 3. MOMENT BASED DENSITY APPROXIMATIONS

\[ \Omega_f(\alpha, \theta, \lambda) = \min_{1 \leq t \leq 100} \int \left| \frac{f_{\text{Approx}, f, t}(s)}{f_{\text{Exact}}(s)} - 1 \right| ds, \]

\[ \Delta'(\alpha, \theta, \lambda) = \min_{1 \leq t \leq 100} \int \left| f_{\text{Exact}}(s) - \sum_{n=0}^{t} \frac{\exp(-\lambda)\lambda^n s^{n\alpha-1} \exp(-s/\theta)}{n! \Gamma(n\alpha) \theta^{n\alpha}} \right| ds, \]

and

\[ \Omega'(\alpha, \theta, \lambda) = \min_{1 \leq t \leq 100} \int \left| \frac{1}{f_{\text{Exact}}(s)} \sum_{n=0}^{t} \frac{\exp(-\lambda)\lambda^n s^{n\alpha-1} \exp(-s/\theta)}{n! \Gamma(n\alpha) \theta^{n\alpha}} - 1 \right| ds \]

for \( f = \text{gamma}, \text{ee}, \text{Weibull}, \text{IG}, \text{LN} \) for \( \alpha = 1, 2, \ldots, 20, \theta = 1, 2, \ldots, 20 \) and \( \lambda = 1, 2, \ldots, 20 \).

Table 3.1 gives the number and percentage of times \( \Delta_f(\alpha, \theta, \lambda) \) was smallest among \( \Delta_{\text{gamma}}(\alpha, \theta, \lambda), \Delta_{\text{ee}}(\alpha, \theta, \lambda), \Delta_{\text{Weibull}}(\alpha, \theta, \lambda), \Delta_{\text{IG}}(\alpha, \theta, \lambda) \) and \( \Delta_{\text{LN}}(\alpha, \theta, \lambda) \). Table 3.1 also gives the number and percentage of times \( \Omega_f(\alpha, \theta, \lambda) \) was smallest among \( \Omega_{\text{gamma}}(\alpha, \theta, \lambda), \Omega_{\text{ee}}(\alpha, \theta, \lambda), \Omega_{\text{Weibull}}(\alpha, \theta, \lambda), \Omega_{\text{IG}}(\alpha, \theta, \lambda) \) and \( \Omega_{\text{LN}}(\alpha, \theta, \lambda) \). Sometimes the discrepancy measure values corresponding to two or more of gamma, ee, Weibull, IG, LN share the smallest value.

In this example and the three subsequent examples, we computed \( \Delta_f, \Omega_f, \Delta'_f \) and \( \Omega'_f \) using values of \( 1 \leq t \leq 100 \) for all possible combinations of parameter values. We note that we did not strictly employ specific convergence criteria for determining the convergence of the errors as we increased the value of \( t \). Instead, the criteria used was to simply select the minimum values of \( \Delta_f, \Omega_f, \Delta'_f \) and \( \Omega'_f \), computed using values of \( t \) in the range of \( 1 \leq t \leq 100 \). In the majority of cases, the smallest values of \( \Delta_f \) and \( \Omega_f \) were found at \( t < 50 \), however, in some cases the smallest values were found at \( t \) as large as 80.

This is in contrast to Jin et al. (2016), who used an alternative method for selecting the value of \( t \). As their examples only considered a single set of parameter values, they suggest plotting the density function of the approximation for a range of values of \( t \) until there is no significant difference between the approximation using \( t \) and \( t + 1 \). Another method that they suggest is choosing the value of \( t \) such that the integral of the squared difference in the density of the approximations using \( t \) and \( t + \Delta t \), for a small positive integer \( \Delta t \), is less than a chosen threshold value \( \epsilon \).

Table 3.1 also gives the order of the smallest values for \( \Delta_f(\alpha, \theta, \lambda) \) and \( \Omega_f(\alpha, \theta, \lambda) \). The orders of \( \Delta'_f(\alpha, \theta, \lambda) \) and \( \Omega'_f(\alpha, \theta, \lambda) \) were computed as \( 10^{-5} \) and \( 10^{-4} \), respectively, so it is pleasing that the moments based approximations perform better. The values of \( t \) needed for \( \Delta'_f(\alpha, \theta, \lambda) \) and \( \Omega'_f(\alpha, \theta, \lambda) \) to be less than \( 10^{-8} \) and \( 10^{-7} \), respectively, were at least 300 and 330, respectively.
Table 3.1: Comparison of the performances of the five approximations under gamma $(\alpha,\theta)$ distributed claim amounts and Poisson $(\lambda)$ distributed claim numbers, shown by the number/percentage of combinations of parameter values tested where the approximation(s) $(f)$ gave the smallest absolute $(\Delta_f)$ or relative $(\Omega_f)$ error.

We see that the Weibull (absolute and relative errors) approximation gives the most number of best fits. The second most number of best fits is by the gamma (absolute error) and exponentiated exponential (relative error) approximations. The third most number of best fits is by the lognormal (absolute error) and gamma (relative error) approximations. The fourth most number of best fits is by the exponentiated exponential (absolute error) and lognormal (relative error) approximations. The fifth most number of best fits is by the inverse Gaussian (absolute and relative errors) approximation.
Figure 3.1: The combinations of the values of parameters ($\alpha, \theta, \lambda$), indicated by dots, for which the approximations based on the gamma (top left), exponentiated exponential (top right), Weibull (middle left), inverse Gaussian (middle right) and lognormal (bottom left) distributions gave the smallest absolute error ($\Delta f$).
Figure 3.2: The combinations of the values of parameters \((\alpha, \theta, \lambda)\), indicated by dots, for which the approximations based on the gamma (top left), exponentiated exponential (top right), Weibull (bottom left) and lognormal (bottom right) distributions gave the smallest relative error \((\Omega_f)\).

Figure 3.1 plots the values of \((\alpha, \theta, \lambda) \in \{1,2,\ldots,20\}^3\) for which \(\Delta_f\) is smallest for \(f = \text{gamma, exponentiated exponential, Weibull, inverse Gaussian and lognormal}\). Figure 3.2 plots the values of \((\alpha, \theta, \lambda) \in \{1,2,\ldots,20\}^3\) for which \(\Omega_f\) is smallest for \(f = \text{gamma, exponentiated exponential, Weibull and lognormal}\). The best fits of the gamma approximation correspond mainly to large values of \(\lambda\) (absolute error) and small values of \(\lambda\) (relative error). The best fits of the exponentiated exponential approximation correspond mainly to \(\lambda\) being close to zero (absolute and relative errors). The best fits of the Weibull approximation correspond mainly to not large values of \(\lambda\) (absolute error) and not small values of \(\alpha\) and \(\lambda\) (relative error). The best fits of the inverse Gaussian and lognormal approximations correspond mainly to \(\lambda\) being close to zero and \(\alpha\) being large (absolute and relative
3.3.2 Example 2

In this example, $X_i$ are independent and identical inverse Gaussian $(\eta, \theta)$ random variables and $N$ is a Poisson ($\lambda$) random variable independent of the $X_i$. So, the exact probability density function of $S$ is

$$f_{\text{Exact}}(s) = \sum_{n=0}^{\infty} \frac{\exp(-\lambda)\lambda^n}{n!} \sqrt{\frac{n^2\theta}{2\pi s^3}} \exp \left[ -\frac{n^2\theta(s-n\eta)^2}{2n^2\eta^2s} \right]. \quad (3.9)$$

The approximations given by (3.1), (3.4), (3.5), (3.6), (3.7) can be computed by knowing

$$\mu(a) = \frac{1}{1-\exp(-\lambda)} \frac{d^n}{dt^n} \left[ \exp \left\{ \lambda \left[ \exp \left\{ \frac{\theta}{\eta} \left[ 1 - \sqrt{1 - \frac{2t\eta^2}{\theta}} \right] \right] - 1 \right\} \right\} \right]_{t=0}.$$ 

We computed the discrepancy measures

$$\Delta_f(\eta, \theta, \lambda) = \min_{1 \leq t \leq 100} \int |f_{\text{Exact}}(s) - f_{\text{Approx},f,t}(s)| \, ds,$$

$$\Omega_f(\eta, \theta, \lambda) = \min_{1 \leq t \leq 100} \int \left| \frac{f_{\text{Approx},f,t}(s)}{f_{\text{Exact}}(s)} - 1 \right| \, ds,$$

$$\Delta'(\eta, \theta, \lambda) = \min_{1 \leq t \leq 100} \int \left| f_{\text{Exact}}(s) - \sum_{n=0}^{t} \frac{\exp(-\lambda)\lambda^n}{n!} \sqrt{\frac{n^2\theta}{2\pi s^3}} \exp \left[ -\frac{n^2\theta(s-n\eta)^2}{2n^2\eta^2s} \right] \right| \, ds$$

and

$$\Omega'(\eta, \theta, \lambda) = \min_{1 \leq t \leq 100} \int \left| \frac{1}{f_{\text{Exact}}(s)} \sum_{n=0}^{t} \frac{\exp(-\lambda)\lambda^n}{n!} \sqrt{\frac{n^2\theta}{2\pi s^3}} \exp \left[ -\frac{n^2\theta(s-n\eta)^2}{2n^2\eta^2s} \right] - 1 \right| \, ds$$

for $f = \text{gamma}, \text{ee}, \text{Weibull}, \text{IG}, \text{LN}$ for $\eta = 1, 2, \ldots, 20$, $\theta = 1, 2, \ldots, 20$ and $\lambda = 1, 2, \ldots, 20$.

Table 3.2 gives the number and percentage of times $\Delta_f(\eta, \theta, \lambda)$ was smallest among $\Delta_{\text{gamma}}(\eta, \theta, \lambda)$, $\Delta_{\text{ee}}(\eta, \theta, \lambda)$, $\Delta_{\text{Weibull}}(\eta, \theta, \lambda)$, $\Delta_{\text{IG}}(\eta, \theta, \lambda)$ and $\Delta_{\text{LN}}(\eta, \theta, \lambda)$. Table 3.2 also gives the number and percentage of times $\Omega_f(\eta, \theta, \lambda)$ was smallest among $\Omega_{\text{gamma}}(\eta, \theta, \lambda)$, $\Omega_{\text{ee}}(\eta, \theta, \lambda)$, $\Omega_{\text{Weibull}}(\eta, \theta, \lambda)$, $\Omega_{\text{IG}}(\eta, \theta, \lambda)$ and $\Omega_{\text{LN}}(\eta, \theta, \lambda)$.

Table 3.2 also gives the order of the smallest values for $\Delta_f(\eta, \theta, \lambda)$ and $\Omega_f(\eta, \theta, \lambda)$. The orders
of $\Delta'(\eta, \theta, \lambda)$ and $\Omega'(\eta, \theta, \lambda)$ were computed as $10^{-4}$ and $10^{-4}$, respectively, so it is pleasing that the moments based approximations again perform better. The values of $t$ needed for $\Delta'(\eta, \theta, \lambda)$ and $\Omega'(\eta, \theta, \lambda)$ to be less than $10^{-8}$ and $10^{-7}$, respectively, were at least 300 and 360, respectively.

<table>
<thead>
<tr>
<th>$f$</th>
<th>No of times of $\Delta_f$ is smallest</th>
<th>Percentage of times of $\Delta_f$ is smallest</th>
<th>Order of $\Delta_f$</th>
<th>No of times of $\Omega_f$ is smallest</th>
<th>Percentage of times of $\Omega_f$ is smallest</th>
<th>Order of $\Omega_f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>gamma</td>
<td>3956</td>
<td>49.5</td>
<td>$10^{-8}$</td>
<td>1515</td>
<td>18.9</td>
<td>$10^{-7}$</td>
</tr>
<tr>
<td>exponentiated exponential</td>
<td>985</td>
<td>12.3</td>
<td>$10^{-9}$</td>
<td>670</td>
<td>8.4</td>
<td>$10^{-7}$</td>
</tr>
<tr>
<td>Weibull</td>
<td>2704</td>
<td>33.8</td>
<td>$10^{-8}$</td>
<td>5814</td>
<td>72.7</td>
<td>$10^{-8}$</td>
</tr>
<tr>
<td>inverse Gaussian</td>
<td>35</td>
<td>0.4</td>
<td>$10^{-9}$</td>
<td>0</td>
<td>0.0</td>
<td>-</td>
</tr>
<tr>
<td>lognormal</td>
<td>320</td>
<td>4.0</td>
<td>$10^{-9}$</td>
<td>1</td>
<td>0.0</td>
<td>$10^{-7}$</td>
</tr>
</tbody>
</table>

Table 3.2: Comparison of the performances of the five approximations under inverse Gaussian ($\eta, \theta$) distributed claim amounts and Poisson ($\lambda$) distributed claim numbers, shown by the number/percentage of combinations of parameter values tested where the approximation(s) ($f$) gave the smallest absolute ($\Delta_f$) or relative ($\Omega_f$) error.

We see that the gamma (absolute error) and Weibull (relative error) approximations give the most number of best fits. The second most number of best fits is by the Weibull (absolute error) and gamma (relative error) approximations. The third most number of best fits is by the exponentiated exponential (absolute and relative errors) approximation. The fourth most number of best fits is by the lognormal (absolute and relative errors) approximation. The fifth most number of best fits is by the inverse Gaussian (absolute and relative errors) approximation.
Figure 3.3: The combinations of the values of parameters ($\eta, \theta, \lambda$), indicated by dots, for which the approximations based on the gamma (top left), exponentiated exponential (top right), Weibull (middle left), inverse Gaussian (middle right) and lognormal (bottom left) distributions gave the smallest absolute error ($\Delta f$).
Figure 3.4: The combinations of the values of parameters \((\eta, \theta, \lambda)\), indicated by dots, for which the approximations based on the gamma (top left), exponentiated exponential (top right) and Weibull (bottom left) distributions gave the smallest relative error \((\Omega_f)\).

Figure 3.3 plots the values of \((\eta, \theta, \lambda) \in \{1, 2, \ldots, 20\}^3\) for which \(\Delta_f\) is smallest for \(f = \text{gamma, exponentiated exponential, Weibull, inverse Gaussian and lognormal}\). Figure 3.4 plots the values of \((\eta, \theta, \lambda) \in \{1, 2, \ldots, 20\}^3\) for which \(\Omega_f\) is smallest for \(f = \text{gamma, exponentiated exponential and Weibull}\). The best fits of the gamma approximation correspond mainly to not small values of \(\lambda\) (absolute error) and small values of \(\eta\) (relative error). The best fits of the exponentiated exponential approximation correspond mainly to small values of \(\theta\) (absolute and relative errors). The best fits of the Weibull approximation correspond mainly to small values of \(\lambda\) (absolute error) and \(\lambda\) and \(\eta\) not being close to zero (relative error). The best fits of the inverse Gaussian approximation correspond
mainly to $\lambda$ being close to zero (absolute error). The best fits of the lognormal approximation correspond mainly to $\theta$ being close to zero (absolute error).

### 3.3.3 Example 3

In this example, $X_i$ are independent and identical gamma ($\alpha, \theta$) random variables and $N$ is a negative binomial ($r, \beta$) random variable independent of the $X_i$. So, the exact probability density function of $S$ is

$$f_{\text{Exact}}(s) = \sum_{n=0}^{\infty} \binom{n + r - 1}{n} \left( \frac{1}{1 + \beta} \right)^n \left( \frac{\beta}{1 + \beta} \right)^{n\alpha - 1} \frac{\exp(-s/\theta)}{\Gamma(n\alpha)\theta^n}.$$

(3.10)

The approximations given by (3.1), (3.4), (3.5), (3.6), (3.7) can be computed by knowing

$$\mu(a) = \frac{1}{1 - (1 + \beta)^{-r}} \frac{d^a}{dt^a} \left\{ [1 + \beta - \beta(1 - \theta t)^{-\alpha}]^{-r} \right\} \bigg|_{t=0}.$$

We computed the discrepancy measures

$$\Delta_f(\alpha, \theta, r, \beta) = \min_{1 \leq t \leq 100} \int |f_{\text{Exact}}(s) - f_{\text{Approx}, f, t}(s)| \, ds,$$

$$\Omega_f(\alpha, \theta, r, \beta) = \min_{1 \leq t \leq 100} \int \left| \frac{f_{\text{Approx}, f, t}(s)}{f_{\text{Exact}}(s)} - 1 \right| \, ds,$$

$$\Delta'(\alpha, \theta, r, \beta) = \min_{1 \leq t \leq 100} \int \left| f_{\text{Exact}}(s) - \sum_{n=0}^{t} \binom{n + r - 1}{n} \left( \frac{1}{1 + \beta} \right)^n \left( \frac{\beta}{1 + \beta} \right)^{n\alpha - 1} \frac{\exp(-s/\theta)}{\Gamma(n\alpha)\theta^n} \right| \, ds$$

and

$$\Omega'(\alpha, \theta, r, \beta) = \min_{1 \leq t \leq 100} \int \left| \frac{1}{f_{\text{Exact}}(s)} \sum_{n=0}^{t} \binom{n + r - 1}{n} \left( \frac{1}{1 + \beta} \right)^n \left( \frac{\beta}{1 + \beta} \right)^{n\alpha - 1} \frac{\exp(-s/\theta)}{\Gamma(n\alpha)\theta^n} - 1 \right| \, ds$$

for $f = \text{gamma, ec, Weibull, IG, LN}$ for $\alpha = 1, 2, \ldots, 20$, $\theta = 1, 2, \ldots, 20$, $r = 1, 2, \ldots, 20$ and $\beta = 1, 2, \ldots, 20$. Table 3.3 gives the number and percentage of times $\Delta_f(\alpha, \theta, r, \beta)$ was smallest among $\Delta_{\text{gamma}}(\alpha, \theta, r, \beta)$, $\Delta_{\text{ec}}(\alpha, \theta, r, \beta)$, $\Delta_{\text{Weibull}}(\alpha, \theta, r, \beta)$, $\Delta_{\text{IG}}(\alpha, \theta, r, \beta)$ and $\Delta_{\text{LN}}(\alpha, \theta, r, \beta)$. Table 3.3 also gives the number and percentage of times $\Omega_f(\alpha, \theta, r, \beta)$ was smallest among $\Omega_{\text{gamma}}(\alpha, \theta, r, \beta)$, $\Omega_{\text{ec}}(\alpha, \theta, r, \beta)$, $\Omega_{\text{Weibull}}(\alpha, \theta, r, \beta)$, $\Omega_{\text{IG}}(\alpha, \theta, r, \beta)$ and $\Omega_{\text{LN}}(\alpha, \theta, r, \beta)$. The last four rows of Table 3.3 account for two or more of gamma, ec, Weibull, IG, LN sharing the smallest discrepancy measure.
value.

Table 3.3 also gives the order of the smallest values for $\Delta_f(\alpha, \theta, r, \beta)$ and $\Omega_f(\alpha, \theta, r, \beta)$. The orders of $\Delta'(\alpha, \theta, r, \beta)$ and $\Omega'(\alpha, \theta, r, \beta)$ were computed as $10^{-6}$ and $10^{-5}$, respectively, so it is pleasing that the moments based approximations again perform better. The values of $t$ needed for $\Delta'(\alpha, \theta, r, \beta)$ and $\Omega'(\alpha, \theta, r, \beta)$ to be less than $10^{-8}$ and $10^{-7}$, respectively, were at least 310 and 350, respectively.

<table>
<thead>
<tr>
<th>$f$</th>
<th>No of times $\Delta_f$ is smallest</th>
<th>Percentage $\Delta_f$ is smallest</th>
<th>Order $\Delta_f$</th>
<th>No of times $\Omega_f$ is smallest</th>
<th>Percentage $\Omega_f$ is smallest</th>
<th>Order $\Omega_f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>gamma</td>
<td>72087</td>
<td>45.1</td>
<td>$10^{-8}$</td>
<td>6072</td>
<td>3.8</td>
<td>$10^{-7}$</td>
</tr>
<tr>
<td>exponentiated exponential</td>
<td>16156</td>
<td>10.1</td>
<td>$10^{-8}$</td>
<td>7522</td>
<td>4.7</td>
<td>$10^{-8}$</td>
</tr>
<tr>
<td>Weibull</td>
<td>20580</td>
<td>12.9</td>
<td>$10^{-9}$</td>
<td>143239</td>
<td>89.5</td>
<td>$10^{-8}$</td>
</tr>
<tr>
<td>inverse Gaussian</td>
<td>2849</td>
<td>1.8</td>
<td>$10^{-9}$</td>
<td>37</td>
<td>0.0</td>
<td>$10^{-8}$</td>
</tr>
<tr>
<td>lognormal</td>
<td>6176</td>
<td>3.9</td>
<td>$10^{-9}$</td>
<td>3099</td>
<td>1.9</td>
<td>$10^{-7}$</td>
</tr>
<tr>
<td>IG, LN</td>
<td>432</td>
<td>0.3</td>
<td>$10^{-8}$</td>
<td>0</td>
<td>0.0</td>
<td>-</td>
</tr>
<tr>
<td>gamma, IG, LN</td>
<td>10427</td>
<td>6.5</td>
<td>$10^{-9}$</td>
<td>0</td>
<td>0.0</td>
<td>-</td>
</tr>
<tr>
<td>ee, IG, LN</td>
<td>15</td>
<td>0.0</td>
<td>$10^{-9}$</td>
<td>0</td>
<td>0.0</td>
<td>-</td>
</tr>
<tr>
<td>gamma, ee, IG, LN</td>
<td>31033</td>
<td>19.4</td>
<td>$10^{-9}$</td>
<td>0</td>
<td>0.0</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 3.3: Comparison of the performances of the five approximations under gamma ($\alpha, \theta$) distributed claim amounts and negative binomial ($r, \beta$) distributed claim numbers, shown by the number/percentage of combinations of parameter values tested where the approximation(s) ($f$) gave the smallest absolute ($\Delta_f$) or relative ($\Omega_f$) error.

We see that the gamma (absolute error) and Weibull (relative error) approximations give the most number of best fits. The second most number of best fits is by the Weibull (absolute error) and exponentiated exponential (relative error) approximations. The third most number of best fits is by the exponentiated exponential (absolute error) and gamma (relative error) approximations. The fourth most number of best fits is by the lognormal (absolute and relative errors) approximation. The fifth most number of best fits is by the inverse Gaussian (absolute and relative errors) approximation. Counting also cases where the smallest value is shared by more than one approximation, the gamma, exponentiated exponential, inverse Gaussian and lognormal approximations together give a better fit than the Weibull approximation with respect to absolute errors.
Figure 3.5: The combinations of the values of parameters \((\alpha, r, \beta)\) from \((\alpha, \theta, r, \beta)\), indicated by dots, for which the approximations based on the exponentiated exponential (top left), Weibull (top right), inverse Gaussian (bottom left) and lognormal (bottom right) distributions gave the smallest absolute error \((\Delta f)\).
Figure 3.6: The combinations of the values of parameters $(\alpha, r, \beta)$ from $(\alpha, \theta, r, \beta)$, indicated by dots, for which the approximations based on the gamma (top left), exponentiated exponential (top right), inverse Gaussian (bottom left) and lognormal (bottom right) distributions gave the smallest relative error ($\Omega_f$).

Figure 3.5 plots the values of $(\alpha, r, \beta) \in \{1, 2, \ldots, 20\}^3$ for which $\Delta_f$ is smallest for $f = \text{exponentiated exponential, Weibull, inverse Gaussian and lognormal}$. Figure 3.6 plots the values of $(\alpha, r, \beta) \in \{1, 2, \ldots, 20\}^3$ for which $\Omega_f$ is smallest for $f = \text{gamma, exponentiated exponential, inverse Gaussian and lognormal}$. Some plots were omitted due to a lack of distinguishable pattern, and plots of other three-dimensional marginals of $(r, \beta, \alpha, \theta) \in \{1, 2, \ldots, 20\}^4$ showed similar results. The best fits of the gamma approximation correspond mainly to small values of $r$ (relative error). The best fits of the exponentiated exponential approximation correspond mainly to not large values of $r$ (absolute and relative errors). The best fits of the Weibull approximation correspond mainly to either small
values of $r$ or small values of $\beta$ (absolute error). The best fits of the inverse Gaussian and lognormal approximations correspond mainly to small values of $r$ (absolute and relative errors).

3.3.4 Example 4

In this example, $X_i$ are independent and identical inverse Gaussian $(\eta, \theta)$ random variables and $N$ is a negative binomial $(r, \beta)$ random variable independent of the $X_i$. So, the exact probability density function of $S$ is

$$f_{\text{Exact}}(s) = \sum_{n=0}^{\infty} \binom{n+r-1}{n} \left( \frac{1}{1+\beta} \right)^r \left( \frac{\beta}{1+\beta} \right)^n \sqrt{\frac{n^2\theta}{2\pi s^3}} \exp \left[ -\frac{n^2\theta(s-n\eta)^2}{2n^2\eta^2s} \right]. \quad (3.11)$$

The approximations given by (3.1), (3.4), (3.5), (3.6), (3.7) can be computed by knowing

$$\mu(a) = \frac{1}{1 - (1 + \beta)^{-r}} \frac{d^n}{dt^n} \left\{ \left[ 1 + \beta - \beta \exp \left( \frac{\theta}{\eta} \left[ 1 - \sqrt{1 - 2\eta^2t^{\frac{1}{\theta}}} \right] \right) \right]^{-r} \right\} \bigg|_{t=0}.$$

We computed the discrepancy measures

$$\Delta f(\eta, \theta, r, \beta) = \min_{1 \leq t \leq 100} \int \left| f_{\text{Exact}}(s) - f_{\text{Approx}, f, t}(s) \right| ds,$$

$$\Omega f(\eta, \theta, r, \beta) = \min_{1 \leq t \leq 100} \int \left| \frac{f_{\text{Approx}, f, t}(s)}{f_{\text{Exact}}(s)} - 1 \right| ds,$$

$$\Delta'(\eta, \theta, r, \beta) = \min_{1 \leq t \leq 100} \int \left| f_{\text{Exact}}(s) - \sum_{n=0}^{t} \binom{n+r-1}{n} \left( \frac{1}{1+\beta} \right)^r \right. \left. \cdot \left( \frac{\beta}{1+\beta} \right)^n \sqrt{\frac{n^2\theta}{2\pi s^3}} \exp \left[ -\frac{n^2\theta(s-n\eta)^2}{2n^2\eta^2s} \right] \right| ds$$

and

$$\Omega'(\eta, \theta, r, \beta) = \min_{1 \leq t \leq 100} \int \left| \frac{1}{f_{\text{Exact}}(s)} \sum_{n=0}^{t} \binom{n+r-1}{n} \left( \frac{1}{1+\beta} \right)^r \right. \left. \cdot \left( \frac{\beta}{1+\beta} \right)^n \sqrt{\frac{n^2\theta}{2\pi s^3}} \exp \left[ -\frac{n^2\theta(s-n\eta)^2}{2n^2\eta^2s} \right] - 1 \right| ds.$$
for \( f = \text{gamma, ee, Weibull, IG, LN} \) for \( \eta, \theta, r, \beta = 1, 2, \ldots, 20 \). Table 3.4 gives the number and percentage of times \( \Delta_f(\eta, \theta, r, \beta) \) was smallest among \( \Delta_{\text{gamma}}(\eta, \theta, r, \beta), \Delta_{\text{ee}}(\eta, \theta, r, \beta), \Delta_{\text{Weibull}}(\eta, \theta, r, \beta), \Delta_{\text{IG}}(\eta, \theta, r, \beta) \) and \( \Delta_{\text{LN}}(\eta, \theta, r, \beta) \). Table 3.4 also gives the number and percentage of times \( \Omega_f(\eta, \theta, r, \beta) \) was smallest among \( \Omega_{\text{gamma}}(\eta, \theta, r, \beta), \Omega_{\text{ee}}(\eta, \theta, r, \beta), \Omega_{\text{Weibull}}(\eta, \theta, r, \beta), \Omega_{\text{IG}}(\eta, \theta, r, \beta) \) and \( \Omega_{\text{LN}}(\eta, \theta, r, \beta) \). The last two rows of Table 3.4 account for two or more of gamma, ee, Weibull, IG, LN sharing the smallest discrepancy measure value.

Table 3.4 also gives the order of the smallest values for \( \Delta_f(\eta, \theta, r, \beta) \) and \( \Omega_f(\eta, \theta, r, \beta) \). The orders of \( \Delta'_f(\eta, \theta, r, \beta) \) and \( \Omega'_f(\eta, \theta, r, \beta) \) were computed as \( 10^{-6} \) and \( 10^{-4} \), respectively, so it is pleasing that the moments based approximations again perform better. The values of \( t \) needed for \( \Delta'_f(\eta, \theta, r, \beta) \) and \( \Omega'_f(\eta, \theta, r, \beta) \) to be less than \( 10^{-9} \) and \( 10^{-6} \), respectively, were at least 380 and 400, respectively.

<table>
<thead>
<tr>
<th>( f )</th>
<th>( \Delta_f ) is smallest</th>
<th>Percentage</th>
<th>Order</th>
<th>( \Omega_f ) is smallest</th>
<th>Percentage</th>
<th>Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>gamma</td>
<td>112877</td>
<td>70.1</td>
<td>( 10^{-9} )</td>
<td>4545</td>
<td>2.8</td>
<td>( 10^{-7} )</td>
</tr>
<tr>
<td>exponentiated exponential</td>
<td>3362</td>
<td>2.1</td>
<td>( 10^{-9} )</td>
<td>3094</td>
<td>1.9</td>
<td>( 10^{-7} )</td>
</tr>
<tr>
<td>Weibull</td>
<td>41993</td>
<td>26.1</td>
<td>( 10^{-9} )</td>
<td>151745</td>
<td>94.8</td>
<td>( 10^{-7} )</td>
</tr>
<tr>
<td>inverse Gaussian</td>
<td>908</td>
<td>0.6</td>
<td>( 10^{-9} )</td>
<td>542</td>
<td>0.3</td>
<td>( 10^{-6} )</td>
</tr>
<tr>
<td>lognormal</td>
<td>752</td>
<td>0.5</td>
<td>( 10^{-9} )</td>
<td>74</td>
<td>0.0</td>
<td>( 10^{-6} )</td>
</tr>
<tr>
<td>ee, LG, LN</td>
<td>8</td>
<td>0.0</td>
<td>( 10^{-9} )</td>
<td>0</td>
<td>0.0</td>
<td>-</td>
</tr>
<tr>
<td>gamma, ee, IG, LN</td>
<td>100</td>
<td>0.1</td>
<td>( 10^{-9} )</td>
<td>0</td>
<td>0.0</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 3.4: Comparison of the performances of the five approximations under inverse Gaussian \((\eta, \theta)\) distributed claim amounts and negative binomial \((r, \beta)\) distributed claim numbers, shown by the number/percentage of combinations of parameter values tested where the approximation(s) \((f)\) gave the smallest absolute \((\Delta_f)\) or relative \((\Omega_f)\) error.

We see that the gamma (absolute error) and Weibull (relative error) approximations give the most number of best fits. The second most number of best fits is by the Weibull (absolute error) and gamma (relative error) approximations. The third most number of best fits is by the exponentiated exponential (absolute and relative errors) approximation. The fourth most number of best fits is by the inverse Gaussian (absolute and relative errors) approximation. The fifth most number of best fits is by the lognormal (absolute and relative errors) approximation.
Figure 3.7: The combinations of the values of parameters \((\eta, r, \beta)\) from \((\eta, \theta, r, \beta)\), indicated by dots, for which the approximations based on the exponentiated exponential (top left), Weibull (top right), inverse Gaussian (bottom left) and lognormal (bottom right) distributions gave the smallest absolute error \((\Delta_f)\).
Figure 3.8: The combinations of the values of parameters $(\eta, r, \beta)$ from $(\eta, \theta, r, \beta)$, indicated by dots, for which the approximations based on the gamma (top left), exponentiated exponential (top right), inverse Gaussian (bottom left) and lognormal (bottom right) distributions gave the smallest relative error ($\Omega_f$).

Figure 3.7 plots the values of $(\eta, r, \beta) \in \{1, 2, \ldots, 20\}^3$ for which $\Delta_f$ is smallest for $f =$ exponentiated exponential, Weibull, inverse Gaussian and lognormal. Figure 3.8 plots the values of $(\eta, r, \beta) \in \{1, 2, \ldots, 20\}^3$ for which $\Omega_f$ is smallest for $f =$ gamma, exponentiated exponential, inverse Gaussian and lognormal. Some plots were omitted due to a lack of distinguishable pattern, and plots of other three-dimensional marginals of $(r, \beta, \eta, \theta) \in \{1, 2, \ldots, 20\}^4$ showed similar results. The best fits of the gamma approximation correspond mainly to small values of $r$ (relative error). The best fits of the exponentiated exponential approximation correspond mainly to either $r$ being close to zero or $\theta$ being close to zero (absolute and relative errors). The best fits of the Weibull approximation
correspond mainly to small values of $r$ (absolute error). The best fits of the inverse Gaussian and lognormal approximations correspond mainly to either $r$ being close to zero or $\beta$ being close to zero (absolute and relative errors).

3.4 Conclusions

We have compared the performances of five moments based approximations for the compound sum of independent and identical random variables. One of these five is that based on the gamma distribution due to Jin et al. (2016), whilst the other four were introduced in this paper. They are based on the two-parameter exponentiated exponential, two-parameter Weibull, two-parameter inverse Gaussian and two-parameter lognormal distributions.

The comparison uses four of the six examples discussed in Jin et al. (2016). In all but example 1, the gamma approximation performs the best with respect to absolute error: 35 percent of the best fits for example 1, 50 percent of the best fits for example 2, 45 percent of the best fits for example 3 and 70 percent of the best fits for example 4 are given by the gamma approximation. The Weibull approximation gives the best performance for example 1. The second best performance for examples 2 to 4 is also by the Weibull approximation.

In all of the examples, the Weibull approximation performs the best with respect to relative error: 83 percent of the best fits for example 1, 73 percent of the best fits for example 2, 90 percent of the best fits for example 3 and 95 percent of the best fits for example 4 are given by the Weibull approximation. In two of the four examples, the exponentiated exponential approximation gives the second best performance. In the remaining two examples, the gamma approximation gives the second best performance.

Hence, the approximations based on the Weibull and exponentiated exponential distributions are good competitors to the gamma approximation.

The moments based approximations based on the gamma, exponentiated exponential, Weibull, inverse Gaussian and lognormal distributions have also been shown to be more accurate than the truncated versions of (3.8)-(3.11).

In terms of parameter values, the gamma approximation performs the best when the mean of $N$ is large (absolute error) or when the mean of $N$ is small (relative error). The Weibull approximation performs the best when the mean of $N$ is not large (absolute error) or when the mean of $N$ is large (relative error). The other approximations perform best when the mean of $N$ is close to zero (absolute error).
and relative errors). These findings can help in choosing one of the five approximations for a given real data set.

We note that unlike the gamma distribution, the Weibull and exponentiated exponential distributions have closed form expressions for their cumulative distribution and hazard rate functions, and they also have more useful hazard rate properties. Moreover, their moment expressions are as simple as those of the gamma distribution.

A future work is to extend the comparisons presented for compound sums of independent and identical bivariate/multivariate random vectors.
Bibliography


Chapter 4

Bias reduction in the size estimation of large populations using small data sets

4.1 Introduction

Big data can be defined as being “data of a very large size, typically to the extent that its manipulation and management presents significant logistical challenges” (Oxford English Dictionary, 2015). It is common for these types of data sets to be represented and modelled as large graphs or networks. In the analysis of such data sets, a significant problem of the sheer size of the data is that many traditional methods and measures used to obtain representative samples for real graphs become problematic. Thus, the challenge becomes how to get the best and smallest sample of data which gives a representative estimate of the true size of a real graph (Leskovec and Faloutsos, 2006).

Population or size estimation can be a useful tool both directly and indirectly in finance. A common problem that firms often face is estimating the size of a market, which is essential in determining whether the introduction of new businesses or products into a market may be feasible. By obtaining an accurate estimate of the market size, firms can then, for example, use this in computing future sales forecasts to evaluate potential revenue, profit, or loss. For foreign markets, knowing the size of a market may give an indication of growth opportunities for firms.
With online social networks, it is often the case that when the number of users is reported it is a very general figure and in some cases it is not even reported. Advertisers may wish to estimate the total number of users, or users in a certain demographic, to gauge which platforms to launch advertising campaigns or social applications on, in order to get the greatest coverage and generate the largest financial benefit (Katzir, et al., 2011).

Many ideas relating to the accuracy and bias of population size estimation have been proposed since the 1990s. Chao et al. (1992) used a non-parametric estimation method for population size relating to the capture-recapture model in ecology. Lawrence and Lee Giles (1998) estimated the size of the internet by using the volume of documents returned by search engines from typical queries performed by scientists. Similarly, Lu and Li (2010) proposed a novel estimation method for the size of deep web data sources, using random queries and returned document IDs. They proposed two estimators, which were for uniformly distributed samples and samples obtained from queries – both of which were verified to have small bias and variance. With the rise of online social networks (OSN), much of the recent work has been focused in this area, and many studies have concentrated on methods utilising random walk sampling. Gjoka et al. (2009) used two methods to obtain unbiased estimators for population size, which are the Metropolis Hastings random walk and a re-weighted random walk. Their study used real Facebook user data for comparisons, and the random walk methods were shown to perform better. Lu and Li (2012) used the method of simple random walk to estimate properties of OSNs, and applied this to a dataset of 41.7 million Twitter users. They also showed that the method of random walk sampling performed better than those based on uniform random sampling.

Let \( n \) denote the number of random nodes sampled, \( \gamma \) the coefficient of variation of the degrees of the network, \( C \) the number of collisions, and \( p \) the probability of a collision. As noted in Lu and Li (2013), in a sample of size \( n \) there are \( \binom{n}{2} \) pairs of nodes, so it is reasonable to suppose that \( C \) is a binomial random variable with size parameter \( \binom{n}{2} \) and probability parameter \( p \). Based on this assumption, Katzir et al. (2011) proposed the following population size estimator

\[
\hat{N} = (\gamma^2 + 1) \binom{n}{2} \frac{1}{C}.
\]  

(4.1)

Based on (4.1), Lu and Li (2013) proposed the following bias-corrected estimator

\[
\hat{N}^* = (\gamma^2 + 1) \binom{n}{2} \frac{1}{C + 1}.
\]  

(4.2)
CHAPTER 4. BIAS REDUCTION IN POPULATION SIZE ESTIMATION

If \( \{c_1, c_2, \ldots, c_m\} \) are independent observations on \( C \), then practical versions of (4.1) and (4.2) are

\[
\hat{N} = \frac{1}{m} \left( \gamma^2 + 1 \right) \left( \frac{n}{2} \right) \sum_{i=1}^{m} \frac{1}{c_i},
\]

(4.3)

and

\[
\hat{N}^* = \frac{1}{m} \left( \gamma^2 + 1 \right) \left( \frac{n}{2} \right) \sum_{i=1}^{m} \frac{1}{c_i + 1},
\]

(4.4)

respectively. Through simulation studies and an application to real Twitter data, Lu and Li (2013) showed that (4.4) produced significantly less biased population estimates than those produced by (4.3).

This chapter is motivated by Lu and Li (2013). We propose a new population size estimator, based on Zellner (1978). We show that the estimator performs better than (4.3) and (4.4) in a simulation study and when applied to two real datasets, especially in the case of small samples.

The contents of this chapter are as follows. In Section 4.2, we propose the new estimator. In Section 4.3, we compare the performance of the three estimators in a simple simulation study. In Section 4.4, we compare the performance of the three estimators using two real datasets. In Section 4.5, we provide some concluding remarks.

4.2 Proposed estimator

Consider a normal population with mean \( \mu \) and variance \( \sigma^2 \). The need for the estimation of \( 1/\mu \) arises in many practical situations: in experimental nuclear Physics, a charged particle's momentum is \( p = 1/\mu \) when \( \mu \) is the track curvature of a particle (Lamanna et al., 1981; Treadwell, 1982); in the one-dimensional special case of the single period control problem, as discussed by Zellner (1971), and Zaman (1981b); in the estimation of structural parameters of a simultaneous equation as recognized in Zellner (1978) and described in Zaman (1981a); and so on.

There are a number of known estimators for \( 1/\mu \). If \( x_1, x_2, \ldots, x_m \) are independent observations from the normal population, then the maximum likelihood estimator for \( 1/\mu \) is:

\[
\frac{1}{\bar{x}}
\]
an estimator with infinite variance, where

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

Zellner (1978) provided the following improved estimator:

$$\frac{\bar{x}}{\bar{x}^2 + \sigma^2/m}. \quad (4.5)$$

The most recent estimator for $1/\mu$ was proposed by Withers and Nadarajah (2012), however, this estimator is much more complicated. Furthermore, evidence not presented here showed that it does not improve much on (4.5) at least for big data sets.

A binomial distribution with parameters \(\binom{n}{2}\) and \(p\) can be approximated by a normal distribution with mean \(\binom{n}{2}p\) and variance \(\binom{n}{2}p(1-p)\). Hence, (4.5) suggests that the population size of big data sets can be estimated by

$$\hat{N}^+ = (\gamma^2 + 1) \binom{n}{2} \frac{\bar{c}}{\bar{x}^2 + \frac{1}{m(m-1)} \sum_{i=1}^{m} (c_i - \bar{c})^2}, \quad (4.6)$$

where

$$\bar{c} = \frac{1}{m} \sum_{i=1}^{m} c_i.$$  

In the subsequent sections, we shall see how (4.6) performs compared to (4.3) and (4.4).

The derivation of the estimator, (4.5), is fully explained in Section 2 of Zellner (1978). Srivastava and Bhatnagar (1981, equations (12) and (13)) showed that (4.5) has bias of the order of $O(n^{-2})$. Hence, (4.6) has bias of the order of $O(n^{-2})$. On the other hand, the biases of (4.3) and (4.4) are of the order of $O(n^{-1})$. Hence, (4.6) will be expected to be the most accurate estimator.

In Sections 4.3 and 4.4, we have used replications of samples to compare the estimators (4.3), (4.4) and (4.6). In some practical situations, only a single sample may be available. But having a single sample or being able to replicate multiple samples are essentially equivalent. In the former case, multiple samples can be simulated from the empirical distribution function of the single sample, using resampling techniques like jackknifing and bootstrapping. Efron and Tibshirani (1994) provide an excellent account of resampling techniques.
4.3 Simulation study

For an initial comparison, we perform a simulation study to compare the performances between the estimators $\hat{N}$ and $\hat{N}^*$ in Lu and Li (2013), and our proposed estimator $\hat{N}^+$. In our simulations, the total population is set at $N = 10^6$ nodes or vertices, with the nodes being uniformly distributed such that the probability $p_i = 1/N$, of each node being selected, is identical. It follows that the probability of a collision occurring in all the nodes is $p = \sum_{i=1}^{N} p_i^2 = 1/N$. As there is no heterogeneity in the sampling probability, $\gamma = 0$ in the expressions for $\hat{N}$, $\hat{N}^*$ and $\hat{N}^+$.

Sample sizes of between 100 and $10^4$ are tested and we do not enforce a restriction that requires at least one collision for every test. For each sample size, $10^4$ tests are conducted generating $10^4$ collision values, and we calculate the relative biases (RBs) of the three estimators using these values. They are

$$RB(\hat{N}) = \frac{E(\hat{N}) - N}{N}, \quad RB(\hat{N}^*) = \frac{E(\hat{N}^*) - N}{N}, \quad RB(\hat{N}^+) = \frac{\hat{N}^+ - N}{N},$$

respectively. Note that the calculation of the estimator $\hat{N}^+$ incorporates the expectation into its calculation, thus the relative bias can be calculated by using the estimator directly.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$RB(\hat{N})$</th>
<th>$RB(\hat{N}^*)$</th>
<th>$RB(\hat{N}^+)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>Inf</td>
<td>-0.99506</td>
<td>-0.10052</td>
</tr>
<tr>
<td>200</td>
<td>Inf</td>
<td>-0.98030</td>
<td>-0.00505</td>
</tr>
<tr>
<td>300</td>
<td>Inf</td>
<td>-0.95619</td>
<td>-0.04980</td>
</tr>
<tr>
<td>400</td>
<td>Inf</td>
<td>-0.92337</td>
<td>-0.02685</td>
</tr>
<tr>
<td>500</td>
<td>Inf</td>
<td>-0.88301</td>
<td>-0.03964</td>
</tr>
<tr>
<td>600</td>
<td>Inf</td>
<td>-0.83530</td>
<td>0.01583</td>
</tr>
<tr>
<td>700</td>
<td>Inf</td>
<td>-0.78286</td>
<td>0.00638</td>
</tr>
<tr>
<td>800</td>
<td>Inf</td>
<td>-0.72644</td>
<td>-0.00343</td>
</tr>
<tr>
<td>900</td>
<td>Inf</td>
<td>-0.66749</td>
<td>-0.00382</td>
</tr>
<tr>
<td>1000</td>
<td>Inf</td>
<td>-0.60666</td>
<td>0.00828</td>
</tr>
</tbody>
</table>

$N = 10^6$. Sample size $n$ between 100 and 1000, repeated $10^4$ times.

Table 4.1: Relative biases of $\hat{N}$, $\hat{N}^*$ and $\hat{N}^+$ in a simulation study using small sample sizes.
Table 4.1 gives the results for simulations over sample sizes $n$ from 100 to 1000. It is immediately clear that the relative bias of $\hat{N}$ is infinite for the whole range of sample sizes shown, whilst the bias for $\hat{N}^*$ is negative indicating underestimation of the true population size (when considering small samples). The size of the bias of our new estimator is small and both negative and positive for different sample sizes, indicating that the estimated value, $\hat{N}^+$, fluctuates closely around the true population value $N$. Therefore, for small sample sizes (less than 1000), the relative bias of the new estimator $\hat{N}^+$ is significantly less than those of $\hat{N}$ and $\hat{N}^*$.

A clear problem with the estimator $\hat{N}$ is that the bias is infinite for small sample sizes. We have that

$$\hat{N} = \left(\frac{n}{2}\right) \frac{1}{C}, \quad \text{when } \gamma = 0,$$

where $C$ is a vector of $10^4$ collision values generated from the binomial distribution. Therefore, when calculating the relative bias and the expectation $E\left(\hat{N}\right)$, we have that

$$E\left(\hat{N}\right) = \left(\frac{n}{2}\right) E\left(\frac{1}{C}\right),$$

where $E\left(\frac{1}{C}\right)$ is the mean of the reciprocal of the $10^4$ collision values. However, since we do not impose a restriction on the minimum number of collisions per test, the number of collisions generated in any one of the $10^4$ tests can take a value of zero. Thus, if at least one test generates a collision value of zero, then $E\left(\frac{1}{C}\right) = Inf$ and $RB\left(\hat{N}\right) = Inf$. 
Relative bias

<table>
<thead>
<tr>
<th>$n$</th>
<th>$RB(\hat{N})$</th>
<th>$RB(\hat{N}^*)$</th>
<th>$RB(\hat{N}^+)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>Inf</td>
<td>-0.60666</td>
<td>0.00828</td>
</tr>
<tr>
<td>2000</td>
<td>Inf</td>
<td>-0.14002</td>
<td>-0.00403</td>
</tr>
<tr>
<td>3000</td>
<td>Inf</td>
<td>-0.00974</td>
<td>-0.00158</td>
</tr>
<tr>
<td>4000</td>
<td>Inf</td>
<td>0.00455</td>
<td>0.00898</td>
</tr>
<tr>
<td>5000</td>
<td>0.09318</td>
<td>-0.00351</td>
<td>-0.00287</td>
</tr>
<tr>
<td>6000</td>
<td>0.06282</td>
<td>-0.00024</td>
<td>0.00009</td>
</tr>
<tr>
<td>7000</td>
<td>0.04171</td>
<td>-0.00266</td>
<td>-0.00229</td>
</tr>
<tr>
<td>8000</td>
<td>0.03105</td>
<td>-0.00216</td>
<td>-0.00111</td>
</tr>
<tr>
<td>9000</td>
<td>0.02647</td>
<td>0.00043</td>
<td>0.00012</td>
</tr>
<tr>
<td>10000</td>
<td>0.02441</td>
<td>0.00341</td>
<td>0.00302</td>
</tr>
</tbody>
</table>

$N = 10^6$. Sample size $n$ between 1000 and 10,000. Repeated $10^4$ times.

Table 4.2: Relative biases of $\hat{N}, \hat{N}^*$ and $\hat{N}^+$ in a simulation study using large sample sizes.

Similarly, Table 4.2 shows the results of the simulations for larger sample sizes of $n$ from 1000 to $10^4$. Again, the relative bias of the original estimator, $RB(\hat{N})$, is infinite until the sample size reaches 5000, indicating that when $n = 5000$, all $10^4$ collision values generated are greater than zero. Where the bias of $\hat{N}$ can be calculated, it is positive indicating an overestimation of the true population size. From a sample size of 2000 upwards, the biases for both $\hat{N}^*$ and $\hat{N}^+$ are small, and both positive and negative, indicating that their estimated values fluctuate closely around the true population value $N$. It can be seen that when using a sample size in excess of 2000, the bias of $\hat{N}^*$ decreases significantly and becomes much closer to the bias of $\hat{N}^+$. Both estimators $\hat{N}^*$ and $\hat{N}^+$ perform better than $\hat{N}$ over the whole range of sample sizes, and $\hat{N}^+$ generates lower bias (better performance) than $\hat{N}^*$ for all sample sizes except for $n = 4000$. 
Figure 4.1 plots the relative bias against the sample size $n/10$ on a logarithmic scale. It is clear that for small sample sizes, the bias of $\hat{N}^+$ is much smaller than those of $\hat{N}$ and $\hat{N}^*$. With sample size around 250, the bias of $\hat{N}^+$ is already significantly small and within the range of -0.05 to 0.05, whilst the bias for $\hat{N}^*$ is still large and greater than 0.95 in absolute terms (from Table 4.1). When sample size is approximately 2500, the relative bias of $\hat{N}^*$ appears to become significantly smaller and starts to converge quickly towards the relative bias of $\hat{N}^+$ as sample size continues to increase. The relative bias of $\hat{N}$ only becomes significantly small when sample size is greater than 5000. Thus, from our simulation study, it appears that $\hat{N}^+$ gives the best performance out of the three estimators.
CHAPTER 4. BIAS REDUCTION IN POPULATION SIZE ESTIMATION

4.4 Real data applications

In this section, we provide the results of the bias of the three estimators when applied to two real data sets. The first data set is the Twitter data used by Lu and Li (2013) and Kwak et al. (2010), which captured the whole Twitter network at July 2009. This consists of approximately 41.7 million nodes or users and 1.47 billion edges. The second data set captures the LiveJournal social network, where users can declare friendships for each other, similar to how Twitter users can follow and be followed. This data is given in a similar form to that of the Twitter data, however, the LiveJournal data is smaller in size, with approximately 4 million nodes and 34.7 million edges. The LiveJournal data was obtained from the Stanford Network Analysis Project (SNAP) website, which provides large network dataset collections of over 50 large networks including social networks, web graphs and internet networks. Note that both the Twitter and LiveJournal data were obtained in an undirected form and analysed using R (R Development Core Team, 2015).

4.4.1 Twitter network analysis

For the comparison of $\hat{N}^+$ with the original estimators $\hat{N}$ and $\hat{N}^*$ on the Twitter data, we followed closely the method presented in Section 5.2 of Lu and Li (2013). First, a subset of 2 million nodes were obtained from the full data set via simple random walk. We then chose sample sizes of $n$ from 400 to 3600, in steps of 400, and ran 500 random walks (at each sample size) on the subset of Twitter data, taking samples at every few steps in order to reduce the probability of spurious collisions (Lu and Li, 2013). From these samples, we were able to calculate the parameter $\gamma^2$ using the method in Section 5.1 of Lu and Li (2013). With real data, the probabilities of each node being sampled are no longer equal, so we have that $\gamma^2 \neq 0$. We then calculated the relative biases $RB\left(\hat{N}\right)$, $RB\left(\hat{N}^*\right)$ and $RB\left(\hat{N}^+\right)$, using the same expressions as in the simulation study. By using the same approach as Lu and Li (2013), we were able to reproduce similar results for $\hat{N}$ and $\hat{N}^*$, as shown by the values for $E(C)$, $1/E(C)$, $\hat{N}$ and $\hat{N}^*$ in Table 4.3.
Relative bias

<table>
<thead>
<tr>
<th>$n$</th>
<th>$E(C)$</th>
<th>$1/E(C)$</th>
<th>$RB(\hat{N})$</th>
<th>$RB(\hat{N}^*)$</th>
<th>$RB(\hat{N}^+)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>400</td>
<td>9.57114</td>
<td>0.10448</td>
<td>0.20250</td>
<td>0.04954</td>
<td>0.04117</td>
</tr>
<tr>
<td>800</td>
<td>21.96794</td>
<td>0.04552</td>
<td>0.09806</td>
<td>0.04124</td>
<td>0.01593</td>
</tr>
<tr>
<td>1200</td>
<td>39.44689</td>
<td>0.02535</td>
<td>0.06019</td>
<td>0.03120</td>
<td>0.01507</td>
</tr>
<tr>
<td>1600</td>
<td>61.15431</td>
<td>0.01635</td>
<td>0.04780</td>
<td>0.02978</td>
<td>0.01795</td>
</tr>
<tr>
<td>2000</td>
<td>87.90782</td>
<td>0.01138</td>
<td>0.03961</td>
<td>0.02731</td>
<td>0.01718</td>
</tr>
<tr>
<td>2400</td>
<td>120.60521</td>
<td>0.00829</td>
<td>0.02810</td>
<td>0.01928</td>
<td>0.01022</td>
</tr>
<tr>
<td>2800</td>
<td>156.22645</td>
<td>0.00640</td>
<td>0.03408</td>
<td>0.02725</td>
<td>0.01840</td>
</tr>
<tr>
<td>3200</td>
<td>198.40281</td>
<td>0.00504</td>
<td>0.02793</td>
<td>0.02261</td>
<td>0.01445</td>
</tr>
<tr>
<td>3600</td>
<td>244.69539</td>
<td>0.00409</td>
<td>0.02407</td>
<td>0.01978</td>
<td>0.01308</td>
</tr>
</tbody>
</table>

$N = 4.17 \times 10^7$. Sample size $n$ between 400 and 3600, with 500 random walks performed at each sample size.

Table 4.3: Comparison of mean collisions, reciprocal of mean collisions, and relative biases of $\hat{N}$, $\hat{N}^*$ and $\hat{N}^+$ from estimating the size of the Twitter network using different sample sizes.

Table 4.3 shows the results of the relative biases of the three estimators when applied to the Twitter data set. It is clear that the results are consistent with those from our simulation study, in that our proposed estimator $\hat{N}^+$ has the smallest bias at all sample sizes tested. This appears to hold true for small and larger sample sizes. As sample size increases above 400, the bias of $\hat{N}^+$ decreases and converges to values within the range of 0.01 to 0.02. Whilst the biases of $\hat{N}$ and $\hat{N}^*$ are initially relatively larger than $\hat{N}^+$, they also decrease with an increasing sample size. We notice also that the bias of the three estimators is positive for all sample sizes, indicating an overestimation of the true population size $N$. 
Figure 4.2: Mean estimates of the true size of the Twitter network (solid horizontal line) generated by $\hat{N}$, $\hat{N}^*$ and $\hat{N}^+$ at different sample sizes.

Figure 4.2 plots the mean estimates of $N$, generated by the three estimators, at each of the sample sizes. We find that the relative bias of the three estimators converge towards zero as the sample size increases. Whilst the bias of $\hat{N}$ is initially large, it becomes within the range of $\hat{N}^*$ and $\hat{N}^+$ as sample size increases beyond 1000. The biases of $\hat{N}^*$ and $\hat{N}^+$, however, are initially small and become closer to zero as sample size increases. The population size estimates of the three estimators converge towards the true value of $N$ as sample size increases. Over all sample sizes, the estimates produced by $\hat{N}^+$ are closer to $N$ than those by $\hat{N}$ and $\hat{N}^*$.

4.4.2 LiveJournal network analysis

We again follow a similar method to that in Section 5.2 of Lu and Li (2013), for the comparison of $\hat{N}^+$, $\hat{N}$ and $\hat{N}^*$ for the LiveJournal data. Comparing the size of the LiveJournal data to the Twitter data, it is approximately 10 times smaller, however, we still perform an initial random walk to obtain a subset of the original data. We obtain a subset of 80,000 nodes and their degrees by sampling the nodes via random walk, taking a sample every three steps apart. For the LiveJournal data, we chose
two sets of sample sizes: i) sample sizes of $n$ from 100 to 900, in steps of 100; ii) sample sizes of $n$ from 1000 to 9000, in steps of 1000. Due to the random walk method used in taking samples of size $n$ from the subset of data, we are restricted to performing a maximum of 80 random walks in the first set of sample sizes and 8 in the second set. We proceeded to calculate $\gamma^2$ and the relative biases of the three estimators as in Section 4.4.1.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$E(C)$</th>
<th>$1/E(C)$</th>
<th>$RB(\hat{N})$</th>
<th>$RB(\hat{N}^*)$</th>
<th>$RB(\hat{N}^+)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.21519</td>
<td>4.64706</td>
<td>Inf</td>
<td>-0.96368</td>
<td>-0.82792</td>
</tr>
<tr>
<td>200</td>
<td>0.27848</td>
<td>3.59091</td>
<td>Inf</td>
<td>-0.92588</td>
<td>-0.71330</td>
</tr>
<tr>
<td>300</td>
<td>0.34177</td>
<td>2.92593</td>
<td>Inf</td>
<td>-0.87599</td>
<td>-0.58754</td>
</tr>
<tr>
<td>400</td>
<td>0.45570</td>
<td>2.19444</td>
<td>Inf</td>
<td>-0.81543</td>
<td>-0.51732</td>
</tr>
<tr>
<td>500</td>
<td>0.55696</td>
<td>1.79545</td>
<td>Inf</td>
<td>-0.74338</td>
<td>-0.42617</td>
</tr>
<tr>
<td>600</td>
<td>0.60759</td>
<td>1.64583</td>
<td>Inf</td>
<td>-0.65827</td>
<td>-0.28609</td>
</tr>
<tr>
<td>700</td>
<td>0.87342</td>
<td>1.14493</td>
<td>Inf</td>
<td>-0.59794</td>
<td>-0.33693</td>
</tr>
<tr>
<td>800</td>
<td>1.20253</td>
<td>0.83158</td>
<td>Inf</td>
<td>-0.54515</td>
<td>-0.37899</td>
</tr>
<tr>
<td>900</td>
<td>1.37975</td>
<td>0.72477</td>
<td>Inf</td>
<td>-0.48817</td>
<td>-0.32514</td>
</tr>
</tbody>
</table>

$N = 4.03 \times 10^6$. Sample size $n$ between 100 and 900, with 80 random walks performed at each sample size. Table 4.4: Comparison of mean collisions, reciprocal of mean collisions, and relative biases of $\hat{N}$, $\hat{N}^*$ and $\hat{N}^+$ from estimating the size of the LiveJournal network using small sample sizes.

Table 4.4 gives the results for the three estimators over small sample sizes of $n$ from 100 to 900. Similar to the results for the Twitter data, these results are also consistent with our initial simulations. We can observe directly from the table that the relative bias of $\hat{N}$ is infinite for the whole range of sample sizes, whilst the biases for both $\hat{N}^*$ and $\hat{N}^+$ are negative indicating underestimation of the true population size. However, we note that the bias of $\hat{N}^+$ is significantly smaller than those of $\hat{N}$ and $\hat{N}^*$ for all sample sizes. This difference in bias is greatest at a sample size of $n = 600$, but decreases as $n$ increases beyond 600.
Figure 4.3: Plot of relative biases of \( \hat{N}^* \) and \( \hat{N}^+ \) from the estimation of the LiveJournal network size using small sample sizes.

Figure 4.3 plots the trend of the biases for both \( \hat{N}^* \) and \( \hat{N}^+ \), and it is clear that the biases of both estimators show a decreasing trend towards zero. As sample size increases from 100 and 600, the bias of \( \hat{N}^+ \) decreases at a faster rate than \( \hat{N}^* \), but as \( n \) increases beyond 600, the bias increases slightly. However, at each sample size the bias of \( \hat{N}^+ \) is always (relatively) closer to zero than \( \hat{N}^* \). Therefore, for small samples of \( n \leq 900 \), our estimator gives the best estimate of the population size among the three estimators.

Using the set of larger sample sizes, we expected to see a greater number of collisions occurring from random walks on the data. Therefore, the estimated bias of the original estimator \( \hat{N} \) should no longer be infinite, but instead be some real value.
**CHAPTER 4. BIAS REDUCTION IN POPULATION SIZE ESTIMATION**

Relative bias

<table>
<thead>
<tr>
<th>n</th>
<th>$E(C)$</th>
<th>$1/E(C)$</th>
<th>$RB(\hat{N})$</th>
<th>$RB(\hat{N}^*)$</th>
<th>$RB(\hat{N}^+)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>27.28571</td>
<td>0.03665</td>
<td>-0.84094</td>
<td>-0.84742</td>
<td>-0.85456</td>
</tr>
<tr>
<td>2000</td>
<td>36.85714</td>
<td>0.02713</td>
<td>-0.75636</td>
<td>-0.76362</td>
<td>-0.77427</td>
</tr>
<tr>
<td>3000</td>
<td>47.00000</td>
<td>0.02128</td>
<td>-0.67323</td>
<td>-0.68063</td>
<td>-0.68935</td>
</tr>
<tr>
<td>4000</td>
<td>58.85714</td>
<td>0.01699</td>
<td>-0.60738</td>
<td>-0.61418</td>
<td>-0.61600</td>
</tr>
<tr>
<td>5000</td>
<td>71.71429</td>
<td>0.01394</td>
<td>-0.53702</td>
<td>-0.54355</td>
<td>-0.54440</td>
</tr>
<tr>
<td>6000</td>
<td>87.57143</td>
<td>0.01142</td>
<td>-0.48774</td>
<td>-0.49361</td>
<td>-0.49213</td>
</tr>
<tr>
<td>7000</td>
<td>116.85714</td>
<td>0.00856</td>
<td>-0.49250</td>
<td>-0.49687</td>
<td>-0.49645</td>
</tr>
<tr>
<td>8000</td>
<td>144.57143</td>
<td>0.00692</td>
<td>-0.47183</td>
<td>-0.47551</td>
<td>-0.47639</td>
</tr>
<tr>
<td>9000</td>
<td>171.42857</td>
<td>0.00583</td>
<td>-0.44807</td>
<td>-0.45130</td>
<td>-0.45033</td>
</tr>
</tbody>
</table>

$N = 4.03 \times 10^6$. Sample size $n$ between 1000 and 9000, with 8 random walks performed at each sample size.

Table 4.5: Comparison of mean collisions, reciprocal of mean collisions, and relative biases of $\hat{N}$, $\hat{N}^*$ and $\hat{N}^+$ from estimating the size of the LiveJournal network using large sample sizes.

From Table 4.5, in the set of large sample sizes, we can see that although the three estimators show a clear reduction in bias as the sample size increases, our estimator does not show an advantage in reducing the bias between the estimate and the real population value. All three estimators produce very similar values for the bias at every sample size. With the exception of the proposed estimator, among larger sample sizes, $\hat{N}$ and $\hat{N}^*$ give a better performance compared with that in the set of small sample sizes. This result is still partly consistent with the results from the simulation study in Section 4.3.

From Table 4.2 and Figure 4.1 in the simulation study, we observe that with a population of $N = 10^6$ and using large sample sizes of $n \geq 2000$, the bias of $\hat{N}^*$ converges rapidly towards $\hat{N}^+$. In particular, with $n \geq 6000$ the biases of both are very close. We can see a similar result here using larger sample sizes with the LiveJournal data, as for sample sizes of $n \geq 1000$ the biases of $\hat{N}^*$, $\hat{N}^+$ and even $\hat{N}$ are indeed very similar.
Figure 4.4: Plot of relative biases of $\hat{N}$, $\hat{N}^*$ and $\hat{N}^+$ from the estimation of the LiveJournal network size using large sample sizes.

Figure 4.4 shows the trend of the biases of the three estimators for the range of sample sizes $n$ from 1000 to 9000. In general, the biases of the three estimators have a similar trend as the sample size increases. As mentioned previously, this shows the tendency for bias to reduce as sample size increases, however, there is a small fluctuation between sample sizes of 6000 and 8000. From the plot, it is clear that our estimator does not show its advantage in the case of large sample sizes, with all three estimators producing nearly the same values.

A possible explanation for the poor performance of our estimator in the large sample size group may be due to the size of the subsample taken from the full data set. The total population of the LiveJournal data is approximately 4 million, around 10 times smaller than that of the Twitter data. Based on the same ratio in the Twitter case, we first generated a subsample of 80000 nodes, also around 10 times smaller than the subsample in the Twitter case. When we performed a random walk using large sample sizes, we had to reduce our skip step (in the random walk process) to a very
CHAPTER 4. BIAS REDUCTION IN POPULATION SIZE ESTIMATION

low value as we required a larger sample of nodes, but from a fixed size subsample. This may lead to an increase in the probability (and number) of spurious collisions, thus the number of collisions may be overestimated. From (4.3) and (4.4), this may result in an underestimation of the population size, in other words, the bias of the estimator increases. Therefore, further study on even larger data sets is needed to investigate more properties of our estimator.

4.5 Conclusions

The aim of this chapter was to propose a method for generating accurate population size estimates based on relatively small samples from big data sets. We have proposed a new population estimator, based on Zellner (1978), and compared its performance with two estimators presented in Lu and Li (2013).

From the results of our initial simulation study, we found that our estimator performs significantly better than the two in Lu and Li (2013) for small sample sizes, with very small bias. However, for large sample sizes, the bias of the bias corrected estimator (Lu and Li, 2013) becomes very close to ours, indicating a very narrow advantage for our estimator.

We also compared our estimator using two real big data sets from the Twitter and LiveJournal social networks. Our results showed that with the Twitter data, our estimator performs better than the other two, over the same range of sample sizes tested in Lu and Li (2013). The results for the LiveJournal data showed a similar trend, with our estimator giving the best performance out of the three over small sample sizes. However, for large sample sizes, the performances of all three are very similar, with no clear advantage for our estimator.

Future work could be extended to test more data sets, since the Twitter and LiveJournal data are both social media related, however, there are many other types of big data sets that the estimator could be applied to. Since both sets of data used were undirected (no distinction between the direction of the connection between nodes or users), this somewhat simplifies the connection between nodes. Thus, applying these estimators to big data sets which are directed may yield similar or different results.
Bibliography


Chapter 5

A statistical analysis of UK financial networks


5.1 Introduction

Large datasets have useful applications in many different areas, for example, science, engineering, and computer science, to name just a few. Such large data sets can often be represented in terms of large graphs, comprising vertices (or nodes) connected by edges. In the most simple case, we can consider an undirected graph, say \( G(V, E) \), which is defined by the set of vertices \( V \) and edges \( E \subset V \times V \) connecting pairs of vertices (Boginski et al., 2005).

The applications of large graphs or networks has been studied greatly, and also spans across a number of fields. For example, in chemistry, Chou (1990) applies large graphs to enzyme kinetics and protein folding mechanisms; in ecology, Bunn et al. (2000) apply large graphs to landscape and habitat connectivity; in engineering, Dobrjanskyj and Freundenstein (1967) apply large graphs to the structural analysis of mechanisms. In recent years, there has been a rise in the application of large graphs and networks to large datasets in the area of finance.

Networks can be used to represent many types of financial structures and processes based on variables such as price volatility of stocks, interest rates, hedge fund asset net values, trading activity
of financial agents, stock ownership etc. Some of the most influential are bank level financial networks, for example inter-bank or bank-firm networks which model money flows and banking transactions between agents – commonly in terms of loans (Battiston et al., 2010). At the most basic level, the modelling of financial networks can reveal a network structure which can offer insights into the general organisation of the banking sector. More importantly, these financial networks are some of the many measures that can be used to quantify systemic risk in the financial system (Elminejad, 2017).

One network statistic, although it may seem trivial, is the degree distribution of a network, which captures the difference between nodes with many versus few edges. This is important in that it affects the topology of a network, in other words whether a network has a small or large degree distribution will influence the characterisation of its structure – an example being the power law model. In financial and economic networks, connections between nodes are commonly based on transaction costs meaning that nodes connect with others that are easiest to connect to, and that give the greatest return for making that connection.

Different types of financial networks have been shown to have differing characterisations and this is especially apparent in the tail of the degree distribution (Rendón de la Torre et al., 2016). Peer-to-peer financial markets, where many small lenders are matched with many small borrowers, have been found to possess a distributed network structure where nearly all nodes have similar connectivity tending towards the average number of degrees. On the other hand, stock market pricing networks and inter-bank networks have been shown to be characterised by the power law model where there are many nodes with few connections and few nodes with many connections. The fat tail nature of the degree distribution indicates that there are few banks interacting with many others, possibly suggesting that these are banks which are too big to fail (Battiston et al., 2010).

We can also use information from the degree distribution to infer other properties of networks. For example, scale free degree distributions may indicate a growing network and linear preferential attachment (Vázquez, 2003). They are also shown to be more robust against attacks and breakdowns in nodes, which could be useful in “determining the criticality of financial systems” (Battiston et al., 2010).

The structure of stock markets has been examined by many researchers, see for example Kullman et al. (2002), Jung et al. (2006), Ping and Binghong (2006), Zhuang et al. (2007), Tabak et al. (2010), Zhang et al. (2010) and Vizgunov et al. (2014). In particular, networks constructed based on the relationships between stock prices have been shown to follow a common distributional model.
Kim and Kim (2002) studied cross-correlations in stock price changes among S&P 500 companies by a weighted random graph. They found that the influence-strength distribution in absolute terms follows a power law. Huang et al. (2009) used a threshold model to construct China’s stock correlation network and studied the structural properties. After conducting a statistical analysis on the network, they also showed that it follows a power law model. Tse et al. (2010) created a complex network of the US stock market and studied the correlations between closing prices of the stocks. Their results found that the network shows a scale free distribution, and variations in the stocks were heavily influenced by a small number of stocks. Boginski et al. (2005) also studied the characteristics of the graph representing the structure of the US stock market, using a network representation based on cross-correlations of stock price fluctuations.

Let $P_i(t)$ denote the price of stock $i, i = 1, \ldots, N$, on day $t, t = 1, \ldots, T$, where $N$ is the total number of stocks in the sample, and $T$ is the total number of trading days in the time period considered. Also, let $R_i(t) = \ln \left( \frac{P_i(t)}{P_i(t-1)} \right)$ be the log one-period returns of stock $i$, from time $t - 1$ to $t$. It follows that the correlation coefficient, $C_{i,j} \in [-1, 1]$ for all $i,j = 1, \ldots, N$, can be calculated as:

$$C_{i,j} = \frac{E(R_i R_j) - E(R_i) E(R_j)}{\sqrt{Var(R_i) Var(R_j)}},$$

(5.1)

where $E(R_i)$ is the average log returns of stock $i$ over $T$ days, $E(R_i R_j)$ is the average of the product of the log returns of stocks $i, j$ over $T$ days, and $Var(R_i)$ is the variance of the log returns of stock $i$ over $T$ days (Boginski et al., 2005).

Define the market graph and its complement to be graphs with $N$ vertices represented by each of the stocks. Let $\theta \in [-1, 1]$ denote a threshold value, such that an edge connects a pair of stocks $(i, j)$ if $C_{i,j} \geq \theta$ (market), or $C_{i,j} \leq \theta$ (complement). Boginski et al. (2005) used data for 6546 financial instruments traded on the US stock markets over 500 consecutive trading days in 2000-2002 and constructed graphs using various threshold values $\theta$. They showed that under certain conditions, the degree distribution of the vertices of the market graphs follows a power law model.

This chapter is motivated by Boginski et al. (2005). Whilst Boginski et al. (2005) consider the degree distribution of network graphs of the US stock market, we construct network graphs of the UK stock market, and fit various distributions to the degree density of the vertices in these graphs. In our analysis, we find that for the market graphs of the UK stock market, no particular model (including the Pareto I) gives a significantly better fit compared with the others. However, complements of market graphs, constructed using a threshold of $\theta$ negative and close to zero, can be fitted well with the lognormal or Fréchet distributions.
The contents of this chapter are as follows. In Section 5.2, we give a brief overview of our dataset and the construction of our network graphs. In Section 5.3, we describe the models fitted and the criteria used for assessing the fit of the models. In Section 5.4, we present our results comparing the performances of the fitted models. In Section 5.5, we give a short discussion of our results and their relation to some of the recent literature. In Section 5.6, we provide some concluding remarks.

5.2 Data and construction of the network graphs

The data set considered in our analysis consists of UK stock market data for stocks or financial instruments traded on three UK stock markets: FTSE100, FTSE250, and FTSEAIM, over three different time periods each of two years in length. These are the daily closing prices of each stock listed on any of the three markets, for each trading day between the dates: i) 1st January 2000 - 31st December 2001; ii) 1st January 2006 - 31st December 2007; iii) 1st January 2012 - 31st December 2013. These three periods have totals of $T = 505, 505, 506$ trading days, respectively. The stock data were obtained using the Datastream package (Datastream International, 2016) and all analysis was performed using R (R Development Core Team, 2016). Note that we consider each time period independently, and for each two year period we use only the data for stocks that were traded continuously throughout the time period – stocks for which closing prices were available for all trading days within a two year period. For each of the three periods, we have a total sample of $N = 492, 1052, 991$ stocks, respectively. Any stocks which started trading (on a market) after the first trading day or ceased trading before the last day of the two year period were omitted, and stocks with a significant amount of missing data were also omitted.

We form our network model of the UK stock market by constructing network graphs using the cross correlations of stock price movements, for each two year period. We build the graphs, using (5.1), according to Boginski et al. (2005)’s method as described in Section 5.1, for a range of values of $\theta$. To aid us in our computations, we consider the market graphs and complements in terms of their adjacency matrices – these are $N \times N$ matrices, whose elements represent the edges between stocks. The element $(i, j)$, $i, j = 1, 2, \ldots, N$, refers to the edge (link) between stocks $i$ and $j$, taking a value of zero if $C_{i,j} < \theta$ or one if $C_{i,j} \geq \theta$, for the market graph, and vice versa in the case of the complement.

Whilst Boginski et al. (2005) analysed the degree distribution of the market graphs and their complements, for various values of $\theta$, we consider the distribution of the degree density of the market graphs and their complements. The degree values of the stocks could be calculated directly from
the adjacency matrices, by totalling the number of edges of each stock – in other words, summing up all the elements in each row of the matrices, to obtain $N^* \leq N$ degree values for each market graph and complement. Note that in the case of the market graph, we subtract one edge from the total sum of each row to remove the ‘self-correlation’, correlation values of $C_{i,j} = 1$, where $i = j$, $i, j = 1, \ldots, N$, which would lead to an extra edge being accounted for, regardless of the value of $\theta$ chosen. The number of degree values, $N^*$, is allowed to be less than the number of stocks, as we do not consider a stock if it has a degree value (total number of edges) equal to zero.

5.3 Distributions fitted to the data

The Pareto I distribution has the probability density function (PDF) given by

$$f(x) = \frac{\alpha \lambda^\alpha}{x^{\alpha+1}}$$

for $x > \lambda > 0$ and $\alpha > 0$. Taking a very simple mathematical form, this is not a very flexible heavy tailed distribution. We consider three other heavy tailed distributions that are more flexible: the generalized Pareto distribution (GPD) given by the PDF

$$f(x) = \frac{1}{\lambda p} \left(1 + \alpha \frac{x - u}{\lambda}\right)^{-\frac{\alpha}{\alpha - 1}}$$

for $x > u$ if $\alpha \geq 0$, $u < x < u - \lambda/\alpha$ if $\alpha < 0$ and $\lambda > 0$, where $u$ is a fixed value known as the threshold and $p$ is the probability that the data exceeds $u$; the lognormal distribution given by the PDF

$$f(x) = \frac{1}{\sqrt{2\pi\alpha x}} \exp\left\{-\frac{(\ln x - \lambda)^2}{2\alpha^2}\right\}$$

for $x > 0$, $\lambda > 0$ and $\alpha > 0$; the Fréchet distribution given by the PDF

$$f(x) = \frac{\alpha \lambda^\alpha}{x^{\alpha+1}} \exp\left\{-\left(\frac{\lambda}{x}\right)^\alpha\right\}$$

for $x > 0$, $\lambda > 0$ and $\alpha > 0$. Each of these PDFs can be monotonically decreasing or unimodal, and we have chosen them because of their simple mathematical forms. There are other heavy tailed distributions which are even more flexible (for example, capable of allowing for bimodality), but they take complicated mathematical forms often involving special functions. In addition, they suffer from
estimation issues such as identifiability.

Each distribution was fitted by the method of maximum likelihood, and to avoid possible complications when fitting the distributions, we consider only those stocks which have at least one degree. Suppose \(\{x_1, x_2, \ldots, x_n\}\) are independent observations, then the maximum likelihood estimates for the Pareto I distribution are

\[
\hat{\lambda} = \min (x_1, x_2, \ldots, x_n)
\]

and

\[
\hat{\alpha} = n \left[ \sum_{i=1}^{n} \ln x_i - n \ln \min (x_1, x_2, \ldots, x_n) \right]^{-1}.
\]

The maximum likelihood estimates for the lognormal distribution are

\[
\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} \ln x_i
\]

and

\[
\hat{\alpha} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left( \ln x_i - \frac{1}{n} \sum_{j=1}^{n} \ln x_j \right)^2}.
\]

The maximum likelihood estimate of \(\alpha\) for the Fréchet distribution is the root of

\[
\frac{n}{\alpha} + \left( \frac{1}{n} \sum_{i=1}^{n} x_i^{-\alpha} \right)^{-1} \sum_{i=1}^{n} x_i^{-\alpha} \ln x_i + \frac{1}{\alpha} \ln \left( \frac{1}{n} \sum_{i=1}^{n} x_i^{-\alpha} \right) \left( \frac{1}{n} \sum_{i=1}^{n} x_i^{-\alpha} \right)^{-1} \sum_{i=1}^{n} x_i^{-\alpha} = \frac{n}{\alpha} \ln \left( \frac{1}{n} \sum_{i=1}^{n} x_i^{-\alpha} \right) + \sum_{i=1}^{n} \ln x_i.
\]

The maximum likelihood estimate of \(\lambda\) for the Fréchet distribution is

\[
\hat{\lambda} = \left( \frac{1}{n} \sum_{i=1}^{n} x_i^{-\alpha} \right)^{-\frac{1}{\alpha}}.
\]

The maximum likelihood estimates for the GPD are the simultaneous solutions of

\[
\frac{1 + \alpha}{\alpha^2} \sum_{i=1}^{n} (x_i - u) \left( 1 + \alpha \frac{x_i - u}{\lambda} \right)^{-1} = \frac{n}{\lambda}.
\]
and
\[ 1 + \alpha \sum_{i=1}^{n} (x_i - u) \left( 1 + \alpha \frac{x_i - u}{\lambda} \right)^{-1} = \frac{1}{\alpha^2} \sum_{i=1}^{n} \ln \left( 1 + \alpha \frac{x_i - u}{\lambda} \right). \]

In addition to a visual comparison of the fits of the models to the true degree density, we also assess the fitted distributions by other various criteria:

- the Akaike information criterion due to Akaike (1974) defined by
  \[ AIC = 2k - 2 \ln L(\hat{\Theta}), \]
  where \( \Theta \) is the vector of unknown parameters, \( \hat{\Theta} \) is the maximum likelihood estimate of \( \Theta \) and \( k \) is the number of unknown parameters;

- the Bayesian information criterion due to Schwarz (1978) defined by
  \[ BIC = k \ln n - 2 \ln L(\hat{\Theta}); \]

- the Kolmogorov-Smirnov statistic (Kolmogorov, 1933; Smirnov, 1948) defined by
  \[ KS = \sup_x \left| \frac{1}{n} \sum_{i=1}^{n} I\{x_i \leq x\} - \hat{F}(x) \right|, \]
  where \( I\{\cdot\} \) denotes the indicator function and \( \hat{F}(\cdot) \) the maximum likelihood estimate of \( F(x) \).

The smaller the values of these criteria the better the fit. For more discussion on these criteria, see Burnham and Anderson (2004) and Fang (2011).

We also used the chi-squared goodness of fit statistic to compare the fits of the models, however, its values were similar to those of the Kolmogorov-Smirnov statistic. Therefore, we do not report the values of the chi-squared goodness of fit statistic.

## 5.4 Results

In this section, we provide short graphical comparisons of the fits, for each of the three time periods considered, of the fitted distributions to the true degree density of the market graphs and their complements, generated under various threshold values of \( \theta \). We also compare numerically
the AIC and BIC values, and the results of the KS test for the fitted distributions, for each $\theta$ value. For each of the three time periods, we generated market graphs using threshold values $\theta = 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6$, and the complements of the market graphs using $\theta = -0.125, -0.1, -0.075, -0.05, -0.025$.

![Figure 5.1: Comparison of the distributions of correlation coefficients, between pairs of stocks, in the three time periods: i) 1st January 2000 - 31st December 2001 (solid black line); ii) 1st January 2006 - 31st December 2007 (dashed red line); iii) 1st January 2012 - 31st December 2013 (dotted blue line).](image)

A quick look at the distributions of the correlation coefficients of the samples of stocks (Figure 5.1), in each of the three time periods, reveals a similar shape to that of the equivalent distribution in Figure 1 of Boginski et al. (2005). We see that all three distributions are not strictly symmetric or centred around zero, but are instead skewed slightly towards the right tail. The distributions of the coefficients for the two periods covering 2006 - 2008 and 2012 - 2014 have a higher peak density.
than that for the earliest period of 2000 - 2002. The density of correlations in the range of 0 to 0.2 appear higher for the earliest period, whilst the density of correlations between -0.1 and 0, and in excess of 0.3, appears slightly greater for the most recent two periods.

One of the key assumptions of the distributions in Section 5.3 is that the data are independently and identically distributed and have no serial correlation. We tested for no serial correlation using Durbin and Watson (1950, 1951, 1971)'s method. This gave the \( p \)-values of 0.122, 0.056 and 0.066 for the three data periods. We tested for the independent and identical hypothesis using the difference sign, turning point and the rank tests (Brockwell and Davis (2002), Chapter 1). The \( p \)-values of the difference sign test for the three periods were 0.122, 0.056 and 0.066. The \( p \)-values of the turning point test for the three periods were 0.064, 0.195 and 0.120. The \( p \)-values of the rank test for the three periods were 0.072, 0.174 and 0.123.

For the plots in the following subsections, the blue line denotes the Pareto I distribution; the green line denotes the Fréchet distribution; the yellow line denotes the lognormal distribution; the purple line denotes the GPD.

5.4.1 Period 1: 1st January 2000 - 31st December 2001

The first period contains a total sample of \( N = 492 \) stocks, for which the daily closing prices on each of the \( T = 505 \) trading days were available. We note here that the results and plots produced are for a range of \( \theta \) values smaller than that specified in Section 5.4, \( \theta \in [0.1, 0.4] \) for the market graph and \( \theta \in [-0.125, -0.025] \) for the complements. For the values of \( \theta \) which were omitted, the degree density plots either had no real distinct shape or not enough data to be fitted.
Figure 5.2: Comparison of the four models fitted to the degree density distribution of market graphs constructed using threshold values: $\theta = 0.1$ (top left); $\theta = 0.2$ (top right); $\theta = 0.3$ (bottom left); $\theta = 0.4$ (bottom right), for the period of 1st January 2000 - 31st December 2001. Fitted models are coded as blue for Pareto I, green for Fréchet, yellow for lognormal, and purple for GPD.

In Figure 5.2, for the market graph of $\theta$ close to 0, i.e. $\theta = 0.1$, it is hard to distinguish any real fit between the fitted distributions. As $\theta$ increases to 0.2, it appears that the Fréchet and lognormal distributions give the best overall fit, followed by the Pareto I distribution, and the least best fit is given by the GPD. Increasing $\theta$ further to 0.3, the best overall fit now appears to be given by the Pareto I distribution, and then by the Fréchet and lognormal distributions which appear to show some overestimation. However, the GPD seems to provide the best fit to the upper part of the
density, where the other fitted distributions tail off and decay towards 0. With \( \theta \geq 0.4 \) the results are similar to that of \( \theta = 0.3 \), however, we note that the true density becomes more sparse with some degree ranges having a density of zero. We also found that as the chosen threshold value \( \theta \) increased, the number of degrees of stocks decreased, and the total number of stocks with at least one degree also fell. The upper limit of \( \theta = 0.4 \) was therefore chosen, as it was found that using a threshold value above 0.4 made it difficult get a reasonable fit from any of the fitted distributions. This could be explained by the decrease in the number of nodes with at least one degree when \( \theta \) exceeds 0.4, due to the low density of stocks with correlation coefficients greater than 0.4.

<table>
<thead>
<tr>
<th>( \theta )</th>
<th>Pareto I</th>
<th>Fréchet</th>
<th>Lognormal</th>
<th>GPD</th>
<th>Pareto I</th>
<th>Fréchet</th>
<th>Lognormal</th>
<th>GPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>6621.787</td>
<td>6008.05</td>
<td>5794.663</td>
<td>5653.215</td>
<td>6630.184</td>
<td>6016.447</td>
<td>5803.06</td>
<td>5661.612</td>
</tr>
<tr>
<td>0.15</td>
<td>5472.051</td>
<td>5175.919</td>
<td>5035.272</td>
<td>4820.501</td>
<td>5480.407</td>
<td>5184.275</td>
<td>5043.628</td>
<td>4828.857</td>
</tr>
<tr>
<td>0.2</td>
<td>3934.483</td>
<td>3891.716</td>
<td>3831.735</td>
<td>3531.416</td>
<td>3942.558</td>
<td>3899.792</td>
<td>3839.81</td>
<td>3539.492</td>
</tr>
<tr>
<td>0.25</td>
<td>2617.496</td>
<td>2671.312</td>
<td>2657.229</td>
<td>2358.229</td>
<td>2625.039</td>
<td>2678.855</td>
<td>2664.771</td>
<td>2365.772</td>
</tr>
<tr>
<td>0.3</td>
<td>1705.176</td>
<td>1794.992</td>
<td>1801.067</td>
<td>1519.499</td>
<td>1712.061</td>
<td>1801.877</td>
<td>1807.952</td>
<td>1526.383</td>
</tr>
<tr>
<td>0.35</td>
<td>1145.129</td>
<td>1198.558</td>
<td>1192.365</td>
<td>965.3112</td>
<td>1151.11</td>
<td>1204.539</td>
<td>1198.345</td>
<td>971.2921</td>
</tr>
<tr>
<td>0.4</td>
<td>859.4584</td>
<td>903.899</td>
<td>893.8395</td>
<td>644.3275</td>
<td>864.8594</td>
<td>909.2999</td>
<td>899.2404</td>
<td>649.7284</td>
</tr>
</tbody>
</table>

Table 5.1: AIC and BIC values for the four models fitted to the degree density distribution of market graphs constructed using threshold values \( \theta \in [0.1, 0.4] \), for the period of 1st January 2000 - 31st December 2001.

From Table 5.1, we observe that the lowest AIC and BIC values over all \( \theta \) are given by the GPD, suggesting that the GPD gives the best fit out of all the fitted distributions. For \( \theta \leq 0.2 \), the Fréchet and lognormal distributions give the next lowest AIC and BIC values after the GPD, whereas for \( \theta > 0.2 \), the Pareto I distribution gives the next lowest AIC and BIC values after the GPD.
The corresponding p-values are given in brackets below the KS statistics.

Table 5.2: KS test statistics for the four models fitted to the degree density distribution of market graphs constructed using threshold values \( \theta \in [0.1, 0.4] \), for the period of 1st January 2000 - 31st December 2001.

The results of the KS test (Table 5.2) show that for \( \theta \) close to zero, the GPD and lognormal distributions produce the lowest KS statistics. For \( \theta \geq 0.2 \), the Fréchet and lognormal distributions give either the lowest or second lowest KS statistics. Looking closer at the p-values, we find that the majority are very small and close to zero – evidence (in these cases) at all significance levels to reject the null hypothesis that the data comes from the respective fitted model. However, when \( \theta = 0.1 \) the GPD gives the lowest KS statistic with a p-value of 0.09174; when \( \theta = 0.2 \) the lognormal distribution gives the lowest KS statistic with a p-value of 0.009342; when \( \theta = 0.25 \) the Fréchet distribution gives the lowest KS statistic with a p-value of 0.001063; when \( \theta = 0.4 \) the lognormal distribution gives the lowest KS statistic with a p-value of 0.001055, respectively. Therefore, we conclude that for \( \theta = 0.1 \), at the 5% significance level we fail to reject the null hypothesis that the sample degree data is drawn from a GPD.
Figure 5.3: Comparison of the four models fitted to the degree density distribution of complements of market graphs constructed using threshold values: $\theta = -0.025$ (top left); $\theta = -0.05$ (top right); $\theta = -0.075$ (bottom left); $\theta = -0.1$ (bottom right), for the period of 1st January 2000 - 31st December 2001. Fitted models are coded as blue for Pareto I, green for Fréchet, yellow for lognormal, and purple for GPD.

Moving on to the complement of the market graph, in Figure 5.3, we see that the differences between fits are more pronounced. It can be observed from the plots that for $\theta$ very close to zero, i.e. $\theta = -0.025, -0.05, -0.075$, the shape of the true degree density of the complement of the market graph differs from that of the original market graph. The peaks in the distributions instead occur at a ‘low’ number of degrees slightly greater than the minimum. For $\theta = -0.025$, the best overall fit appears to be given by the lognormal distribution, the second best fit by the Fréchet distribution,
the third best fit by the GPD, with the Pareto I distribution giving the least best fit overall. These results hold true as we decrease the value (increase in absolute value) of $\theta$ to -0.05, and -0.075. The best overall fits continue to be given by the lognormal and Fréchet distributions, whilst the Pareto I distribution underestimates the density for low degree values. At $\theta = -0.1$, for low degree values the Pareto I distribution seems to provide the best fit to the true distribution. However, the best overall fit may arguably still be given by the Fréchet and lognormal distributions, whilst the GPD gives the least best fit – overestimating the density at low degree values. As was the case with the market graph, we find that the number of stocks with at least one degree decreases, but more significantly as the value of $\theta$ decreases (increases in absolute value), which may be explained by the skew towards the right tail in the distribution of the correlation coefficients (Figure 5.1).

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pareto I</td>
<td>Fréchet</td>
</tr>
<tr>
<td>-0.025</td>
<td>5538.171</td>
<td>4801.396</td>
</tr>
<tr>
<td>-0.05</td>
<td>4673.118</td>
<td>4027.107</td>
</tr>
<tr>
<td>-0.075</td>
<td>3029.539</td>
<td>2883.193</td>
</tr>
<tr>
<td>-0.1</td>
<td>1420.476</td>
<td>1565.515</td>
</tr>
<tr>
<td>-0.125</td>
<td>435.294</td>
<td>534.822</td>
</tr>
</tbody>
</table>

Table 5.3: AIC and BIC values for the four models fitted to the degree density distribution of complements of market graphs constructed using threshold values $\theta \in [-0.125, -0.025]$, for the period of 1st January 2000 - 31st December 2001.
The corresponding $p$-values are given in brackets below the KS statistics.

Table 5.4: KS test statistics for the four models fitted to the degree density distribution of complements of market graphs constructed using threshold values $\theta \in [-0.125, -0.025]$, for the period of 1st January 2000 - 31st December 2001.

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>Pareto I</th>
<th>Fréchet</th>
<th>Lognormal</th>
<th>GPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.025</td>
<td>0.4105</td>
<td>0.1177</td>
<td>0.05124</td>
<td>0.1689</td>
</tr>
<tr>
<td></td>
<td>(0.0000)</td>
<td>(2.395 × 10^{-06})</td>
<td>(0.1510)</td>
<td>(1.289 × 10^{-12})</td>
</tr>
<tr>
<td>-0.05</td>
<td>0.3785</td>
<td>0.1294</td>
<td>0.06823</td>
<td>0.1179</td>
</tr>
<tr>
<td></td>
<td>(0.0000)</td>
<td>(1.502 × 10^{-07})</td>
<td>(0.02086)</td>
<td>(2.412 × 10^{-06})</td>
</tr>
<tr>
<td>-0.075</td>
<td>0.2923</td>
<td>0.1233</td>
<td>0.06540</td>
<td>0.09016</td>
</tr>
<tr>
<td></td>
<td>(0.0000)</td>
<td>(2.221 × 10^{-06})</td>
<td>(0.04222)</td>
<td>(0.001309)</td>
</tr>
<tr>
<td>-0.1</td>
<td>0.2849</td>
<td>0.1915</td>
<td>0.1599</td>
<td>0.2849</td>
</tr>
<tr>
<td></td>
<td>(0.0000)</td>
<td>(3.698 × 10^{-11})</td>
<td>(6.518 × 10^{-08})</td>
<td>(0.0000)</td>
</tr>
<tr>
<td>-0.125</td>
<td>0.3972</td>
<td>0.2601</td>
<td>0.2222</td>
<td>0.3972</td>
</tr>
<tr>
<td></td>
<td>(0.0000)</td>
<td>(1.030 × 10^{-08})</td>
<td>(1.799 × 10^{-06})</td>
<td>(0.0000)</td>
</tr>
</tbody>
</table>

In comparison with Table 5.1 for the market graph, Table 5.3 shows that for $\theta$ negative and very close to 0 ($\theta \geq -0.05$), the fitted lognormal distribution produces the lowest AIC and BIC values, indicating that at these values it provides the best fit of all the distributions. However, as $\theta$ decreases (increases in absolute value) the GPD gives the lowest AIC and BIC values (and best fit), followed by the Pareto I giving the second lowest, with the lognormal and Fréchet distributions giving the highest values.

The results of the KS test (Table 5.4) give a clearer picture, as for all values of $\theta$ the lognormal distribution gives the lowest KS statistic. In general, the next lowest values are given by the Fréchet distribution, then the GPD, with the Pareto I distribution producing the largest KS statistics for all $\theta$. These results are supported by the corresponding $p$-values from the test, as for values of $\theta = -0.025, -0.05, -0.075$ the $p$-values are 0.1510, 0.02086 and 0.04222, respectively, for the KS test on the fitted lognormal distribution. This indicates that at the 10% (1%) significance level for $\theta = -0.025 (\theta = -0.05, -0.075)$ we fail to reject the null hypothesis that the sample degree data is drawn from the lognormal distribution.
5.4.2 Period 2: 1st January 2006 - 31st December 2007

The second period that we examine contains a total sample of \( N = 1052 \) stocks, for which the closing daily prices on each of the \( T = 505 \) trading days are available. For this two year period, we computed and plotted the results for the whole range of \( \theta \) values, as specified in Section 5.4.

![Comparison of the four models fitted to the degree density distribution of market graphs constructed using threshold values: \( \theta = 0.1 \) (top left); \( \theta = 0.2 \) (top right); \( \theta = 0.3 \) (middle left); \( \theta = 0.4 \) (middle right); \( \theta = 0.5 \) (bottom left); \( \theta = 0.6 \) (bottom right), for the period of 1st January 2006 - 31st December 2007. Fitted models are coded as blue for Pareto I, green for Fréchet, yellow for lognormal, and purple for GPD.](image)

From Figure 5.4, we can immediately see a difference with the distributions for the first sample
period (Figure 5.2) – the degree density here is bimodal at low and high degree values, for $\theta$ close to zero. However, none of the fitted distributions captures this, each fitting just one of the two peaks well. For the degree density of the market graph with $\theta = 0.1$, the best fit to the lower half of the distribution appears to be given by the Fréchet and lognormal distributions, followed by the Pareto I distribution, with the GPD giving the best fit to the upper half of the distribution. With $\theta = 0.2, 0.3$, the Pareto I, lognormal and Fréchet distributions jointly give the best fit to the lower half of the distribution, whilst the GPD continues to give the best fit over the upper part of the distribution at large degree values. At $\theta = 0.4$, the distribution is no longer bimodal, and the overall best fit seems to be given by the GPD. For $\theta = 0.5$, at low degree values the best fit appears to be given by the Fréchet, lognormal and Pareto I distributions. When $\theta$ reaches our upper limit of 0.6 we find that the GPD may give the best visual fit out of all the fitted distributions.

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>Pareto I</th>
<th>Fréchet</th>
<th>Lognormal</th>
<th>GPD</th>
<th>Pareto I</th>
<th>Fréchet</th>
<th>Lognormal</th>
<th>GPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>14803.99</td>
<td>13966.86</td>
<td>13752.03</td>
<td>13438.74</td>
<td>14813.9</td>
<td>13976.78</td>
<td>13448.66</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>8412.983</td>
<td>8745.172</td>
<td>8746.506</td>
<td>7907.239</td>
<td>8422.435</td>
<td>8754.624</td>
<td>7916.691</td>
<td></td>
</tr>
<tr>
<td>0.3</td>
<td>5177.502</td>
<td>5175.583</td>
<td>5046.303</td>
<td>3962.849</td>
<td>5185.569</td>
<td>5183.649</td>
<td>5054.369</td>
<td>3970.915</td>
</tr>
<tr>
<td>0.4</td>
<td>3693.377</td>
<td>3543.815</td>
<td>3415.152</td>
<td>3021.877</td>
<td>3700.771</td>
<td>3551.209</td>
<td>3422.546</td>
<td>3029.271</td>
</tr>
<tr>
<td>0.5</td>
<td>2182.37</td>
<td>2136.479</td>
<td>2089.138</td>
<td>1901.512</td>
<td>2189.255</td>
<td>2143.364</td>
<td>2096.023</td>
<td>1908.396</td>
</tr>
<tr>
<td>0.6</td>
<td>869.4814</td>
<td>902.8605</td>
<td>892.2891</td>
<td>721.8855</td>
<td>875.138</td>
<td>908.517</td>
<td>897.9457</td>
<td>727.5421</td>
</tr>
</tbody>
</table>

Table 5.5: AIC and BIC values for the four models fitted to the degree density distribution of market graphs constructed using threshold values $\theta \in [0.1, 0.6]$, for the period of 1st January 2006 – 31st December 2007.

The comparison of the AIC and BIC values from the fitted distributions are shown in Table 5.5. Again, for all values of $\theta$ the GPD produces the lowest AIC and BIC values, indicating that the GPD may offer the best fit out of all the distributions fitted. In general, the next lowest AIC and BIC values are given by the Fréchet and lognormal distributions, with the Pareto I distribution producing the most number of largest values across the range of $\theta$. 

### Table 5.6: KS test statistics for the four models fitted to the degree density distribution of market graphs constructed using threshold values $\theta \in [0.5, 0.6]$, for the period of 1st January 2006 - 31st December 2007.

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>Pareto I</th>
<th>Fréchet</th>
<th>Lognormal</th>
<th>GPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.2276</td>
<td>0.1203</td>
<td>0.09034</td>
<td>0.1724</td>
</tr>
<tr>
<td></td>
<td>$(8.092 \times 10^{-11})$</td>
<td>$(0.002506)$</td>
<td>$(0.04606)$</td>
<td>$(2.174 \times 10^{-06})$</td>
</tr>
<tr>
<td>0.55</td>
<td>0.2033</td>
<td>0.1365</td>
<td>0.1210</td>
<td>0.2788</td>
</tr>
<tr>
<td></td>
<td>$(5.855 \times 10^{-07})$</td>
<td>$(0.002275)$</td>
<td>$(0.009663)$</td>
<td>$(1.028 \times 10^{-12})$</td>
</tr>
<tr>
<td>0.6</td>
<td>0.2320</td>
<td>0.1600</td>
<td>0.1486</td>
<td>0.2698</td>
</tr>
<tr>
<td></td>
<td>$(2.865 \times 10^{-06})$</td>
<td>$(0.003335)$</td>
<td>$(0.008037)$</td>
<td>$(2.514 \times 10^{-08})$</td>
</tr>
</tbody>
</table>

The corresponding $p$-values are given in brackets below the KS statistics.

Table 5.6 shows the results for the KS test for the fitted distributions using the market graphs. We note that output for $\theta < 0.5$ is omitted as the $p$-values corresponding to the KS statistics were found to be either zero or very small and insignificant. The largest $p$-values, relative to all others, can be found at $\theta = 0.5, 0.55, 0.6$ for the smallest KS statistics given by the lognormal distribution, where the respective $p$-values are 0.04606, 0.009663 and 0.008037. From this, we conclude that at the 1% significance level, for $\theta = 0.5$, we fail to reject the null hypothesis that the sample degree data is drawn from the lognormal distribution.
Figure 5.5: Comparison of the four models fitted to the degree density distribution of complements of market graphs constructed using threshold values: $\theta = -0.025$ (top left); $\theta = -0.05$ (top right); $\theta = -0.075$ (bottom left); $\theta = -0.1$ (bottom right), for the period of 1st January 2006 - 31st December 2007. Fitted models are coded as blue for Pareto I, green for Fréchet, yellow for lognormal, and purple for GPD.

In Figure 5.5, we notice that the results for the degree densities of the complements of the market graph, in this second period, are similar to those from the first period. Again, with $\theta$ very close to 0, the shape of the distribution differs from that when considering the market graph; the peak in the distribution of the density occurs close to but not exactly at the minimum number of degrees. With $\theta = -0.025$, it appears that the best fit overall is given by the Fréchet distribution, the second best by the lognormal distribution, the third best by the GPD, with the Pareto I distribution
giving the least best fit. These results continue to hold as we decrease $\theta$ (increase in absolute value) further to -0.05 and -0.075. At $\theta = -0.1$, the fit of all the fitted distributions is much more similar, however, the lognormal and Fréchet distributions still appear to provide the best overall fit. The Pareto I distribution and GPD give the least best fit overall and show signs of underestimation and overestimation, respectively.

A possible explanation for the similarity in results (relating to the complement of the market graph) between the two periods, is that the distributions of the negative correlation coefficients in all periods (Figure 5.1) appear to be quite similar. Thus, due to the similarity in the distribution of the negative correlation coefficients, we may expect the complements of the market graphs (using negative $\theta$ values) in this (and the third) period to be similar to that of the first. Therefore, we may hypothesise that the respective degree densities and the shape of the fitted distributions to be very similar too.

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>AIC</th>
<th>BIC</th>
<th>AIC</th>
<th>BIC</th>
<th>AIC</th>
<th>BIC</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.025</td>
<td>12865.72</td>
<td>11131.17</td>
<td>11108.53</td>
<td>11688.57</td>
<td>12875.64</td>
<td>11141.09</td>
<td>11118.45</td>
<td>11698.49</td>
</tr>
<tr>
<td>-0.05</td>
<td>10756.18</td>
<td>9271.068</td>
<td>9201.071</td>
<td>9662.023</td>
<td>10766.1</td>
<td>9280.983</td>
<td>9210.986</td>
<td>9671.938</td>
</tr>
<tr>
<td>-0.075</td>
<td>8412.428</td>
<td>7449.771</td>
<td>7212.625</td>
<td>7157.561</td>
<td>8422.322</td>
<td>7459.665</td>
<td>7222.519</td>
<td>7167.455</td>
</tr>
<tr>
<td>-0.1</td>
<td>4114.2</td>
<td>4188.687</td>
<td>4177.259</td>
<td>3575.537</td>
<td>4123.638</td>
<td>4198.125</td>
<td>4186.697</td>
<td>3584.975</td>
</tr>
<tr>
<td>-0.125</td>
<td>1421.347</td>
<td>1656.218</td>
<td>1704.641</td>
<td>1249.465</td>
<td>1429.511</td>
<td>1664.382</td>
<td>1712.805</td>
<td>1257.63</td>
</tr>
</tbody>
</table>

Table 5.7: AIC and BIC values for the four models fitted to the degree density distribution of complements of market graphs constructed using threshold values $\theta \in [-0.125, -0.025]$, for the period of 1st January 2006 - 31st December 2007.

From Table 5.7, we can also see a similar pattern to the first period in the information criteria values for the complements of the market graphs. At $\theta$ very close to 0, the lognormal distribution produces the lowest AIC and BIC values, indicating the best fit out of the fitted distributions. This is followed closely, in second, by the Fréchet distribution. As $\theta$ decreases (increases in absolute value), the GPD gives the lowest values for both the AIC and BIC, the Pareto I distribution follows with the second lowest values, whilst the lognormal and Fréchet distributions give the largest values for both.
Table 5.8: KS test statistics for the four models fitted to the degree density distribution of complements of market graphs constructed using threshold values $\theta \in [-0.075, -0.025]$, for the period of 1st January 2006 - 31st December 2007.

The corresponding $p$-values are given in brackets below the KS statistics.

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>Pareto I</th>
<th>Fréchet</th>
<th>Lognormal</th>
<th>GPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.025</td>
<td>0.4093</td>
<td>0.03931</td>
<td>0.06791</td>
<td>0.2368</td>
</tr>
<tr>
<td></td>
<td>(0.0000)</td>
<td>(0.07745)</td>
<td>(0.0001221)</td>
<td>(0.0000)</td>
</tr>
<tr>
<td>-0.05</td>
<td>0.3882</td>
<td>0.05385</td>
<td>0.05500</td>
<td>0.2100</td>
</tr>
<tr>
<td></td>
<td>(0.0000)</td>
<td>(0.004511)</td>
<td>(0.003464)</td>
<td>(0.0000)</td>
</tr>
<tr>
<td>-0.075</td>
<td>0.3543</td>
<td>0.1245</td>
<td>0.06521</td>
<td>0.1250</td>
</tr>
<tr>
<td></td>
<td>(0.0000)</td>
<td>(1.987 \times 10^{-14})</td>
<td>(0.0002885)</td>
<td>(1.577 \times 10^{-14})</td>
</tr>
</tbody>
</table>

For the KS test in Table 5.8, we omit the results for $\theta < -0.075$, as the $p$-values corresponding to the KS statistics were found to be either zero or very small and insignificant. Those that were significant showed a slight departure from the results for the test in the first period (Table 5.4). For $\theta > -0.075$, we find that the Fréchet distribution produces the most number of lowest KS statistics, with the lognormal distribution giving the second most. Across the whole range of $\theta$ values, the highest KS statistics are given by the Pareto I distribution. The largest $p$-values are produced at the $\theta$ values closest to zero (-0.025 and -0.05), however, these correspond to the lowest KS statistics given by the Fréchet distribution, as opposed to the lognormal distribution. We note that for $\theta = -0.025$, -0.05, these $p$-values are 0.07745 and 0.004511, respectively, thus we conclude that for $\theta = -0.025$ at the 5% significance level, we fail to reject the null hypothesis that the sample degree data is drawn from the Fréchet distribution.

5.4.3 Period 3: 1st January 2012 - 31st December 2013

The final period considered contains a total sample of $N = 991$ stocks, for which the closing daily prices on each of the $T = 506$ trading days are available. For this final two year period, we also computed and plotted the results for the whole range of $\theta$ values, as specified in Section 5.4.
The comparisons of the fitted distributions, in Figure 5.6, show very similar results to those in Figure 5.4 for the second period. Again, the degree density is bimodal for \( \theta \) values close to 0 but loses this property when \( \theta \geq 0.4 \). With \( \theta = 0.1 \), the best fit to the lower half of the distribution appears to be given by the Fréchet and lognormal distributions, followed by the Pareto I distribution, whilst the GPD gives the best fit to upper half of the distribution. For \( \theta = 0.2, 0.3 \), the Pareto I, lognormal and Fréchet distributions jointly give the best fit to the lower half of the distribution, and the GPD
continues to give the best fit to the upper part of the distribution. When \( \theta \geq 0.4 \), the bimodal property disappears, with the overall best fit appearing to be given by the GPD. As \( \theta \) approaches the upper limit of 0.6, the GPD may give the best visual fit out of all the fitted distributions.

\[
\begin{array}{cccccccc}
\theta & \text{Pareto I} & \text{Fréchet} & \text{Lognormal} & \text{GPD} & \text{Pareto I} & \text{Fréchet} & \text{Lognormal} & \text{GPD} \\
0.1 & 12940.75 & 12222.75 & 12145.49 & 12124.28 & 12950.54 & 12323.54 & 12155.29 & 12134.08 \\
0.2 & 7353.04 & 7541.859 & 7496.998 & 6733.124 & 7362.142 & 7550.962 & 7506.101 & 6742.226 \\
0.3 & 4721.579 & 4739.022 & 4653.492 & 4052.203 & 4729.693 & 4747.136 & 4661.606 & 4060.316 \\
0.4 & 2922.24 & 2886.202 & 2819.461 & 2505.441 & 2929.48 & 2893.443 & 2803.443 & 2826.702 & 2512.682 \\
0.5 & 1626.955 & 1634.281 & 1602.635 & 1382.236 & 1633.46 & 1640.785 & 1609.14 & 1388.74 \\
0.6 & 673.4181 & 696.9042 & 691.9642 & 578.6943 & 678.7638 & 702.2499 & 697.3098 & 584.04 \\
\end{array}
\]

Table 5.9: AIC and BIC values for the four models fitted to the degree density distribution of market graphs constructed using threshold values \( \theta \in [0.1, 0.6] \), for the period of 1st January 2012 - 31st December 2013.

Over all values of \( \theta \) tested, the GPD again gives the lowest AIC and BIC values (Table 5.9), leading us to conclude that this result holds true for the market graphs in all three time periods. This indicates that according to the information criteria, the GPD would be the preferred distribution relative to the other fitted distributions. Overall, the second lowest AIC and BIC values are given by the lognormal distribution, whilst the highest values are given by the Pareto I and Fréchet distributions.
From the KS test, we find that in all but two of the cases, for $\theta = 0.1, 0.15$ (not shown) the lognormal distribution gives the lowest KS statistics, whilst the Fréchet distribution gives the second lowest. In general, the highest values were found to be given by the GPD for $\theta \leq 0.35$ and the Pareto I distribution for $\theta > 0.35$. However, we omit the results for $\theta < 0.5$ as the $p$-values of the corresponding KS statistics were found to be either zero or very small and insignificant. Where $\theta$ is large (Table 5.10) we find the largest $p$-values, as in the equivalent KS test for the second period, Table 5.6. For $\theta = 0.5$ the lognormal distribution gives the lowest KS statistic with a $p$-value of 0.01337; for $\theta = 0.55$ the lognormal and Fréchet distributions give the lowest KS statistics with $p$-values of 0.06664 and 0.02595, respectively; for $\theta = 0.6$ the lognormal and Fréchet distributions give the lowest KS statistics with $p$-values of 0.04325 and 0.02880, respectively. Thus, for $\theta = 0.55, 0.6$ at the 1% level of significance, we fail to reject the null hypothesis that the sample degree data is drawn from the Fréchet distribution; for $\theta = 0.5, 0.6$ ($\theta = 0.55$) at the 1% (5%) level of significance, we fail to reject the null hypothesis that the sample degree data is drawn from the lognormal distribution.
Figure 5.7: Comparison of the four models fitted to the degree density distribution of complements of market graphs constructed using threshold values: $\theta = -0.025$ (top left); $\theta = -0.05$ (top right); $\theta = -0.075$ (bottom left); $\theta = -0.1$ (bottom right), for the period of 1st January 2012 - 31st December 2013. Fitted models are coded as blue for Pareto I, green for Fréchet, yellow for lognormal, and purple for GPD.

As was the case for the market graph, the distributions of the complements of the market graph in the third period, Figure 5.7, are also very similar to those in the first and second periods over all $\theta$. For $\theta$ close to 0, the best visual fit is given by the Fréchet distribution, second by the lognormal distribution, third by the GPD, and the least best by the Pareto I distribution. As $\theta$ decreases (increases in absolute value) these results still hold, with the Pareto I and GPD still showing some underestimation and overestimation, respectively. However, the overall fit of all the distributions
becomes much more similar. This result would seem to fit our expectations, due to the similarity in
the distribution of the negative correlation coefficients in each of the three time periods, as shown in
Figure 5.1.

<table>
<thead>
<tr>
<th>θ</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.025</td>
<td>12100.16</td>
<td>10986.85</td>
</tr>
<tr>
<td>-0.05</td>
<td>11073.94</td>
<td>9550.712</td>
</tr>
<tr>
<td>-0.075</td>
<td>9359.642</td>
<td>8051.588</td>
</tr>
<tr>
<td>-0.1</td>
<td>6167.447</td>
<td>5980.643</td>
</tr>
<tr>
<td>-0.125</td>
<td>2797.995</td>
<td>3100.486</td>
</tr>
</tbody>
</table>

Table 5.11: AIC and BIC values for the four models fitted to the degree density distribution
of complements of market graphs constructed using threshold values θ ∈ [−0.125, −0.025],
for the period of 1st January 2012 - 31st December 2013.

This similarity in results continues for the information criteria, as Table 5.11 shows that for
θ ≥ −0.075 the lognormal distribution gives the lowest AIC and BIC values followed by the Fréchet
distribution, whilst for θ < −0.075 the GPD gives the lowest AIC and BIC values. This indicates (as
for the second period) that according to the information criteria, at θ negative and very close to zero,
the lognormal distribution gives the best fit of all models, but away from zero the GPD gives the
best fit.

<table>
<thead>
<tr>
<th>θ</th>
<th>KS test</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.025</td>
<td>0.3183 (0.000)</td>
</tr>
<tr>
<td>-0.05</td>
<td>0.3917 (0.000)</td>
</tr>
<tr>
<td>-0.075</td>
<td>0.3787 (0.000)</td>
</tr>
</tbody>
</table>

The corresponding p-values are given in brackets below the KS statistics.

Table 5.12: KS test statistics for the four models fitted to the degree density distribution of
complements of market graphs constructed using threshold values θ ∈ [−0.075, −0.025], for
the period of 1st January 2012 - 31st December 2013.
The results of the KS test are shown in Table 5.12, with the output for \( \theta < -0.075 \) being omitted due to the \( p \)-values of the corresponding KS statistics either being zero or very small and insignificant. We note that the results again show that the largest \( p \)-values correspond to the lowest KS statistics given by the Fréchet and lognormal distributions. In particular, for \( \theta = -0.025 \) the lognormal distribution gives the smallest KS statistic with a \( p \)-value of 0.02237; for \( \theta = -0.05 \) the Fréchet and lognormal distributions give the smallest KS statistics with \( p \)-values of 0.08329 and 0.01677, respectively. Thus, for \( \theta = -0.025, -0.05 \) at the 1% significance level, we fail to reject the null hypothesis that the sample degree data is drawn from a lognormal distribution; for \( \theta = -0.05 \) at the 5% significance level, we fail to reject the null hypothesis that the sample degree data is drawn from the Fréchet distribution.

5.5 Discussion

The differences between the true distributions of degree density of the market graphs using data from period 1 compared with periods 2 and 3, could be explained by the differences in the original distribution of correlation coefficients in Figure 5.1. The number of stocks considered in period 1 was approximately half the number in periods 2 and 3, thus the number of correlation coefficients for period 1 was approximately a quarter of that in periods 2 and 3. Further to this, the distribution of the coefficients in period 1 has a smaller range, whilst those for periods 2 and 3 have a higher peak density and a higher density of large correlation values above 0.3. This may also explain the bimodal nature of the degree densities of the market graphs in periods 2 and 3 (Figures 5.4 and 5.6).

It appears that, in the case of the market graphs, no definitive conclusion could be made about the best fitting distribution to the degree density. Visually, although we note that this may be misleading, the fitted distributions to the degree density of market graphs across the three periods show a similar pattern. For market graphs constructed with a low threshold \( \theta \), the Fréchet, lognormal and Pareto I distributions give a reasonable fit to the lower part of the distribution, whilst the GPD gives a reasonable fit to the upper part of the distribution. For market graphs constructed with a high threshold \( \theta \), the GPD appears to give the best fit overall.

However, for all threshold values \( \theta \) in the three periods, the GPD gives the lowest AIC and BIC values indicating that according to the information criteria, this would be the preferred model out of those that were fitted. In terms of the KS test, we find that the majority of tests on the fitted distributions are insignificant due to very small \( p \)-values. Although, for \( \theta = 0.1 \) in period 1 the GPD gives the smallest KS statistic with a \( p \)-value of almost 0.1 - agreeing with the result from the
information criteria, whilst for \( \theta = 0.5 \) in period 2 the lognormal distribution gives the smallest KS statistic with a \( p \)-value of almost 0.05 - partially agreeing with the information criteria result (with the lognormal distribution giving the second lowest AIC and BIC values in this case).

On the other hand, we find that the true distributions of degree density of the complement of market graphs are much more similar, for each threshold value \( \theta \), across the three periods. This may be due to the similarity in the shape of the distributions of the negative correlation coefficients for each of the three periods analysed (Figure 5.1). Visually, the Fréchet and lognormal distributions appear to give the best fit overall to the true distributions of degree density. For \( \theta \) negative and close to (further from) zero, the lognormal distribution (GPD) gives the lowest AIC and BIC values. Indeed, in periods 2 and 3 the lowest AIC and BIC values from the lognormal distribution are followed by the Fréchet distribution giving the second lowest values for both criteria, showing agreement with the visual fit.

The KS test also appears to show agreement with these results, as in period 1 the lognormal distribution gives the smallest KS statistics with statistically significant \( p \)-values, as high as 0.15, at \( \theta \) values negative and close to zero. In the second period, the Fréchet distribution gives the lowest KS statistics with significant \( p \)-values, for \( \theta \) negative and close to zero. In the third and final period, for \( \theta \) negative and close to zero, the lognormal and Fréchet distributions give the lowest KS statistics (in comparison with the GPD and Pareto I distribution) with significant \( p \)-values.

Many recent papers that have studied financial networks and the construction of network graphs have explored the alternative statistical route of multiple decision procedures. For example, in Koldanov et al. (2013) the construction of sample market graphs for a financial network (as in our analysis) is the procedure of the identification of the true market graph, and this identification problem is treated as a multiple decision problem of the selection of one from a set of hypotheses.

Koldanov et al. (2013) compute a simulation study in which the random vector of random variables representing the daily returns of a stock are distributed according to a multivariate normal distribution or a multivariate Student-t distribution. The main results from the study are that the method of market graph construction is i) “optimal in the class of all unbiased statistical procedures if, for the generating hypotheses testing, one uses the tests of the Neyman structures”; ii) “optimal in a restricted class of unbiased statistical procedures if, for the generating hypotheses testing, one uses the classical Pearson correlation tests” (Koldanov et al., 2013). In the former case, edges connect two stocks \((i, j)\) if the sample correlation \(s_{i,j}\) exceeds a predetermined threshold \(c_{i,j}\). However, whereas our threshold value \( \theta \) was chosen arbitrarily, the threshold \( c_{i,j} \) utilises the conditional distribution of
$s_{i,j}$ from the Wishart distribution. In the latter case, edges connect two stocks $(i,j)$ if the sample correlation $s_{i,j}$ exceeds a predetermined threshold $c$. Indeed, our method appears more similar to this case and is in fact optimal in this class for the special choice of a significance level $\alpha = 0.5$.

The statistical procedure of stock selection by using the Sharpe ratio is considered in Koldanov et al. (2015). The problem of stock selection is presented as a multiple decision problem of choosing (by observations) one hypothesis from a set of hypotheses. In the study, two well known multiple testing statistical procedures are compared, these are the Holm step down procedure and the Hochberg step up procedure. Given a set of observations of a random variable describing a characteristic of stocks in a financial market, stocks are selected if the condition $Sh_i > Sh_0$ is satisfied, that is, the Sharpe ratio of a stock ($Sh_i$) exceeds a threshold value ($Sh_0$). The conditional risk of the two methods as a function of the threshold is explored. For a real financial market which shows a concentration of Sharpe ratios, it is shown to be more appropriate to use the Hochberg procedure for stock selection; in all other cases, the Holm procedure may be better suited. Arguably, this could be viewed as an alternative method for stock selection to what could be achieved through the method used in this chapter. By constructing the complements of the market graphs for threshold values of $\theta$ which are negative and less than zero, one could select stocks and build a portfolio of stocks which are interconnected in these graphs. For example, this may enable one to choose stocks whose returns move in opposite directions, which would offset each other and help to diversify risk in an investment portfolio.

Kalyagin et al. (2016) propose a general method for multivariate network construction based on measures of association. In relation to Kalyagin et al. (2016), our analysis can be categorised as the construction of threshold graphs for UK financial networks, with the aim of recovering the reference threshold graph from observations. This provides the simplest statistical procedure for the identification of the reference threshold graph (a subgraph of the reference network). Due to the measure of association chosen in our method, the network graphs constructed can be classed as Pearson correlation networks. Kalyagin et al. (2016) show that the problem of identifying the reference threshold graph, from observations of attributes of nodes in a multivariate network, can be solved in terms of a multiple decision problem – again, as the selection of one from many hypotheses. In their study, they conclude that for a random vector of nodes in a network with a fixed multivariate distribution, it is possible to construct various reference networks which relate to a particular measure of association, e.g. Pearson correlation network, Kendall correlation network, Spearman correlation network, etc.
5.6 Conclusions

In our analysis, we have fitted four distributions: i) Pareto I distribution; ii) Fréchet distribution; iii) lognormal distribution; iv) generalised Pareto distribution, to the degree density of financial network graphs constructed from correlation coefficients of pairs of stocks in the UK stock market, according to various threshold values of $\theta \in [-1, 1]$, over three different time periods. Our data set consisted of the daily closing prices of stocks listed on any one of three UK stock markets: FTSE100; FTSE250; FTSEAIM, over three time periods each of two years in length covering: 2000 - 2002; 2006 - 2008; 2012 - 2014.

We followed Boginski et al. (2005)’s method to construct the network graphs, using positive threshold values of $\theta$ for market graphs and negative threshold values $\theta$ for the complement of the market graphs, respectively. Although our data set consists of UK stock data, as opposed to Boginski et al. (2005) who considered US stock data, our analysis covers the same time period as Boginski et al. (2005) in addition to two more recent periods.

Our results show that for the degree density of market graphs, constructed using positive values of the threshold $\theta$, none of the fitted models performs significantly better than the others. Observing the plots of the distributions, it appears that (in general) for low values of $\theta$ the Fréchet, lognormal and Pareto I distributions give a reasonable fit to the lower part of the true distribution, whilst the GPD gives a reasonable fit to the upper part. For higher values of $\theta$ the GPD may give the best fit overall. However, according to the information criteria, the GPD would be the preferred model out of the fitted distributions, for all $\theta$ across the three time periods.

On the contrary, for the degree density of the complement of the market graphs, constructed using negative values of the threshold $\theta$, the Fréchet and lognormal distributions appear to give the best fit with $\theta$ close to zero. This is supported by the fact that they give the lowest and second lowest AIC and BIC values when $\theta$ is close to zero. Furthermore, the lowest (and significant) KS statistics are also either given by the lognormal and Fréchet distributions for $\theta$ close to zero.

This work could be extended to analyse stock market data for other countries, in particular for the US. For example, we note that for our data set the distributions of the degree density of market graphs in the second and third periods differ from those in the first. In particular, when using a low threshold $\theta$ the true distributions in periods 2 and 3 appear to be bimodal. It would therefore be interesting to see whether this pattern can be found, over the same time periods, when using data from other international stock markets, such as the US, or whether this may be due to the dependence and correlation between stocks in the UK markets only. In addition, it would be of
interest to test whether the Fréchet and lognormal distributions fit well to the degree density of the complement of market graphs, for other stock data too.

Also, one measure of the goodness of fit used was the information criteria, more specifically the Akaike and Bayesian information criteria. Both of these measures are dependent on the number of estimated parameters in the model, $k$, however, in our analysis our four fitted models each have $k = 2$ estimated parameters. A natural extension to this could be to extend the number of distributions fitted (including those with a different number of parameters) and utilise a greater number of goodness of fit tests.

Finally, we acknowledge that our results only scratch the surface in terms of network analysis of stock market graphs. With regards to the UK stock market network, the next step would be to evaluate and analyse some of the many network measures, such as network disparity; graph centrality; closeness centrality; domination power; clustering coefficients; minimum spanning tree etc. This would allow for a deeper understanding of the complexities of the UK stock market network. Another step is to use more flexible distributions to account for bimodality of some of the data sets.
Bibliography


Chapter 6

A statistical analysis of cryptocurrencies


6.1 Introduction

Bitcoin, the first decentralized cryptocurrency, has gained a large following from the media, academics and the finance industry since its inception in 2009. Built upon blockchain technology, it has established itself as the leader of cryptocurrencies and shows no signs of slowing down. Instead of being based on traditional trust, the currency is based on cryptographic proof which provides many advantages over traditional payment methods (such as Visa and Mastercard) including high liquidity, lower transaction costs, and anonymity, to name just a few.

Indeed, the global interest in Bitcoin has spiked once again in recent months, for example, the UK government is considering paying out research grants in Bitcoin; an increasing number of IT companies are stockpiling Bitcoin to defend against ransomware; growing numbers in China are buying into Bitcoin and seeing it as an investment opportunity. Perhaps most significantly, the Chair of the Board of Governors of the US Federal Reserve has been encouraging central bankers to study new innovations in the financial industry. In particular, the Chair expressed a need to learn more
about financial innovations including Bitcoin, Blockchain, and distributed ledger technologies. With this recent surge in interest, we believe that now is the time to start studying Bitcoin (and other major cryptocurrencies) as key pieces of financial technology.

Since 2009, numerous cryptocurrencies have been developed, with, as of February 2017, 720 in existence. Bitcoin is the largest and most popular representing over 81% of the total market of cryptocurrencies (CoinMarketCap, 2017). However, as the statistics show, many have not garnered the same level of interest. The combined market capitalization of all cryptocurrencies is approximately USD $19 billion (as of February 2017), with the top 15 currencies representing over 97% of the market, and the top seven of these accounting for 90% of the total market capitalization. In our analysis, we focus on these seven cryptocurrencies which fall into the category of having existed for more than two years and are within the top 15 currencies by market capitalization. These are Bitcoin, Ripple, Litecoin, Monero, Dash, MaidSafeCoin, and Dogecoin.

There exists much research on Bitcoin – the most popular cryptocurrency, and we briefly discuss the analysis and results of some of the most popular studies. Hencic and Gourieroux (2014) model and predict the exchange rate of Bitcoin versus the US Dollar, using a noncausal autoregressive process with Cauchy errors. Their results show that the daily Bitcoin/USD exchange rate shows local trends which could indicate periods of speculative behaviour from online trading. Sapuric and Kokkinaki (2014) investigate the volatility of Bitcoin, using data from July 2010 to April 2014, by comparing it to the volatility of the exchange rates of major global currencies. Their analysis indicates that the exchange rate of Bitcoin has high annualised volatility, however, it can be considered more stable when transaction volume is taken into consideration. Briere et al. (2015) use weekly data from 2010 to 2013 to analyse diversified investment portfolios and find that Bitcoin is extremely volatile and shows large average returns. Perhaps surprisingly, the results indicate that Bitcoin offers little correlation with other assets, although it can help to diversify investment portfolios. In Kristoufek (2015), the influencing factors of the price of Bitcoin are investigated and applied to the Chinese Bitcoin market. Short and long term links are found, and Bitcoin is shown to exhibit the properties of both standard financial assets but also speculative assets, which fuel further discussion on whether Bitcoin should be classed as a currency, asset or an investment vehicle. Chu et al. (2015) give the first statistical analysis of the exchange rate of Bitcoin. They fit fifteen of the most common distributions used in finance to the log returns of the exchange rate of Bitcoin versus the US Dollar. Using data from 2011 to 2014 they show that the generalized hyperbolic distribution gives the best fit.

There are thousands of papers published on exchange rates. Hence, it is impossible to provide a review of all papers. Here, we mention seven of the most recent papers: Corlu and Corlu (2015)
compare the performance of the generalized lambda distribution against other flexible distributions such as the skewed $t$ distribution, unbounded Johnson family of distributions, and the normal inverse Gaussian distribution, in capturing the skewness and peakedness of the returns of exchange rates. They conclude that for the Value-at-Risk and Expected Shortfall, the generalised lambda distribution gives a similar performance, and in general it can be used as an alternative for fitting the heavy tail behaviour in financial data; Nadarajah et al. (2015) revisit the study of exchange rate returns in Corlu and Corlu (2015), and show that the Student’s $t$ distribution can give a similar performance to those of the distributions tested in Corlu and Corlu (2015); Bruneau and Moran (2017) investigate the effect of exchange rate fluctuations on labour market adjustments in Canadian manufacturing industries; Dai et al. (2017) examine the role of exchange rates on economic growth in east Asian countries; Parlapiano et al. (2017) examine exchange rate risk exposure on the value of European firms; Schroeder (2017) investigates the macroeconomic performance in developing countries with respect to exchange rates; Seyyedi (2017) provides an analysis of the interactive linkages between gold prices, oil prices, and the exchange rate in India.

One of the aspects we are looking at is the volatility of cryptocurrencies. There are numerous definitions and methods for computing volatility in the literature. For example, in Sapuric and Kokkinaki (2014) volatility is defined as the annualised volatility (or the standard deviation) representing the daily volatility of an exchange rate. This is calculated by multiplying the standard deviation of the exchange rate by the square root of the number of trading days per year. Briere et al. (2015) use annualised returns in their analysis and also compute the annualised volatility. Other types of volatility in the literature can be broadly categorised as future, historical, forecast, and implied volatility (Natenberg, 2007). In the remainder of this chapter, the term volatility is defined as the spread of the daily exchange rates and log returns of the exchange rates of the cryptocurrencies over the time period considered.

This chapter is organised as follows. In Section 6.2, we give an overview of the data used in our analysis including descriptions and sources. Section 6.3 examines the statistical properties of the cryptocurrencies by fitting a wide range of parametric distributions to the data. Section 6.4 provides a discussion of our results. Finally, Section 6.5 concludes and summarizes our findings. Throughout this chapter, 0.000 should not be interpreted as an absolute zero, it merely indicates that the first three decimal places are equal to zero.
6.2 Data

The data used in this chapter are the historical global price indices of cryptocurrencies, and were obtained from the BNC2 database from Quandl. The global price indices were used as they represent a weighted average of the price of the respective cryptocurrencies using prices from multiple exchanges. For our analysis, we choose to use daily data from 23 June 2014 until the end of February 2017. A start date of June 2014 was deliberately chosen so that we can analyze seven of the top fifteen cryptocurrencies, ranked by their market capitalization, as of February 2017 — see CoinMarketCap (2017) for the current rankings of cryptocurrencies by market capitalization. The seven cryptocurrencies chosen to be part of our analysis are: Bitcoin, Dash, LiteCoin, MaidSafeCoin, Monero, DogeCoin, and Ripple. It should be noted that we omitted Ethereum, arguably the second largest cryptocurrency at present, and Ethereum Classic, as those two only started trading in 2015 and 2016, respectively. Other notable cryptocurrencies such as Agur and NEM were also omitted due to the lack of data. We believe that our choice of cryptocurrencies covers the most prominent currencies, and indeed they represent 90% of the market capitalization as of February 2017 (CoinMarketCap, 2017). In the following, we provide a brief introduction of the seven cryptocurrencies chosen.

**Bitcoin** is undoubtedly the most popular and prominent cryptocurrency. It was the first realization of the idea of a new type of money, mentioned over two decades ago, that “uses cryptography to control its creation and transactions, rather than a central authority” (Bitcoin Project, 2017). This decentralization means that the Bitcoin network is controlled and owned by all of its users, and as all users must adhere to the same set of rules, there is a great incentive to maintain the decentralized nature of the network. Bitcoin uses blockchain technology, which keeps a record of every single transaction, and the processing and authentication of transactions are carried out by the network of users (Bitcoin Project, 2017). Although the decentralized nature offers many advantages, such as being free from government control and regulation, critics often argue that apart from its users, there is nobody overlooking the whole system and that the value of Bitcoin is unfounded. In return for contributing their computing power to the network to carry out some of the tasks mentioned above, also known as “mining”, users are rewarded with Bitcoins. These properties set Bitcoin apart from traditional currencies, which are controlled and backed by a central bank or governing body.

**Dash** (formerly known as Darkcoin and XCoin) is a “privacy-centric digital currency with instant transactions” (The Dash Network, 2017). Although it is based upon Bitcoin’s foundations and shares similar properties, Dash’s network is two-tiered, improving upon that of Bitcoin’s. In contrast with Bitcoin, Dash is overseen by a decentralized network of servers — known as “Masternodes” (The
Dash Network, 2017) which alleviates the need for a third party governing body, and allows for functions such as financial privacy and instant transactions. On the other hand, users or “miners” in the network provide the computing power for basic functions such as sending and receiving currency, and the prevention of double spending. The advantage of utilizing Masternodes is that transactions can be confirmed almost in real time (compared with the Bitcoin network) because Masternodes are separate from miners, and the two have non-overlapping functions (The Dash Network, 2017). Dash utilizes the X11 chained proof-of-work hashing algorithm which helps to distribute the processing evenly across the network while maintaining a similar coin distribution to Bitcoin. Using eleven different hashes increases security and reduces the uncertainty in Dash. Dash operates “Decentralized Governance by Blockchain” (The Dash Network, 2017) which allows owners of Masternodes to make decisions, and provides a method for the platform to fund its own development.

LiteCoin (LTC) was created in 2011 by Charles Lee with support from the Bitcoin community. Based on the same peer-to-peer protocol used by Bitcoin, it is often cited as Bitcoin’s leading rival as it features improvements over the current implementation of Bitcoin. It has two main features which distinguish it from Bitcoin, its use of scrypt as a proof-of-work algorithm and a significantly faster confirmation time for transactions. The former enables standard computational hardware to verify transactions and reduces the incentive to use specially designed hardware, whilst the latter reduces transaction confirmation times to minutes rather than hours and is particularly attractive in time-critical situations (LiteCoin Project, 2017).

MaidSafeCoin is a digital currency which powers the peer-to-peer Secure Access For Everyone (SAFE) network, which combines the computing power of all its users, and can be thought of as a “crowd-sourced internet” (MaidSafe, 2017a). Each MadeSafe coin has a unique identity and there exists a hard upper limit of 4.3 billion coins as opposed to Bitcoin’s 21 million. As the currency is used to pay for services on the SAFE network, the currency will be recycled meaning that in theory the amount of MaidSafe coins will never be exhausted. The process of generating new currency is similar to other cryptocurrencies and in the case of the SAFE network it is known as “farming” (MaidSafe, 2017b). Users contribute their computing power and storage space to the network and are rewarded with coins when the network accesses data from their store (MaidSafe, 2017b).

Monero (XMR) is a “secure, private, untraceable currency” (Monero, 2017) centred around decentralization and scalability that was launched in April 2014. The currency itself is completely donation-based, community driven and based entirely on proof-of-work. Whilst transactions in the network are private by default, users can set their level of privacy allowing as much or as little access to their transactions as they wish. Although it employs a proof-of-work algorithm, Monero is more
similar to LiteCoin in that mining of the currency can be done by any modern computer and is not restricted to specially designed hardware. It arguably holds some advantages over other Bitcoin-based cryptocurrencies such as having a dynamic block size (overcoming the problem of scalability), and being a disinflationary currency meaning that there will always exist an incentive to produce the Monero currency (Monero, 2017).

Dogecoin (Dogecoin, 2017) originated from a popular internet meme in December 2013. Created by an Australian brand and marketing specialist, and a programmer in Portland, Oregon, it initially started off as a joke currency but quickly gained traction. It is a variation on Litecoin, running on the cryptographic scrypt enabling similar advantages over Bitcoin such as faster transaction processing times. Part of the attraction of Dogecoin is its light-hearted culture and lower barriers to entry to investing in or acquiring cryptocurrencies. One of the most popular uses for Dogecoin is the tipping of others on the internet who create or share interesting content, and can be thought of as the next level up from a “like” on social media or an “upvote” on internet forums. This in part has arisen from the fact that it has now become too expensive to tip using Bitcoin.

Ripple was originally developed in 2012 and is the first global real-time gross settlement network (RTGS) which “enables banks to send real-time international payments across networks” (Ripple, 2017). The Ripple network is a blockchain network which incorporates a payment system, and a currency system known as XRP which is not based on proof-of-work like Monero and Dash. A unique property of Ripple is that XRP is not compulsory for transactions on the network, although it is encouraged as a bridge currency for more competitive cross border payments (Ripple, 2017). The Ripple protocol is currently used by companies such as UBS, Santander, and Standard Chartered, and increasingly being used by the financial services industry as technology in settlements. Compared with Bitcoin, it has advantages such as greater control over the system as it is not subject to the price volatility of the underlying currencies, and it has a more secure distributed authentication process.
### Table 6.1: Summary statistics of daily exchange rates of Bitcoin, Dash, Dogecoin, Litecoin, MaidSafeCoin, Monero, Ripple and the Euro, versus the US Dollar from 23 June 2014 until 28 February 2017.

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Bitcoin</th>
<th>Dash</th>
<th>Dogecoin</th>
<th>Litecoin</th>
<th>MaidSafeCoin</th>
<th>Monero</th>
<th>Ripple</th>
<th>Euro</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum</td>
<td>192.700</td>
<td>1.178</td>
<td>0.000</td>
<td>1.269</td>
<td>0.012</td>
<td>0.235</td>
<td>0.003</td>
<td>0.626</td>
</tr>
<tr>
<td>Q1</td>
<td>273.600</td>
<td>2.577</td>
<td>0.000</td>
<td>3.091</td>
<td>0.020</td>
<td>0.491</td>
<td>0.006</td>
<td>0.736</td>
</tr>
<tr>
<td>Median</td>
<td>415.200</td>
<td>3.623</td>
<td>0.000</td>
<td>3.662</td>
<td>0.029</td>
<td>0.811</td>
<td>0.007</td>
<td>0.779</td>
</tr>
<tr>
<td>Mean</td>
<td>447.400</td>
<td>5.385</td>
<td>0.000</td>
<td>3.659</td>
<td>0.046</td>
<td>2.355</td>
<td>0.008</td>
<td>0.830</td>
</tr>
<tr>
<td>Q3</td>
<td>593.000</td>
<td>7.921</td>
<td>0.000</td>
<td>4.021</td>
<td>0.074</td>
<td>1.970</td>
<td>0.008</td>
<td>0.856</td>
</tr>
<tr>
<td>Maximum</td>
<td>1140.000</td>
<td>17.560</td>
<td>0.000</td>
<td>9.793</td>
<td>0.152</td>
<td>17.590</td>
<td>0.028</td>
<td>1.207</td>
</tr>
<tr>
<td>Skewness</td>
<td>0.841</td>
<td>1.006</td>
<td>0.417</td>
<td>1.363</td>
<td>0.849</td>
<td>2.108</td>
<td>2.543</td>
<td>1.127</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>3.096</td>
<td>2.992</td>
<td>3.175</td>
<td>6.621</td>
<td>2.503</td>
<td>6.526</td>
<td>10.693</td>
<td>3.067</td>
</tr>
<tr>
<td>SD</td>
<td>193.241</td>
<td>3.583</td>
<td>0.000</td>
<td>1.433</td>
<td>0.032</td>
<td>3.397</td>
<td>0.004</td>
<td>0.142</td>
</tr>
<tr>
<td>Variance</td>
<td>37342.159</td>
<td>12.838</td>
<td>0.000</td>
<td>2.053</td>
<td>0.001</td>
<td>11.543</td>
<td>0.000</td>
<td>0.020</td>
</tr>
<tr>
<td>CV</td>
<td>0.432</td>
<td>0.665</td>
<td>0.294</td>
<td>0.392</td>
<td>0.695</td>
<td>1.443</td>
<td>0.471</td>
<td>0.171</td>
</tr>
<tr>
<td>Range</td>
<td>946.938</td>
<td>16.385</td>
<td>0.000</td>
<td>8.524</td>
<td>0.140</td>
<td>17.358</td>
<td>0.025</td>
<td>0.581</td>
</tr>
<tr>
<td>IQR</td>
<td>319.400</td>
<td>5.344</td>
<td>0.000</td>
<td>0.930</td>
<td>0.054</td>
<td>1.479</td>
<td>0.002</td>
<td>0.119</td>
</tr>
</tbody>
</table>

### Table 6.2: Summary statistics of daily log returns of the exchange rates of Bitcoin, Dash, Dogecoin, Litecoin, MaidSafeCoin, Monero, Ripple and the Euro, versus the US Dollar from 23 June 2014 until 28 February 2017.

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Bitcoin</th>
<th>Dash</th>
<th>Dogecoin</th>
<th>Litecoin</th>
<th>MaidSafeCoin</th>
<th>Monero</th>
<th>Ripple</th>
<th>Euro</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum</td>
<td>-0.159</td>
<td>-0.580</td>
<td>-0.385</td>
<td>-0.278</td>
<td>-0.404</td>
<td>-0.560</td>
<td>-0.299</td>
<td>-0.046</td>
</tr>
<tr>
<td>Q1</td>
<td>-0.011</td>
<td>-0.019</td>
<td>-0.009</td>
<td>-0.010</td>
<td>-0.026</td>
<td>-0.026</td>
<td>-0.014</td>
<td>-0.004</td>
</tr>
<tr>
<td>Median</td>
<td>-0.001</td>
<td>0.003</td>
<td>0.002</td>
<td>0.000</td>
<td>-0.001</td>
<td>0.002</td>
<td>0.002</td>
<td>0</td>
</tr>
<tr>
<td>Mean</td>
<td>-0.001</td>
<td>-0.001</td>
<td>0.000</td>
<td>0.001</td>
<td>-0.002</td>
<td>-0.001</td>
<td>-0.000</td>
<td>-0.00004</td>
</tr>
<tr>
<td>Q3</td>
<td>0.008</td>
<td>0.020</td>
<td>0.015</td>
<td>0.009</td>
<td>0.023</td>
<td>0.028</td>
<td>0.017</td>
<td>0.003</td>
</tr>
<tr>
<td>Maximum</td>
<td>0.205</td>
<td>0.411</td>
<td>0.188</td>
<td>0.433</td>
<td>0.241</td>
<td>0.277</td>
<td>0.288</td>
<td>0.038</td>
</tr>
<tr>
<td>Skewness</td>
<td>0.758</td>
<td>-1.487</td>
<td>-2.506</td>
<td>0.756</td>
<td>-0.478</td>
<td>-1.414</td>
<td>-0.401</td>
<td>-0.145</td>
</tr>
<tr>
<td>SD</td>
<td>0.028</td>
<td>0.051</td>
<td>0.042</td>
<td>0.042</td>
<td>0.054</td>
<td>0.062</td>
<td>0.046</td>
<td>0.006</td>
</tr>
<tr>
<td>Variance</td>
<td>0.001</td>
<td>0.003</td>
<td>0.002</td>
<td>0.002</td>
<td>0.003</td>
<td>0.004</td>
<td>0.002</td>
<td>0.00004</td>
</tr>
<tr>
<td>CV</td>
<td>-47.976</td>
<td>-84.519</td>
<td>89.782</td>
<td>45.619</td>
<td>-21.499</td>
<td>-54.548</td>
<td>-96.585</td>
<td>-143.498</td>
</tr>
<tr>
<td>Range</td>
<td>0.364</td>
<td>0.991</td>
<td>0.573</td>
<td>0.711</td>
<td>0.645</td>
<td>0.837</td>
<td>0.587</td>
<td>0.858</td>
</tr>
<tr>
<td>IQR</td>
<td>0.019</td>
<td>0.039</td>
<td>0.023</td>
<td>0.019</td>
<td>0.049</td>
<td>0.054</td>
<td>0.030</td>
<td>0.007</td>
</tr>
</tbody>
</table>
CHAPTER 6. A STATISTICAL ANALYSIS OF CRYPTOCURRENCIES

Summary statistics of the exchange rates and log returns of the exchange rates of the seven cryptocurrencies are given in Tables 6.1 and 6.2, respectively. In Table 6.1, the summary statistics for the raw exchange rates of the cryptocurrencies, we see a simple reflection of the “worth” or value of each currency. It can clearly be seen that the exchange rate of Dogecoin is the least significant, and at the time of writing (February 2017), the exchange rate is approximately $0.0002 USD to one Dogecoin. This supports the evidence that Dogecoin is primarily used as a currency for online tipping, rather than as a currency for standard payments. It has the lowest minimum, first quartile, median, mean, third quartile, and maximum values. In contrast, being the most popular cryptocurrency, Bitcoin has the largest minimum, first quartile, median, mean, third quartile, and maximum values, which show its greater significance and higher “value” to those with a vested interest in cryptocurrencies. The exchange rates of all seven currencies are positively skewed, with Litecoin, Monero, and Ripple being the most skewed. In terms of kurtosis, MaidSafeCoin shows less peakedness than that of the normal distribution; Bitcoin, Dash, and Dogecoin show levels similar to the normal distribution; Litecoin, Monero, and Ripple have significantly greater peakedness than the normal distribution. The exchange rates of Dogecoin, MaidSafeCoin, and Ripple have the smallest variances and standard deviations, indicating that their low volatility can perhaps be explained by the low values of the exchange rates coupled with the fact that their range and interquartile ranges are very limited. On the other hand, Bitcoin, Dash, and Litecoin’s exchange rates show the greatest variance and standard deviation.

Table 6.2 gives the summary statistics for the log returns of the exchange rates of the seven cryptocurrencies. Here, the log returns show some slightly different results. Dash, MaidSafeCoin, and Monero have the lowest minimum values, whilst Dash and Litecoin have the largest maximums. The means and medians of the log returns of all seven currencies are similar and almost equal to zero. Only the log returns of Bitcoin and Litecoin are positively skewed, all others are negatively skewed with Dogecoin being the most significant. Log returns of all seven currencies have a peakedness significantly greater than that of the normal distribution, with the most peaked being those of Dash, Dogecoin, and Litecoin. Noteworthy, as much has been discussed about the volatility of Bitcoin returns, the log returns of Bitcoin have the lowest variance and standard deviation of the seven cryptocurrencies. Those with the highest variation are Dash, and perhaps unexpectedly, MaidSafeCoin and Monero.

Also shown in Tables 6.1 and 6.2 are the summary statistics of the exchange rates of the Euro and the summary statistics of its log returns. For the exchange rates, the summary statistics for the Euro appear much smaller than those of Bitcoin but comparable to those of other cryptocurrencies. For the log returns of the exchange rates, the summary statistics for the Euro generally appear smaller.
compared to all the cryptocurrencies. An exception is the coefficient of variation. The magnitude of this statistic for the Euro appears largest compared to all the cryptocurrencies.

Fitting of a statistical distribution usually assumes that the data are independent and identically distributed (i.e., randomness), have no serial correlation, and have no heteroskedasticity. We tested for randomness using the difference sign and rank tests. We tested for no serial correlation using Durbin and Watson (1950, 1951, 1971)’s method. We tested for no heteroskedasticity using Breusch and Pagan (1979)’s test. These tests showed that the log returns of the exchange rates of the seven cryptocurrencies can be assumed to be approximately independent and identically distributed, have no serial correlation, and have no heteroskedasticity.

6.3 Distributions fitted

Having briefly examined the summary statistics for both the exchange rates and the log returns of the exchange rates of the seven cryptocurrencies, we provide a visual representation of the distribution of the log returns. Figure 6.1 shows the histograms of the daily log returns of the exchange rate (versus the US Dollar) for all seven cryptocurrencies. From the plots, we find that the log returns in the cases of all seven cryptocurrencies show significant deviation from the normal distribution. In the following, we proceed to fit the parametric distributions to the data.
Figure 6.1: Histograms of daily log returns of the exchange rates of Bitcoin, Dash, Dogecoin, Litecoin, MaidSafecoin, Monero, and Ripple versus the US Dollar from 23 June 2014 until 28 February 2017.

Let $X$ denote a continuous random variable representing the log returns of the exchange rate of the cryptocurrency of interest. Let $f(x)$ denote the probability density function (pdf) of $X$. Let $F(x)$ denote the cumulative distribution function (cdf) of $X$. We suppose $X$ follows one of eight possible distributions, the most popular parametric distributions used in finance. They are specified as follows:

- the Student’s $t$ distribution (Gosset, 1908) with

$$f(x) = \frac{K(\nu)}{\sigma} \left[ 1 + \frac{(x - \mu)^2}{\sigma^2 \nu} \right]^{-(1+\nu)/2}$$

for $-\infty < x < \infty$, $-\infty < \mu < \infty$, $\sigma > 0$ and $\nu > 0$, where $K(\nu) = \sqrt{\nu}B(\nu/2, 1/2)$ and $B(\cdot, \cdot)$ denotes the beta function defined by

$$B(a, b) = \int_0^1 t^{a-1}(1 - t)^{b-1} dt;$$
• the Laplace distribution (Laplace, 1774) with
\[
f(x) = \frac{1}{2\sigma} \exp\left(-\frac{|x - \mu|}{\sigma}\right)
\]
for \(-\infty < x < \infty, -\infty < \mu < \infty\) and \(\sigma > 0\);

• the skew \(t\) distribution (Azzalini and Capitanio, 2003) with
\[
f(x) = \frac{K(\nu)}{\sigma} \left[1 + \frac{(x - \mu)^2}{\sigma^2 \nu}\right]^{-\frac{1+\nu}{2}} + \frac{2K^2(\nu)\lambda(x - \mu)}{\sigma^2} \sum_{k=0}^{\infty} \binom{\nu}{2k} \binom{1+\nu/2}{k} \frac{\lambda^k(x - \mu)^{2k}}{\nu^{3k/2}}
\]
for \(-\infty < x < \infty, -\infty < \mu < \infty, -\infty < \lambda < \infty, \sigma > 0\) and \(\nu > 0\), where \(2F_1(a, b; c; x)\) denotes the Gauss hypergeometric function defined by
\[
2F_1(a, b; c; x) = \sum_{k=0}^{\infty} \frac{(a)_k (b)_k x^k}{(c)_k k!}
\]
where \((e)_k = e(e + 1) \cdots (e + k - 1)\) denotes the ascending factorial;

• the generalized \(t\) distribution (McDonald and Newey, 1988) with
\[
f(x) = \frac{\tau}{2\sigma \nu^{1/\nu} B(\nu, 1/\nu)} \left[1 + \frac{1}{\nu} \left|\frac{x - \mu}{\sigma}\right|^\nu\right]^{-\frac{1}{\nu} + 1/\nu}
\]
for \(-\infty < x < \infty, -\infty < \mu < \infty, \sigma > 0, \nu > 0\) and \(\tau > 0\);

• the skewed Student’s \(t\) distribution (Zhu and Galbraith, 2010) with
\[
f(x) = \frac{K(\nu)}{\sigma} \begin{cases} \left\{ 1 + \frac{1}{\nu} \left[\frac{x - \mu}{2\sigma \alpha}\right]^2 \right\}^{-\frac{\nu+1}{2}}, & \text{if } x \leq \mu, \\ \left\{ 1 + \frac{1}{\nu} \left[\frac{x - \mu}{2\sigma (1-\alpha)}\right]^2 \right\}^{-\frac{\nu+1}{2}}, & \text{if } x > \mu \end{cases}
\]
for \(-\infty < x < \infty, -\infty < \mu < \infty, 0 < \alpha < 1\) and \(\nu > 0\);
• the asymmetric Student’s $t$ distribution (Zhu and Galbraith, 2010) with

\[
f(x) = \frac{1}{\sigma} \begin{cases} \\
\frac{\alpha}{\alpha^*} K(\nu_1) \left\{ 1 + \frac{1}{\nu_1} \left[ \frac{x - \mu}{2\sigma\alpha^*} \right]^2 \right\}^{\frac{\nu_1 + 1}{2}}, & \text{if } x \leq \mu, \\
\frac{1 - \alpha}{1 - \alpha^*} K(\nu_2) \left\{ 1 + \frac{1}{\nu_2} \left[ \frac{x - \mu}{2\sigma(1 - \alpha^*)} \right]^2 \right\}^{\frac{\nu_2 + 1}{2}}, & \text{if } x > \mu
\end{cases}
\]

for $-\infty < x < \infty$, $-\infty < \mu < \infty$, $0 < \alpha < 1$, $\nu_1 > 0$ and $\nu_2 > 0$, where

\[\alpha^* = \frac{\alpha K(\nu_1)}{\alpha K(\nu_1) + (1 - \alpha) K(\nu_2)};\]

• the normal inverse Gaussian distribution (Barndorff-Nielsen, 1977) with

\[
f(x) = \frac{(\gamma/\delta)^{\lambda}}{\sqrt{2\pi K_{-1/2}(\delta\gamma)}} \exp \left[ \beta(x - \mu) \left[ \delta^2 + (x - \mu)^2 \right]^{-1} K_{-1/2} \left( \alpha \sqrt{\delta^2 + (x - \mu)^2} \right) \right]
\]

for $-\infty < x < \infty$, $-\infty < \mu < \infty$, $\delta > 0$, $\alpha > 0$ and $\beta > 0$, where $\gamma = \sqrt{\alpha^2 - \beta^2}$ and $K_{\nu}(\cdot)$ denotes the modified Bessel function of the second kind of order $\nu$ defined by

\[K_{\nu}(x) = \begin{cases} \\
\frac{\pi \csc(\pi \nu)}{2} [I_{-\nu}(x) - I_{\nu}(x)], & \text{if } \nu \notin \mathbb{Z}, \\
\lim_{\nu \rightarrow \nu} K_{\nu}(x), & \text{if } \nu \in \mathbb{Z},
\end{cases}
\]

where $I_{\nu}(\cdot)$ denotes the modified Bessel function of the first kind of order $\nu$ defined by

\[I_{\nu}(x) = \sum_{k=0}^{\infty} \frac{1}{\Gamma(k + \nu + 1)k!} \left( \frac{x}{2} \right)^{2k + \nu},\]

where $\Gamma(\cdot)$ denotes the gamma function defined by

\[\Gamma(a) = \int_0^{\infty} t^{a-1} \exp(-t)dt;\]

• the generalized hyperbolic distribution (Barndorff-Nielsen, 1977) with

\[
f(x) = \frac{(\gamma/\delta)^{\lambda} \alpha^{1/2 - \lambda}}{\sqrt{2\pi K_{-1/2}(\delta\gamma)}} \exp \left[ \beta(x - \mu) \left[ \delta^2 + (x - \mu)^2 \right]^{-1/2} K_{-1/2} \left( \alpha \sqrt{\delta^2 + (x - \mu)^2} \right) \right]
\]

for $-\infty < x < \infty$, $-\infty < \mu < \infty$, $-\infty < \lambda < \infty$, $\delta > 0$, $\alpha > 0$ and $\beta > 0$, where
\[ \gamma = \sqrt{\alpha^2 - \beta^2}. \]

Many of these distributions are nested: the skew \( t \) distribution for \( \lambda = 0 \) is the Student’s \( t \) distribution; the generalized \( t \) distribution for \( \tau = 2 \) is the Student’s \( t \) distribution; the skewed Student’s \( t \) distribution for \( \alpha = 1/2 \) is the Student’s \( t \) distribution; the asymmetric Student’s \( t \) distribution for \( \nu_1 = \nu_2 \) is the skewed Student’s \( t \) distribution; the generalized hyperbolic distribution for \( \lambda = -1/2 \) is the normal inverse Gaussian distribution; and so on.

All but one of the distributions (the Laplace distribution) are heavy tailed, and heavy tails are common in financial data. The Student’s \( t \) distribution is perhaps the simplest of the heavy tailed distributions. It does not allow for asymmetry. The skew \( t \) distribution due to Azzalini and Capitanio (2003) is an asymmetric generalization of the Student’s \( t \) distribution. The generalized \( t \) distribution due to McDonald and Newey (1988) has two parameters controlling its heavy tails, adding more flexibility. The skewed Student’s \( t \) distribution due to Zhu and Galbraith (2010) is a generalization of the Student’s \( t \) distribution with the scale allowed to be different on the two sides of \( \mu \). This distribution is useful if positive log returns have a different scale compared to negative log returns.

The asymmetric Student’s \( t \) distribution due to Zhu and Galbraith (2010) is a generalization of the Student’s \( t \) distribution with the scale as well as the tail parameter allowed to be different on the two sides of \( \mu \). This distribution is useful if positive log returns also have a different heavy tail behavior compared to negative log returns. The generalized hyperbolic distribution due to Barndorff-Nielsen (1977) accommodates semi heavy tails. It is popular in finance because it contains several heavy tailed distributions as particular cases.

The method of maximum likelihood was used to fit each distribution. If \( x_1, x_2, \ldots, x_n \) is a random sample of observed values on \( X \) and if \( \Theta = (\theta_1, \theta_2, \ldots, \theta_k)’ \) are parameters specifying the distribution of \( X \), then the maximum likelihood estimates of \( \Theta \) are those maximizing the likelihood

\[ L(\Theta) = \prod_{i=1}^{n} f(x_i; \Theta) \]

or the log likelihood

\[ \ln L(\Theta) = \sum_{i=1}^{n} \ln f(x_i; \Theta), \]

where \( f(\cdot) \) denotes the pdf of \( X \). We shall let \( \hat{\Theta} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k)’ \) denote the maximum likelihood estimate of \( \Theta \). The maximization was performed using the routine \texttt{optim} in the R software package (R Development Core Team, 2017). The standard errors of \( \hat{\Theta} \) were computed by approximating the
covariance matrix of $\hat{\Theta}$ by the inverse of the observed information matrix, i.e.,
\[
\text{cov} \left( \hat{\Theta} \right) \approx \left( \begin{array}{cccc}
\frac{\partial^2 \ln L}{\partial \theta_1^2} & \frac{\partial^2 \ln L}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 \ln L}{\partial \theta_1 \partial \theta_k} \\
\frac{\partial^2 \ln L}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \ln L}{\partial \theta_2^2} & \cdots & \frac{\partial^2 \ln L}{\partial \theta_2 \partial \theta_k} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial^2 \ln L}{\partial \theta_k \partial \theta_1} & \frac{\partial^2 \ln L}{\partial \theta_k \partial \theta_2} & \cdots & \frac{\partial^2 \ln L}{\partial \theta_k^2}
\end{array} \right)^{-1} \Theta = \hat{\Theta}.
\]

Many of the fitted distributions are not nested, and discrimination among them was performed using various criteria:

- the Akaike information criterion (Akaike, 1974) defined by
  \[
  \text{AIC} = 2k - 2 \ln L \left( \hat{\Theta} \right);
  \]
- the Bayesian information criterion (Schwarz, 1978) defined by
  \[
  \text{BIC} = k \ln n - 2 \ln L \left( \hat{\Theta} \right);
  \]
- the consistent Akaike information criterion (CAIC) (Bozdogan, 1987) defined by
  \[
  \text{CAIC} = -2 \ln L \left( \hat{\Theta} \right) + k (\ln n + 1);
  \]
- the corrected Akaike information criterion (AICc) (Hurvich and Tsai, 1989) defined by
  \[
  \text{AICc} = \text{AIC} + \frac{2k(k+1)}{n-k-1};
  \]
- the Hannan-Quinn criterion (Hannan and Quinn, 1979) defined by
  \[
  \text{HQC} = -2 \ln L \left( \hat{\Theta} \right) + 2k \ln \ln n.
  \]

The five discrimination criteria above, used to discriminate between and determine the best fitting distribution, all utilise the maximum likelihood estimate. In all cases, the smaller the values of the criteria the better the fit. The Akaike information criterion comprises two parts: the bias (log likelihood) and variance (parameters) (Hu, 2007), and the larger the log likelihood the better
the goodness of fit. However, the criterion includes a penalty term which is dependent on the number of parameters in the model. This penalty term increases with the number of estimated parameters and discourages overfitting. The Bayesian information criterion is very similar to that of the Akaike information criterion, the only difference being that the penalty term is not twice the number of estimated parameters, but instead is the number of parameters multiplied by the natural logarithm of the number of observed data points. The two criteria possess different properties and thus their use is also dependent on different factors. In addition, the Bayesian information criterion is asymptotically efficient, whilst the Akaike information criterion is not (Vrieze, 2012). The consistent Akaike information criterion and the corrected Akaike information criterion are also very similar to both the Akaike and Bayesian information criteria. The former acts as a direct extension of the Akaike information criterion in that it is asymptotically efficient and still includes a penalty term which penalises overfitting more strictly (Bozdogan, 1987). On the other hand, the latter includes a correction for small sample bias and includes an additional penalty term which is a function of the sample size (Anderson et al., 2010). Finally, an alternative to the Akaike and Bayesian information criteria is the Hannan-Quinn information criterion. The expression for the Hannan-Quinn criterion is the same as that for the Akaike information criterion, however, the parameter term is multiplied by the double logarithm of the sample size. In general, the Hannan-Quinn criterion penalises models with a greater number of parameters more, compared to both the Akaike and Bayesian information criteria. However, it tends to show signs of overfitting when the sample size is small (McNelis, 2005). For a more detailed discussion on these criteria, see Burnham and Anderson (2004) and Fang (2011).

Apart from the five criteria, various other measures could be used to discriminate between non-nested models. These could include

- the Kolmogorov-Smirnov statistic (Kolmogorov, 1933; Smirnov, 1948) defined by

  \[ KS = \sup_x \left| \frac{1}{n} \sum_{i=1}^{n} I \{ x_i \leq x \} - \hat{F}(x) \right|, \]

  where \( I \{ \cdot \} \) denotes the indicator function and \( \hat{F}(\cdot) \) the maximum likelihood estimate of \( F(x) \);

- the Anderson-Darling statistic (Anderson and Darling, 1954) defined by

  \[ AD = -n - \sum_{i=1}^{n} \left\{ \ln \hat{F}(x_{(i)}) + \ln \left[ 1 - \hat{F}(x_{(n+1-i)}) \right] \right\}, \]

  where \( x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)} \) are the observed data arranged in increasing order;
the Cramer-von Mises statistic (Cramer, 1928; von Mises, 1928) defined by
\[
CM = \frac{1}{12n} + \sum_{i=1}^{n} \left[ \frac{2i - 1}{2n} - \hat{F}(x_{(i)}) \right].
\]
Once again, the smaller the values of these statistics the better the fit. The use of these statistics in Section 6.4, instead of the five criteria, also led to the same conclusions.

6.4 Results

In this section, we provide our analysis in terms of the best fitting distributions, and the results, for the log returns of the different cryptocurrencies. The results provided are in terms of log likelihood values, information criteria, goodness of fit tests, probability plots, quantile plots, plots of two important financial risk measures, back-testing using Kupiec’s test and dynamic volatility.

6.4.1 Fitted distributions and results

The eight distributions in Section 6.3 were fitted to the data described in Section 6.2 via the method of maximum likelihood. The log likelihood values, and the values of AIC, AICc, BIC, HQC and CAIC for the fitted distributions (for each of the seven cryptocurrencies) are shown in Tables B.1, B.2, B.3, B.4, B.5, B.6, and B.7. The parameter estimates and their standard errors for the best fitting distributions in each case are given in Table 6.3. We find that there is no one best fitting distribution jointly for all seven cryptocurrencies. However, we find that for Bitcoin (the most popular cryptocurrency) and its nearest rival Litecoin, the generalized hyperbolic distribution gives the best fit. For three out of the seven: Dash, Monero, and Ripple, the normal inverse Gaussian distribution gives the best fit. For Dogecoin and MaidSafeCoin, the best fitting distributions are the generalized $t$ and Laplace distributions, respectively. The adequacy of the best fitting distributions is assessed in terms of Q-Q plots, P-P plots, the one-sample Kolmogorov-Smirnov test, the one-sample Anderson-Darling test and the one-sample Cramer-von Mises test.
<table>
<thead>
<tr>
<th>Cryptocurrency</th>
<th>Best fitting distribution</th>
<th>Parameter estimates and standard errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bitcoin</td>
<td>Generalized hyperbolic</td>
<td>$\hat{\mu} = -0.001$ (0.000),</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\hat{\delta} = 0.003$ (0.001),</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\hat{\alpha} = 29.644$ (3.707),</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\hat{\beta} = 0.530$ (1.305),</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\hat{\lambda} = 0.220$ (0.010).</td>
</tr>
<tr>
<td>Dash</td>
<td>Normal inverse Gaussian</td>
<td>$\hat{\mu} = 0.004$,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\hat{\delta} = 0.025$,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\hat{\alpha} = 10.714$,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\hat{\beta} = -2.100$.</td>
</tr>
<tr>
<td>Dogecoin</td>
<td>Generalized $t$</td>
<td>$\hat{\mu} = 0.002$ (0.000),</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\hat{\sigma} = 0.014$ (0.002),</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\hat{\tau} = 0.893$ (0.094),</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\hat{\nu} = 3.768$ (1.269).</td>
</tr>
<tr>
<td>Litecoin</td>
<td>Generalized hyperbolic</td>
<td>$\hat{\mu} = 0.000$ (0.000),</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\hat{\delta} = 0.006$ (0.001),</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\hat{\alpha} = 10.517$ (2.021),</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\hat{\beta} = 0.412$ (0.801),</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\hat{\lambda} = -0.186$ (0.078).</td>
</tr>
<tr>
<td>MaidSafeCoin</td>
<td>Laplace</td>
<td>$\hat{\mu} = -0.001$ (0.001),</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\hat{\sigma} = 0.0368$ (0.001).</td>
</tr>
<tr>
<td>Monero</td>
<td>Normal inverse Gaussian</td>
<td>$\hat{\mu} = 0.005$,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\hat{\delta} = 0.040$,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\hat{\alpha} = 11.164$,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\hat{\beta} = -1.705$.</td>
</tr>
<tr>
<td>Ripple</td>
<td>Normal inverse Gaussian</td>
<td>$\hat{\mu} = 0.003$,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\hat{\delta} = 0.018$,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\hat{\alpha} = 8.729$,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\hat{\beta} = -1.670$.</td>
</tr>
</tbody>
</table>

Table 6.3: Best fitting distributions to the daily log returns of the exchange rates of Bitcoin, Dash, Dogecoin, Litecoin, MaidSafeCoin, Monero, and Ripple versus the US Dollar from 23 June 2014 until 28 February 2017, and parameter estimates, with standard errors given in brackets.
CHAPTER 6. A STATISTICAL ANALYSIS OF CRYPTOCURRENCIES

The best fitting distributions show the following: log returns of the exchange rates of MaidSafeCoin have light tails; log returns of the exchange rates of Dogecoin have heavy tails; log returns of the exchange rates of Bitcoin, Dash, Litecoin, Monero and Ripple have semi heavy tails. Among the last five, the asymmetry of tails as measured by $\beta$ is negative for Dash, Monero and Ripple. Of these the largest asymmetry is for Dash, followed by Monero and then Ripple. The asymmetry of tails as measured by $\beta$ is positive for Bitcoin and Litecoin. Of these the larger asymmetry is for Bitcoin.

It is surprising that a light tailed distribution (the Laplace distribution) gives the best fit for log returns of the exchange rates of MaidSafeCoin. However, many authors have found that the Laplace distribution can provide adequate fits to financial data: Linden (2001) models stock returns by the Laplace distribution; Linden (2005) models the realized stock return volatility by the Laplace distribution; Aquino (2006) establishes that the Laplace distribution characterizes the price movements for the Philippine stock market; Podobnik et al. (2009) show that the Laplace distribution fits stock growth rates for the Nasdaq Composite and the New York Stock Exchange Composite.

All of the cryptocurrencies show substantially larger volatility than the exchange rate of the Euro. Nevertheless, we also observe substantial differences within the group of cryptocurrencies. The distribution of the tails of their returns ranges from light tailed via semi-heavy tailed to heavy tailed. Given that traditional financial instruments usually exhibit heavy tails, this is a slightly surprising, but also a very important, result.

6.4.2 Q-Q plots

The Q-Q plots for the best fitting distribution for each of the seven cryptocurrencies are shown in Figure B.1. For Bitcoin, the best fitting distribution captures the middle and lower parts of the data well but not the upper tail. For Dash, the best fitting distribution captures the middle, lower and upper parts of the data well. For Dogecoin, the best fitting distribution captures the middle part of the data well but not the lower or upper tails. For Litecoin, the best fitting distribution captures the middle, lower and upper parts of the data well. For MaidSafeCoin, the best fitting distribution captures the middle and upper parts of the data well but not the lower tail. For Monero, the best fitting distribution captures the middle and upper parts of the data well but not the lower tail. For Ripple, the best fitting distribution captures the middle part of the data well but not the lower or upper tails.
6.4.3 P-P plots

The P-P plots for the best fitting distribution for each of the seven cryptocurrencies are shown in Figure B.2. For each cryptocurrency, the best fitting distribution captures the middle, lower and upper parts of the data well.

6.4.4 Goodness of fit tests

The \( p \)-values of the one-sample Kolmogorov-Smirnov test for the best fitting distributions listed in Table 6.3 are 0.095, 0.124, 0.164, 0.094, 0.119, 0.162 and 0.056. The corresponding \( p \)-values of the one-sample Anderson-Darling test are 0.103, 0.144, 0.154, 0.120, 0.051, 0.157 and 0.176. The corresponding \( p \)-values of the one-sample Cramer-von Mises test are 0.082, 0.116, 0.196, 0.085, 0.168, 0.088 and 0.122. Hence, all of the best fitting distributions are adequate at the five percent significance level.

6.4.5 VaR and ES plots

Value at Risk (VaR) and Expected Shortfall (ES) are the two most popular financial risk measures (Kinateder, 2015, 2016). If \( \hat{F}(\cdot) \) denotes the cdf of the best fitting distribution, then the VaR and ES corresponding to probability \( q \) can be defined by

\[
\text{VaR}(q) = \hat{F}^{-1}(q)
\]

and

\[
\text{ES}(q) = \frac{1}{q} \int_0^q \text{VaR}(u) du,
\]

respectively, for \( 0 < q < 1 \). Plots of VaR \( (q) \) and ES \( (q) \) for the best fitting distributions for the seven cryptocurrencies are shown in Figures B.3 and B.4. Also shown in the figures are estimates of VaR \( (q) \) and ES \( (q) \) for daily log returns of the Euro computed using the same best fitting distributions. It is clear that each cryptocurrency is riskier than the Euro. With respect to the upper tail of VaR, Litecoin, MainSafecoin and Monero appear to have the largest risks. With respect to the lower tail of VaR, Monero appears to have the largest risk. Monero also has the largest risk with respect to the lower tail of ES.
6.4.6 Kupiec’s test

The $p$-values of Kupiec’s test for the best fitting distribution for each of the seven cryptocurrencies are shown in Figure B.5. The $p$-values (points above 0.05 in Figure B.5) suggest that the out of sample performance of VaR can be considered accurate at the corresponding values of $q$ for any of the best fitting distributions. The motivation for using the Kupiec’s test is that we can give predictions of the exchange rate of the different cryptocurrencies, including predictions for the extreme worst case scenario and the extreme best case scenario.

6.4.7 Dynamic volatility

Throughout our analysis thus far, we have supposed that volatility is represented by a fixed parameter of a distribution. Often in financial series, volatility varies throughout time. This could be accommodated in various ways, for example one is to treat volatility itself as a random variable. Another is to let volatility depend on some covariates including time as in GARCH models for example. In the first case, the interest will be on the distribution of volatility.

For the seven cryptocurrencies and the Euro, we computed the standard deviations of the daily log returns of the exchange rates over rolling windows of 20 days. Histograms of these standard deviations are shown in Figure B.6. We see that the distribution is skewed for the seven cryptocurrencies and the Euro. The range of volatility appears smallest for the Euro and second smallest for Bitcoin, whilst the range appears largest for Dogecoin, Litecoin and Monero.

6.5 Conclusions

We have analyzed the exchange rate of the top seven cryptocurrencies versus the US Dollar using eight of the most popular parametric distributions in finance. From our analysis of over two and a half years of data, it is clear that most cryptocurrencies exhibit heavy tails. Using the discrimination criteria of the log likelihood, AIC, AICc, BIC, HQC and CAIC, the results obtained show that none of the distributions used give the best fit jointly across the data for all of the cryptocurrencies. Instead, we find that the generalized hyperbolic distribution gives the best fit for Bitcoin and Litecoin; the normal inverse Gaussian distribution gives the best fit for Dash, Monero, and Ripple; the generalized $t$ distribution gives the best fit to Dogecoin; the Laplace distribution gives the best fit to MaidSafeCoin.

Implications of these results are in the area of risk management, where one may need to compute
the VaR and ES for risk, but also for investment purposes. To the best of our knowledge, this is the first study investigating the statistical properties of cryptocurrencies, going beyond just Bitcoin and traditional currencies. Indeed, there is much scope for future work, and possible extensions could include: i) using GARCH type processes to model the log returns of cryptocurrencies, for example the distributions mentioned in Section 6.3 can be used for modeling the innovation processes; ii) using multivariate processes to model the joint distribution of the log returns of cryptocurrencies; iii) using nonparametric or semiparametric distributions to analyze the exchange rates of cryptocurrencies.
Bibliography


Appendix A

Appendix to Chapter 2

Case 1: $a_i$ simulated from uniform $[0, 3]$ and $b_i$ simulated from uniform $[0, 3]$ independently as

- $a = (1.106042, 2.481823, 2.416046, 2.168359, 2.797924, 0.05945489)$ and $b = (0.3928571, 2.127725, 2.369427, 2.193387, 1.449108, 2.832263),$
- $a = (0.0123726, 1.689808, 1.007412, 0.4843852, 2.837575, 2.736775, 1.532353, 0.6757434)$ and $b = (2.675403, 0.04858216, 0.4397163, 0.4186979, 1.249646, 2.654463, 1.033633, 2.708868),$
- $a = (0.085536, 1.469983, 0.2673928, 1.697669, 2.352464, 1.279901, 0.7425467, 2.126125, 0.6656652, 2.022727) and $b = (0.6477089, 0.1334479, 1.980812, 0.969125, 0.02878696, 2.015512, 2.544392, 1.833014, 1.819751, 1.874621),$
- $a = (1.735091, 2.861762, 0.2198963, 0.2210587, 0.934081, 1.446621, 0.4881861, 0.3566665, 1.46175, 2.392575, 2.406489, 0.7938465) and $b = (0.1573016, 1.383633, 0.5100128, 2.069284, 1.74952, 2.404543, 0.9456252, 1.886564, 1.911445, 0.09226389, 2.415659, 1.058335).

Case 2: $a_i$ simulated from uniform $[0, 3]$ and $b_i$ simulated from Weibull (3, 5) independently as

- $a = (1.423429, 1.082625, 1.655053, 1.152481, 0.05196177, 1.001952) and b = (4.57264, 3.603475, 3.998249, 4.704911, 3.569123, 4.039699),$
- $a = (2.017251, 1.206258, 0.1953755, 0.6045331, 2.117807, 0.1142521, 0.347635, 2.052791) and b = (8.192137, 5.166402, 4.029354, 5.332025, 4.137933, 4.482512, 4.022804, 6.569997),$
- $a = (2.799903, 0.6656203, 2.566126, 2.952712, 1.740536, 1.740826, 2.063482, 2.419983, 0.4464243, 0.705589) and b = (5.651544, 4.950088, 3.827902, 4.637529, 2.837013, 5.226642, 6.278133, 5.515519, 5.312168, 3.058361),$

140
Case 3: $a_i$ simulated from uniform $[0, 3]$ and $b_i$ simulated from chisquared $(3)$ independently as

- $a = (1.848669, 1.378527, 0.4404986, 2.662331, 0.22705, 1.40141)$ and $b = (3.931913, 2.185307, 5.446896, 0.7627859, 1.424711, 1.790901),$ 
- $a = (2.453322, 1.789157, 1.381468, 0.5512082, 2.52837, 1.338784, 0.7274495, 1.176251)$ and $b = (5.445, 2.990886, 0.7683813, 2.293982, 0.107612, 1.503014, 2.009508, 1.429635),$ 
- $a = (0.2176901, 2.136661, 0.6973105, 0.9692848, 1.001722, 1.00783, 1.203301, 0.6380348, 0.3121062, 1.854234)$ and $b = (2.64449, 0.2922778, 3.351485, 0.8765912, 1.531889, 11.25758, 2.687317, 3.132109, 4.735183, 1.318886),$ 
- $a = (2.38574, 2.684072, 1.289495, 1.741308, 0.5535498, 1.734817, 2.362223, 2.157232, 1.069317, 0.7457933, 1.355856, 1.967968)$ and $b = (3.413016, 2.300216, 1.19086, 2.635635, 2.746352, 13.43873, 0.8506862, 0.3576513, 0.3178849, 3.8022, 1.544177, 3.714247).$

Case 4: $a_i$ simulated from uniform $[0, 3]$ and $b_i$ simulated from lognormal $(1, 3)$ independently as

- $a = (2.609709, 1.793544, 1.160203, 0.8257406, 0.206197, 2.952884)$ and $b = (0.1215165, 0.07059568, 0.4970423, 4.410984, 1.642273, 0.6080638),$ 
- $a = (2.384237, 0.1646325, 1.568915, 1.926518, 1.392569, 2.051041, 0.1912039, 2.934475)$ and $b = (20.18051, 3.425202, 0.7037836, 23.0033, 2.444438, 0.7719453, 95.2414, 0.3458288),$ 
- $a = (0.6758847, 0.5023112, 0.2281325, 0.2013796, 2.92679, 0.5091848, 0.3981476, 1.744489, 2.895798, 2.988323)$ and $b = (0.5469263, 1.802245, 6.915945, 0.1820084, 4.972993, 12.66339, 12.17815, 24.35371, 2.439753, 7.967482),$ 
- $a = (1.945457, 2.64646, 1.347509, 2.407616, 1.426061, 2.824233, 0.2837708, 0.3755052, 2.874297, 0.5403967, 0.771879, 2.152608)$ and $b = (0.02375104, 0.1568451, 311.1175, 234.1375, 0.89355, 0.01641649, 0.01312351, 0.1023698, 13.21694, 0.001039766, 17.06562, 1073.126).$

Case 5: $a_i$ simulated from Weibull $(3, 5)$ and $b_i$ simulated from uniform $[0, 3]$ independently as

- $a = (5.273956, 4.610141, 2.389271, 5.037316, 3.589953, 3.904395)$ and $b = (2.944036, 2.475881, 1.531461, 2.132437, 2.43523, 2.290124),$
Case 7: $a_i$ simulated from Weibull (3, 5) and $b_i$ simulated from Weibull (3, 5) independently as

- $a = (3.871054, 3.535526, 4.604096, 3.68539, 5.01035, 2.581044, 2.818227, 4.386177)$ and $b = (0.4182519, 0.3534002, 2.927693, 2.262205, 1.37849, 1.310595, 2.820153, 1.218306),$
- $a = (4.132925, 3.017805, 6.04032, 5.369273, 1.231136, 6.477537, 7.598634, 4.980612, 1.06686, 3.710756)$ and $b = (2.843196, 2.870392, 0.556707, 0.06754689, 1.761655, 0.6621931, 0.5155944, 0.8233932, 0.5012798, 1.140807),$
- $a = (5.841154, 6.967267, 5.638253, 5.914464, 3.263699, 6.466435, 5.552804, 3.702132, 3.109833, 6.8105, 4.527539, 4.980747)$ and $b = (2.196082, 2.635602, 0.7778389, 2.576844, 0.4539989, 0.6913288, 0.9882215, 2.895813, 1.819437, 0.9296329, 2.361362, 1.92344).

Case 6: $a_i$ simulated from Weibull (3, 5) and $b_i$ simulated from Weibull (3, 5) independently as

- $a = (6.224859, 2.383691, 1.246752, 5.070905, 7.395576, 4.638046)$ and $b = (4.555933, 4.48759, 5.527347, 3.321585, 4.148188, 6.623512),$
- $a = (3.939015, 5.966001, 3.2889, 7.532086, 3.934509, 4.442056, 5.646698, 6.857782)$ and $b = (3.101816, 4.080915, 2.621501, 4.617988, 7.478792, 2.352317, 4.518471, 5.115784),$
- $a = (3.998188, 6.18554, 3.106791, 5.855275, 5.544518, 5.293995, 4.920894, 5.825265, 3.535835, 6.614369)$ and $b = (7.090504, 7.797076, 3.002553, 4.390184, 4.235928, 6.03972, 4.584893, 6.282284, 3.349656, 6.432941),$

Case 7: $a_i$ simulated from Weibull (3, 5) and $b_i$ simulated from chisquared (3) independently as

- $a = (3.531276, 4.032538, 8.647575, 5.305838, 6.13223, 4.146321)$ and $b = (0.3451049, 8.414199, 8.511304, 4.731345, 4.525146, 2.921905),$
- $a = (5.497293, 5.053544, 2.3919, 2.81458, 1.854095, 3.739478, 4.568025, 2.905965)$ and $b = (1.344997, 2.133204, 2.525747, 3.731614, 1.179878, 3.649183, 2.018736, 1.631682),$
- $a = (4.007546, 4.593861, 7.430956, 5.921879, 6.103289, 5.923999, 2.104499, 5.155392, 3.82353, 6.210309)$ and $b = (3.169027, 3.183852, 1.698119, 3.536053, 1.345952, 1.642581, 4.116989, 0.3984947, 1.578258, 0.1749832),$
- $a = (9.15205, 2.353533, 2.500374, 5.099444, 3.275363, 6.887862, 3.844962, 4.649605, 2.593291, 5.560257, 5.733111, 5.655299)$ and $b = (4.088533, 0.1538948, 0.7835756, 2.399516, 1.089951, 2.314194, 2.353753, 0.7981358, 0.8876652, 2.871523, 3.204572, 2.061364).$
Case 8: $a_i$ simulated from Weibull (3, 5) and $b_i$ simulated from lognormal (1, 3) independently as

- $a = (6.154684, 3.835334, 3.495301, 4.34435, 5.703149, 6.548319)$ and $b = (0.1219968, 0.2810046, 53.34952, 2.345394, 42.28274, 0.03966336),$
- $a = (3.93218, 4.580583, 4.849474, 3.723203, 3.644307, 2.38203, 4.926058, 5.550437)$ and $b = (1.042865, 0.05577661, 0.03030607, 0.7670033, 0.05761375, 719.2797, 0.00794546, 0.8062949),$
- $a = (3.923938, 1.224206, 5.478915, 7.552481, 3.570537, 4.12205, 6.51977, 1.911965, 4.931234, 7.543953)$ and $b = (0.02151777, 80.21761, 296.9334, 10.87638, 0.2601826, 1.850425, 0.1796827, 52.21101, 0.06905063, 0.4560928),$
- $a = (6.981487, 5.406974, 4.369757, 5.433589, 3.983309, 6.139726, 5.229663, 2.901443, 6.130317, 4.973267, 2.857522, 5.550283)$ and $b = (57.00482, 1626.797, 8.827263, 0.1146537, 124.1582, 104.4949, 1.144981, 0.2294726, 62.22093, 2.116839, 2883.756, 6.570516),$

Case 9: $a_i$ simulated from chisquared (3) and $b_i$ simulated from uniform [0, 3] independently as

- $a = (3.650032, 1.893852, 0.4376349, 0.3953226, 0.362869, 3.980955)$ and $b = (2.526928, 0.5476575, 2.336397, 0.2255485, 2.054914, 2.061315),$
- $a = (1.3165, 1.170919, 3.502502, 3.127233, 1.924567, 0.2008392, 1.18299, 2.432505)$ and $b = (2.399344, 2.321608, 2.110988, 0.03579314, 0.5994146, 0.648734, 0.487861, 1.700262),$
- $a = (2.64154, 2.505644, 6.785222, 0.8188809, 1.818816, 2.345548, 5.708997, 3.491265, 1.154884, 1.680403)$ and $b = (2.104083, 2.849131, 2.330287, 0.02010747, 0.5658654, 0.2356123, 1.562978, 1.447809, 2.187167, 1.647641),$
- $a = (7.544003, 3.568231, 8.036778, 6.525969, 5.416202, 5.55678, 1.370177, 3.130449, 2.259869, 1.969572, 2.000751, 0.7588828)$ and $b = (0.4790278, 0.06675777, 1.395189, 2.117694, 1.995748, 1.069335, 2.929283, 0.2243011, 1.476158, 0.6044525, 1.889666, 1.790853),$

Case 10: $a_i$ simulated from chisquared (3) and $b_i$ simulated from Weibull (3, 5) independently as

- $a = (3.789622, 2.199284, 4.321561, 0.3378058, 5.877808, 0.257535)$ and $b = (4.177478, 4.432327, 1.714618, 4.853774, 5.143581, 0.989115),$
- $a = (8.245037, 0.1211343, 3.872118, 1.42064, 0.0289729, 4.590378, 5.002248, 1.54276)$ and $b = (4.542927, 3.167449, 4.734756, 4.487352, 2.616288, 5.386692, 4.360018, 3.888895),$
- $a = (3.574066, 2.780634, 1.007904, 5.343683, 1.750636, 4.850619, 1.966682, 0.6103026, 3.475134, 6.71659)$ and $b = (5.714227, 5.070929, 4.514497, 3.353621, 7.051511, 1.810287, 1.335049, 4.21586, 5.288201, 6.6341),$
Case 11: $a_i$ simulated from chisquared (3) and $b_i$ simulated from chisquared (3) independently as

- $a = (0.5284351, 3.041898, 3.402771, 1.357026, 3.257756, 5.823867)$ and $b = (3.662309, 3.709705, 10.76515, 9.691956, 4.468251, 6.882667)$,
- $a = (7.827955, 0.2090685, 3.791955, 1.050267, 2.648588, 5.047332, 1.190545, 3.958732)$ and $b = (1.541705, 0.6007765, 1.435225, 2.372091, 7.364606, 0.8101704, 1.952104, 2.565142)$,
- $a = (1.774117, 2.106949, 6.422483, 6.111116, 5.991397, 1.276793, 0.5070477, 0.3032951, 0.4325267, 3.526915)$ and $b = (6.348678, 0.1185307, 7.880857, 1.419293, 3.039167, 12.35556, 3.835319, 2.56608, 1.835552, 1.938718)$,
- $a = (1.841189, 3.240125, 2.807235, 6.665391, 0.5135182, 0.8503191, 5.72202, 2.213338, 4.911463, 2.214239, 4.888353, 2.907367)$ and $b = (2.275809, 6.549404, 0.4924177, 1.823462, 1.966943, 0.4144944, 3.913419, 5.44639, 0.8300679, 4.235747, 3.649807, 10.28398)$.

Case 12: $a_i$ simulated from chisquared (3) and $b_i$ simulated from lognormal (1, 3) independently as

- $a = (1.162989, 0.963536, 2.495387, 3.854517, 3.059907, 4.121989)$ and $b = (8.192419, 2.754437, 213.0764, 0.1981606, 11.17125, 0.3870891)$,
- $a = (1.742243, 1.304452, 1.15187, 1.114471, 1.451578, 7.196846, 1.000837, 3.80333)$ and $b = (0.003437644, 1.231358, 10.69447, 540.4851, 1.285839, 1.696417, 0.6426382, 404.5152)$,
- $a = (0.01868152, 4.6015, 3.529983, 3.435674, 3.068682, 0.7357939, 2.670088, 2.464441, 3.637356, 1.408243)$ and $b = (116.9773, 0.03367166, 6.805159, 24.10298, 72.84418, 143.1154, 5.345697, 0.4986386, 1.286501, 0.2071632)$,
- $a = (3.57359, 3.843265, 3.685503, 3.3834, 2.67798, 4.893151, 0.8193489, 5.814517, 6.852481, 1.353179, 3.208474, 3.344128)$ and $b = (450.6367, 8777.122, 89.05154, 1.223546, 2.957545, 2.569918, 242.7423, 4.297934, 67.70167, 0.06203204, 231.932, 11.4393)$.

Case 13: $a_i$ simulated from lognormal (1, 3) and $b_i$ simulated from uniform [0, 3] independently as

- $a = (0.2376846, 47.19699, 71.74766, 1.779527, 30.33478, 71.72048)$ and $b = (0.1766171, 0.974097, 2.128963, 0.19879, 1.626444, 2.575717)$,
Case 14: $a_i$ simulated from lognormal $(1,3)$ and $b_i$ simulated from Weibull $(3,5)$ independently as

- $a = (15.06097, 0.4628673, 2.663943, 3.833967, 0.05403325, 0.3830108)$ and $b = (4.238042, 2.767669, 3.133979, 5.135852, 5.344516, 2.130748)$,
- $a = (0.9974214, 28.83534, 645.333, 0.09914743, 219.4095, 3.257483, 0.6751812, 4.709165)$ and $b = (2.588262, 7.523441, 6.531016, 4.375314, 7.986907, 6.381553, 3.821439, 4.493831)$,
- $a = (205.2537, 3.435555, 323.7743, 75.92212, 0.4969098, 2.960647, 7.311558, 0.001785917, 0.00830147, 1.368675)$ and $b = (3.375639, 3.396994, 3.134946, 2.524961, 2.640427, 7.297835, 1.226491, 6.137272, 6.093392, 4.800603)$,
- $a = (0.01990273, 74.95541, 869.592, 23.72196, 29.30186, 3.343322, 0.9182575, 0.3653378, 13.21507, 20.78439, 2.770618, 4.841228)$ and $b = (5.400236, 4.205336, 7.458074, 3.104021, 4.913819, 6.775869, 6.157434, 5.392798, 6.034453, 2.222937, 8.749919, 3.467849)$.

Case 15: $a_i$ simulated from lognormal $(1,3)$ and $b_i$ simulated from chisquared $(3)$ independently as

- $a = (0.1981479, 0.4892375, 1.749151, 0.7551661, 4.382262, 1.140357)$ and $b = (2.656272, 0.92555, 1.892894, 3.568952, 3.333083, 3.739689)$,
- $a = (0.5948598, 13.75449, 0.01539715, 0.3507828, 9.317147, 8.482252, 3225.614, 7.336529)$ and $b = (2.501029, 5.04406, 2.713488, 4.585566, 6.222398, 3.439231, 0.2364512, 1.736901)$,
- $a = (11.3165, 5.026143, 50.26044, 10.80203, 0.3510156, 0.2260627, 0.1352482, 0.317119, 0.7910012, 0.1559992)$ and $b = (2.001264, 1.047305, 4.565139, 2.460229, 9.836424, 0.9610001, 1.939174, 0.5635076, 0.08525531, 0.2284625)$,
- $a = (0.8133675, 1.242386, 363.8553, 2.276828, 0.1774438, 0.3783875, 0.09207683, 1.216534, 1.077944, 33.47727, 892.6558, 9.869482)$ and $b = (4.368883, 1.712376, 6.183354, 1.612633, 1.637485, 1.228592, 1.291486, 11.51825, 0.8319545, 0.2504083, 1.73909, 0.2175308)$.
Case 16: $a_i$ simulated from lognormal (1, 3) and $b_i$ simulated from lognormal (1, 3) independently as

- $a = (2.613738, 2.656848, 2.888233, 1.560111, 1.071585, 4788.66)$ and $b = (1.186717, 0.3329809, 18.08782, 5.016621, 0.3703533, 0.4333734)$,

- $a = (4.77529, 23.56734, 9.888781, 1.484049, 54.92544, 16.35135, 1.175349, 9.042096)$ and $b = (12.25761, 0.524581, 48.30502, 5.176546, 0.1305331, 0.5444215, 0.694618, 0.01514411)$,

- $a = (2.554265, 9.806766, 5.404822, 10.41458, 1.938993, 0.4978821, 17.41773, 3397.616, 6.180577, 7.022354)$ and $b = (0.6086207, 0.1107876, 0.0512613, 8.690013, 9.46448, 3.254687, 0.04277155, 3.652914, 0.249186, 22.24216)$,

- $a = (52.66761, 1.785833, 0.1465873, 323.8589, 1.265566, 0.0743344, 22.93243, 2.554163, 14.77695, 0.5469197, 4.264674, 0.00397322)$ and $b = (2.143064, 2.887734, 0.2374735, 14.93014, 7.531026, 0.4232416, 1.879243, 1.874091, 0.01817543, 0.4578481, 12.30416, 16.02691)$.
## Appendix B

### Appendix to Chapter 6

<table>
<thead>
<tr>
<th>Distribution</th>
<th>– ln L</th>
<th>AIC</th>
<th>AICC</th>
<th>BIC</th>
<th>HQC</th>
<th>CAIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student t</td>
<td>-2303.9</td>
<td>-4601.8</td>
<td>-4601.8</td>
<td>-4587.3</td>
<td>-4596.3</td>
<td>-4584.3</td>
</tr>
<tr>
<td>Laplace</td>
<td>-2289.3</td>
<td>-4574.7</td>
<td>-4574.7</td>
<td>-4565.0</td>
<td>-4571.0</td>
<td>-4563.0</td>
</tr>
<tr>
<td>Skew t</td>
<td>-2304.0</td>
<td>-4600.0</td>
<td>-4599.9</td>
<td>-4580.5</td>
<td>-4592.6</td>
<td>-4576.5</td>
</tr>
<tr>
<td>GT</td>
<td>-2325.7</td>
<td>-4643.3</td>
<td>-4643.3</td>
<td>-4623.8</td>
<td>-4635.9</td>
<td>-4619.8</td>
</tr>
<tr>
<td>SST</td>
<td>-2304.0</td>
<td>-4600.0</td>
<td>-4599.9</td>
<td>-4580.5</td>
<td>-4592.6</td>
<td>-4576.5</td>
</tr>
<tr>
<td>AST</td>
<td>-2304.3</td>
<td>-4598.6</td>
<td>-4598.6</td>
<td>-4574.3</td>
<td>-4589.4</td>
<td>-4569.3</td>
</tr>
<tr>
<td>NIG</td>
<td>-2316.0</td>
<td>-4623.9</td>
<td>-4623.9</td>
<td>-4604.5</td>
<td>-4616.5</td>
<td>-4600.5</td>
</tr>
<tr>
<td>GH</td>
<td>-2325.7</td>
<td>-4641.5</td>
<td>-4641.4</td>
<td>-4617.1</td>
<td>-4632.2</td>
<td>-4612.1</td>
</tr>
</tbody>
</table>

Table B.1: Fitted distributions and information criteria results for daily log returns of the exchange rates of Bitcoin from 23 June 2014 until 28 February 2017.
### Table B.2: Fitted distributions and information criteria results for daily log returns of the exchange rates of Dash from 23 June 2014 until 28 February 2017.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$-\ln L$</th>
<th>AIC</th>
<th>AICC</th>
<th>BIC</th>
<th>HQC</th>
<th>CAIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student t</td>
<td>-1704.8</td>
<td>-3403.6</td>
<td>-3403.6</td>
<td>-3389.0</td>
<td>-3398.1</td>
<td>-3386.0</td>
</tr>
<tr>
<td>Laplace</td>
<td>-1689.4</td>
<td>-3374.8</td>
<td>-3374.7</td>
<td>-3365.0</td>
<td>-3371.1</td>
<td>-3363.0</td>
</tr>
<tr>
<td>Skew t</td>
<td>-1707.4</td>
<td>-3406.8</td>
<td>-3406.8</td>
<td>-3387.3</td>
<td>-3399.4</td>
<td>-3383.3</td>
</tr>
<tr>
<td>GT</td>
<td>-1708.9</td>
<td>-3409.7</td>
<td>-3409.7</td>
<td>-3390.3</td>
<td>-3402.3</td>
<td>-3386.3</td>
</tr>
<tr>
<td>SST</td>
<td>-1707.5</td>
<td>-3407.0</td>
<td>-3406.9</td>
<td>-3387.5</td>
<td>-3399.6</td>
<td>-3383.5</td>
</tr>
<tr>
<td>AST</td>
<td>-1707.5</td>
<td>-3405.1</td>
<td>-3405.0</td>
<td>-3380.8</td>
<td>-3395.8</td>
<td>-3375.8</td>
</tr>
<tr>
<td>NIG</td>
<td>-1710.4</td>
<td>-3412.8</td>
<td>-3412.8</td>
<td>-3393.4</td>
<td>-3405.4</td>
<td>-3389.4</td>
</tr>
<tr>
<td>GH</td>
<td>-1710.5</td>
<td>-3410.9</td>
<td>-3410.8</td>
<td>-3386.6</td>
<td>-3401.6</td>
<td>-3381.6</td>
</tr>
</tbody>
</table>

### Table B.3: Fitted distributions and information criteria results for daily log returns of the exchange rates of Dogecoin from 23 June 2014 until 28 February 2017.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$-\ln L$</th>
<th>AIC</th>
<th>AICC</th>
<th>BIC</th>
<th>HQC</th>
<th>CAIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student t</td>
<td>-2037.7</td>
<td>-4069.4</td>
<td>-4069.4</td>
<td>-4054.8</td>
<td>-4063.8</td>
<td>-4051.8</td>
</tr>
<tr>
<td>Laplace</td>
<td>-1985.9</td>
<td>-3967.8</td>
<td>-3967.8</td>
<td>-3958.1</td>
<td>-3964.1</td>
<td>-3956.1</td>
</tr>
<tr>
<td>Skew t</td>
<td>-2037.7</td>
<td>-4067.8</td>
<td>-4067.8</td>
<td>-4058.1</td>
<td>-4064.1</td>
<td>-4056.1</td>
</tr>
<tr>
<td>GT</td>
<td>-2051.9</td>
<td>-4095.8</td>
<td>-4095.8</td>
<td>-4076.4</td>
<td>-4084.8</td>
<td>-4072.4</td>
</tr>
<tr>
<td>SST</td>
<td>-2037.7</td>
<td>-4067.4</td>
<td>-4067.3</td>
<td>-4047.9</td>
<td>-4060.0</td>
<td>-4043.9</td>
</tr>
<tr>
<td>AST</td>
<td>-2039.6</td>
<td>-4069.1</td>
<td>-4069.1</td>
<td>-4044.8</td>
<td>-4059.9</td>
<td>-4039.8</td>
</tr>
<tr>
<td>NIG</td>
<td>-2048.1</td>
<td>-4088.1</td>
<td>-4088.1</td>
<td>-4068.7</td>
<td>-4080.7</td>
<td>-4064.7</td>
</tr>
<tr>
<td>GH</td>
<td>-2052.2</td>
<td>-4094.3</td>
<td>-4094.2</td>
<td>-4070.0</td>
<td>-4085.0</td>
<td>-4065.0</td>
</tr>
</tbody>
</table>
APPENDIX B. APPENDIX TO CHAPTER 6

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$-\ln L$</th>
<th>AIC</th>
<th>AICC</th>
<th>BIC</th>
<th>HQC</th>
<th>CAIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student t</td>
<td>-2113.3</td>
<td>-4220.7</td>
<td>-4220.7</td>
<td>-4206.1</td>
<td>-4215.1</td>
<td>-4203.1</td>
</tr>
<tr>
<td>Laplace</td>
<td>-2020.0</td>
<td>-4036.0</td>
<td>-4036.0</td>
<td>-4026.3</td>
<td>-4032.3</td>
<td>-4024.3</td>
</tr>
<tr>
<td>Skew t</td>
<td>-2113.6</td>
<td>-4219.3</td>
<td>-4219.3</td>
<td>-4199.8</td>
<td>-4211.9</td>
<td>-4195.8</td>
</tr>
<tr>
<td>GT</td>
<td>-2125.7</td>
<td>-4243.4</td>
<td>-4243.4</td>
<td>-4223.9</td>
<td>-4236.0</td>
<td>-4219.9</td>
</tr>
<tr>
<td>SST</td>
<td>-2113.5</td>
<td>-4219.1</td>
<td>-4219.1</td>
<td>-4199.6</td>
<td>-4211.7</td>
<td>-4195.6</td>
</tr>
<tr>
<td>AST</td>
<td>-2113.6</td>
<td>-4217.2</td>
<td>-4217.1</td>
<td>-4192.8</td>
<td>-4207.9</td>
<td>-4187.8</td>
</tr>
<tr>
<td>NIG</td>
<td>-2123.9</td>
<td>-4239.9</td>
<td>-4239.9</td>
<td>-4220.4</td>
<td>-4232.5</td>
<td>-4216.4</td>
</tr>
<tr>
<td>GH</td>
<td>-2130.6</td>
<td>-4251.2</td>
<td>-4251.1</td>
<td>-4226.9</td>
<td>-4241.9</td>
<td>-4221.9</td>
</tr>
</tbody>
</table>

Table B.4: Fitted distributions and information criteria results for daily log returns of the exchange rates of Litecoin from 23 June 2014 until 28 February 2017.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$-\ln L$</th>
<th>AIC</th>
<th>AICC</th>
<th>BIC</th>
<th>HQC</th>
<th>CAIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student t</td>
<td>-1533.2</td>
<td>-3060.3</td>
<td>-3060.3</td>
<td>-3045.7</td>
<td>-3054.8</td>
<td>-3042.7</td>
</tr>
<tr>
<td>Laplace</td>
<td>-1540.4</td>
<td>-3076.8</td>
<td>-3076.8</td>
<td>-3067.1</td>
<td>-3073.1</td>
<td>-3065.1</td>
</tr>
<tr>
<td>Skew t</td>
<td>-1533.3</td>
<td>-3058.6</td>
<td>-3058.5</td>
<td>-3039.1</td>
<td>-3051.2</td>
<td>-3035.1</td>
</tr>
<tr>
<td>GT</td>
<td>-1541.9</td>
<td>-3075.9</td>
<td>-3075.8</td>
<td>-3056.4</td>
<td>-3068.5</td>
<td>-3052.4</td>
</tr>
<tr>
<td>SST</td>
<td>-1533.3</td>
<td>-3058.5</td>
<td>-3058.5</td>
<td>-3039.1</td>
<td>-3051.1</td>
<td>-3035.1</td>
</tr>
<tr>
<td>AST</td>
<td>-1533.4</td>
<td>-3056.7</td>
<td>-3056.7</td>
<td>-3032.4</td>
<td>-3047.5</td>
<td>-3027.4</td>
</tr>
<tr>
<td>NIG</td>
<td>-1539.0</td>
<td>-3070.0</td>
<td>-3070.0</td>
<td>-3050.6</td>
<td>-3062.6</td>
<td>-3046.6</td>
</tr>
<tr>
<td>GH</td>
<td>-1542.4</td>
<td>-3074.9</td>
<td>-3074.8</td>
<td>-3050.6</td>
<td>-3065.6</td>
<td>-3045.6</td>
</tr>
</tbody>
</table>

Table B.5: Fitted distributions and information criteria results for daily log returns of the exchange rates of MaidSafeCoin from 23 June 2014 until 28 February 2017.
### APPENDIX B. APPENDIX TO CHAPTER 6

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$-\ln L$</th>
<th>AIC</th>
<th>AICC</th>
<th>BIC</th>
<th>HQC</th>
<th>CAIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student t</td>
<td>-1438.6</td>
<td>-2871.2</td>
<td>-2871.2</td>
<td>-2856.6</td>
<td>-2865.6</td>
<td>-2853.6</td>
</tr>
<tr>
<td>Laplace</td>
<td>-1432.2</td>
<td>-2860.5</td>
<td>-2860.5</td>
<td>-2850.7</td>
<td>-2856.8</td>
<td>-2848.7</td>
</tr>
<tr>
<td>Skew t</td>
<td>-1440.0</td>
<td>-2871.9</td>
<td>-2871.9</td>
<td>-2852.5</td>
<td>-2864.5</td>
<td>-2848.5</td>
</tr>
<tr>
<td>GT</td>
<td>-1440.5</td>
<td>-2872.9</td>
<td>-2872.9</td>
<td>-2853.5</td>
<td>-2865.5</td>
<td>-2849.5</td>
</tr>
<tr>
<td>SST</td>
<td>-1439.7</td>
<td>-2871.4</td>
<td>-2871.4</td>
<td>-2852.0</td>
<td>-2864.0</td>
<td>-2848.0</td>
</tr>
<tr>
<td>AST</td>
<td>-1440.6</td>
<td>-2871.2</td>
<td>-2871.1</td>
<td>-2846.9</td>
<td>-2861.9</td>
<td>-2841.9</td>
</tr>
<tr>
<td>NIG</td>
<td>-1441.9</td>
<td>-2875.8</td>
<td>-2875.7</td>
<td>-2856.3</td>
<td>-2868.4</td>
<td>-2852.3</td>
</tr>
<tr>
<td>GH</td>
<td>-1442.0</td>
<td>-2874.0</td>
<td>-2874.0</td>
<td>-2849.7</td>
<td>-2864.8</td>
<td>-2844.7</td>
</tr>
</tbody>
</table>

Table B.6: Fitted distributions and information criteria results for daily log returns of the exchange rates of Monero from 23 June 2014 until 28 February 2017.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$-\ln L$</th>
<th>AIC</th>
<th>AICC</th>
<th>BIC</th>
<th>HQC</th>
<th>CAIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student t</td>
<td>-1867.6</td>
<td>-3729.1</td>
<td>-3729.1</td>
<td>-3714.5</td>
<td>-3723.6</td>
<td>-3711.5</td>
</tr>
<tr>
<td>Laplace</td>
<td>-1831.9</td>
<td>-3659.9</td>
<td>-3659.9</td>
<td>-3650.2</td>
<td>-3656.2</td>
<td>-3648.2</td>
</tr>
<tr>
<td>Skew t</td>
<td>-1869.5</td>
<td>-3731.1</td>
<td>-3731.0</td>
<td>-3711.6</td>
<td>-3723.7</td>
<td>-3707.6</td>
</tr>
<tr>
<td>GT</td>
<td>-1870.6</td>
<td>-3733.1</td>
<td>-3733.1</td>
<td>-3713.7</td>
<td>-3725.7</td>
<td>-3709.7</td>
</tr>
<tr>
<td>SST</td>
<td>-1869.3</td>
<td>-3730.5</td>
<td>-3730.5</td>
<td>-3711.1</td>
<td>-3723.1</td>
<td>-3707.1</td>
</tr>
<tr>
<td>AST</td>
<td>-1869.7</td>
<td>-3729.4</td>
<td>-3729.4</td>
<td>-3705.1</td>
<td>-3720.2</td>
<td>-3700.1</td>
</tr>
<tr>
<td>NIG</td>
<td>-1875.6</td>
<td>-3743.3</td>
<td>-3743.2</td>
<td>-3723.8</td>
<td>-3735.8</td>
<td>-3719.8</td>
</tr>
<tr>
<td>GH</td>
<td>-1875.9</td>
<td>-3741.8</td>
<td>-3741.8</td>
<td>-3717.5</td>
<td>-3732.6</td>
<td>-3712.5</td>
</tr>
</tbody>
</table>

Table B.7: Fitted distributions and information criteria results for daily log returns of the exchange rates of Ripple from 23 June 2014 until 28 February 2017.
Figure B.1: The Q-Q plots of the best fitting distributions for daily log returns of the exchange rates of Bitcoin (first row, left), Dash (first row, right), Dogecoin (second row, left), Litecoin (second row, right), MaidSafeCoin (third row, left), Monero (third row, right) and Ripple (last row) from 23 June 2014 until 28 February 2017.
Figure B.2: The P-P plots of the best fitting distributions for daily log returns of the exchange rates of Bitcoin (first row, left), Dash (first row, right), Dogecoin (second row, left), Litecoin (second row, right), MaidSafeCoin (third row, left), Monero (third row, right) and Ripple (last row) from 23 June 2014 until 28 February 2017.
Figure B.3: Value at risk for the best fitting distributions for daily log returns of the exchange rates of Bitcoin (first row, left), Dash (first row, right), Dogecoin (second row, left), Litecoin (second row, right), MaidSafeCoin (third row, left), Monero (third row, right) and Ripple (last row) from 23 June 2014 until 28 February 2017. Also shown are the value at risk for daily log returns of the exchange rates of the Euro, computed using the same best fitting distributions.
Figure B.4: Expected shortfall for the best fitting distributions for daily log returns of the exchange rates of Bitcoin (first row, left), Dash (first row, right), Dogecoin (second row, left), Litecoin (second row, right), MaidSafeCoin (third row, left), Monero (third row, right) and Ripple (last row) from 23 June 2014 until 28 February 2017. Also shown are the expected shortfall values for daily log returns of the exchange rates of the Euro, computed using the same best fitting distributions.
Figure B.5: Kupiec’s p-values for the best fitting distributions for daily log returns of the exchange rates of Bitcoin (first row, left), Dash (first row, right), Dogecoin (second row, left), Litecoin (second row, right), MaidSafeCoin (third row, left), Monero (third row, right) and Ripple (last row) from 23 June 2014 until 28 February 2017.
Figure B.6: Histograms of standard deviations of daily log returns of the exchange rates of the seven cryptocurrencies over windows of width 20 days. Also shown is the histogram of standard deviations of daily log returns of the exchange rates of the Euro over windows of width 20 days.