EM Algorithms for Multivariate Skewed Variance Gamma Distribution with Unbounded Densities and Applications

Thanakorn Nitithumbundit

A thesis submitted in fulfillment of
the requirements for the degree of
Doctor of Philosophy

School of Mathematics and Statistics
Faculty of Science
University of Sydney

April 2018

Abstract

The multivariate skewed variance gamma (VG) distribution is useful for modelling data with heavy-tails and high density around the location parameter. When the shape parameter is sufficiently small, the density function is unbounded at the location parameter. Not much research have been conducted to deal with distributions with unbounded density at the location parameter especially under the multivariate case.

In this thesis, we proposed three modifications to appropriately bound the likelihood function so that the maximum is well-defined. These modified likelihoods are the capped, leave-one-out (LOO), and weighted LOO likelihoods. Moreover, we present expectation/conditional maximisation (ECM) algorithms to accurately estimate parameters of the VG distribution using its normal mean-variance mixture representation and the three proposed likelihoods.

Apart from parameter estimation, we also calculate standard errors (SEs) to assess the significance of the parameter estimates. However, the SE calculation requires calculation of the observed information matrix for the VG distribution which is tedious as it involves the second order derivative of the log-likelihood function with respect to vector/matrices. We derive these formulas to efficiently compute the observed and Fisher information matrices for the VG distribution by applying new matrix differentiation formulas.

These SE calculations rely on asymptotic properties of the maximum likelihood estimator (MLE) which have been extensively studied under the smooth likelihood case. For the cusp/unbounded case, proving these asymptotic properties are a challenge as they do not satisfy the smoothness regularity condition. We numerically investigate these asymptotic properties for the location estimator when the likelihood function has cusp or unbounded points. We demonstrated its super-efficient rate of convergence and found the double generalised gamma distribution provides a good approximation to the asymptotic distribution of the location parameter.

Lastly, the ECM algorithms are applied to vector autoregressive moving average model with VG and Student's t innovations to capture serial correlation, leptokurtosis, skewness, and cross dependence of return data from high frequency stock indices and cryptocurrencies.

Acknowledgements

Firstly, I would like to thank my supervisor Assoc. Prof. Jennifer Chan for her ongoing support and motivation throughout my PhD. Her guidance and hard work helped me work to the best of my ability while giving me plenty of opportunities to explore and discover places that was thought to be unreachable.

I would like to extend my gratitude to Professor Eugene Seneta and Dr. Lamiae Azizi for their helpful advice and insightful discussions even though it was difficult to organise meetings during our busy schedule.

I want to give my special thanks to my fellow colleagues in level 8. In particular, thanks to Matt, Eddie, Joshua, Chong, William, Weichang, Hongxuan, Mark, Kevin, Lucy, Emi, Elynor, Kerry, Helen and James for providing me with countless hours of simulating discussions and puzzles which are a good distraction during my prolonged period of procrastination. And a big thanks to Andrew for supplying me with his invaluable Bitcoin data set.

Lastly, I am indebted to my whole family especially my parents and brother for their unbounded support and encouragement throughout my life.

Table of Contents

Abstract ... iii

Acknowledgements ... v

List of Figures ... xiii

List of Tables ... xv

Chapter 1. Introduction ... 1

1.1 Background ... 1

1.2 Maximum likelihood estimation ... 7

1.2.1 Likelihood function ... 7

1.2.2 Information matrix ... 8

1.2.3 Newton-Raphson method ... 9

1.2.4 Properties of MLE ... 10

1.3 EM algorithm ... 11

1.3.1 Introduction ... 11

1.3.2 Convergence of EM algorithm ... 13

1.3.3 Score function with missing data ... 16

1.3.4 Information matrix with missing data ... 16

1.3.5 Rate of convergence ... 17

1.4 Extensions to EM algorithm ... 18

1.4.1 ECM algorithm ... 19

1.4.2 MCECM algorithm ... 21

1.4.3 ECME algorithm ... 22

1.5 Normal mean-variance mixture representation ... 23

1.5.1 Generalised inverse Gaussian distribution ... 23

1.5.2 Generalised hyperbolic distribution ... 26

1.5.3 Variance gamma distribution ... 28

1.6 Contributions and structure of the thesis ... 31

Chapter 2. EM Algorithms for Variance Gamma Distribution **35**
2.1 ECM algorithm for VG distribution 36
2.1.1 E-step 37
2.1.2 CM-step for $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ and $\boldsymbol{\gamma}$ 38
2.1.3 CM-step for ν 39
2.2 Alternating ECM algorithm for skewed VG distribution 40
2.3 Capped likelihood method for dealing with unbounded likelihood 41
2.4 Observed information matrix 43
2.4.1 Hessian matrix by numerical differentiation 43
2.4.2 Louis' method 43
2.4.3 Fisher information matrix 45
2.4.4 Singularity of the information matrix 45
2.5 Simulation studies 46
2.5.1 Comparing EM algorithms 46
2.5.2 Optimal choice of capping level 48
2.5.3 Comparing standard error calculations 49
2.6 Application 52
2.7 Conclusion 54

Chapter 3. Estimation using Leave-one-out Likelihood **57**
3.1 Introduction 57
3.2 Maximum leave-one-out likelihood 58
3.2.1 Leave-one-out likelihood 58
3.2.2 Properties of maximum LOO likelihood estimator 59
3.3 AECM algorithm using LOO likelihood 61
3.3.1 E-step 62
3.3.2 CM-step 62
3.3.3 AECM algorithm 66
3.3.4 Convergence of AECM algorithm using LOO likelihood 67
3.4 Simulation studies 69
3.4.1 Accuracy of estimates for AECM algorithm 69
3.4.2 Asymptotic properties for the location estimates of VG distribution 71
3.5 Conclusion 79

Chapter 4. Weighted Leave-one-out Likelihood for data multiplicity ... 81
4.1 Introduction ... 81
4.2 Leave-multiple-out and weighted LOO likelihoods ... 82
4.3 Examples ... 84
4.3.1 Example 1: data multiplicity at a single location ... 84
4.3.2 Example 2: data multiplicities at two locations ... 87
4.3.3 Example 3: general data multiplicities at two locations ... 89
4.4 Simulation study ... 89
4.4.1 Results for data multiplicity due to repetition ... 91
4.4.2 Results for data multiplicity due to rounding ... 93
4.5 Conclusion ... 96

Chapter 5. Applications to Financial Time Series ... 99
5.1 Introduction ... 99
5.2 Estimation of VAR-VG model ... 101
5.2.1 VAR-VG model ... 102
5.2.2 Likelihood functions for VAR-VG model ... 103
5.2.3 E-step ... 103
5.2.4 CM-step for $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\gamma}$ ... 104
5.2.5 CM-step for ν ... 105
5.2.6 Summary of ECME algorithm ... 105
5.3 Estimation of VARMA-VG model ... 106
5.3.1 VARMA-VG model ... 106
5.3.2 Properties ... 107
5.3.3 ECM algorithm for VARMA-VG model ... 110
5.3.4 Forecasting using VARMA-VG model ... 112
5.4 Estimation of VARMA-t model ... 114
5.4.1 VARMA-t model ... 114
5.4.2 E-step ... 115
5.4.3 CM-step for υ ... 116
5.5 Simulation study ... 116
5.5.1 Identifiable VARMA-VG model ... 116
5.5.2 Non-identifiable VARMA-VG model ... 123
5.6 Applications ... 127

5.6.1 Application to cryptocurrency . . . 127
5.6.2 Stock market indices . . . 136
5.6.3 High frequency stock indices . . . 141
5.7 Conclusion . . . 151

Chapter 6. Conclusion . . . 153

Appendix A. Matrix Differentiation . . . 157
A1 Introduction . . . 157
A2 Matrix operators . . . 157
A3 Definitions and basic rules . . . 159
A4 Derivative of products . . . 161
A5 Derivative of trace . . . 164
A6 Derivative of vectorisation . . . 165
A7 Derivative with respect to structured matrix . . . 167
A7.1 Derivatives with respect to symmetric matrix . . . 167
A7.2 Second order derivatives with respect to symmetric matrix . . . 168
A8 Derivatives of complete data log-likelihood for VG distribution . . . 169
A8.1 First order derivatives . . . 170
A8.2 Second order derivatives . . . 170
A8.3 Cross derivatives . . . 171
A9 Derivative of complete data log-likelihood for VARMA-VG model . . . 171
A9.1 First order derivatives . . . 172
A9.2 Second order derivatives . . . 174
A9.3 Cross derivatives . . . 175

Appendix B. Fisher Information Matrix of VG Distribution . . . 177
B1 Preliminary results . . . 177
B2 Matrix representation of first order derivatives . . . 180
B3 Simplification of missing information matrix calculation . . . 182
B4 Conditional normal moment results . . . 183
B5 Multidimensional integration . . . 185
B5.1 Higher order spherical moments . . . 186
B6 Fisher information matrix . . . 191
B6.1 First term of Fisher information matrix . . . 192
B6.2 Second term of Fisher information matrix . . . 195

B6.3 Third term of Fisher information matrix 201

Appendix C. Other Related Functions and Distributions 206

C1 Modified Bessel function of the second kind 206

C2 Student's t distribution 208

C3 Generalised Gumbel distribution 209

C4 Double generalised gamma distribution 210

Bibliography 212

List of Figures

1.1 Contour and 3D plots of bivariate skewed VG distribution 30

2.1 Log of the optimal capping level 49
2.2 Time series plots for the five daily return series 53

3.1 Comparison of full and LOO log-likelihood 60
3.2 Vioplots of parameter estimates of unbounded bivariate VG using LOO likelihood 70
3.3 Contour and 3D plot of the LOO log-likelihood 70
3.4 Relative error of the estimated optimal rate of convergence index 73
3.5 Density plots of standardised $\hat{\mu}_n$ using LOO likelihood 73
3.6 Density plots of $\log|\hat{\mu}_n|$ using LOO likelihood 74
3.7 Estimates of generalised Gumbel fitted to $\log|\hat{\mu}_n|$ 76
3.8 P-P plots of GG approximation vs. $\log|\hat{\mu}_n|$ samples for $0.02 \leq \nu \leq 0.5$ 77
3.9 P-P plots for $0.52 \leq \nu \leq 1$. 78

4.1 Plots of full, LOO, LMO and WLOO log-likelihood for data set {-1,0,1,0} . 85
4.2 Plot of full, LOO, LMO and WLOO log-likelihood for data set {-1,0,1,0,0,1} 88
4.3 Accuracy for `ghyp`, full, adaptive capping level, LOO and WLOO likelihood method for repeated data points 94
4.4 Accuracy for rounded data points 95

5.1 Vioplots for case 1 with $\boldsymbol{\gamma} = (0.2, 0.3)$ and $\nu = 3$ 119
5.2 Vioplots for case 2 with $\boldsymbol{\gamma} = (0.2, 0.3)$ and $\nu = 0.7$ 120
5.3 Vioplots for case 3 with $\boldsymbol{\gamma} = (1, 2)$ and $\nu = 3$ 121
5.4 Vioplots for case 4 with $\boldsymbol{\gamma} = (1, 2)$ and $\nu = 0.7$ 122
5.5 Vioplots for non-identifiable case with $\boldsymbol{\gamma} = (0.2, 0.3)$ and $\nu = 3$ 125
5.6 Density plots of the square root of the resultant 126

5.7 Time series plots for the returns of cryptocurrencies. 130
5.8 ACF plots of returns of cryptocurrency 131
5.9 Density plots of errors of VARMA(2,0)-VG fitted to cryptocurrency data . . 137
5.10 Density plots of errors of VARMA(2,0)-t fitted to cryptocurrency data 138
5.11 Plot of ACF for DAX, S&P 500, FTSE 100, AORD and CAC 40 daily returns. 139
5.12 Density plots of errors of VAR(1)-VG fitted to daily returns of DAX, S&P 500, FTSE 100, AORD and CAC 40. 142
5.13 Fitted contour plot of VAR(1)-VG for DAX and FTSE 100 data sets after filtering the mean function. 143
5.14 Forecasting daily returns 144
5.15 ACFs of high frequency returns 146
5.16 Density plots of residuals of VARMA-VG model fitted to high frequency returns 150

C.1 Density plot of generalised Gumbel distribution 209

List of Tables

2.1 Median of SAE, computation time, and number of iterations for each ECM algorithm when applied to simulated VG samples with $d = 2$ and $\nu = 0.5$... 47

2.2 Median of SAE, computation time, and number of iterations for each ECM algorithm when applied to simulated VG samples with $d = 2$ and $\nu = 0.04$.. 47

2.3 Median SE estimates based on various SE methods for comparison..... 51

2.4 Summary statistics for DAX, S&P 500, FTSE 100, AORD and CAC 40 daily returns. 52

2.5 Parameter estimates and its SEs of the VG model using DAX, S&P 500, FTSE 100, AORD and CAC 40 daily returns. 54

4.1 Summary of `ghyp`, full, adaptive Δ, LOO and WLOO likelihood methods. . 90

4.2 Median of accuracy measures of parameter estimates across five likelihood methods with no data multiplicity. 92

5.1 SEs based on simulated estimates and calculation using Louis' method for cases 1 and 3..... 118

5.2 SEs based on simulated estimates and calculation using Louis' method and the double generalised gamma approximation for cases 2 and 4. 123

5.3 Numerical summaries of the daily returns of cryptocurrencies along with p-values of Box-Pierce test for serial correlation..... 129

5.4 Correlation matrix of the daily returns of cryptocurrencies..... 129

5.5 P-values for the Box-Pierce test of serial conditional heteroscedasticity and serial conditional correlation of cryptocurrencies. 129

5.6 AICc of VARMA-VG model for different p's and q's. 134

5.7 AICc of VARMA-t model for different p's and q's..... 134

5.8 Computational time and number of iterations until convergence using different ECM algorithms for VARMA(2,0)-VG and VARMA(2,0)-t models. 134

5.9 Parameter estimates and SEs for the VARMA(2,0)-VG model using the first method in Table 5.8. 135

5.10 Parameter estimates and SEs for the VARMA(2,0)-t model using the first method in Table 5.8. 135

5.11 Estimates, SEs and correlation matrix $\boldsymbol{\rho}$ for the VAR(1)-VG model using DAX, S&P 500, FTSE 100, AORD and CAC 40 daily return series. 140

5.12 Numerical summaries of 1 hour, 15 min, 5 min and 1 min sampling frequencies of ASX 200, CAC 40, FTSE 100 and S&P 500 returns. 145

5.13 Robust numerical summaries and p-values of the Box-Pierce test for different sampling frequencies and indices. 145

5.14 AICc of VARMA-VG model for different p and q's and different sampling frequencies. 147

5.15 Parameter estimates and correlation matrix $\boldsymbol{\rho}$ of VARMA(p, q)-VG model for 1hr and 15min high frequency returns. 148

5.16 Parameter estimates and correlation matrix $\boldsymbol{\rho}$ of VARMA(p, q)-VG model for 5 min and 1 min high frequency returns. 149

CHAPTER 1

Introduction

1.1 Background

The Variance gamma (VG) distribution (also called generalised Laplace distribution [60]) proposed by Madan and Seneta [71] is widely used to model financial time series data. This distribution is particularly useful to model the increment of log-prices (also called returns) which often display high concentration of data points around the centre and occasional outliers. However, to accommodate for the extreme kurtosis, the density function of the VG distribution can be cusp or even unbounded. As a result, the likelihood function may contain many unbounded points which poses great difficulties in the estimation procedure, especially since many popular estimation techniques relies on the smoothness of the likelihood function.

There is a rich literature in estimation methodologies for the VG distribution. These include Chebyshev polynomial expansion of characteristic function [70], method of moments [71, 102], product-density maximum likelihood estimator [34], minimum χ^2 estimator, Bayesian approach using `WinBUGS` [35] and expectation/maximisation (EM) algorithms [51, 75]. However, these methods encounter some significant issues when the density of the VG distribution becomes unbounded, so the literature typically avoid the cases of cusp and unbounded densities in their simulation studies and real applications.

The problem of unbounded density does not only exist in the VG distribution. Other examples includes the finite mixtures of normals [4, 100] and mixtures of two Weibull distributions [3] when one of the scale parameters approaches to zero. More examples includes the three-parameter lognormal [24, 36] and gamma distribution with threshold

parameter [21]. Cheng and Traylor [22] and Liu et al. [67] attempted to classify these models with unbounded likelihoods into different categories. The development of the maximum likelihood (ML) estimation methods for these examples are still limited and so further research is required in this area.

The ML estimation methodology has been extensively studied in literature and it possesses many desirable asymptotic properties under some regularity conditions. However, most of these properties rely on the assumption that the likelihood function is differentiable but this assumption might be violated. For the case when the density has a cusp at its mode with respect to the location parameter, Rao [91] and Ibragimov and Khasminskii [53, 54] showed that under some regularity conditions, the ML and Bayesian estimators of the location parameter are consistent, super-efficient, and have a limiting distribution with no simple expression. They also showed that this estimation problem is asymptotically equivalent to the estimation of the location of a non-stationary process.

For the case when the likelihood is unbounded, many of the desirable properties does not hold and even the maximum likelihood estimator (MLE) is not well-defined. Specifically, the likelihood becomes unbounded whenever the location parameter approaches to any data point and this problem is exacerbated when there are repeated data points. This unbounded likelihood is the source of many numerical errors and can hinder the performance of an estimator when such problem is not properly handled. Some examples of these numerical issues include failing to converge to the local maximum as many algorithms rely on the derivative of the likelihood function which becomes problematic if the likelihood is unbounded. Another numerical issue is the overflow (or underflow) when calculating the ratio of some extremely large (or small) values arised from the unbounded likelihood.

Apart from the ML approach, the Bayesian paradigm is getting popular in recent years as it has some advantages over the ML approach. Firstly, it replaces the problem of maximising a log-likelihood function for some complicated models by posterior sampling making use of some hierarchical structures. Secondly, it can incorporate external information in form of priors in the estimation. Lastly, it provides a posterior distribution for all parameters of interest. However, it also holds some drawbacks like the choice of priors and the expensive numerical computation. In applications, Bayesian models are implemented by performing posterior simulation using sampling techniques

such as Markov chain Monte Carlo (MCMC) and Gibbs sampler. When the likelihood function is unbounded with respect to the location parameter, its posterior distribution may also be unbounded and multimodal which can cause slow convergence [17, 42] or even non-convergence as well as other numerical instabilities issues especially if there is no simple sampling scheme for the posterior distribution.

As running Bayesian MCMC is known to be computationally expensive, some researchers have directed their efforts to solve the unbounded likelihood problem in the ML approach by modifying the likelihood function so that the maximum is well-defined. Giesbrecht and Kempthorne [36] and Cheng and Iles [21] proposed the rounded likelihood approach by discretising the continuous density function so that the densities become probabilities. Cheng and Amin [20] considered the maximum product of spacings method to replace the likelihood function by the product of spacings where the spacing is defined by the integral of the density function between two data points. Lastly, Seo and Kim [100] proposed the k-deleted likelihood method by removing the k largest terms in the likelihood. Although these three methods can deal with the unbounded likelihood, they are prone to the following minor drawbacks. The rounded likelihood and maximum product spacing methods require integration of the density function which can be computationally inefficient. Additionally, the rounded likelihood depends on some arbitrary chosen parameters while the maximum product of spacing and k-deleted likelihood methods may encounter problems when there are many repeated data points. Moreover, these likelihood methods only deal with the univariate case and some of these methods do not have a simple multivariate extension.

To address this unbounded likelihood problem for the multivariate case, we propose three different modifications to the classical likelihood function. The first modification is to bound the density function whenever a data point falls within some small neighbourhood around the location parameter. The second modification is to extend the leave-one-out likelihood (LOO) proposed by Podgórski and Wallin [89] to the multivariate case where it leaves out a data point that causes the singularity in the likelihood function. The third modification adds weights to the LOO likelihood to deal with repeated data points. To demonstrate the implementation of our proposed likelihood modifications, we present the EM algorithm [27] and its various extensions [65, 78] to estimate parameters of the multivariate skewed VG distribution.

As previously mentioned, the VG distribution is relevant in modelling the high kurtosis in financial time series. In addition, there are two more important reasons for studying the VG distribution. Firstly, it is an important limiting case that corresponds to the unbounded density case of the generalised hyperbolic (GH) distribution [6]. When the GH distribution approaches the VG distribution, one of its shape parameters approaches to the boundary of the parameter space potentially causing the density to become unbounded. Hence, the regular EM algorithm proposed by Protassov [90] for the GH distribution does not truly capture the unbounded density. Secondly, the VG distribution has a normal mean-variance mixture representation [7] that facilitates the implementation of the expectation-conditional maximisation (ECM) algorithm and its extensions.

Different extensions to the ECM algorithm have been proposed to improve the computational efficiency. One such extension is called the alternating ECM (AECM) algorithm proposed by Meng and van Dyk [79] where the data is allowed to vary within each iteration to improve the convergence rate. Moreover, Liu [63] applied the algorithm to multivariate symmetric Student's t distribution. We extend the application of the AECM algorithm to multivariate asymmetric distributions using the VG distribution as an example.

Apart from deriving estimation methods to obtain parameter estimates, it is also important to assess the significance of these parameter estimates by calculating their standard errors (SEs). This requires calculating the observed information matrix which is obtained from the second order derivative of the observed log-likelihood function. Many authors such as He [46] and Tsay [104] have provided formulas for calculating the second order derivatives for the multivariate Student's t and multivariate time series models. However, none of them have verified these derivative formulas using simulations. Furthermore, they incorrectly differentiated the log-likelihood with respect to the scale matrix $\boldsymbol{\Sigma}$ by not incorporating its symmetric structure into the calculation.

Details of the corrected derivative formula with respect to a symmetric matrix is provided in the appendices. Additionally, we also provide the matrix representation of the derivative formulas to enable efficient implementation in programming, and verify these SE calculations using numerical simulation. All details of these calculations are provided in the appendices and are applicable to both the VG and Student's t distribution.

Apart from the observed information matrix, the Fisher information matrix is also considered as it can provide more stable SE estimates. However, its derivation is extremely tedious for the VG distribution. Kawai [57] derived an asymptotic formula for the Fisher information matrix for the univariate VG distribution as the shape parameter tends towards zero. To the best of our knowledge, there were no formulas derived to numerically compute the Fisher information matrix for the multivariate skewed VG distribution as it requires multidimensional integration which is often numerically infeasible. We take this challenge to derive the formula by taking expectation of Louis' formula [68] and utilising the normal mean-variance mixture structure of the density function for the VG distribution to reduce the multidimensional integral down to one-dimensional integral which is much easier to compute. Additionally, we verify the formulas to numerically calculate the Fisher information matrix of the VG distribution using numerical simulations. Our method to derive these formulas for the observed and Fisher information matrices can also be applied to other distributions with normal mean-variance mixture representation such as the GH distribution. These formulas will sure provide a significant contribution to the literature on normal mean-variance mixture distributions.

As previously mentioned, the second modification to the likelihood function adopts the LOO likelihood from Podgórski and Wallin [89] where they proved the consistency and super-efficiency of the location estimator that maximises the LOO likelihood when the density is unbounded at data points. More precisely, under mild regularity conditions, they found a lower bound for the rate of convergence for the estimator of the location parameter. We extend the AECM algorithm to incorporate the LOO likelihood and perform numerical simulations to investigate the asymptotic properties of the parameter estimates that maximise the LOO likelihood. Currently, there is no literature which provide theoretical results regarding the optimal rate of convergence and asymptotic distribution for the location parameter estimates when the density is unbounded or even cusp at the mode. We believe that this pioneer work will provide insight for further theoretical development.

For the case when there are repeated data points, the LOO likelihood becomes unbounded at these points since leaving out a single data point is not enough to remove the unbounded likelihood. This problem can be circumvented by applying suitable weights to the LOO likelihood so that it leaves out multiple data points if they all

contribute to the unbounded likelihood. Our weighted LOO (WLOO) likelihood not only smooths out the likelihood caused by these unbounded points with data multiplicity but also preserves the overall structure of the likelihood in comparison to the original unbounded likelihood. We perform some simulation studies to compare the performance of different likelihood methods with data multiplicity.

Solving all the previously mentioned technical problems allows for real applications of the VG distribution to high frequency financial time series that often exhibits large kurtosis with some skewness which is difficult to model using the multivariate normal distributions. To also capture the persistence of these time series, we propose the vector autoregressive moving average (VARMA) model with multivariate skewed VG innovations. This flexible distribution can capture some important features such as serial correlation, cross-correlation, heavy-tailedness, positive skewness and high kurtosis. Heracleous [47] and Wang and Tsay [105] have studied multivariate time series models with symmetric Student's t innovation. However, not much research have been directed to the VARMA model with skewed innovations. We derive an AECM algorithm to efficiently estimate parameters for VARMA models with VG and Student's t innovations using the WLOO likelihood and provide formulas to calculate SEs using Louis' method. We also demonstrate applications by analysing returns from high frequency market indices and cryptocurrency market prices including Bitcoin as they both exhibit large kurtosis in the error distribution while comparing the model performance of VARMA models between VG and Student's t innovations.

This chapter is devoted to provide some background information for the topics in this thesis. We begin by providing some basic theories on the ML estimation in Section 1.2. Under the ML approach, Section 1.3 gives an overview of the EM algorithm which is the methodological focus for this thesis. Section 1.4 describes the various extensions of the EM algorithm to improve convergence rate and accuracy. Section 1.5 introduces distributions with normal mean-variance mixture representation which includes the GH, Student's t and VG distributions which can facilitate the implementation of the the EM algorithm. Lastly, Section 1.6 states the contributions and structure of this thesis.

1.2 Maximum likelihood estimation

Maximum likelihood estimation is an estimation method which involves finding the parameter values that maximises the likelihood function given the data. Under the Bayesian context, this is equivalent to finding the maximum of the posterior distribution based on non-informative priors.

In this section, we present some basic theories of the ML estimation for the case where there are no cusp nor unbounded points in the likelihood function so that all the regularity conditions are satisfied.

1.2.1 Likelihood function

Suppose there is no missing data so that $\boldsymbol{y} = (\boldsymbol{y}_1, \dots, \boldsymbol{y}_n)$ represents the complete data, and let $f(\boldsymbol{y}; \boldsymbol{\theta})$ be the joint density function for some parameter vector $\boldsymbol{\theta}$ in the parameter space $\boldsymbol{\Theta}$. Assuming that $\boldsymbol{y}_1, \dots, \boldsymbol{y}_n$ are independent, the likelihood function is given by

$$L(\boldsymbol{\theta}; \boldsymbol{y}) := f(\boldsymbol{y}; \boldsymbol{\theta}) = \prod_{i=1}^{n} f(\boldsymbol{y}_i; \boldsymbol{\theta}).$$

Equivalently, we can also consider the log-likelihood function defined by

$$\ell(\boldsymbol{\theta}; \boldsymbol{y}) := \log L(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{n} \log f(\boldsymbol{y}_i; \boldsymbol{\theta}). \tag{1.1}$$

The likelihood function gives a criterion for parameter estimation whereby $L(\boldsymbol{\theta}_1; \boldsymbol{y}) > L(\boldsymbol{\theta}_2; \boldsymbol{y})$ indicates that the data $\boldsymbol{y}$ is more likely to follow the model with parameter $\boldsymbol{\theta}_1$ than $\boldsymbol{\theta}_2$, so the parameter $\boldsymbol{\theta}_1$ is preferred over $\boldsymbol{\theta}_2$. From this interpretation, it makes sense to choose the parameter that "best" represents the data. We refer to this parameter which maximises $\ell(\boldsymbol{\theta}; \boldsymbol{y})$ over the whole parameter space $\boldsymbol{\Theta}$ as the *maximum likelihood estimator* (MLE) and it is defined as

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmax}}\, \ell(\boldsymbol{\theta}; \boldsymbol{y}).$$

Under certain regularity conditions and assuming the likelihood function is differentiable, the MLE can be obtained by solving the likelihood equation

$$\boldsymbol{S}(\boldsymbol{\theta}; \boldsymbol{y}) = \mathbf{0} \tag{1.2}$$

where the score function is defined by

$$\boldsymbol{S}(\boldsymbol{\theta}; \boldsymbol{y}) = \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \boldsymbol{y}) \tag{1.3}$$

which is the first order (vector) derivative of the log-likelihood function.

A nice property of the score function is that

$$\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{S}(\boldsymbol{\theta}; \boldsymbol{Y})] = \mathbf{0}$$

where the expectation is taken with respect to random variables $\boldsymbol{Y} = (\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n)$ which is distributed based on a certain statistical model with parameters $\boldsymbol{\theta}$.

1.2.2 Information matrix

Assuming the likelihood function is twice differentiable, the (observed) information matrix is defined by

$$\boldsymbol{I}(\boldsymbol{\theta}; \boldsymbol{y}) = -\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \ell(\boldsymbol{\theta}; \boldsymbol{y}) \tag{1.4}$$

which is the negative of the second order derivative of the log-likelihood function, and the Fisher information matrix is defined by

$$\mathcal{I}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{S}(\boldsymbol{\theta}; \boldsymbol{Y}) \boldsymbol{S}(\boldsymbol{\theta}; \boldsymbol{Y})']. \tag{1.5}$$

If $\ell(\boldsymbol{\theta}; \boldsymbol{y})$ is twice differentiable with respect to $\boldsymbol{\theta}$, and satisfies certain regularity conditions, then the Fisher information matrix can be written as

$$\mathcal{I}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{I}(\boldsymbol{\theta}; \boldsymbol{Y})].$$

The information can be thought of as the amount of curvature around the MLE. So a large amount of information gives a sharp peak around the maximum, whereas less information indicates that the peak is more flat.

The asymptotic covariance matrix of the MLE $\hat{\boldsymbol{\theta}}$ can be approximated by inverting the Fisher information matrix evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. Hence, the SE of parameter estimates

can be approximated by

$$\mathrm{SE}(\hat{\theta}_i) \approx \sqrt{\left[\mathcal{I}(\hat{\boldsymbol{\theta}})^{-1}\right]_{ii}}$$

where $\hat{\theta}_i = [\hat{\boldsymbol{\theta}}]_i$, and $[\boldsymbol{A}]_{ij}$ represents the $(i,j)^{\text{th}}$ entry of a matrix $\boldsymbol{A}$.

Typically, calculating the Fisher information matrix is more tedious. So instead, we may use the observed information matrix

$$\mathrm{SE}(\hat{\theta}_i) \approx \sqrt{\left[\boldsymbol{I}(\hat{\boldsymbol{\theta}}; \boldsymbol{y})^{-1}\right]_{ii}}$$

based on the data, avoiding the evaluation of expectation analytically.

1.2.3 Newton-Raphson method

Under certain regularity conditions, the MLE is unique and may even have closed-form solution. However, for most cases, the MLE is not unique and can only be defined locally. Moreover, it may not have a closed-form solution.

In the case where there are no closed-form solution, the Newton-Raphson (NR) method can be used to numerically solve for the likelihood equation in (1.2) by iteratively computing

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \boldsymbol{I}\left(\boldsymbol{\theta}^{(t)}; \boldsymbol{y}\right)^{-1} \boldsymbol{S}\left(\boldsymbol{\theta}^{(t)}; \boldsymbol{y}\right)$$

at iteration t where the iteration is initialised by some suitable starting value $\boldsymbol{\theta}^{(0)}$.

If the likelihood function is concave and unimodal, the iterative sequence $\left\{\boldsymbol{\theta}^{(t)}\right\}$ converges to $\hat{\boldsymbol{\theta}}_{\text{MLE}}$. On the other hand, if the likelihood function is not concave, then the iterative sequence is not guaranteed to converge for arbitrary starting values. Thus certain assumptions needs to be checked to ensure the validity of the estimates using the NR method.

The main advantage of the NR method is its quadratic rate of convergence which is relatively fast for a general optimisation problem. However, there are several major drawbacks. Firstly, the derivatives in (1.3) and (1.4) for the computation of $\boldsymbol{S}(\boldsymbol{\theta}; \boldsymbol{y})$ and $\boldsymbol{I}(\boldsymbol{\theta}; \boldsymbol{y})$ respectively may not be obtained analytically, and so the derivatives need to be approximated numerically. See [19] for an example. Secondly, the inverse of $\boldsymbol{I}(\boldsymbol{\theta}; \boldsymbol{y})$ needs to be computed at each iteration which can be computationally demanding for

large parameter vector. Thirdly, the method heavily relies on a good starting value as it has the tendency to converge towards a saddle point or a local maximum.

Extensions to the NR method have been proposed to mitigate some of these drawbacks. Böhning and Lindsay [10] demonstrated how the NR algorithm can be monotonic with some modification. Shanno [101] proposed a quasi-Newton method where the Hessian matrix is approximated using updates specified by gradient evaluations. Labelle [61] extended the method to have cubic rate of convergence. See Deuflhard [28] and Nocedal and Wright [84] for more information on NR methods.

1.2.4 Properties of MLE

The MLE possesses many desirable properties which is presented in this section. See Robert V. Hogg [94] and Newey and McFadden [83] for more information.

Theorem 1.2.1 (Functional invariance)**.** *Suppose $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$, and let $\boldsymbol{g}(\cdot)$ be a vector function (not necessarily one-to-one) from $\mathbb{R}^d$ to a subset of $\mathbb{R}^k$. Then $\boldsymbol{g}(\hat{\boldsymbol{\theta}})$ is the MLE of $\boldsymbol{g}(\boldsymbol{\theta})$.*

In other words, the MLE does not depend on the parametrisation of $\boldsymbol{\theta}$.

Theorem 1.2.2 (Consistency)**.** *Under some regularity conditions, the MLE is consistent. That is,*

$$\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}$$

where $\xrightarrow{p}$ represents convergence in probability.

The regularity conditions in Theorem 1.2.2 refer to the identification, compactness, continuity and dominance conditions [83, Theorem 2.5] which are also sufficient conditions to establish consistency. The interpretation is that as the sample size gets larger, there is a larger certainty that the MLE will get closer towards the true parameter. Additionally, we can obtain information about the variability of the estimator for large sample of size n using the following theorem.

Theorem 1.2.3 (Asymptotic Normality)**.** *Under some regularity conditions, the MLE is asymptotically normally distributed. That is,*

$$\sqrt{n}\big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\big) \xrightarrow{d} \mathcal{N}\big(\mathbf{0}, \mathcal{I}(\boldsymbol{\theta})^{-1}\big)$$

where $\xrightarrow{d}$ *represents convergence in distribution and* $\mathcal{I}(\boldsymbol{\theta})$ *is the Fisher information matrix defined in (1.5).*

The regularity conditions [83, Theorem 3.3] in Theorem 1.2.3 essentially requires the log-likelihood to be smooth enough so that the Fisher information matrix is well-defined. In particular, $f(\boldsymbol{y};\boldsymbol{\theta})$ needs to be at least twice differentiable with respect to $\boldsymbol{\theta}$. The theorem essentially states that the MLE asymptotically follows a normal distribution and the variance decays at the rate of $1/n$.

1.3 EM algorithm

The expectation/maximisation (EM) algorithm formalised by Dempster et al. [27] is a general iterative algorithm for calculating the ML estimates of a statistical model involving missing data. In this section, we give a brief summary of the EM algorithm while also stating its convergence properties, formulas for calculating the observed information matrix, and some of its extensions. For further information, see McLachlan and Krishnan [74].

1.3.1 Introduction

For the case when there is missing data, let $\boldsymbol{y}_{\text{com}} = (\boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}})$ be the complete data where $\boldsymbol{y}_{\text{obs}}$ and $\boldsymbol{y}_{\text{mis}}$ represent the observed and missing data respectively. We assume that the missing mechanism is missing at random [96] so that

$$\begin{aligned} f(\boldsymbol{y}_{\text{com}};\boldsymbol{\theta}) &= f(\boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}};\boldsymbol{\theta}) \\ &= f(\boldsymbol{y}_{\text{obs}};\boldsymbol{\theta})\, f(\boldsymbol{y}_{\text{mis}}|\boldsymbol{y}_{\text{obs}};\boldsymbol{\theta}). \end{aligned}$$

Taking logarithm and rearranging gives us the observe data log-likelihood

$$\ell_{\text{obs}}(\boldsymbol{\theta};\boldsymbol{y}_{\text{obs}}) = \ell_{\text{com}}(\boldsymbol{\theta};\boldsymbol{y}_{\text{com}}) - \ell_{\text{mis|obs}}(\boldsymbol{\theta};\boldsymbol{y}_{\text{mis}}|\boldsymbol{y}_{\text{obs}})$$

where $\ell_{\text{com}}(\boldsymbol{\theta};\boldsymbol{y}_{\text{com}})$ represents the complete data log-likelihood, and $\ell_{\text{mis|obs}}(\boldsymbol{\theta};\boldsymbol{y}_{\text{mis}}|\boldsymbol{y}_{\text{obs}})$ represents the conditional log-likelihood of the missing data given the observed data.

The usefulness of the EM algorithm comes in when maximising $\ell_{\text{obs}}(\boldsymbol{\theta}; \boldsymbol{y}_{\text{obs}})$ is challenging since it involves integrating out the missing data whereas maximising $\ell_{\text{com}}(\boldsymbol{\theta}; \boldsymbol{y}_{\text{com}})$ is much simpler. The general idea of the EM algorithm is to iteratively compute the MLE by the following procedure:

Step 1: Replace the missing data in the complete data likelihood by their conditional expectations.

Step 2: Estimate the parameters by maximising this conditional expectation of the complete data likelihood.

Step 3: Repeat steps 1 and 2 until parameter estimates converge.

More formally, suppose that $\boldsymbol{\theta}^{(t)}$ is the current parameter estimate, then the EM algorithm is composed of the expectation step (E-step) and maximisation step (M-step) which is described as follows:

E-step: Calculate the expected conditional log-likelihood defined as

$$\begin{aligned} Q\big(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}\big) &= \mathbb{E}_{\boldsymbol{\theta}^{(t)}}[\ell_{\text{com}}(\boldsymbol{\theta}; \boldsymbol{y}_{\text{com}}) | \boldsymbol{y}_{\text{obs}}] \\ &= \int \ell_{\text{com}}(\boldsymbol{\theta}; \boldsymbol{y}_{\text{com}}) f\big(\boldsymbol{y}_{\text{mis}} | \boldsymbol{y}_{\text{obs}}; \boldsymbol{\theta}^{(t)}\big)\, d\boldsymbol{y}_{\text{mis}} \end{aligned} \tag{1.6}$$

where the expectation is computed with respect to conditional distribution $\boldsymbol{y}_{\text{mis}} | \boldsymbol{y}_{\text{obs}}$ given our current estimate $\boldsymbol{\theta}^{(t)}$. This function is also referred to as the *Q-function*.

M-step: Update the parameter estimate to $\boldsymbol{\theta}^{(t+1)}$ by choosing the parameter that maximises $Q\big(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}\big)$. That is,

$$\boldsymbol{\theta}^{(t+1)} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmax}}\, Q\big(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}\big). \tag{1.7}$$

Convergence criterion: These two steps are repeated until the difference of successive log-likelihood values becomes sufficiently small. That is,

$$\ell_{\text{obs}}\big(\boldsymbol{\theta}^{(t+1)}; \boldsymbol{y}_{\text{obs}}\big) - \ell_{\text{obs}}\big(\boldsymbol{\theta}^{(t)}; \boldsymbol{y}_{\text{obs}}\big) \leq \delta \tag{1.8}$$

where we choose $\delta = \max\big\{10^{-7}, 10^{-8}\big|\ell\big(\boldsymbol{\theta}^{(t)}; \boldsymbol{y}_{\text{obs}}\big)\big|\big\}$ for the rest of this thesis. Note that other convergence criterion can be used for the EM algorithm.

Algorithm 1: EM algorithm

Input: Initial value $\boldsymbol{\theta}^{(0)}$
while $\ell_{\text{obs}}(\boldsymbol{\theta}^{(t+1)}; \boldsymbol{y}_{\text{obs}}) - \ell_{\text{obs}}(\boldsymbol{\theta}^{(t)}; \boldsymbol{y}_{\text{obs}}) > \delta$ **do**
  $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}) \leftarrow \mathbb{E}_{\boldsymbol{\theta}^{(t)}}[\ell_{\text{com}}(\boldsymbol{\theta}; \boldsymbol{y}_{\text{com}})|\boldsymbol{y}_{\text{obs}}]$;
  $\boldsymbol{\theta}^{(t+1)} \leftarrow \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\text{argmax}}\, Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)})$;
end

1.3.2 Convergence of EM algorithm

The EM algorithm provides a convenient way to maximise $\ell_{\text{obs}}(\boldsymbol{\theta}; \boldsymbol{y}_{\text{obs}})$ by instead maximising the conditional expectation of $\ell_{\text{com}}(\boldsymbol{\theta}; \boldsymbol{y}_{\text{com}})$. In this section, we show that under some regularity conditions, this algorithm produces iterative values $\boldsymbol{\theta}^{(t)}$ such that it indeed converges to the parameter that maximises $\ell_{\text{obs}}(\boldsymbol{\theta}; \boldsymbol{y}_{\text{obs}})$. To show this, we need the following two fundamental results:

Lemma 1.3.1.

$$\ell_{\text{obs}}(\boldsymbol{\theta}; \boldsymbol{y}_{\text{obs}}) = Q\big(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}\big) - H\big(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}\big)$$

where $Q\big(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}\big)$ *is defined earlier in* (1.6)*, and*

$$H\big(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}\big) = \int \log f(\boldsymbol{y}_{\text{mis}}|\boldsymbol{y}_{\text{obs}}; \boldsymbol{\theta})\, f\big(\boldsymbol{y}_{\text{mis}}|\boldsymbol{y}_{\text{obs}}; \boldsymbol{\theta}^{(t)}\big)\, d\boldsymbol{y}_{\text{mis}}.$$

Proof. The idea is to decompose the complete data log-likelihood into two parts, then apply conditional expectation. This decomposition can be done by applying the Bayes' rule

$$f(\boldsymbol{y}_{\text{com}}; \boldsymbol{\theta}) = f(\boldsymbol{y}_{\text{obs}}; \boldsymbol{\theta}) f(\boldsymbol{y}_{\text{mis}}|\boldsymbol{y}_{\text{obs}}; \boldsymbol{\theta}).$$

Taking logarithm of both sides, we obtain

$$\ell_{\text{com}}(\boldsymbol{\theta}; \boldsymbol{y}_{\text{com}}) = \ell_{\text{obs}}(\boldsymbol{\theta}; \boldsymbol{y}_{\text{obs}}) + \log f(\boldsymbol{y}_{\text{mis}}|\boldsymbol{y}_{\text{obs}}; \boldsymbol{\theta}).$$

Applying conditional expectation over $\boldsymbol{y}_{\text{mis}}$ given $\boldsymbol{y}_{\text{obs}}$ at current estimate $\boldsymbol{\theta}^{(t)}$, and rearranging gives us

$$\begin{aligned}
\ell_{\text{obs}}(\boldsymbol{\theta}; \boldsymbol{y}_{\text{obs}}) &= \int \ell_{\text{com}}(\boldsymbol{\theta}; \boldsymbol{y}_{\text{com}})\, f\big(\boldsymbol{y}_{\text{mis}} \big| \boldsymbol{y}_{\text{obs}}; \boldsymbol{\theta}^{(t)}\big)\, d\boldsymbol{y}_{\text{mis}} \\
&\quad - \int \log\, f(\boldsymbol{y}_{\text{mis}} | \boldsymbol{y}_{\text{obs}}; \boldsymbol{\theta})\, f\big(\boldsymbol{y}_{\text{mis}} \big| \boldsymbol{y}_{\text{obs}}; \boldsymbol{\theta}^{(t)}\big)\, d\boldsymbol{y}_{\text{mis}} \\
&= \mathbb{E}_{\boldsymbol{\theta}^{(t)}}[\ell_{\text{com}}(\boldsymbol{\theta}; \boldsymbol{y}_{\text{com}}) | \boldsymbol{y}_{\text{obs}}] - \mathbb{E}_{\boldsymbol{\theta}^{(t)}}[\log f(\boldsymbol{y}_{\text{mis}} | \boldsymbol{y}_{\text{obs}}; \boldsymbol{\theta}) | \boldsymbol{y}_{\text{obs}}] \\
&= Q\big(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}\big) - H\big(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}\big)
\end{aligned}$$

which completes the proof. □

Lemma 1.3.2. *Given $\boldsymbol{\theta}^{(t)}$, then for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$*

$$H\big(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}\big) \leq H\big(\boldsymbol{\theta}^{(t)}; \boldsymbol{\theta}^{(t)}\big).$$

Proof. We want to show that

$$H\big(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}\big) - H\big(\boldsymbol{\theta}^{(t)}; \boldsymbol{\theta}^{(t)}\big) \leq 0.$$

Simplifying the left hand side gives us

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\theta}^{(t)}}[\log f(\boldsymbol{y}_{\text{mis}} | \boldsymbol{y}_{\text{obs}} \boldsymbol{\theta}) | \boldsymbol{y}_{\text{obs}}] - \mathbb{E}_{\boldsymbol{\theta}^{(t)}}\big[\log f\big(\boldsymbol{y}_{\text{mis}} | \boldsymbol{y}_{\text{obs}} \boldsymbol{\theta}^{(t)}\big) | \boldsymbol{y}_{\text{obs}}\big] \\
&= \mathbb{E}_{\boldsymbol{\theta}^{(t)}}\left[\log\left(\frac{f(\boldsymbol{y}_{\text{mis}} | \boldsymbol{y}_{\text{obs}} \boldsymbol{\theta})}{f(\boldsymbol{y}_{\text{mis}} | \boldsymbol{y}_{\text{obs}} \boldsymbol{\theta}^{(t)})}\right) \middle| \boldsymbol{y}_{\text{obs}}\right].
\end{aligned}$$

By Jensen's inequality, we have that

$$\mathbb{E}_{\boldsymbol{\theta}^{(t)}}\left[\log\left(\frac{f(\boldsymbol{y}_{\text{mis}} | \boldsymbol{y}_{\text{obs}} \boldsymbol{\theta})}{f(\boldsymbol{y}_{\text{mis}} | \boldsymbol{y}_{\text{obs}} \boldsymbol{\theta}^{(t)})}\right) \middle| \boldsymbol{y}_{\text{obs}}\right] \leq \log \mathbb{E}_{\boldsymbol{\theta}^{(t)}}\left[\frac{f(\boldsymbol{y}_{\text{mis}} | \boldsymbol{y}_{\text{obs}}; \boldsymbol{\theta})}{f(\boldsymbol{y}_{\text{mis}} | \boldsymbol{y}_{\text{obs}}; \boldsymbol{\theta}^{(t)})} \middle| \boldsymbol{y}_{\text{obs}}\right].$$

Expressing the expectation as integrals

$$\begin{aligned}
& \mathbb{E}_{\boldsymbol{\theta}^{(t)}}[\log f(\boldsymbol{y}_{\text{mis}} | \boldsymbol{y}_{\text{obs}} \boldsymbol{\theta}) | \boldsymbol{y}_{\text{obs}}] - \mathbb{E}_{\boldsymbol{\theta}^{(t)}}\big[\log f\big(\boldsymbol{y}_{\text{mis}} | \boldsymbol{y}_{\text{obs}} \boldsymbol{\theta}^{(t)}\big) | \boldsymbol{y}_{\text{obs}}\big] \\
\leq\ & \log \int \frac{f(\boldsymbol{y}_{\text{mis}} | \boldsymbol{y}_{\text{obs}}; \boldsymbol{\theta})}{f(\boldsymbol{y}_{\text{mis}} | \boldsymbol{y}_{\text{obs}}; \boldsymbol{\theta}^{(t)})} f\big(\boldsymbol{y}_{\text{mis}} | \boldsymbol{y}_{\text{obs}}; \boldsymbol{\theta}^{(t)}\big)\, d\boldsymbol{y}_{\text{mis}} \\
=\ & \log \underbrace{\int f(\boldsymbol{y}_{\text{mis}} | \boldsymbol{y}_{\text{obs}}; \boldsymbol{\theta})\, d\boldsymbol{y}_{\text{mis}}}_{=1} \\
=\ & 0
\end{aligned}$$

which gives us the result. □

We now have the results to prove the monotonic convergence of the EM algorithm.

Theorem 1.3.3.

$$\ell_{\text{obs}}\big(\boldsymbol{\theta}^{(t+1)}; \boldsymbol{y}_{\text{obs}}\big) \geq \ell_{\text{obs}}\big(\boldsymbol{\theta}^{(t)}; \boldsymbol{y}_{\text{obs}}\big).$$

Proof. We want to show that

$$\ell_{\text{obs}}\big(\boldsymbol{\theta}^{(t+1)}; \boldsymbol{y}_{\text{obs}}\big) - \ell_{\text{obs}}\big(\boldsymbol{\theta}^{(t)}; \boldsymbol{y}_{\text{obs}}\big) \geq 0.$$

Applying Lemma 1.3.1 to both terms on the left hand side gives us

$$= \underbrace{Q\big(\boldsymbol{\theta}^{(t+1)}; \boldsymbol{\theta}^{(t)}\big) - Q\big(\boldsymbol{\theta}^{(t)}; \boldsymbol{\theta}^{(t)}\big)}_{\geq 0} - \underbrace{\big[H\big(\boldsymbol{\theta}^{(t+1)}; \boldsymbol{\theta}^{(t)}\big) - H\big(\boldsymbol{\theta}^{(t)}; \boldsymbol{\theta}^{(t)}\big)\big]}_{\leq 0}$$

where the first inequality is from the definition of $\boldsymbol{\theta}^{(t+1)}$, and the second inequality is from Lemma 1.3.2. Thus applying these inequalities completes the proof. □

This theorem states that the likelihood is non-decreasing after each iteration of the EM algorithm. Additionally, assuming that $L(\boldsymbol{\theta}^{(t)})$ is bounded from above, then this theorem implies that $L(\boldsymbol{\theta}^{(t)})$ converges monotonically to some fixed point $L(\boldsymbol{\theta}^*)$.

To prove that $\boldsymbol{\theta}^{(t)}$ indeed converges to $\boldsymbol{\theta}^*$ and that $\boldsymbol{\theta}^*$ are local maximas of L, the following regularity conditions are necessary:

(i) $\boldsymbol{\Theta}$ is a subset in $\mathbb{R}^d$,
(ii) $\{\boldsymbol{\theta} \in \boldsymbol{\Theta} : L(\boldsymbol{\theta}) \geq L(\boldsymbol{\theta}_0)\}$ is compact for any $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ such that $L(\boldsymbol{\theta}_0) > -\infty$,
(iii) L is continuous in $\boldsymbol{\Theta}$ and differentiable in the interior of $\boldsymbol{\Theta}$.

See Wu [109] for further details on the convergence properties of the EM algorithm.

Generalised EM algorithm:

In the M-step (1.7), the parameter estimate is chosen such that it globally maximises the Q-function which can be difficult for complicated Q-function. Instead, we can choose $\boldsymbol{\theta}^{(t+1)}$ such that it increases the Q-function. That is,

$$Q\big(\boldsymbol{\theta}^{(t+1)}; \boldsymbol{\theta}^{(t)}\big) \geq Q\big(\boldsymbol{\theta}^{(t)}; \boldsymbol{\theta}^{(t)}\big). \tag{1.9}$$

From Lemma 1.3.2, we see that this condition is sufficient to ensure the monotonic convergence of the EM algorithm. We refer to this algorithm as the GEM (i.e. *generalised*

EM) algorithm. This algorithm also shares similar convergence properties as the EM algorithm and was discussed by Wu [109].

Algorithm 2: GEM algorithm

Input: Initial value $\boldsymbol{\theta}^{(0)}$
while $\ell_{\text{obs}}(\boldsymbol{\theta}^{(t+1)}; \boldsymbol{y}_{\text{obs}}) - \ell_{\text{obs}}(\boldsymbol{\theta}^{(t)}; \boldsymbol{y}_{\text{obs}}) > \delta$ **do**
  $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}) \leftarrow \mathbb{E}_{\boldsymbol{\theta}^{(t)}}[\ell_{\text{com}}(\boldsymbol{\theta}; \boldsymbol{y}_{\text{com}})|\boldsymbol{y}_{\text{obs}}]$;
  $\boldsymbol{\theta}^{(t+1)} \leftarrow$ Any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ such that $Q\big(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}\big) \geq Q\big(\boldsymbol{\theta}^{(t)}; \boldsymbol{\theta}^{(t)}\big)$;
end

1.3.3 Score function with missing data

We already looked at the score function in (1.3) under the case where there is no missing data. Under the EM algorithm framework, we have the complete data score function

$$\boldsymbol{S}_{\text{com}}(\boldsymbol{\theta}; \boldsymbol{y}_{\text{com}}) = \frac{\partial}{\partial \boldsymbol{\theta}} \ell_{\text{com}}(\boldsymbol{\theta}; \boldsymbol{y}_{\text{com}})$$

and the observed data score function

$$\boldsymbol{S}_{\text{obs}}(\boldsymbol{\theta}; \boldsymbol{y}_{\text{obs}}) = \frac{\partial}{\partial \boldsymbol{\theta}} \ell_{\text{obs}}(\boldsymbol{\theta}; \boldsymbol{y}_{\text{obs}}).$$

The observed data score function can be expressed by the conditional expectation of the complete data score function given $\boldsymbol{y}$. That is,

$$\boldsymbol{S}_{\text{obs}}(\boldsymbol{\theta}; \boldsymbol{y}_{\text{obs}}) = \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{S}_{\text{com}}(\boldsymbol{\theta}; \boldsymbol{y}_{\text{com}})|\boldsymbol{y}_{\text{obs}}]$$

where the condition for interchanging the operations of differentiation and integration hold. A sufficient condition for the interchangeability is using the dominating convergence theorem.

1.3.4 Information matrix with missing data

The precision of the estimators can be estimated by calculating the observed information matrix using the estimates from the EM algorithm. However, this calculation involves the second order derivatives of the observed log-likelihood which can be extremely complicated when there are missing data. Instead, one can use the complete data log-likelihood to calculate the complete data information matrix as well as the missing data

information matrix. This is more preferable if the EM algorithm is already implemented in the first place.

Louis [68] derived a formula that allows the observed data information matrix to be expressed in terms of the complete data information matrix and missing data information matrix.

$$\boldsymbol{I}_{\text{obs}}(\boldsymbol{\theta}; \boldsymbol{y}_{\text{obs}}) = \mathcal{I}_{\text{com}}(\boldsymbol{\theta}; \boldsymbol{y}_{\text{obs}}) - \mathcal{I}_{\text{mis}}(\boldsymbol{\theta}; \boldsymbol{y}_{\text{obs}}) \tag{1.10}$$

where the conditional expectation of complete data information matrix is

$$\mathcal{I}_{\text{com}}(\boldsymbol{\theta}; \boldsymbol{y}_{\text{obs}}) = -\mathbb{E}_{\boldsymbol{\theta}}\left(\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \ell_{\text{com}}(\boldsymbol{\theta} | \boldsymbol{y}_{\text{com}}) \middle| \boldsymbol{y}_{\text{obs}} \right) \tag{1.11}$$

and the missing data information matrix is

$$\begin{aligned} \mathcal{I}_{\text{mis}}(\boldsymbol{\theta}; \boldsymbol{y}_{\text{obs}}) &= \text{cov}_{\boldsymbol{\theta}}\left(\frac{\partial}{\partial \boldsymbol{\theta}} \ell_{\text{com}}(\boldsymbol{\theta} | \boldsymbol{y}_{\text{com}}) \middle| \boldsymbol{y}_{\text{obs}} \right) \\ &= \mathbb{E}_{\boldsymbol{\theta}}\left(\frac{\partial}{\partial \boldsymbol{\theta}} \ell_{\text{com}}(\boldsymbol{\theta}; \boldsymbol{y}_{\text{com}}) \, \frac{\partial}{\partial \boldsymbol{\theta}'} \ell_{\text{com}}(\boldsymbol{\theta}; \boldsymbol{y}_{\text{com}}) \middle| \boldsymbol{y}_{\text{obs}} \right) \\ &\quad - \mathbb{E}_{\boldsymbol{\theta}}\left(\frac{\partial}{\partial \boldsymbol{\theta}} \ell_{\text{com}}(\boldsymbol{\theta}; \boldsymbol{y}_{\text{com}}) \middle| \boldsymbol{y}_{\text{obs}} \right) \mathbb{E}_{\boldsymbol{\theta}}\left(\frac{\partial}{\partial \boldsymbol{\theta}} \ell_{\text{com}}(\boldsymbol{\theta}; \boldsymbol{y}_{\text{com}}) \middle| \boldsymbol{y}_{\text{obs}} \right)' \end{aligned} \tag{1.12}$$

assuming the conditions for interchanging the operations of expectation and differentiation hold. See [74, equations 3.51, 4.1 and 4.3] for reference.

The equation (1.12) is referred to as the missing information principle [86] and intuitively can be thought of as

$$\texttt{Observed Information} = \texttt{Complete Information} - \texttt{Missing Information}.$$

Other methods to calculate SEs includes Bootstrap [29, 30], Baker's [5] and Oakes' method [85], as well as supplementary EM [77] and conditional normal approximation algorithm [64]. Also see [18] for an application.

1.3.5 Rate of convergence

It is clear that there is a loss of information due to missing data in the calculation of the information matrix. This also affect the convergence rate of EM algorithm. Given the t^{th} iteration of the EM algorithm, the iterative step $\boldsymbol{\theta}^{(t)} \to \boldsymbol{\theta}^{(t+1)}$ can be thought of

as a mapping

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{M}\big(\boldsymbol{\theta}^{(t)}\big), \quad t = 0, 1, 2, \ldots \tag{1.13}$$

for some vector function $\boldsymbol{M} : \boldsymbol{\Theta} \to \boldsymbol{\Theta}$. Let $\boldsymbol{\theta}^*$ be a fixed point such that $\boldsymbol{\theta}^* = \boldsymbol{M}(\boldsymbol{\theta}^*)$. Expanding (1.13) around $\boldsymbol{\theta}^*$ using the Taylor expansion gives us

$$\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^* \approx \boldsymbol{J}(\boldsymbol{\theta}^*)\big(\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^*\big) \tag{1.14}$$

where $\boldsymbol{J}(\boldsymbol{\theta})$ represents the Jacobian matrix of $\boldsymbol{M}(\boldsymbol{\theta})$. Then around the neighbourhood of $\boldsymbol{\theta}^*$, the EM algorithm is essentially a linear iteration with (matrix) rate of convergence $\boldsymbol{J}(\boldsymbol{\theta}^*)$.

The global rate of convergence also called the fractional missing index is given by

$$r := \lim_{t\to\infty} \frac{\big\|\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^*\big\|}{\|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^*\|} \tag{1.15}$$

where $\|\cdot\|$ represents a norm in the Euclidean space. Under certain regularity conditions,

$$r = \lambda_{\max} := \text{the largest eigenvalue of } \boldsymbol{J}(\boldsymbol{\theta}^*). \tag{1.16}$$

Note that larger values of r implies slower convergence.

Dempster et al. [27] showed that the Jacobian in (1.14) can be written as

$$\boldsymbol{J}(\boldsymbol{\theta}^*) = \mathcal{I}_{\text{com}}^{-1}(\boldsymbol{\theta}^*; \boldsymbol{y})\, \mathcal{I}_{\text{mis}}(\boldsymbol{\theta}^*; \boldsymbol{y}). \tag{1.17}$$

The result in (1.16) implies that the rate of convergence of the EM algorithm is given by the largest eigenvalue of the ratio of information matrices. This ratio can be thought of as the proportion of missing information over complete information. In other words, the higher the fraction of missing information, the slower the convergence rate.

The fraction of missing information may vary depending on $\boldsymbol{\theta}$ which suggests that the algorithm converges rapidly to $\boldsymbol{\theta}^*$ for some regions in $\boldsymbol{\Theta}$ and converges slowly for other regions.

1.4 Extensions to EM algorithm

For some problems, the M-step can be difficult to compute as it may involve complicated models with many parameters. A natural extension is to partition the M-step

into several conditional maximisation (CM) steps. This extension is referred to as the expectation/conditional maximisation (ECM) algorithm proposed by Meng and Rubin [78]. This algorithm simplifies the maximisation step for the NMVM model (in Section 1.5) by utilising some standard results of the normal distribution given the mixing variables. As a consequence, although it typically requires more iterations for each CM-step as compared with the EM algorithm, the computation within each iteration can be more efficient.

Meng [76] considered a variation of the ECM algorithm called multicycle ECM (MCECM) algorithm which inserts extra E-steps before each CM-step. Liu and Rubin [65] advanced the ECM algorithm to ECM either (ECME) algorithm by maximising the observed likelihood rather than the expected conditional likelihood to improve the speed of convergence by reducing the number of iterations. Liu and Rubin [66] applied the MCECM and ECME algorithms to obtain the ML estimates for multivariate Student's t distribution with incomplete data. They also found that the ECME algorithm converges much more efficiently than the EM and ECM algorithms in terms of computational time. Hu and Kercheval [52] used the MCECM algorithm with the Student's t distribution for portfolio credit risk measurement. These extensions, namely the ECM, MCECM and ECME algorithms are discussed in Sections 1.4.1 to 1.4.3, respectively.

1.4.1 ECM algorithm

As mentioned in the previous section, the EM algorithm maximises the conditional expectation of the complete data log-likelihood instead of the observed data log-likelihood which is often simpler to compute. For some models, this maximisation can still be computationally challenging. In spite of that, it can be simplified by partitioning the parameter vector and performing several conditional maximisation (CM) steps with over some smaller parameter space. These partitions are typically chosen so that some CM-steps have closed-form solution while the others require numerical optimisation methods. This ECM algorithm can improve the numerical efficiency and stability than the EM algorithm.

Suppose we partition the parameter vector into S subvectors $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_S)$, and the M-step is replaced by $S \geq 1$ CM-steps, and $\boldsymbol{\theta}^{(t+s/S)}$ represents the parameter estimate after the s^{th} CM-step during the t^{th} to $(t+1)^{\text{th}}$ iteration of the ECM algorithm. The

parameter $\boldsymbol{\theta}^{(t+s/S)}$ is estimated by maximising $Q\big(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}\big)$ for some sub-vector of $\boldsymbol{\theta}$ when other parameters are kept fixed. More formally, this can be written as

$$\boldsymbol{\theta}^{(t+s/S)} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}_s}{\operatorname{argmax}}\, Q\big(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}\big) \tag{1.18}$$

where the expected conditional log-likelihood $Q\big(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}\big)$ is defined in (1.6),

$$\boldsymbol{\Theta}_s := \big\{\boldsymbol{\theta}\in\boldsymbol{\Theta} : g_s(\boldsymbol{\theta}) = g_s\big(\boldsymbol{\theta}^{(t+(s-1)/S)}\big)\big\}$$

and $g_s(\cdot)$ represents the vector function that consists of all subvectors of $\boldsymbol{\theta}$ except $\boldsymbol{\theta}_s$. Specifically, $g_s(\cdot)$ represents the pre-selected vector functions of $\boldsymbol{\theta}$ (see [78]). For the S^{th} CM-step, $\boldsymbol{\theta}^{(t+S/S)} = \boldsymbol{\theta}^{(t+1)}$ is taken to be the final estimate for the $(t+1)^{\text{th}}$ iteration, and used for the next iteration. The following theorem shows that the ECM algorithm preserves the monotonic convergence property as described in Section 1.3.2.

Theorem 1.4.1. *The ECM algorithm is a GEM algorithm described in Algorithm 2.*

Proof. From the definition of $\boldsymbol{\theta}^{(t+s/S)}$ in equation (1.18), this can also be written as

$$Q\big(\boldsymbol{\theta}^{(t+s/S)};\boldsymbol{\theta}^{(t)}\big) \geq Q\big(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}\big)$$

for all $\boldsymbol{\theta}\in\boldsymbol{\Theta}_s$. Applying this for each CM-step during the t^{th} to $(t+1)^{\text{th}}$ iteration gives us,

$$\begin{aligned} Q\big(\boldsymbol{\theta}^{(t+1)};\boldsymbol{\theta}^{(t)}\big) &\geq Q\big(\boldsymbol{\theta}^{(t+(S-1)/S)};\boldsymbol{\theta}^{(t)}\big) \\ &\vdots \\ &\geq Q\big(\boldsymbol{\theta}^{(t)};\boldsymbol{\theta}^{(t)}\big). \end{aligned}$$

This implies the ECM is a GEM algorithm since it satisfies equation in (1.9). □

In other words, the ECM algorithm preserves the monotonic convergence properties from the GEM algorithm. Similar arguments for the GEM algorithm can be applied to each CM-step as well. Instead of globally maximising the Q-function from (1.18), we can instead choose any $\boldsymbol{\theta}\in\boldsymbol{\Theta}_s$ such that

$$Q\big(\boldsymbol{\theta}^{(t+s/S)};\boldsymbol{\theta}^{(t)}\big) \geq Q\big(\boldsymbol{\theta}^{(t+(s-1)/S)};\boldsymbol{\theta}^{(t)}\big) \tag{1.19}$$

which is computationally more feasible if the CM-step is complicated.

Algorithm 3: ECM algorithm

Input: Initial value $\boldsymbol{\theta}^{(0)}$
while $\ell_{\text{obs}}(\boldsymbol{\theta}^{(t+1)}; \boldsymbol{y}_{\text{obs}}) - \ell_{\text{obs}}(\boldsymbol{\theta}^{(t)}; \boldsymbol{y}_{\text{obs}}) > \delta$ **do**
$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}) \leftarrow \mathbb{E}_{\boldsymbol{\theta}^{(t)}}[\ell_{\text{com}}(\boldsymbol{\theta}; \boldsymbol{y}_{\text{com}})|\boldsymbol{y}_{\text{obs}}]$;
$\boldsymbol{\theta}^{(t+1/S)} \leftarrow \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}_1}{\operatorname{argmax}} Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)})$;
$\vdots$
$\boldsymbol{\theta}^{(t+1)} \leftarrow \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}_S}{\operatorname{argmax}} Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t+(S-1)/S)})$;
end

1.4.2 MCECM algorithm

For the case when the E-step is easy to compute, additional E-steps can be added before each CM-step to potentially speed up the convergence rate of the ECM algorithm. This procedure proposed by Meng and Rubin [78] is called the multicycle ECM (MCECM) algorithm. In general, the E-step can be added to selected CM-steps. For simplicity, we consider the case when the E-step is performed before each CM-step

During the s^{th} CM-step of the t^{th} to $(t+1)^{\text{th}}$ iteration of ECM algorithm, $\boldsymbol{\theta}^{(t+s/S)}$ is calculated by maximising $Q\big(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}\big)$. However, for the MCECM algorithm, $Q\big(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t+(s-1)/S)}\big)$ is maximised instead. Since the Q-function is changing after each CM-step, the MCECM algorithm may not be a GEM algorithm. Instead we have that

$$Q\big(\boldsymbol{\theta}^{(t+s/S)}; \boldsymbol{\theta}^{t+(s-1)/S}\big) \geq Q\big(\boldsymbol{\theta}^{t+(s-1)/S}; \boldsymbol{\theta}^{t+(s-1)/S}\big)$$

which is a sufficient condition to prove that

$$\ell\big(\boldsymbol{\theta}^{(t+s/S)}; \boldsymbol{y}_{\text{obs}}\big) \geq \ell\big(\boldsymbol{\theta}^{(t+(s-1)/S)}; \boldsymbol{y}_{\text{obs}}\big).$$

Thus the MCECM monotonically increases the log-likelihood after each iteration.

Additionally, the convergence result applies since the MCECM algorithm can be thought of as S different EM algorithms combined into one big algorithm. More generally, for the case when the E-step is added to selected CM-steps, then the MCECM is just a combination of R different ECM algorithms where R is the number of E-steps in the MCECM algorithm.

One iteration of MCECM requires more computation than one iteration of ECM due to the extra E-steps. Intuitively, one might expect the MCECM algorithm to converge

faster than ECM since the missing values (or Q-function) are constantly being updated. However, Meng and Rubin [78] remarked that in some cases when applied to real data, the MCECM algorithm may in fact converges slower than ECM algorithm. Despite this, the MCECM algorithm usually converges faster than ECM algorithm.

Algorithm 4: MCECM algorithm

Input: Initial value $\boldsymbol{\theta}^{(0)}$
while $\ell_{\text{obs}}(\boldsymbol{\theta}^{(t+1)}; \boldsymbol{y}_{\text{obs}}) - \ell_{\text{obs}}(\boldsymbol{\theta}^{(t)}; \boldsymbol{y}_{\text{obs}}) > \delta$ **do**
$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}) \leftarrow \mathbb{E}_{\boldsymbol{\theta}^{(t)}}[\ell_{\text{com}}(\boldsymbol{\theta}; \boldsymbol{y}_{\text{com}})|\boldsymbol{y}_{\text{obs}}]$;
$\boldsymbol{\theta}^{(t+1/S)} \leftarrow \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}_1}{\operatorname{argmax}} Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)})$;
$\vdots$
$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t+(S-1)/S)}) \leftarrow \mathbb{E}_{\boldsymbol{\theta}^{(t+(S-1)/S)}}[\ell_{\text{com}}(\boldsymbol{\theta}; \boldsymbol{y}_{\text{com}})|\boldsymbol{y}_{\text{obs}}]$;
$\boldsymbol{\theta}^{(t+1)} \leftarrow \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}_S}{\operatorname{argmax}} Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t+(S-1)/S)})$;
end

1.4.3 ECME algorithm

The ECM either (ECME) algorithm is an extension to the ECM algorithm proposed by Liu and Rubin [65] where the "either" refers to either maximising the Q-function or the observed log-likelihood $\ell_{\text{obs}}(\boldsymbol{\theta}; \boldsymbol{y}_{\text{obs}})$ for the CM-step.

Typically, the maximisation of the observed log-likelihood is more complicated. However, the reward is dramatically faster convergence rate. This is because calculating the observed log-likelihood does not require estimating the missing values, and that the speed of convergence is inversely proportional to the fractional missing index in (1.15). As a result, each iteration of the ECME algorithm is computationally slower than the ECM algorithm. However, the faster convergence rate dramatically reduces the overall computation time. In the example of Liu and Rubin [65], the computation time is reduced by a factor of seven.

The monotonic convergence for the ECME algorithm was proved by Liu and Rubin [65], but was later noted by Meng and van Dyk [79] that the monotonic convergence holds only if all the CM-steps applied to the Q-functions are performed before the CM-step applied to the observed log-likelihood (see [74] in §5.7). Liu and Rubin [65] studied the ECME algorithm and found that it has faster global speed of convergence than the

ECM algorithm. Moreover, they noted that there are some rare situations when the global speed of convergence is slower than the ECM algorithm.

Algorithm 5: ECME algorithm

Input: Initial value $\boldsymbol{\theta}^{(0)}$
while $\ell_{\text{obs}}(\boldsymbol{\theta}^{(t+1)}; \boldsymbol{y}_{\text{obs}}) - \ell_{\text{obs}}(\boldsymbol{\theta}^{(t)}; \boldsymbol{y}_{\text{obs}}) > \delta$ **do**
  $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}) \leftarrow \mathbb{E}_{\boldsymbol{\theta}^{(t)}}[\ell_{\text{com}}(\boldsymbol{\theta}; \boldsymbol{y}_{\text{com}})|\boldsymbol{y}_{\text{obs}}]$;
  $\boldsymbol{\theta}^{(t+1/S)} \leftarrow \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}_1}{\operatorname{argmax}}\, Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)})$;
  $\vdots$
  $\boldsymbol{\theta}^{(t+1)} \leftarrow \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}_S}{\operatorname{argmax}}\, \ell_{\text{obs}}(\boldsymbol{\theta}; \boldsymbol{y}_{\text{obs}})$;
end

1.5 Normal mean-variance mixture representation

The EM algorithms in Sections 1.3 and 1.4 can be applied to the VG distribution via the normal mean-variance mixture (NMVM) representation where the mixing variable can be treated as unobserved data. This representation can also be interpreted as a hierarchical state-space model which facilitates the Bayesian approach.

The NMVM representation preserves some nice properties from the normal distribution such as closure under linear transformation and infinite divisibility. Other types of mixtures include the scale mixture of uniform. See [14] for more examples of variance mixture distributions.

In this section, we discuss about the generalised inverse Gaussian (GIG) distribution which is the mixing distribution of the GH distribution. We note that the VG distribution is the limiting case of the GH distribution that can have unbounded density.

1.5.1 Generalised inverse Gaussian distribution

The GIG distribution [8, 31, 56] is the mixing distribution of the GH distribution in the NMVM representation. Properties of the GIG distribution is presented in this section.

Definition 1.5.1 (Generalised Inverse Gaussian Distribution)**.** *The random variable U follows a generalised inverse Gaussian (GIG) distribution if its probability density*

function (pdf) is

$$f(u) = \frac{(\psi/\chi)^{\lambda/2}}{2K_\lambda\big(\sqrt{\chi\psi}\big)} u^{\lambda-1} \exp\left(-\frac{1}{2}\Big(\frac{\chi}{u} + \psi u\Big)\right), \ x > 0 \tag{1.20}$$

where $K_\lambda(\cdot)$ represents the modified Bessel function of the second kind with index λ (see Appendix C1) and the parameters (λ, χ, ψ) satisfy the conditions

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

The GIG random variable is denoted by $U \sim \mathcal{GIG}(\lambda, \chi, \psi)$.

The pdf is unimodal and the mode is located at

$$\begin{cases} \frac{\lambda-1+\sqrt{(\lambda-1)^2+\chi\psi}}{\psi} & \text{if } \psi > 0, \\ \frac{\chi}{2(1-\lambda)} & \text{if } \psi = 0. \end{cases}$$

In other words, λ can be considered as a parameter that controls the location of the mode, and focuses the weighting on specific regions on the real line.

Looking at the tail of the pdf in (1.20) as $u \to \infty$, the factor $\exp\big(-\frac{\chi}{2u}\big)$ becomes negligible, and the factor $u^{\lambda-1}\exp\big(-\frac{\psi u}{2}\big)$ dominates when $\psi > 0$. So smaller values of ψ puts more weight at the tail probability while the other parameters are fixed.

Approaching the lower region as $u \to 0$, the factor $\exp\big(-\frac{\psi u}{2}\big)$ becomes negligible, and the factor $u^{\lambda-1}\exp\big(-\frac{\chi}{2u}\big)$ dominates when $\chi > 0$. So smaller values of χ puts more weight at the zero probability while the other parameters are fixed. In fact, the pdf is unbounded when $\chi = 0$ and $0 < \lambda < 1$.

A useful parametrisation is given by

$$\omega = \sqrt{\chi\psi}, \quad \eta = \sqrt{\frac{\chi}{\psi}}.$$

Setting $\omega = 0$ encaptures the limiting cases of either $\chi \to 0, \psi > 0$ and $\psi \to 0, \chi > 0$. For the case when $\omega > 0$, then the pdf in (1.20) takes an alternate form of

$$f(u) = \frac{\eta^{-\lambda}}{2K_\lambda(\omega)} u^{\lambda-1} \exp\left(-\frac{\omega}{2}\left(\frac{\eta}{u} + \frac{u}{\eta}\right)\right).$$

Increasing ω increases the probability around the mean while also decreasing the variance. For this reason, ω is referred to as the concentration parameter while η is referred to as the scale parameter.

The general moments and log-moments of the GIG random variable for the non-limiting case ($\chi > 0$ and $\psi > 0$) is given by

$$\mathbb{E}[U^m] = \eta^m \frac{K_{\lambda+m}(\omega)}{K_\lambda(\omega)}, \tag{1.21}$$

$$\mathbb{E}[U^m \log U] = \frac{d}{ds}\mathbb{E}[U^s]\Big|_{s=m} = \eta^m \frac{K_{\lambda+m}(\omega)\log\eta + K^{(1,0)}_{\lambda+m}(\omega)}{K_\lambda(\omega)}, \tag{1.22}$$

$$\mathbb{E}\big[(\log U)^2\big] = \frac{d^2}{ds^2}\mathbb{E}[U^s]\Big|_{s=0} = (\log\eta)^2 + \frac{2K^{(1,0)}_\lambda(\omega)\log\eta + K^{(2,0)}_\lambda(\omega)}{K_\lambda(\omega)} \tag{1.23}$$

for $m \in \mathbb{R}$ where $K^{(1,0)}_\lambda(\omega) = \frac{\partial}{\partial\alpha}K_\alpha(\omega)\big|_{\alpha=\lambda}$ and $K^{(2,0)}_\lambda(\omega) = \frac{\partial^2}{\partial^2\alpha}K_\alpha(\omega)\big|_{\alpha=\lambda}$. The moment generating function (MGF) is given by

$$\mathbb{M}_U(t) = \mathbb{E}\big[e^{tU}\big] = \left(1 - \frac{2t}{\psi}\right)\frac{K_\lambda\Big(\omega\Big(1-\frac{2t}{\psi}\Big)\Big)}{K_\lambda(\omega)} \text{ for } \psi > 2t. \tag{1.24}$$

The GIG distribution contains the following special cases:

(i) *Inverse Gaussian distribution* when $\lambda = -0.5$,

(ii) *Inverse gamma distribution* when $\psi = 0$ and $\lambda < 0$ such that by setting $\lambda = -\alpha$, $\chi = 2\beta$, the pdf becomes

$$f(u) = \frac{\beta^\alpha}{\Gamma(\alpha)}u^{-\alpha-1}\exp\left(-\frac{\beta}{u}\right) \quad \text{for } u > 0, \tag{1.25}$$

and is denoted by $\mathcal{IG}(\alpha, \beta)$.

(iii) *Gamma distribution* when $\chi = 0$, $\lambda > 0$ such that by setting $\lambda = \alpha$, $\psi = 2\beta$, the pdf becomes

$$f(u) = \frac{\beta^\alpha}{\Gamma(\alpha)}u^{\alpha-1}\exp(-\beta u), \quad \text{for } u > 0\,.$$

and is denoted by $\mathcal{G}(\alpha, \beta)$. Note that the pdf is unbounded at 0 for $0 < \alpha < 1$.

The general moments and log-moments of $U \sim \mathcal{G}(\alpha, \beta)$ are given by

$$\mathbb{E}[U^m] = \frac{\Gamma(\alpha+m)}{\beta^m \Gamma(\alpha)} \quad \text{for } \alpha + m > 0, \tag{1.26}$$

$$\mathbb{E}[U^m \log U] = \frac{\Gamma(\alpha+m)}{\beta^m \Gamma(\alpha)} \big(\psi(\alpha+m) - \log\beta\big) \quad \text{for } \alpha + m > 0, \tag{1.27}$$

$$\mathbb{E}[(\log U)^2] = \big(\psi(\alpha) - \log\beta\big)^2 + \psi'(\alpha) \tag{1.28}$$

where $\psi(x) = \frac{d}{dx}\log\Gamma(x)$ represents a digamma, and $\psi'(x)$ represents a trigamma function.

We remark that $\mathcal{G}(\alpha, \beta)$ and $\mathcal{IG}(\alpha, \beta)$ are the mixing distributions of VG and Student's t distributions respectively from the NMVM representation. See Jørgensen [56], Embrechts [31], and Barndorff-Nielsen and Stelzer [8] for other properties of the GIG distribution.

1.5.2 Generalised hyperbolic distribution

Definition 1.5.2 (Normal Mean-Variance Mixture)**.** *A random variable $\boldsymbol{Y}$ is said to have a normal mean-variance mixture (NMVM) representation if it can be expressed as*

$$\boldsymbol{Y} \overset{d}{=} \boldsymbol{\mu} + U\boldsymbol{\gamma} + \sqrt{U}\boldsymbol{A}\boldsymbol{Z}, \tag{1.29}$$

where $\boldsymbol{Z} \sim \mathcal{N}_k(\boldsymbol{0}, \boldsymbol{I}_k)$, U is a non-negative random variable independent of $\boldsymbol{Z}$, $\boldsymbol{\mu} \in \mathbb{R}^d$, $\boldsymbol{\gamma} \in \mathbb{R}^d$ and $\boldsymbol{A} \in \mathbb{R}^{d\times k}$.

The random variable U is referred to as the mixing variable, $\boldsymbol{\mu}$ as the location parameter, $\boldsymbol{\gamma}$ as the skewness parameter and $\boldsymbol{A}$ as the scale parameter. When $\boldsymbol{A}$ is a square matrix, it can be thought of as the Cholesky decomposition of the scale matrix $\boldsymbol{\Sigma}$ (i.e. $\boldsymbol{A}\boldsymbol{A}' = \boldsymbol{\Sigma}$).

Another interpretation of the mixture representation is that the conditional distribution of $\boldsymbol{Y}$ given U is

$$\boldsymbol{Y}|U \sim \mathcal{N}_d(\boldsymbol{\mu} + U\boldsymbol{\gamma}, U\boldsymbol{\Sigma}). \tag{1.30}$$

Thus, this mixture representation allows us to easily generate random variables $\boldsymbol{Y}$ by first generating U, then generating $\boldsymbol{Y}$ from the conditional normal distribution.

We also can easily obtain the following formulas for the mean and covariance matrix using the mixture representation

$$\mathbb{E}(\boldsymbol{Y}) = \boldsymbol{\mu} + \mathbb{E}(U)\boldsymbol{\gamma}, \tag{1.31}$$

$$\operatorname{cov}(\boldsymbol{Y}) = \mathbb{E}(U)\boldsymbol{\Sigma} + \operatorname{var}(U)\boldsymbol{\gamma}\boldsymbol{\gamma}'. \tag{1.32}$$

It is common to set $\mathbb{E}(U) = 1$ so that the scale parameter $\boldsymbol{\Sigma}$ corresponds to the covariance matrix for the symmetric case.

The expression for the MGF can be easily obtained from the mixture representation with

$$\begin{aligned}
\mathbb{M}_{\boldsymbol{Y}}(\boldsymbol{t}) &= \mathbb{E}_U\big[\mathbb{E}_{\boldsymbol{Y}|U}[\exp(\boldsymbol{t}'\boldsymbol{Y})|U]\big] \\
&= e^{\boldsymbol{t}'\boldsymbol{\mu}}\,\mathbb{E}_U\big[\exp\big(U\big(\boldsymbol{t}'\boldsymbol{\gamma} + \tfrac{1}{2}\boldsymbol{t}'\boldsymbol{\Sigma}\boldsymbol{t}\big)\big)\big] \\
&= e^{\boldsymbol{t}'\boldsymbol{\mu}}\,\mathbb{M}_U\big(\boldsymbol{t}'\boldsymbol{\gamma} + \tfrac{1}{2}\boldsymbol{t}'\boldsymbol{\Sigma}\boldsymbol{t}\big)
\end{aligned}$$

where $\mathbb{M}_U$ represents the MGF of the mixing variable. Similarly, the expression for the characteristic function is given by

$$\phi_{\boldsymbol{Y}}(\boldsymbol{t}) = e^{i\boldsymbol{t}'\boldsymbol{\mu}}\,\mathbb{M}_U\big(i\boldsymbol{t}'\boldsymbol{\gamma} + \tfrac{1}{2}\boldsymbol{t}'\boldsymbol{\Sigma}\boldsymbol{t}\big).$$

Another useful property is that

$$\mathbb{E}\left[\frac{1}{U}(\boldsymbol{Y} - \boldsymbol{\mu})\right] = \mathbb{E}\left[\boldsymbol{\gamma} + \frac{1}{\sqrt{U}}\boldsymbol{A}\boldsymbol{Z}\right] = \boldsymbol{\gamma}. \tag{1.33}$$

See Barndorff-Nielsen et al. [7] for other properties of distributions with NMVM representation.

Definition 1.5.3 (Generalised Hyperbolic Distribution)**.** *The random variable $\boldsymbol{Y}$ has a d-dimensional generalised hyperbolic (GH) distribution if it has a normal mean-variance mixture representation with mixing variable $U \sim \mathcal{GIG}(\lambda, \chi, \psi)$, and has pdf*

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{\left(\frac{\psi}{\chi}\right)^{\frac{\lambda}{2}}(\psi + \boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma})^{\frac{d}{2}-\lambda}}{(2\pi)^{\frac{d}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}K_\lambda\big(\sqrt{\chi\psi}\big)} \times \frac{K_{\lambda-\frac{d}{2}}\Big(\sqrt{(\chi + z^2)(\psi + \boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma})}\Big)e^{(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}}}{\Big(\sqrt{(\chi + z^2)(\psi + \boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma})}\Big)^{\frac{d}{2}-\lambda}} \tag{1.34}$$

where

$$z^2 = (\boldsymbol{y} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}) \tag{1.35}$$

is the Mahalanobis distance, and is denoted by $\boldsymbol{Y} \sim \mathcal{GH}_d(\lambda, \chi, \psi, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\gamma})$.

The GH distribution is closed under linear transformations which can be expressed using the following proposition.

Proposition 1.5.4 (Linear Transformation)**.** *If* $\boldsymbol{Y} \sim \mathcal{GH}_d(\lambda, \chi, \psi, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\gamma})$*, then*

$$\boldsymbol{BY} + \boldsymbol{a} \sim \mathcal{GH}_k(\lambda, \chi, \psi, \boldsymbol{B\mu} + \boldsymbol{a}, \boldsymbol{B\Sigma B}', \boldsymbol{B\gamma})$$

where $\boldsymbol{B} \in \mathbb{R}^{k \times d}$ *and* $\boldsymbol{a} \in \mathbb{R}^k$.

The parametrisation used in the pdf in (1.34) has an identification problem since $\mathcal{GH}(\lambda, \chi, \psi, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\gamma})$ and $\mathcal{GH}(\lambda, \chi/k, k\psi, \boldsymbol{\mu}, k\boldsymbol{\Sigma}, k\boldsymbol{\gamma})$ both produce the same pdf for $k > 0$. This becomes problematic when estimating the parameters of the GH distribution using this parametrisation, and so extra constraints are needed resulting in multiple parametrisations. See Breymann and Lüthi [16], McNeil et al. [75] for other parametrisations of the GH distribution.

The GH distribution contains the following special cases:

(i) *Hyperbolic distribution* when $\lambda = \frac{d+1}{2}$,
(ii) *Normal inverse Gaussian distribution* when $\lambda = -\frac{1}{2}$,
(iii) *Multivariate skew Student's t distribution* when $\lambda = -\upsilon/2$, $\chi = \upsilon$, and $\psi = 0$ (see Appendix C2),
(iv) *Multivariate skew VG distribution* when $\lambda = \nu$, $\psi = 2\nu$, and $\chi = 0$.

1.5.3 Variance gamma distribution

Definition 1.5.5 (Variance Gamma Distribution)**.** *The pdf of a d-dimensional multivariate skewed* variance gamma *(VG) distribution is given by*

$$f_{VG}(\boldsymbol{y}) = \frac{2\nu^{\nu}}{(2\pi)^{\frac{d}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}} \Gamma(\nu)} \times \frac{K_{\nu - \frac{d}{2}}\left(\sqrt{(2\nu + \boldsymbol{\gamma}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\gamma}) z^2}\right) e^{(\boldsymbol{y} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} \boldsymbol{\gamma}}}{\left(\sqrt{z^2 / (2\nu + \boldsymbol{\gamma}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\gamma})}\right)^{\frac{d}{2} - \nu}} \tag{1.36}$$

where $\nu > 0$, z^2 *in* (1.35)*, and the distribution is denoted by* $\mathcal{VG}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}, \nu)$.

Using the NMVM representation, the VG distribution can be represented by

$$\boldsymbol{Y}|U \sim \mathcal{N}_d(\boldsymbol{\mu} + \boldsymbol{\gamma}U, U\boldsymbol{\Sigma}), \quad U \sim \mathcal{G}(\nu, \nu) \tag{1.37}$$

and so the mean and covariance matrix of a VG random vector $\boldsymbol{Y}$ are given by

$$\mathbb{E}(\boldsymbol{Y}) = \boldsymbol{\mu} + \boldsymbol{\gamma} \quad \text{and} \quad \text{cov}(\boldsymbol{Y}) = \boldsymbol{\Sigma} + \tfrac{1}{\nu}\boldsymbol{\gamma}\boldsymbol{\gamma}' \tag{1.38}$$

respectively from the mean and covariance formulas in (1.31) and (1.32).

Using the asymptotic properties of the modified Bessel function of the second kind in Appendix C1, the pdf in (1.36) as $\boldsymbol{y} \to \boldsymbol{\mu}$ is given by

$$f_{\text{VG}}(\boldsymbol{y}) \sim \begin{cases} \dfrac{2^{-\nu}\pi^{-\frac{d}{2}}\nu^{\nu}}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}\Gamma(\nu)}\left(\dfrac{2^{2\nu-d}\Gamma\left(\nu-\frac{d}{2}\right)}{(2\nu+\boldsymbol{\gamma}'\boldsymbol{\Sigma}\boldsymbol{\gamma})^{\nu-\frac{d}{2}}} + \Gamma\left(\tfrac{d}{2}-\nu\right)z^{2\nu-d}\right) & \text{if } \nu \neq \tfrac{d}{2}, \\ \dfrac{2^{-\nu}\pi^{-\frac{d}{2}}\nu^{\nu}}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}\Gamma(\nu)}(-2\log(z)) & \text{if } \nu = \tfrac{d}{2}. \end{cases} \tag{1.39}$$

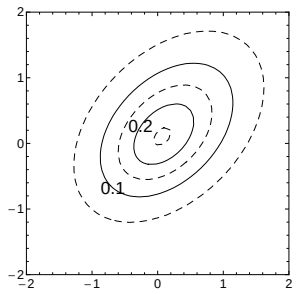
Looking at the index of z in the asymptotic expressions above, the pdf is

Case 1: differentiable when $2\nu - d > 1 \Rightarrow \nu > \frac{d+1}{2}$,
Case 2: cusped when $0 < 2\nu - d < 1 \Rightarrow \nu \in (\frac{d}{2}, \frac{d+1}{2}]$, and
Case 3: unbounded when $2\nu - d \leq 0 \Rightarrow \nu \leq \frac{d}{2}$.

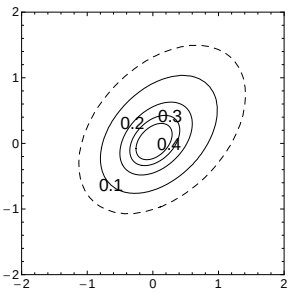
To visualise the shape of a bivariate VG distribution, Figure 1.1 gives four pairs of contour and three-dimensional plots for various parameters of the bivariate VG distribution. The first pair of plots is based on parameters

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.4 \\ 0.4 & 1 \end{pmatrix}, \quad \boldsymbol{\gamma} = \begin{pmatrix} 0.2 \\ 0.3 \end{pmatrix}, \text{ and } \quad \nu = 3. \tag{1.40}$$

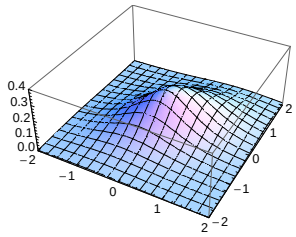
Based on the distribution for the first pair of plots, three other pairs of plots demonstrate the changes in pdf when the shape parameter decreases to $\nu = 0.6$, the skewness parameter increases to $\boldsymbol{\gamma} = (0.5, 2)$, and the correlation coefficient in $\boldsymbol{\Sigma}$ increases to 0.8, respectively, while keeping other parameters fixed. Plots (b) and (d) display high central density indicating unbounded density when the shape parameter drops to $\nu = 0.6$ (since $\nu \leq \frac{d}{2}$). Plots (e) and (g) show that the centres of the contours are skewed to one side and move away from the origin of (0,0) when the two skewness increase and differ more. Lastly, plots (f) and (h) show that the contours are more elliptical than rounded as the correlation between the two dimensions increases.



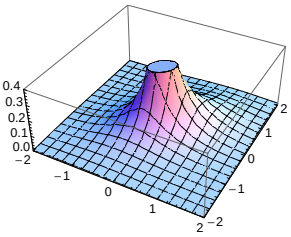
(a) standard contour plot



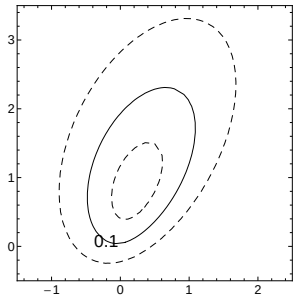
(b) small ν contour plot



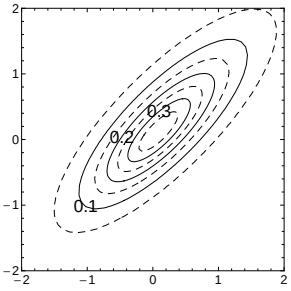
(c) standard 3D plot



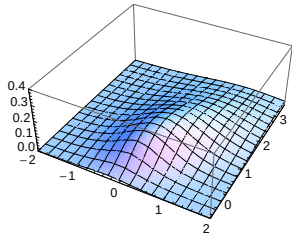
(d) small ν 3D plot



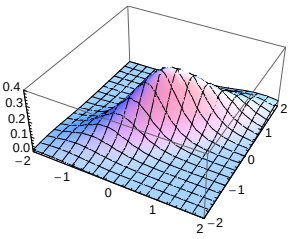
(e) large $\boldsymbol{\gamma}$ contour plot



(f) large correlation contour plot



(g) large $\boldsymbol{\gamma}$ 3D plot



(h) large correlation 3D plot

Figure 1.1. Various contour and 3D plots of bivariate skewed VG distribution for different parameters. In the contour plots, the bold lines represent level sets $\{0.1, 0.2, 0.3, 0.4\}$, and the dashed lines represent level sets $\{0.05, 0.15, 0.25, 0.35\}$. The density for the 3D plots is kept between 0 and 0.4

1.6 Contributions and structure of the thesis

As discussed in Section 1.1, there are a number of significant research gaps that this thesis aims to address and are summarised below.

Our first and most important contribution is to derive an efficient estimation method to implement the VG distribution. It is the limiting case of the GH distribution when the parameter χ approaches to zero which lies on the boundary of the parameter space. In general, the density of the GH distribution in (1.34) with shape parameter $\chi > 0$ is bounded since $\chi + z^2 > 0$ is always satisfied. However when $\chi = 0$, the pdf can be unbounded and it involves a ratio which has the form of $\frac{\infty}{\infty}$ when $\boldsymbol{y}$ approaches to $\boldsymbol{\mu}$, and so the VG distribution behave differently from the GH distribution. Thus the VG distribution is not a simple sub-member of the GH distribution, so methodologies developed for GH distribution cannot simply be applied to the VG distribution. We present different estimation methods within the EM framework that address the cusp and unbounded density problem associated with the VG distribution.

Our second contribution is to develop EM algorithms to address specifically the issue of unbounded likelihood with respect to the location parameter. We review in Section 1.2 desirable properties of MLE and remark that these properties fail for the location estimates when the likelihood is cusped or unbounded. Furthermore, even the estimator become problematic as the derivatives of the likelihood may also be unbounded. We present three modifications to the classical likelihood, namely the capped, LOO and weighted LOO likelihoods. For the capped likelihood method, we study the optimal choice of capping level for different shape parameters of the VG distribution and propose an algorithm where the capping level updates after each iteration. We compare the performance of these methodologies in the simulation study in Section 4.4. To the best of our knowledge, there is no literature that has successfully developed and implemented methods to estimate parameters when the likelihood is unbounded with respect to the location parameter and so this work is pioneer in the field.

Our third contribution is to study the properties of the LOO estimator for the location parameter designed to solve the problem of unbounded likelihood. As previously mentioned, research on the parameter estimation involving cusp and unbounded likelihood is very limited. Podgórski and Wallin [89] proved the consistency and the lower bound

on the rate of convergence for the location estimate using the LOO likelihood for the unbounded likelihood case. To get a better understanding of the behaviour of the location estimator that maximises the LOO likelihood, we find that the double generalised gamma distribution seems to provide a good approximation to the distribution of the location estimator. We believe that our findings provides useful insight for further theoretical development for the properties of the location estimator when the likelihood has cusp or unbounded points at the mode.

Our fourth contribution is to provide efficient methods to compute the SEs of the VG distribution. Currently, there are no explicit formulas available for the SE calculation for the VG distribution. We derive formulas to calculate the observed and Fisher information matrices for all parameters using Louis' method in (1.10). These formulas are expressed in matrix form to facilitate implementation through programming. Our empirical result from the simulation study is able to demonstrate the successful implementation of these methods to calculate SE estimates for the VG distribution and its extension to multivariate time series models.

Our fifth contribution is to extend the VARMA model to have VG or Student's t innovations to model multivariate financial time series. Data sets such as Bitcoin and high frequency financial returns display large kurtosis with some skewness and persistence. This suggest the need to adopt a time series model like the VARMA with VG innovations. We first extend the VAR model to adopt VG or Student's t innovations which is called the VAR-VG and VAR-t model respectively. This extension can be easily implemented utilising the NMVM representation to obtain a closed-form solution for the CM-step. However, upon adding MA terms into the model, there is no close-form solution for the CM-step. So instead, we consider an approximation using a higher order VAR type model for the CM-step. This model is applied to fit high frequency financial stocks and daily cryptocurrency return series. Model performance is assessed and forecast is performed. To the best of our knowledge, there is no research work on multivariate financial time series models with VG or Student's t innovations to capture the extreme kurtosis. We believe that this work makes a significant contribution to the time series modelling and investment portfolio settings.

The remaining part of the thesis is structured in the following way: Chapter 2 develops the ECM algorithm to estimate parameters of the VG distribution for the unbounded density case. We propose the likelihood with an optimal capping level and present

the alternating ECM (AECM) algorithm along with the calculation of SEs. Chapter 3 introduces the LOO likelihood method and present the theory for the maximum LOO likelihood estimators. Moreover, we discuss some approximation methods for the implementation of the AECM algorithm when using the LOO likelihood, and numerically investigate asymptotic properties of the location estimator using the LOO likelihood when the density of VG distribution is cusped or unbounded at the mode. Chapter 4 motivates the weighted LOO likelihood to deal with repeated data points and compares different likelihood methods when applied to data sets with data multiplicity. Chapter 5 extends the AECM algorithm to accommodate the VARMA-VG and VARMA-t models, and applies the algorithm to model daily and high frequency stock indices, and daily cryptocurrency returns including the emerging Bitcoin index. Finally, a brief conclusion with discussion of future research is given in Chapter 6. The appendices present details about the derivatives of the log-likelihood applied to calculating the observed and Fisher information matrices. It also summarises results on related functions and distributions.

CHAPTER 2

EM Algorithms for Variance Gamma Distribution

The VG distribution has applications in many areas such as finance, signal processing and quality control. See Kotz et al. [60] and Madan and Seneta [71] for other applications. This chapter aims to develop ECM algorithms to estimate parameters of the VG distribution.

An outline of the MCECM algorithm for estimating the parameters of the GH distribution have been presented by Hu [51] and McNeil et al. [75]. They claimed their algorithm applies to the VG distribution as it is a limiting case of the GH distribution when the shape parameter χ approaches zero. However, they did not address two issues in their algorithm. Firstly, the VG distribution can have unbounded density which can lead to instabilities in the ECM algorithm since some expectations in the E-step diverge to infinity. Secondly, there is no guarantee that the ECM algorithm monotonically converge since the unbounded likelihood violates the differentiability regularity condition as discussed in Section 1.3.2. Moreover, the compactness regularity condition is also violated, particularly, if we set $\boldsymbol{\theta}_0$ to be any point in $\boldsymbol{\Theta}$ such that $\boldsymbol{\mu}$ is at any data point.

Our extensive literature review found limited research on methodologies addressing the unbounded likelihood problem. Podgórski and Wallin [89] considered this problem by developing the leave-one-out (LOO) likelihood where the likelihood is unbounded with respect to the location parameter. They showed the consistency and super-efficiency of the maximum LOO likelihood estimator for the location parameter and discussed the applicability of the LOO likelihood method using the EM algorithm. However, their focus was not on the numerical implementation of their algorithm. We see the need to address this issue by providing computationally efficient and accurate methodology for parameter estimation applied to a wide range of data sets.

For the remaining part of this chapter, Section 2.1 constructs an ECM algorithm for the VG distribution. Section 2.2 extends the ECM algorithm to the AECM algorithm to improve computational efficiency. Section 2.3 analyses issues regarding the unbounded likelihood and proposes the capped likelihood method. Section 2.4 illustrates the calculation of the observed information matrix using Louis' method, Hessian matrix using second order numerical differentiation and Fisher information matrix. Section 2.5 conducts three different simulation studies: the first one evaluates the performance of three ECM algorithms; the second one studies the optimal choice of capping level Δ and the last one compares the SE calculation using the three methods. Section 2.6 presents an application to daily financial returns, and finally the chapter is concluded in Section 2.7.

2.1 ECM algorithm for VG distribution

The MLE of parameters $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}, \nu)$ from the VG distribution in the parameter space $\boldsymbol{\Theta}$ maximises the observed data log-likelihood function

$$\ell(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{n} \log f_{VG}(\boldsymbol{y}_i; \boldsymbol{\theta}) \tag{2.1}$$

where we let $f_{VG}(\cdot)$ be the pdf of the VG distribution in (1.36) and $\boldsymbol{y} = (\boldsymbol{y}_1, \dots, \boldsymbol{y}_n)$ be the observed data. Using the NMVM representation of the VG distribution in (1.37) and letting $\boldsymbol{u} = \{u_1, \dots, u_n\}$ to represent the unobserved or missing data and $\{\boldsymbol{y}, \boldsymbol{u}\}$ to represent the complete data, the complete data likelihood function can be written as

$$L(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{u}) = f(\boldsymbol{y}, \boldsymbol{u}; \boldsymbol{\theta}) = \prod_{i=1}^{n} f_N(\boldsymbol{y}_i | u_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}) f_G(u_i; \nu). \tag{2.2}$$

The complete data log-likelihood function can be factorised into two distinct log-likelihood functions

$$\ell(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{u}) = \ell_N(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}; \boldsymbol{y}, \boldsymbol{u}) + \ell_G(\nu; \boldsymbol{u}) \tag{2.3}$$

where the log-likelihood of the conditional normal distribution ignoring additive constants is given by

$$\begin{aligned}
&\ell_N(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}; \boldsymbol{y}, \boldsymbol{u}) \qquad (2.4)\\
&= -\frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n}\frac{1}{u_i}(\boldsymbol{y}_i - \boldsymbol{\mu} - u_i\boldsymbol{\gamma})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu} - u_i\boldsymbol{\gamma})\\
&= -\frac{1}{2}\Bigg[n\log|\boldsymbol{\Sigma}| + \sum_{i=1}^{n}\frac{1}{u_i}(\boldsymbol{y}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}) + \sum_{i=1}^{n}u_i\boldsymbol{\gamma}'\boldsymbol{\gamma}\\
&\qquad\qquad - \sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{\mu})'\boldsymbol{\gamma} - \sum_{i=1}^{n}\boldsymbol{\gamma}'(\boldsymbol{y}_i - \boldsymbol{\mu})\Bigg]
\end{aligned}$$

and the log-likelihood of the gamma distribution is given by

$$\ell_G(\nu; \boldsymbol{u}) = n\nu\log\nu - n\log\Gamma(\nu) + (\nu - 1)\sum_{i=1}^{n}\log u_i - \nu\sum_{i=1}^{n}u_i. \qquad (2.5)$$

The idea of the estimation procedure of the ECM algorithm is to first estimate the mixing variables $\boldsymbol{u}$ by its conditional expectation given the observed data $\boldsymbol{y}$. Then condition on $\boldsymbol{u}$, the estimation of the parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}, \nu)$ can be separated in two blocks: the conditional maximisation of the conditional normal log-likelihood function with respect to $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\gamma})$ and the conditional maximisation of the gamma log-likelihood function with respect to ν. Details of the estimation procedures are described below.

2.1.1 E-step

Suppose $\boldsymbol{\theta}^{(t)} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}, \nu)$ are the current parameter estimates, then the calculation of the Q-function in (1.6) requires taking the conditional expectation of (2.4) and (2.5) given $\boldsymbol{y}$. Equivalently, it is sufficient to calculate the following conditional expectations:

$$\mathbb{E}_{\boldsymbol{\theta}^{(t)}}\left[\frac{1}{u_i}\middle|\boldsymbol{y}_i\right], \qquad \mathbb{E}_{\boldsymbol{\theta}^{(t)}}[u_i|\boldsymbol{y}_i], \qquad \mathbb{E}_{\boldsymbol{\theta}^{(t)}}[\log u_i|\boldsymbol{y}_i].$$

To derive the conditional expectations of u_i given $\boldsymbol{y}_i$, we need the conditional distribution of u_i given $\boldsymbol{y}_i$ which has density function as:

$$\begin{aligned}
f\big(u_i|\boldsymbol{y}_i; \boldsymbol{\theta}^{(t)}\big) &\propto f\big(u_i, \boldsymbol{y}_i; \boldsymbol{\theta}^{(t)}\big)\\
&\propto u_i^{\nu - \frac{d}{2} - 1}\exp\left[-\frac{z_i^2}{2u_i} - \frac{u_i}{2}\big(2\nu + \boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}\big)\right] \qquad (2.6)
\end{aligned}$$

where $z_i^2 = (\boldsymbol{y}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu})$ which corresponds to a $\mathcal{GIG}(\nu - d/2, z_i^2, 2\nu + \boldsymbol{\gamma}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\gamma})$ distribution (1.20). Using this distribution, we can calculate the following conditional expectations:

$$\widehat{u_i} = \mathbb{E}_{\boldsymbol{\theta}^{(t)}}[u_i | \boldsymbol{y}_i] = \frac{\eta_i \, K_{\nu - \frac{d}{2} + 1}(\omega_i)}{K_{\nu - \frac{d}{2}}(\omega_i)}, \tag{2.7}$$

$$\widehat{1/u_i} = \mathbb{E}_{\boldsymbol{\theta}^{(t)}}\left[\frac{1}{u_i} \middle| \boldsymbol{y}_i \right] = \frac{K_{\nu - \frac{d}{2} - 1}(\omega_i)}{\eta_i \, K_{\nu - \frac{d}{2}}(\omega_i)}, \tag{2.8}$$

$$\widehat{\log u_i} = \mathbb{E}_{\boldsymbol{\theta}^{(t)}}[\log u_i | \boldsymbol{y}_i] = \log \eta_i + \frac{K^{(1,0)}_{\nu - \frac{d}{2}}(\omega_i)}{K_{\nu - \frac{d}{2}}(\omega_i)} \tag{2.9}$$

where $\eta_i = z_i / \sqrt{2\nu + \boldsymbol{\gamma}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\gamma}}$, $\omega_i = z_i \sqrt{2\nu + \boldsymbol{\gamma}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\gamma}}$ and $K^{(1,0)}_{\lambda}(z) = \frac{\partial}{\partial \alpha} K_\alpha(z) \big|_{\alpha = \lambda}$ which can be approximated using the second order central difference approximation

$$K^{(1,0)}_{\lambda}(z) \approx \frac{K_{\lambda + h}(z) - K_{\lambda - h}(z)}{2h}$$

where we let $h = 10^{-5}$.

2.1.2 CM-step for $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ and $\boldsymbol{\gamma}$

Suppose $\boldsymbol{u}$ is given, the MLE of $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\gamma})$ is obtained by maximising $\ell_N(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}; \boldsymbol{y}, \boldsymbol{u})$ in (2.4) with respect to $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\gamma})$ by equating each component of the partial derivatives of $\ell_N(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}; \boldsymbol{y}, \boldsymbol{u})$ to zero. This gives us the following estimates:

$$\hat{\boldsymbol{\mu}} = \frac{S_{\boldsymbol{y}/u} S_u - n S_{\boldsymbol{y}}}{S_{1/u} S_u - n^2}, \tag{2.10}$$

$$\hat{\boldsymbol{\gamma}} = \frac{S_{\boldsymbol{y}} - n \hat{\boldsymbol{\mu}}}{S_u}, \tag{2.11}$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{u_i} (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}})' - \frac{1}{n} \hat{\boldsymbol{\gamma}} \hat{\boldsymbol{\gamma}}' S_u \tag{2.12}$$

where the complete data sufficient statistics are

$$S_{\boldsymbol{y}} = \sum_{i=1}^{n} \boldsymbol{y}_i, \quad S_{\boldsymbol{y}/u} = \sum_{i=1}^{n} \frac{1}{u_i} \boldsymbol{y}_i, \quad S_u = \sum_{i=1}^{n} u_i, \quad S_{1/u} = \sum_{i=1}^{n} \frac{1}{u_i}. \tag{2.13}$$

2.1.3 CM-step for ν

Given the mixing variables $\boldsymbol{u}$, the MLE of ν can be obtained by maximising the log-likelihood of the gamma distribution,

$$\ell_G(\nu; \boldsymbol{u}) = n\nu \log \nu - n \log \Gamma(\nu) + (\nu - 1) S_{\log u} - \nu S_u \tag{2.14}$$

with respect to ν using numerical optimisation techniques where

$$S_{\log u} = \sum_{i=1}^{n} \log u_i. \tag{2.15}$$

This maximisation corresponds to the MCECM algorithm in Section 1.4.2. Alternatively, maximising the observed log-likelihood $\ell_{VG}(\boldsymbol{\theta}; \boldsymbol{y})$ in (2.1) with respect to ν corresponds to the ECME algorithm in Section 1.4.3 and can dramatically improve the convergence rate of the algorithm.

Algorithm 6: MCECM algorithm for VG distribution

Input: Initial value $\boldsymbol{\theta}^{(0)}$
while $\ell(\boldsymbol{\theta}^{(t+1)}; \boldsymbol{y}) - \ell(\boldsymbol{\theta}^{(t)}; \boldsymbol{y}) > \delta$ **do**
  $Q\big(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}\big) \leftarrow \mathbb{E}_{\boldsymbol{\theta}^{(t)}}[\ell(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{u}) | \boldsymbol{y}]$;
  $\boldsymbol{\theta}^{(t+1/2)} \leftarrow \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}_1}{\operatorname{argmax}} Q\big(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}, \nu^{(t)}; \boldsymbol{\theta}^{(t)}\big)$;
  $Q\big(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t+1/2)}\big) \leftarrow \mathbb{E}_{\boldsymbol{\theta}^{(t+1/2)}}[\ell(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{u}) | \boldsymbol{y}]$;
  $\boldsymbol{\theta}^{(t+1)} \leftarrow \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}_2}{\operatorname{argmax}} Q\big(\boldsymbol{\mu}^{(t+1/2)}, \boldsymbol{\Sigma}^{(t+1/2)}, \boldsymbol{\gamma}^{(t+1/2)}, \nu; \boldsymbol{\theta}^{(t+1/2)}\big)$;
end

Algorithm 7: ECME algorithm for VG distribution

Input: Initial value $\boldsymbol{\theta}^{(0)}$
while $\ell(\boldsymbol{\theta}^{(t+1)}; \boldsymbol{y}) - \ell(\boldsymbol{\theta}^{(t)}; \boldsymbol{y}) > \delta$ **do**
  $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}) \leftarrow \mathbb{E}_{\boldsymbol{\theta}^{(t)}}[\ell(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{u}) | \boldsymbol{y}]$;
  $\boldsymbol{\theta}^{(t+1/2)} \leftarrow \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}_1}{\operatorname{argmax}} Q(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}, \nu^{(t)}; \boldsymbol{\theta}^{(t)})$;
  $\boldsymbol{\theta}^{(t+1)} \leftarrow \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}_2}{\operatorname{argmax}} \ell(\boldsymbol{\mu}^{(t+1/2)}, \boldsymbol{\Sigma}^{(t+1/2)}, \boldsymbol{\gamma}^{(t+1/2)}, \nu; \boldsymbol{y})$;
end

2.2 Alternating ECM algorithm for skewed VG distribution

The ECM algorithm utilises the NMVM representation in (1.37) as a conventional data augmentation scheme. To improve the rate of convergence of the ECM algorithm, we consider a more general data augmentation scheme called the alternating ECM (AECM) algorithm [79] which is a generalisation of the ECME algorithm. Let $u_i = v_i/a(\boldsymbol{\theta})$ where $a(\boldsymbol{\theta})$ is any positive function of $\boldsymbol{\theta}$. Then (1.37) becomes

$$\boldsymbol{y}_i|u_i \sim \mathcal{N}_d\left(\boldsymbol{\mu} + v_i\frac{\boldsymbol{\gamma}}{a(\boldsymbol{\theta})}, v_i\frac{\boldsymbol{\Sigma}}{a(\boldsymbol{\theta})}\right), \quad \frac{v_i}{a(\boldsymbol{\theta})} \sim \mathcal{G}(\nu,\nu). \tag{2.16}$$

The purpose of the data augmentation is to choose a positive function $a(\boldsymbol{\theta})$ such that it allows the fractional missing index in (1.15) to vary according to $a(\boldsymbol{\theta})$. One popular choice is $a(\boldsymbol{\theta}) = |\boldsymbol{\Sigma}|^\alpha$ where α is a working parameter [79]. However, for a general function $a(\boldsymbol{\theta})$, the parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ in the factorisation $f(\boldsymbol{y}_i, u_i|\boldsymbol{\theta}) = f(\boldsymbol{y}_i|u_i, \boldsymbol{\theta}_1)f(u_i|\boldsymbol{\theta}_2)$ in (2.2) may be dependent, making the implementation complicated with possibly no closed-form solution. To simplify the implementation procedure, Liu [63] considered $a(\boldsymbol{\theta})$ itself as a parameter denoted by κ where $\kappa = 1$ corresponds to the conventional data augmentation. He proposed an updating formula for κ by maximising the observed log-likelihood of the multivariate symmetric Student's t distribution given ν as well as a procedure that estimates (κ, ν) together. In this thesis, we consider updating κ for the multivariate skewed VG distribution by choosing

$$\hat{\kappa} = \underset{\kappa>0}{\operatorname{argmax}}\ \ell_{VG}\left(\hat{\boldsymbol{\mu}}, \frac{\widehat{\boldsymbol{\Sigma}}}{\kappa}, \frac{\widehat{\boldsymbol{\gamma}}}{\kappa}, \hat{\nu}\right) \tag{2.17}$$

using numerical optimisation techniques given the current estimates $(\hat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}, \widehat{\boldsymbol{\gamma}}, \hat{\nu})$. Then we update the new parameter estimates as

$$\widehat{\boldsymbol{\gamma}}^* = \frac{\widehat{\boldsymbol{\gamma}}}{\hat{\kappa}}, \quad \text{and} \quad \widehat{\boldsymbol{\Sigma}}^* = \frac{\widehat{\boldsymbol{\Sigma}}}{\hat{\kappa}}. \tag{2.18}$$

In summary, the AECM algorithm involves the following steps:

Initialisation step: Choose suitable starting values $(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0, \boldsymbol{\gamma}_0, \nu_0)$. It is recommended to choose starting values $(\bar{\boldsymbol{y}}, \mathrm{cov}(\boldsymbol{y}), \boldsymbol{0}, d+3)$ where $\bar{\boldsymbol{y}}$ and $\mathrm{cov}(\boldsymbol{y})$ denote the sample mean and sample covariance matrix of $\boldsymbol{y}$ respectively.

At the t^{th} iteration with current estimates $(\boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}, \boldsymbol{\gamma}^{(t)}, \nu^{(t)})$:

E-step 1: Calculate $\widehat{u}_i$ and $\widehat{1/u_i}$ for $i = 1, \ldots, n$ in (2.7) and (2.8), respectively, using $(\boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}, \boldsymbol{\gamma}^{(t)}, \nu^{(t)})$. Calculate also the sufficient statistics $S_{\boldsymbol{y}/u}$, S_u and $S_{1/u}$ in (2.13).

CM-step 1: Update the parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\gamma})$ in (2.10) to (2.12) respectively using the sufficient statistics.

CM-step 2: Estimate κ to update the parameters $(\boldsymbol{\Sigma}, \boldsymbol{\gamma})$ using (2.18).

CM-step 3: Update the parameter ν by maximising the observed log-likelihood $\ell(\boldsymbol{\theta}; \boldsymbol{y})$ in (2.1).

Stopping rule: Repeat the procedures until the relative increment of log-likelihood function is sufficiently small as in (1.8).

Algorithm 8: AECM algorithm for VG distribution

Input: Initial value $\boldsymbol{\theta}^{(0)}$

while $\ell(\boldsymbol{\theta}^{(t+1)}; \boldsymbol{y}) - \ell(\boldsymbol{\theta}^{(t)}; \boldsymbol{y}) > \delta$ **do**

$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}) \leftarrow \mathbb{E}_{\boldsymbol{\theta}^{(t)}}[\ell(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{u})|\boldsymbol{y}]$;

$\boldsymbol{\theta}^{(t+1/3)} \leftarrow \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}_1}{\operatorname{argmax}}\, Q(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}, \nu^{(t)}; \boldsymbol{\theta}^{(t)})$;

$\boldsymbol{\theta}^{(t+2/3)} \leftarrow \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}_2}{\operatorname{argmax}}\, \ell(\boldsymbol{\mu}^{(t+1/3)}, \frac{1}{\kappa}\boldsymbol{\Sigma}^{(t+1/3)}, \frac{1}{\kappa}\boldsymbol{\gamma}^{(t+1/3)}, \nu^{(t+1/3)}; \boldsymbol{y})$;

$\boldsymbol{\theta}^{(t+1)} \leftarrow \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}_3}{\operatorname{argmax}}\, \ell(\boldsymbol{\mu}^{(t+2/3)}, \boldsymbol{\Sigma}^{(t+2/3)}, \boldsymbol{\gamma}^{(t+2/3)}, \nu; \boldsymbol{y})$;

end

2.3 Capped likelihood method for dealing with unbounded likelihood

Numerical problems may occur when dealing with small shape parameter such that $\nu \leq \frac{d}{2}$ since $f_{VG}(\boldsymbol{y})$ in (2.1) at $\boldsymbol{\mu}$ is unbounded which was shown in (1.39). See Figure 3.1 for a graphical illustration of the unbounded log-likelihood function with respect to the location parameter. Using the asymptotic properties of the modified Bessel function

of the second kind in Appendix C1, we can show that as $\boldsymbol{\mu} \to \boldsymbol{y}_i$,

$$\mathbb{E}_{\boldsymbol{\theta}}[u_i|\boldsymbol{y}_i] \sim \begin{cases} \frac{2\nu-d}{2\nu+\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}} & \text{if } \nu > \frac{d}{2}, \\ -\frac{1}{(2\nu+\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma})\log\left(\sqrt{2\nu+\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}}z_i\right)} & \text{if } \nu = \frac{d}{2}, \\ \frac{\Gamma(\nu-\frac{d}{2}+1)}{\Gamma(\frac{d}{2}-\nu)}2^{2\nu-d+1}(2\nu+\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma})^{\frac{d}{2}-\nu-1}z_i^{d-2\nu} & \text{if } \nu \in \left(\frac{d}{2}-1, \frac{d}{2}\right), \\ -\log\left(\sqrt{2\nu+\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}}z_i\right)z_i^2 & \text{if } \nu = \frac{d}{2}-1, \\ \frac{z_i^2}{d-2(\nu+1)} & \text{if } \nu < \frac{d}{2}-1, \end{cases}$$

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{1}{u_i}\middle|\boldsymbol{y}_i\right] \sim \begin{cases} \frac{2\nu+\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}}{2\nu-d-2} & \text{if } \nu > \frac{d}{2}+1, \\ -(2\nu+\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma})\log\left(\sqrt{2\nu+\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}}z_i\right) & \text{if } \nu = \frac{d}{2}+1, \\ \frac{\Gamma(1-\nu+\frac{d}{2})}{\Gamma(\nu-\frac{d}{2})}2^{1-2\nu+d}(2\nu+\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma})^{\nu-\frac{d}{2}}z_i^{2\nu-d-2} & \text{if } \nu \in (\frac{d}{2}, \frac{d}{2}+1), \\ -\frac{1}{\log\left(\sqrt{2\nu+\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}}z_i\right)z_i^2} & \text{if } \nu = \frac{d}{2}, \\ \frac{d-2\nu}{z_i^2} & \text{if } \nu < \frac{d}{2}, \end{cases}$$

$$\mathbb{E}_{\boldsymbol{\theta}}[\log u_i|\boldsymbol{y}_i] \sim \begin{cases} \psi(\nu-\frac{d}{2}) - \log\left(\frac{2\nu+\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}}{2}\right) & \text{if } \nu > \frac{d}{2}, \\ \log z_i - \frac{1}{2}\log(2\nu+\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}) & \text{if } \nu = \frac{d}{2}, \\ -\psi(\frac{d}{2}-\nu) - \log 2 + 2\log z_i & \text{if } \nu < \frac{d}{2}. \end{cases}$$

where $z_i = \sqrt{(\boldsymbol{y}_i-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_i-\boldsymbol{\mu})}$. Thus for the case when $\nu \leq \frac{d}{2}$, the main source of numerical problem for the ECM algorithm comes from calculating $\mathbb{E}_{\boldsymbol{\theta}}[\frac{1}{u_i}|\boldsymbol{y}_i]$ since it diverges to infinity at a hyperbolic rate as the estimate for $\boldsymbol{\mu}$ approaches to one of the data points which is where the maximum of the likelihood function occur. This leads to numerical problems when calculating $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ in (2.10) and (2.12) respectively.

One solution to this problem is to bound the conditional expectations around $\boldsymbol{\mu}$ by a region such that if

$$z_i < \Delta \tag{2.19}$$

where Δ is some small fixed positive constant and z_i is defined in Section 2.1.1, then we compute the conditional expectations in (2.7) to (2.9) by replacing z_i with $z_i^* = \max(z_i, \Delta)$ which helps mitigate numerical problems. Moreover, this method can be

applied to the observed log-likelihood function to avoid the unbounded likelihood. We denote the region in (2.19) to be the capping region and Δ to be the capping level. We perform simulation studies in Section 2.5.1 to assess the performance of the capping approach and choose a suitable value of Δ.

2.4 Observed information matrix

The observed information matrix can be calculated using these three methods:

Method 1: Hessian matrix by direct numerical differentiation,
Method 2: Louis' method, and
Method 3: Fisher information matrix.

We describe each of these methods in more detail and compare their accuracy later using Monte Carlo simulations in Section 2.5.3.

2.4.1 Hessian matrix by numerical differentiation

The Hessian matrix defined as the second order derivative of the observed log-likelihood function in (2.1) can be computed directly by numerical differentiation. This can be implemented using the `hessian` function in the `R` package called `numDeriv` which uses Richardson extrapolation method [93].

2.4.2 Louis' method

Let $\boldsymbol{y}_{\text{com}} = (\boldsymbol{y}, \boldsymbol{u})$ be the complete data, and $\boldsymbol{y}_{\text{obs}} = \boldsymbol{y}$ be the observed data. Then the observed information matrix can be expressed in terms of the conditional expectation of the derivatives of the complete data log-likelihood using Louis' formula in (1.10) which

is given by

$$\begin{aligned}\boldsymbol{I}_{\text{obs}}(\boldsymbol{\theta};\boldsymbol{y}_{\text{obs}}) &= -\mathbb{E}_{\boldsymbol{\theta}}[\ell''(\boldsymbol{\theta};\boldsymbol{y}_{\text{com}})|\boldsymbol{y}_{\text{obs}}] - \text{cov}[\ell'(\boldsymbol{\theta};\boldsymbol{y}_{\text{com}})|\boldsymbol{y}_{\text{obs}}] \\ &= \mathbb{E}_{\boldsymbol{\theta}}\left[-\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\ell_{\text{com}}(\boldsymbol{\theta};\boldsymbol{y}_{\text{com}})\middle|\boldsymbol{y}_{\text{obs}}\right] \\ &\quad - \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\ell_{\text{com}}(\boldsymbol{\theta};\boldsymbol{y}_{\text{com}})\,\frac{\partial}{\partial\boldsymbol{\theta}^\top}\ell_{\text{com}}(\boldsymbol{\theta};\boldsymbol{y}_{\text{com}})\middle|\boldsymbol{y}_{\text{obs}}\right] \\ &\quad + \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\ell_{\text{com}}(\boldsymbol{\theta};\boldsymbol{y}_{\text{com}})\middle|\boldsymbol{y}_{\text{obs}}\right]\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\ell_{\text{com}}(\boldsymbol{\theta};\boldsymbol{y}_{\text{com}})\middle|\boldsymbol{y}_{\text{obs}}\right]^\top \end{aligned} \tag{2.20}$$

where the first order and second order derivatives of the complete data log-likelihood of the VG distribution are given in Appendix A8.

Calculating the second term in (2.20) directly is not straight forward since it requires taking expectation of the product of two summations. This calculation can be simplified by representing the summations of the first order derivatives in terms of matrices in Appendix B2 and using the mutual independence of the $\boldsymbol{u}_i$'s to simplify the missing information matrix in Section B3. This matrix representation allows the second and third term to be easily calculated using (B.14) and (B.15) respectively.

Since the conditional distribution of u_i given $\boldsymbol{y}_i$ follows $\mathcal{GIG}(\nu - \frac{d}{2}, z_i^2, 2\nu + \boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma})$, the conditional expectations is given by

$$\mathbb{E}_{\boldsymbol{\theta}}[u_i^m|\boldsymbol{y}_i] = \eta_i^m \frac{K_{\lambda+m}(\omega_i)}{K_\lambda(\omega_i)}, \tag{2.21}$$

$$\mathbb{E}_{\boldsymbol{\theta}}[u_i^m \log u_i|\boldsymbol{y}_i] = \eta_i^m \frac{K_{\lambda+m}(\omega_i)\log\eta_i + K_{\lambda+m}^{(1,0)}(\omega_i)}{K_\lambda(\omega_i)}, \tag{2.22}$$

$$\mathbb{E}_{\boldsymbol{\theta}}\left[(\log u_i)^2|\boldsymbol{y}_i\right] = (\log\eta_i)^2 + \frac{2K_\lambda^{(1,0)}(\omega_i)\log\eta_i + K_\lambda^{(2,0)}(\omega_i)}{K_\lambda(\omega_i)} \tag{2.23}$$

where $\lambda = \nu - d/2$, $\eta_i = z_i/\sqrt{2\nu + \boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}}$, $\omega_i = z_i\sqrt{2\nu + \boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}}$, and $K_\lambda^{(2,0)}(z) = \frac{\partial^2}{\partial\alpha^2}K_\alpha(z)\big|_{\alpha=\lambda}$ which is approximated using second order approximation

$$K_\lambda^{(2,0)}(z) \approx \frac{K_{\lambda+h}(z) - 2K_\lambda(z) + K_{\lambda-h}(z)}{h^2} \tag{2.24}$$

and setting $h = 10^{-5}$.

Since the expectations in (2.21) to (2.23) have the same numerical problem as in Section 2.3, we bound these conditional expectations using the same capping region as in (2.19).

2.4.3 Fisher information matrix

The Fisher information matrix of the VG distribution can be obtained by integrating the observed information matrix in (2.20) with respect to $\boldsymbol{y}_i$ over $\mathbb{R}^d$ which is evaluated in Appendix B. The first and second term of (2.20) can be simplified by swapping the order of integration using Lemma B1.1, then equating the higher order moments of the conditional normal distribution given the missing variables in B4. This procedure is then applied to each block of the matrix for the first and second term in Appendix B6.1 and B6.2 respectively.

The third term of (2.20) is the most challenging as order of integration cannot be interchanged, and so we are required to integrate over $\mathbb{R}^d$ with respect to $\boldsymbol{y}_i$. However, the integral can be partitioned into its spherical and radial parts using spherical coordinates. The spherical integral consists of spherical moments of the VG distribution which can be derived exactly using Theorem B5.1 and matrix derivative results in Appendix A. What remains is the integral of the radial part which can be evaluated numerically using the `integrate` function in `R`. This construction essentially reduces the dimension of the integral evaluated on $\mathbb{R}^+$ instead of $\mathbb{R}^d$ which is much more feasible to compute.

The formulas for the first, second and third term are given in Sections B6.1, B6.2 and B6.3 respectively. Combining these terms together gives us Fisher information matrix for the VG distribution.

2.4.4 Singularity of the information matrix

The asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$ can be approximated by the inverse of the observed information matrix $\boldsymbol{I}_{\text{obs}}(\hat{\boldsymbol{\theta}})$. This gives us a way to approximate the SE of $\hat{\theta}_i = (\hat{\boldsymbol{\theta}})_i$ by calculating

$$SE(\hat{\theta}_i) \approx \sqrt{\left[\boldsymbol{I}_{\text{obs}}(\hat{\boldsymbol{\theta}}; \boldsymbol{y}_{\text{obs}})^{-1}\right]_{ii}}. \tag{2.25}$$

However, the observed information with respect to $\boldsymbol{\mu}$ is not well-defined for $\nu < \frac{d}{2}$ due to the unbounded likelihood. This issue has been discussed by Kawai [57] for the univariate VG distribution where he showed that for $\nu < 1/2$,

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\left(\frac{\partial}{\partial \mu} \log f(Y; \boldsymbol{\theta})\right)^2\right] = \infty \tag{2.26}$$

where Y is an univariate VG random variable with density function f, and the expectation is taken with respect to Y which depends on $\boldsymbol{\theta}$. Thus for the unbounded density case, we omit the location parameter in the information matrix and SE calculation.

2.5 Simulation studies

2.5.1 Comparing EM algorithms

To assess the performance of our proposed algorithms, we compare the accuracy and computational efficiency of the MCECM, ECME and AECM algorithms for two different choices of Δ:

(i) $\Delta =$ `sqrt(.Machine$double.xmin)` $\approx$ `1.5e-154` where `double.xmin` represents the smallest non-zero normalised floating-point number in `R`.

(ii) $\Delta =$ `sqrt(.Machine$double.eps)` $\approx$ `1.5e-8` where `double.eps` represents the smallest positive floating-point number `x` such that `1+x`$\neq$`1` in `R`.

The procedure for the simulation study is described below:

Step 1: We set the dimension d to be one of the values from 1 to 5. For each dimension, we choose some parameter value for $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\gamma}$. For example, the true values are $\boldsymbol{\mu} = (0, 0)$, $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & -0.4 \\ -0.4 & 1 \end{pmatrix}$ and $\boldsymbol{\gamma} = (-1.2, -0.2)$ when $d = 2$.

Step 2: For each dimension, we set the shape parameter ν to be either one of the smaller values $\{0.01, 0.02, 0.03, 0.04\}$ or regular values $\{0.05, 0.1, \ldots, 1.95, 2\}$.

Step 3: For each pair of (d, ν), we generate $M = 200$ different sets of sample each from VG distribution with dimension d, shape parameter ν, and sample size $n = 2000$.

We present the accuracy of each parameter by reporting the median of the sum of the absolute errors (SAE) over all elements in a vector or lower triangular matrix. We also present the median computation time and number of iterations required for the convergence of the EM algorithms. The results are tabulated in Table 2.1 and 2.2 when the shape parameters are $\nu = 0.5$ and $\nu = 0.04$ respectively for $d = 2$.

From Table 2.1 and Table 2.2, each of the EM algorithms gives fairly similar results for the two levels of ν. Generally, the AECM algorithm requires less number of iterations and computation time while it can still give reasonably accurate estimates. However,

Table 2.1. Median of SAE, computation time, and number of iterations for each ECM algorithm when applied to simulated VG samples with $d = 2$ and $\nu = 0.5$.

Capping level	$\Delta \approx$ `1.5e-154`			$\Delta \approx$ `1.5e-8`		
Algorithm	MCECM	ECME	AECM	MCECM	ECME	AECM
$SAE(\hat{\boldsymbol{\mu}})$	4.2e-3	3.4e-3	3.4e-3	4.2e-3	3.4e-3	3.4e-3
$SAE(\hat{\boldsymbol{\Sigma}})$	0.15	0.17	0.17	0.12	0.11	0.11
$SAE(\hat{\boldsymbol{\gamma}})$	0.05	0.05	0.05	0.05	0.05	0.05
$\hat{\nu}$	0.39	0.39	0.39	0.50	0.50	0.50
Time (sec)	1.5	1.4	1.6	0.4	0.7	0.5
Iterations	62	41	33	43	43	17

Table 2.2. Median of SAE, computation time, and number of iterations for each ECM algorithm when applied to simulated VG samples with $d = 2$ and $\nu = 0.04$.

Capping level	$\Delta \approx$ `1.5e-154`			$\Delta \approx$ `1.5e-8`		
Algorithm	MCECM	ECME	AECM	MCECM	ECME	AECM
$SAE(\hat{\boldsymbol{\mu}})$	1.4e-38	2.5e-38	1.6e-38	6.2e-11	6.2e-11	6.5e-11
$SAE(\hat{\boldsymbol{\Sigma}})$	0.22	0.23	0.21	2.20	2.20	2.21
$SAE(\hat{\boldsymbol{\gamma}})$	0.09	0.10	0.10	1.11	1.11	1.12
$\hat{\nu}$	0.040	0.040	0.040	0.047	0.047	0.046
Time (sec)	8.8	13.8	3.0	3.8	6.8	1.2
Iterations	274	259	65	314	318	20

as the trade-off, the computational time of the ECME algorithm is higher than the MCECM algorithm as each iteration requires more numerical computation. When comparing the performance of the capping levels, we see that $\Delta \approx$ `1.5e-154` performs better for $\nu = 0.04$, while $\Delta \approx$ `1.5e-8` performs better for $\nu = 0.5$.

In summary, the AECM algorithm performs better than the MCECM and ECME algorithms in terms of accuracy and computational efficiency. Additionally, smaller Δ performs better for smaller ν which suggests that choosing suitable Δ can improve accuracy for different ν.

Algorithm 9: AECM algorithm for VG distribution with adaptive Δ

Input: Initial value $\boldsymbol{\theta}^{(0)}$, and $\Delta^{(0)} = \hat{g}(1, d)$
while $\ell(\boldsymbol{\theta}^{(t+1)}; \boldsymbol{y}) - \ell(\boldsymbol{\theta}^{(t)}; \boldsymbol{y}) > \delta$ **do**
$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}) \leftarrow \mathbb{E}_{\boldsymbol{\theta}^{(t)}}[\ell(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{u})|\boldsymbol{y}]$;
$\boldsymbol{\theta}^{(t+1/3)} \leftarrow \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}_1}{\operatorname{argmax}}\, Q(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}, \nu^{(t)}; \boldsymbol{\theta}^{(t)})$;
$\boldsymbol{\theta}^{(t+2/3)} \leftarrow \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}_2}{\operatorname{argmax}}\, \ell(\boldsymbol{\mu}^{(t+1/3)}, \frac{1}{\kappa}\boldsymbol{\Sigma}^{(t+1/3)}, \frac{1}{\kappa}\boldsymbol{\gamma}^{(t+1/3)}, \nu^{(t+1/3)}; \boldsymbol{y})$;
$\boldsymbol{\theta}^{(t+1)} \leftarrow \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}_3}{\operatorname{argmax}}\, \ell(\boldsymbol{\mu}^{(t+2/3)}, \boldsymbol{\Sigma}^{(t+2/3)}, \boldsymbol{\gamma}^{(t+2/3)}, \nu; \boldsymbol{y})$;
$\Delta^{(t+1)} \leftarrow \hat{g}(\nu^{(t+1)}, d)$;
end

2.5.2 Optimal choice of capping level

To determine the optimal capping level for a wide range of shape parameters, we perform the following simulation study:

Step 1: Choose ν out of $\{0.02, 0.04, \dots, 1.18, 1.2\}$, and d out of $\{1, \dots, 30\}$.

Step 2: Apply the `R` function `optimise` to find the optimal Δ such that it minimises $f(\Delta; \nu, d) = \sum_{k=1}^{r} \left|\log \hat{\nu}_{k,d}^{\Delta} - \log \nu\right|$ where for each $k = 1, \dots, r$ (where we set $r = 50$), we simulate from standard VG distribution ($\boldsymbol{\mu} = \boldsymbol{0}, \boldsymbol{\Sigma} = \boldsymbol{I}_d, \boldsymbol{\gamma} = \boldsymbol{0}$) with chosen ν and sample size $n = 2000$ and estimate $\hat{\nu}_{k,d}^{\Delta}$ by maximising the observed likelihood (with capping level Δ) with respect to ν while fixing all other parameters.

Step 3: Repeat step 2 to obtain 200 optimal Δ estimates for each ν and d.

The results depicted in Figure 2.1 shows that as the shape parameter decreases, the median of the optimal Δ decreases and the variability of the optimal Δ increases. As we increase the dimension, the median of the optimal Δ slightly increases. This optimal Δ can be applied in the AECM algorithm by first fitting the median of the optimal Δ in Figure 2.1 with a cubic spline represented as $\hat{g}(\nu, d)$, then after the $t \mapsto t+1$ iteration of the AECM algorithm, update $\Delta^{(t+1)} = \hat{g}\big(\nu^{(t+1)}, d\big)$.

Since Δ changes after each iteration, the log-likelihood also changes. Thus the convergence results in Section 1.3.2 does not apply for this algorithm. Nevertheless, as long as the likelihood improves after each iteration, then the AECM algorithm with adaptive Δ in each iteration can still be implemented. We refer this algorithm as the AECM algorithm with adaptive Δ.

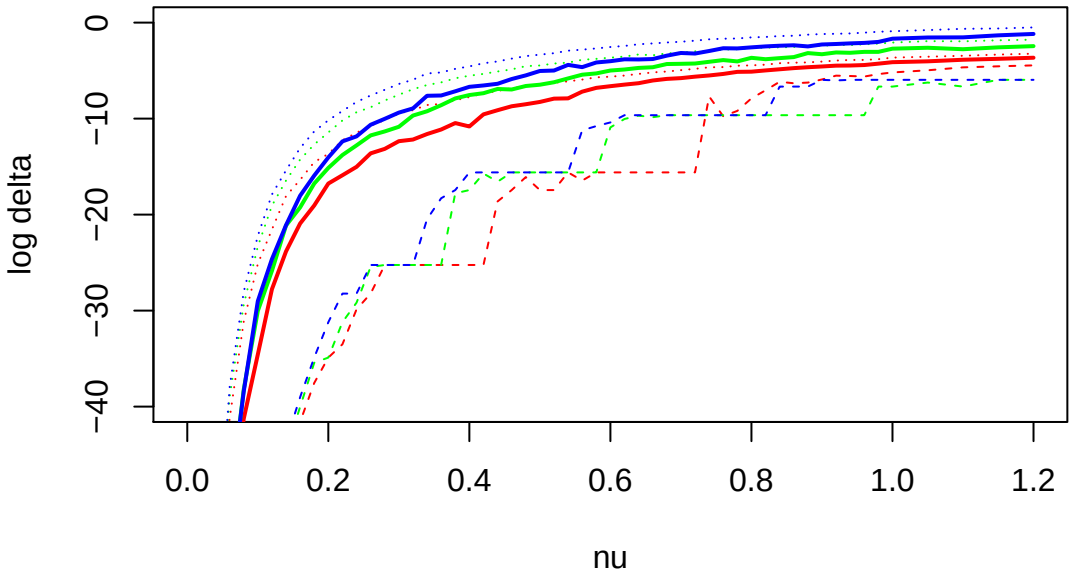


Figure 2.1. plotting the median (thick solid), 95% quantile (dotted) and 5% quantile (dashed) of the optimal $\log \Delta$ estimates for each ν and dimensions 1 (red), 5 (green), and 30 (blue).

2.5.3 Comparing standard error calculations

The aim of this section is to verify the calculation of SE by comparing the estimated SE from simulated data sets with the theoretical SE from Fisher information matrix and the following two methods for calculating SE:

Numerical Hessian method: calculate the Hessian matrix using numerical differentiation evaluated at $\hat{\boldsymbol{\theta}}$ in Section 2.4.1.

Louis' method: calculate the complete and missing information matrices evaluated at $\hat{\boldsymbol{\theta}}$ using the formulas (2.20) in Section 2.4.2. See Appendix A8 for the derivatives.

We calculate the theoretical SE based on the Fisher information matrix evaluated at the true parameter values in Section 2.4.3. See Appendix B6 for the calculation of the Fisher information matrix for the VG distribution.

For this simulation study, the true parameter values are chosen to be

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}, \quad \boldsymbol{\gamma} = \begin{pmatrix} 0.8 \\ 1 \end{pmatrix}. \tag{2.27}$$

Then the procedure of the simulation is as follows:

Step 1: Choose ν out of $\{0.7, 1.2, 1.7, 3, 5\}$.

Step 2: Sample $n = 1000$ data points from the VG distribution with parameters in (2.27) and ν chosen in step 1.

Step 3: Apply the AECM algorithm with adaptive capping level to obtain parameter estimates for the VG distribution.

Step 4: Use the parameter estimates to calculate the SEs using numerical Hessian and Louis' methods.

Step 5: For each ν, repeat steps 2 and 4 to get 500 different SEs.

The median of the SE estimates based on simulation along with the SEs from Louis' method, numerical Hessian method and Fisher information matrix are displayed in Table 2.3. The first column labelled "Simulated" is the standard derivation of estimates over $r = 500$ replications. The last column labelled "Fisher" is calculated using the formulas in Appendix B6. Since the information corresponding to $\hat{\boldsymbol{\mu}}$ is not well-defined when $\nu < 1$, we write `NA`. For each ν, the SE estimates based on simulation is consistent with the SEs from numerical Hessian and Louis' methods. The SE from the Fisher information matrix evaluated at the true parameters is consistent with the other SEs for each ν except for $\nu = 5$. This slight inconsistency possibly suggests that the performance of the algorithm can be improved for larger ν. Note that many authors such as in [46, 104] do not provide simulation results to confirm the consistency of the SE estimates since they do not account for the correction factor for derivatives involving $\boldsymbol{\Sigma}$ which is discussed in Section A7.1.

In conclusion, the numerical Hessian and Louis' methods both provide accurate SE estimates for each parameter. While both methods use second order numerical differentiation for $K_\lambda(z)$ such as (2.24), Louis' method is often more numerically stable as the differentiation is evaluated to each term of the log-likelihood of the conditional normal and gamma distribution which has closed-form expression, whereas for numerical Hessian method, it was applied to the observed log-likelihood directly. The SE from

Table 2.3. Median SE estimates based on various SE methods for comparison.

true ν	SE	Simulated	Louis	Hessian	Fisher
$\nu = 0.7$	$SE(\hat{\boldsymbol{\mu}}')$	$\begin{pmatrix}0.014 & 0.014\end{pmatrix}$	NA	NA	NA
	$SE(\hat{\boldsymbol{\Sigma}})$	$\begin{pmatrix}0.08 & 0.07\\ & 0.8\end{pmatrix}$	$\begin{pmatrix}0.08 & 0.07\\ & 0.07\end{pmatrix}$	$\begin{pmatrix}0.08 & 0.07\\ & 0.07\end{pmatrix}$	$\begin{pmatrix}0.07 & 0.07\\ & 0.08\end{pmatrix}$
	$SE(\hat{\boldsymbol{\gamma}}')$	$\begin{pmatrix}0.05 & 0.05\end{pmatrix}$	$\begin{pmatrix}0.05 & 0.04\end{pmatrix}$	$\begin{pmatrix}0.05 & 0.04\end{pmatrix}$	$\begin{pmatrix}0.04 & 0.05\end{pmatrix}$
	$SE(\hat{\nu})$	0.035	0.035	0.035	0.035
$\nu = 1.2$	$SE(\hat{\boldsymbol{\mu}}')$	$\begin{pmatrix}0.04 & 0.04\end{pmatrix}$	$\begin{pmatrix}0.05 & 0.05\end{pmatrix}$	$\begin{pmatrix}0.05 & 0.05\end{pmatrix}$	$\begin{pmatrix}0.03 & 0.03\end{pmatrix}$
	$SE(\hat{\boldsymbol{\Sigma}})$	$\begin{pmatrix}0.07 & 0.06\\ & 0.07\end{pmatrix}$	$\begin{pmatrix}0.08 & 0.07\\ & 0.07\end{pmatrix}$	$\begin{pmatrix}0.08 & 0.07\\ & 0.07\end{pmatrix}$	$\begin{pmatrix}0.07 & 0.06\\ & 0.07\end{pmatrix}$
	$SE(\hat{\boldsymbol{\gamma}}')$	$\begin{pmatrix}0.06 & 0.05\end{pmatrix}$	$\begin{pmatrix}0.06 & 0.06\end{pmatrix}$	$\begin{pmatrix}0.06 & 0.06\end{pmatrix}$	$\begin{pmatrix}0.05 & 0.06\end{pmatrix}$
	$SE(\hat{\nu})$	0.09	0.09	0.09	0.09
$\nu = 1.7$	$SE(\hat{\boldsymbol{\mu}}')$	$\begin{pmatrix}0.07 & 0.06\end{pmatrix}$	$\begin{pmatrix}0.07 & 0.06\end{pmatrix}$	$\begin{pmatrix}0.07 & 0.06\end{pmatrix}$	$\begin{pmatrix}0.06 & 0.06\end{pmatrix}$
	$SE(\hat{\boldsymbol{\Sigma}})$	$\begin{pmatrix}0.08 & 0.06\\ & 0.07\end{pmatrix}$	$\begin{pmatrix}0.08 & 0.07\\ & 0.07\end{pmatrix}$	$\begin{pmatrix}0.08 & 0.07\\ & 0.07\end{pmatrix}$	$\begin{pmatrix}0.07 & 0.07\\ & 0.07\end{pmatrix}$
	$SE(\hat{\boldsymbol{\gamma}}')$	$\begin{pmatrix}0.08 & 0.07\end{pmatrix}$	$\begin{pmatrix}0.08 & 0.07\end{pmatrix}$	$\begin{pmatrix}0.08 & 0.07\end{pmatrix}$	$\begin{pmatrix}0.07 & 0.07\end{pmatrix}$
	$SE(\hat{\nu})$	0.18	0.17	0.17	0.16
$\nu = 3$	$SE(\hat{\boldsymbol{\mu}}')$	$\begin{pmatrix}0.14 & 0.13\end{pmatrix}$	$\begin{pmatrix}0.13 & 0.12\end{pmatrix}$	$\begin{pmatrix}0.13 & 0.12\end{pmatrix}$	$\begin{pmatrix}0.12 & 0.13\end{pmatrix}$
	$SE(\hat{\boldsymbol{\Sigma}})$	$\begin{pmatrix}0.08 & 0.07\\ & 0.07\end{pmatrix}$	$\begin{pmatrix}0.08 & 0.07\\ & 0.07\end{pmatrix}$	$\begin{pmatrix}0.08 & 0.07\\ & 0.07\end{pmatrix}$	$\begin{pmatrix}0.07 & 0.06\\ & 0.07\end{pmatrix}$
	$SE(\hat{\boldsymbol{\gamma}}')$	$\begin{pmatrix}0.15 & 0.13\end{pmatrix}$	$\begin{pmatrix}0.14 & 0.12\end{pmatrix}$	$\begin{pmatrix}0.14 & 0.12\end{pmatrix}$	$\begin{pmatrix}0.12 & 0.13\end{pmatrix}$
	$SE(\hat{\nu})$	0.51	0.49	0.49	0.46
$\nu = 5$	$SE(\hat{\boldsymbol{\mu}}')$	$\begin{pmatrix}0.26 & 0.22\end{pmatrix}$	$\begin{pmatrix}0.26 & 0.23\end{pmatrix}$	$\begin{pmatrix}0.26 & 0.23\end{pmatrix}$	$\begin{pmatrix}0.20 & 0.22\end{pmatrix}$
	$SE(\hat{\boldsymbol{\Sigma}})$	$\begin{pmatrix}0.08 & 0.07\\ & 0.07\end{pmatrix}$	$\begin{pmatrix}0.08 & 0.07\\ & 0.07\end{pmatrix}$	$\begin{pmatrix}0.08 & 0.07\\ & 0.07\end{pmatrix}$	$\begin{pmatrix}0.07 & 0.06\\ & 0.07\end{pmatrix}$
	$SE(\hat{\boldsymbol{\gamma}}')$	$\begin{pmatrix}0.26 & 0.23\end{pmatrix}$	$\begin{pmatrix}0.26 & 0.23\end{pmatrix}$	$\begin{pmatrix}0.26 & 0.23\end{pmatrix}$	$\begin{pmatrix}0.20 & 0.22\end{pmatrix}$
	$SE(\hat{\nu})$	1.49	1.44	1.44	1.17

the Fisher information matrix evaluated at true parameters are consistent with the simulated results, and also indicate that the AECM algorithm with adaptive capping level performs well for smaller ν. Moreover, these results verify the matrix derivatives in Appendix A8 and multidimensional integration results in Appendix B6 used to calculate the observed information matrix from (2.20) and the Fisher information matrix respectively. The formulas for Louis' and Fisher's methods can also be used to calculate the SE of other distributions with NMVM representation such as the multivariate Student's t and GH distributions.

Table 2.4. Summary statistics for DAX, S&P 500, FTSE 100, AORD and CAC 40 daily return series.

Indices	Mean	SD	Skewness	Kurtosis	Correlation matrix
DAX	2.9e-4	0.015	0.02	9.5	$\begin{pmatrix} 1 & 0.64 & 0.87 & 0.37 & 0.36 \\ & 1 & 0.61 & 0.15 & 0.63 \\ & & 1 & 0.41 & 0.31 \\ & & & 1 & -0.01 \\ & & & & 1 \end{pmatrix}$
S&P 500	1.1e-4	0.014	−0.29	12.9	
FTSE 100	1.2e-4	0.013	−0.09	11.0	
AORD	1.6e-4	0.011	−0.73	10.5	
CAC 40	1.8e-4	0.029	0.15	11.7	

2.6 Application

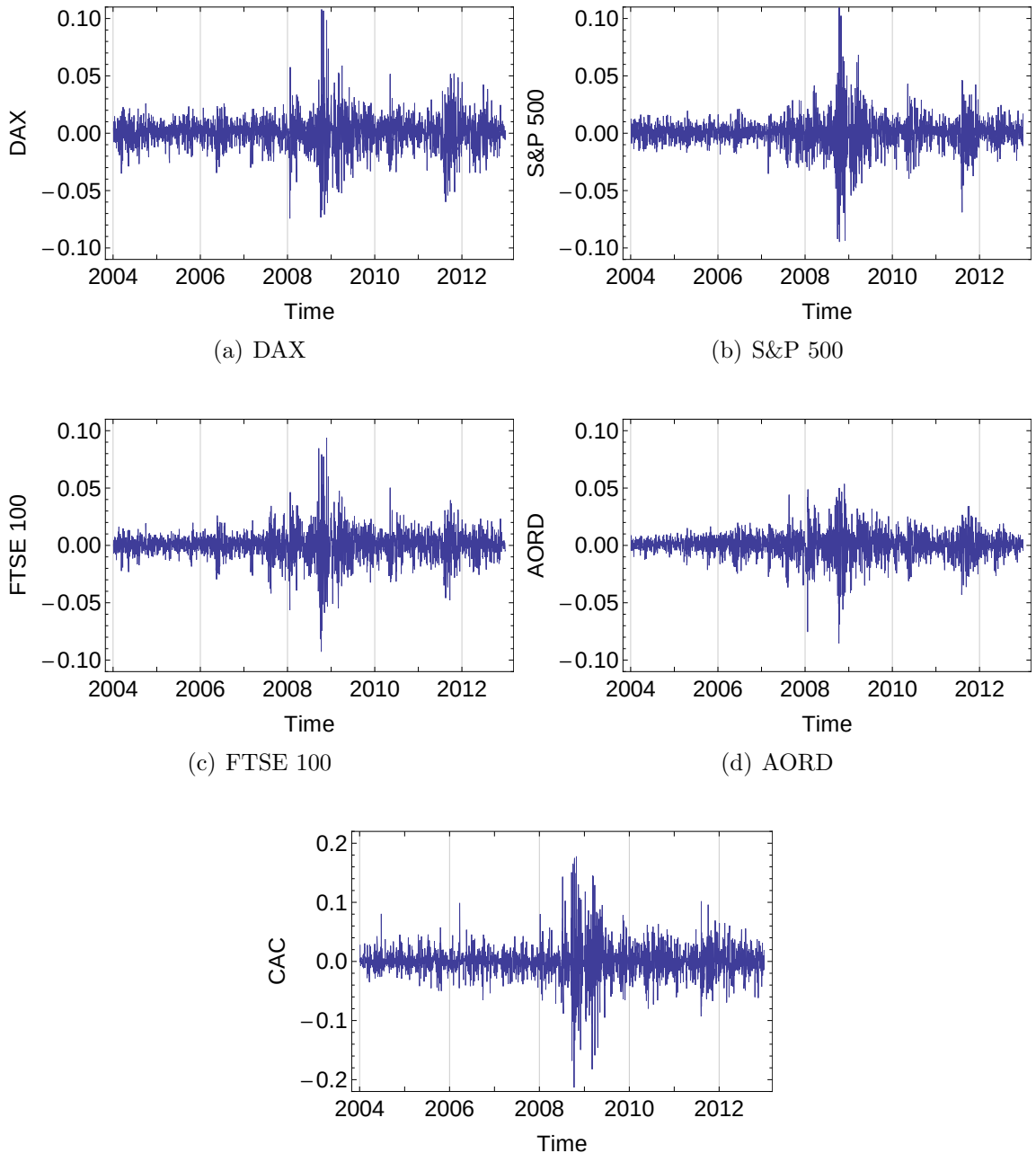
To illustrate the applicability of the AECM algorithm using VG distribution, we consider the returns of the five daily closing price indices, namely, Deutscher Aktien (DAX), Standard & Poors 500 (S&P 500), Financial Times Stock Exchange 100 (FTSE 100), All Ordinaries (AORD) and Cotation Assistée en Continu 40 (CAC 40) from 1$^{\text{st}}$ January 2004 to 31$^{\text{st}}$ December 2012. The return of market indices is defined as

$$r_t = \log(p_t) - \log(p_{t-1}) \tag{2.28}$$

for $t = 2, 3, \ldots$ where p_t refers to the closing price at time t. After filtering the data with missing closing prices, we obtain the data size of $n = 2188$. Plots of the five time series are given in Figure 2.2. They all show low autocorrelation and high volatility during the financial crisis in 2008. As the summary statistics in Table 2.4 show that the data exhibit considerable skewness and kurtosis, we begin our analysis with the VG distribution to capture the skewness and kurtosis.

The results for the estimated parameters and their SEs using Louis' method are given in Table 2.5 as well as the estimated correlation matrix $\boldsymbol{\rho}$ based on the estimated covariance of $\boldsymbol{Y}$ given by $\hat{\boldsymbol{\Sigma}} + \frac{1}{\hat{\nu}}\hat{\boldsymbol{\gamma}}\hat{\boldsymbol{\gamma}}'$.

Not surprisingly, the scale estimate of $\boldsymbol{\Sigma}$ for CAC 40 is the largest as it has the largest sample standard derivation. Moreover, the positive skewness estimate is also in agreement the sample skewness. After allowing for the skewness, the location estimate of CAC 40 is lower compared with other indices. Regarding the correlation based on the model, the pair of DAX and FTSE 100 has the strongest whereas AORD and CAC 40 has the lowest. This seems to agree with the geographical locations for these indices.



(a) DAX

(b) S&P 500

(c) FTSE 100

(d) AORD

(e) CAC 40

Figure 2.2. Time series plots for the five daily return series

Table 2.5. Parameter estimates and its SEs using Louis' method of the VG model using DAX, S&P 500, FTSE 100, AORD and CAC 40 daily return series.

	Estimates	Standard errors
$\boldsymbol{\mu}'$	$10^{-4}\begin{pmatrix}18.3 & 9.6 & 12.1 & 20.1 & -7.4\end{pmatrix}$	NA
$\boldsymbol{\Sigma}$	$10^{-5}\begin{pmatrix}18.8 & 9.7 & 13.5 & 4.5 & 11.7\\ & 13.9 & 7.9 & 1.6 & 17.3\\ & & 13.3 & 4.2 & 8.8\\ & & & 11.7 & 0.5\\ & & & & 65.9\end{pmatrix}$	$10^{-6}\begin{pmatrix}7.4 & 4.9 & 5.6 & 3.8 & 9.0\\ & 5.4 & 4.0 & 3.1 & 8.8\\ & & 5.2 & 3.3 & 7.5\\ & & & 4.7 & 6.6\\ & & & & 26.0\end{pmatrix}$
$\boldsymbol{\gamma}'$	$10^{-4}\begin{pmatrix}-15.4 & -8.5 & -10.9 & -18.6 & 9.1\end{pmatrix}$	$10^{-4}\begin{pmatrix}4.5 & 3.7 & 3.8 & 3.6 & 8.4\end{pmatrix}$
ν	1.40	0.054
$\boldsymbol{\rho}$	$\begin{pmatrix}1 & 0.60 & 0.86 & 0.31 & 0.33\\ & 1 & 0.58 & 0.13 & 0.57\\ & & 1 & 0.35 & 0.29\\ & & & 1 & 0.01\\ & & & & 1\end{pmatrix}$	

In summary, our proposed AECM algorithm can be applied to fit the VG distribution and the SE can be calculated using Louis' method in Section 2.4. We note that the SE for $\hat{\boldsymbol{\mu}}$ is not provided when $\hat{\nu} < d/2$ since the information is not well-defined from (2.26) which is the case for this analysis. This motivating analysis illustrates the need to consider LOO and WLOO likelihoods in Chapters 3 and 4 to improve the parameter estimation and SE approximation for $\hat{\boldsymbol{\mu}}$ when the likelihood becomes unbounded.

2.7 Conclusion

We proposed various extensions to the ECM algorithm to estimate parameters of the VG distribution. We improve the efficiency and stability of the ECM algorithm by implementing the AECM algorithm. This algorithm with the capped likelihood method can also deal with the unbounded density of the VG distribution when $\nu < d/2$ which may arise when fitting it to high frequency data with high kurtosis. Further details on fitting high frequency data is explored in Section 5.6.

The challenge from unbounded density is that it gives numerically unstable conditional expectations in the E-step when the location parameter tends towards an observation.

We resolved the problem by imposing a bound as in (2.19). From the simulation studies, the effect of bounding the conditional expectations allows for more numerically stable parameter estimates and AECM algorithm with adaptive capping level also performs better than MCECM and ECME algorithms in terms of accuracy and computational efficiency. We also studied the optimal choices of Δ for dimensions $d = 1, \ldots, 5$ using the AECM algorithm. We propose the adaptive Δ method to update Δ after each iteration. The third simulation study also confirms the accuracy of the SE calculation using both numerical Hessian and Louis' methods when comparing to the estimates based entirely on simulation as well as the theoretical Fisher information matrix using true values.

However, despite the good performance of the AECM algorithm and SE calculation, there are some limitations. Both numerical Hessian and Louis' methods fail to provide SE estimates for the location parameter when $\nu < d/2$ since the likelihood function is unbounded and so the information matrix is not well-defined from (2.26). Moreover, the choice of Δ may subject to debate and the optimal Δ needs to be estimated using simulations such as in Section 2.5.2. In the next chapter, we explore the properties of the LOO likelihood method as an alternative way to deal with the unbounded likelihood and numerically investigate the distribution of the location estimator using LOO likelihood which can be applied to calculate the SE of location estimates.

CHAPTER 3

Estimation using Leave-one-out Likelihood

3.1 Introduction

In Chapter 2, we propose a method by choosing the optimal capping level to bound the density in order to avoid the unbounded likelihood. A major drawback to this method is that simulations are required to estimate the optimal capping level. Furthermore, the optimal capping level can change for different dimensions and different distributions. In this chapter, we consider the leave-one-out (LOO) likelihood to leave out the data point that causes the likelihood to become unbounded. This construction removes the dependency of an arbitrary capping level which is a desirable property.

The main objective of this chapter is three-folded. Our first objective is to extend the definition of the LOO likelihood in [89] to accommodate for multivariate data sets while also dealing with the unbounded likelihood. Our second objective is to propose an AECM algorithm to obtain the maximum LOO estimates for the parameters of the VG distributions when densities are cusped or unbounded with respect to the location parameter. We also remark that our methodology is general enough to apply to other distributions with NMVM representation including the Student's t and GH distribution. Our third objective is to analyse the asymptotic behaviour including the optimal convergence rate and asymptotic distribution of the maximum LOO likelihood estimator for the location parameter through simulation studies using data simulated from the VG distribution with different samples sizes and shape parameters.

The remaining chapter is structured as follows. Section 3.2 formulates the maximum LOO likelihood framework for parameter estimation of multivariate distributions with

unbounded densities with respect to the location parameter and states some properties of the estimator. Section 3.3 introduces the AECM algorithm using the LOO likelihood to estimate parameters of the VG distribution. Section 3.4 presents two simulation studies. The first study assess the accuracy of our estimator while the second study analyse the asymptotic behaviour of the maximum LOO likelihood estimator for the location parameter of the VG distribution. Lastly, Section 3.5 concludes the chapter with some remarks.

3.2 Maximum leave-one-out likelihood

Let $\boldsymbol{y} = (\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n)$ be observed data from the VG distribution with corresponding mixing variables $\boldsymbol{u} = (u_1, \cdots, u_n)$, and $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}, \nu)$ be parameters of the VG distribution in the parameter space $\boldsymbol{\Theta}$. The density of the VG distribution is unbounded at $\boldsymbol{\mu}$ when $\nu \leq \frac{d}{2}$. Consequently, the MLE is not well-defined since there are multiple unbounded points at each data point in the likelihood function. Kawai [57] has shown that for the univariate case, and the Fisher information matrix with respect to μ is also not well-defined which was briefly discussed in Section 2.4.4.

3.2.1 Leave-one-out likelihood

The classical likelihood function needs to be modified so that the maximum is well-defined even with the unbounded likelihoods. Podgórski and Wallin [89] proposed the observed leave-one-out (LOO) likelihood function defined as

$$L^{\text{LOO}}(\boldsymbol{\theta}; \boldsymbol{y}) = \prod_{i \neq k(\boldsymbol{\mu})} f(\boldsymbol{y}_i; \boldsymbol{\theta}) \tag{3.1}$$

for some density function f where the LOO index is defined as

$$k(\boldsymbol{\mu}) = \underset{k \in \{1,\ldots,n\}}{\operatorname{argmin}} (\boldsymbol{y}_k - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}_k - \boldsymbol{\mu}). \tag{3.2}$$

Note that we slightly modify the convention in Podgórski and Wallin [89]: *"if there are two indices we take the one for which corresponding $y_{k(\mu)}$ is on the right side of μ"* as it only deals with the univariate case and cannot be easily extended to the multivariate setting.

We remark that when considering asymmetric distributions, the LOO likelihood function is discontinuous. For the VG distribution with skewness, the discontinuity is not an issue since the density is asymptotically symmetric as $\boldsymbol{y} \to \boldsymbol{\mu}$ from (1.39), and so the effect of the discontinuities is minimised for larger sample size. On the other hand, when using other distributions with different skewness behaviour, the LOO index can alternatively be defined as

$$k(\boldsymbol{\mu}) = \underset{k\in\{1,\dots,n\}}{\operatorname{argmax}} f(\boldsymbol{y}_k; \boldsymbol{\theta}). \tag{3.3}$$

In this thesis, we simply adopt the LOO index in (3.2).

Let the observed LOO log-likelihood function be defined as

$$\ell^{\mathrm{LOO}}(\boldsymbol{\theta}; \boldsymbol{y}) = \log L^{\mathrm{LOO}}(\boldsymbol{\theta}; \boldsymbol{y}) \tag{3.4}$$

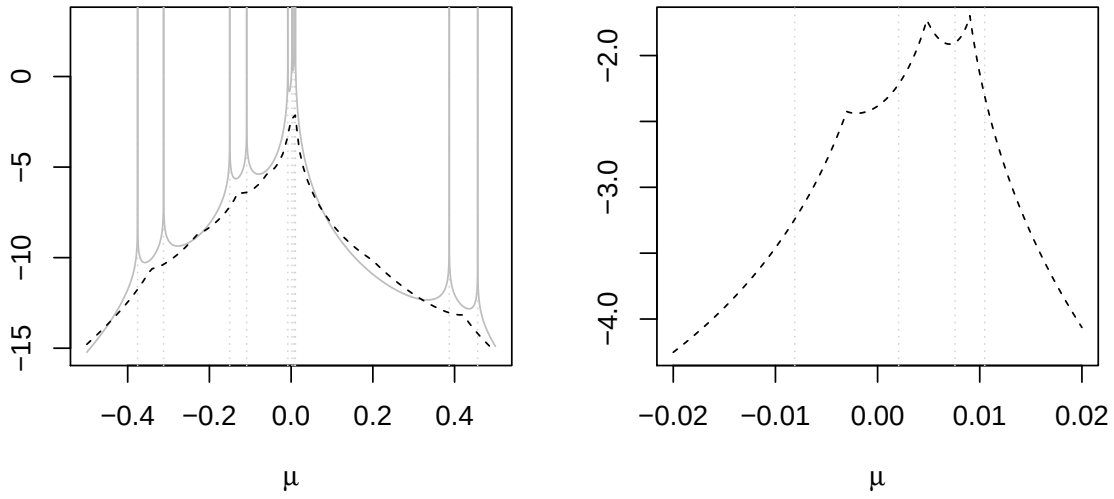
and the maximum LOO likelihood estimator which maximises the LOO likelihood function with respect to $\boldsymbol{\theta}$ be denoted as $\hat{\boldsymbol{\theta}}_n$.

The unbounded density problem is illustrated with a data of 10 observations simulated from the standard VG distribution ($\mu = 0$, $\sigma = 1$, $\gamma = 0$) with shape parameter $\nu = 0.2$. In Figure 3.1, we plot both the full (or classical) log-likelihood function along with the LOO log-likelihood function with respect to the location parameter. We see that leaving the data point out essentially removes the unbounded points of the log-likelihood function so that the maximum can be well-defined. Additionally, if we zoom in at around $\mu = 0$, we observe that non-differentiable points tend to occur between data points. We describe in more detail in Section 3.3.2.3 on how to deal with these non-differentiable points when estimating parameters.

3.2.2 Properties of maximum LOO likelihood estimator

The following proposition shows that the LOO likelihood indeed attains maximum at the midpoints for the cusp or unbounded density cases which was seen from Figure 3.1(b),

Proposition 3.2.1. *Let $\boldsymbol{y} = (y_1, \dots, y_n)$ be univariate symmetric VG random variables, $y_{(1)} < \dots < y_{(n)}$ be ordered values of $\boldsymbol{y}$, and $x_i = \big(y_{(i)} + y_{(i+1)}\big)/2$ for $i = 1, \dots, n-1$. Then the LOO likelihood attains its maximum at one of the $\{x_i\}$ for $\nu < \frac{d}{2} + \frac{1}{2}$.*



(a) log-likelihood comparison

(b) close-up on the LOO log-likelihood

Figure 3.1. Left: Comparing full log-likelihood (solid grey) vs. LOO log-likelihood (dashed black) of simulated data from standardised VG distribution with $\nu = 0.2$ and sample size of ten with vertical dotted grey lines denoting the positions of data points. Right: Close-up of the left figure at around $\mu = 0$ focusing on the LOO log-likelihood.

Proof. The idea of the proof is similar to Hossain et al. [50, Proposition 4.6]. □

For the one-dimensional case, some asymptotic properties of the estimator for the location parameter $\hat{\mu}_n$ such as consistency and super-efficient rate of convergence are proved by Podgórski and Wallin [89]. We state both the assumptions and theorem relating to these asymptotic properties:

Assumptions:

(A1) The pdf $f(y) = p(y)|y|^{\alpha}$ where $\alpha \in (-1, 0)$, p has bounded derivative on $\mathbb{R}\backslash\{0\}$ and, for some $\epsilon > 0$, f is non-zero and continuous either on $[-\epsilon, 0]$ or on $[0, \epsilon]$.

(A2) There exist $\rho > 0$ such that $f(y) = O(|y|^{-\rho-1})$ when $|y| \to \infty$.

(A3) For all $\epsilon > 0$, the incomplete Fisher information is finite. That is,

$$\mathcal{I}_{\epsilon}(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\theta}}\left[\left(\frac{\partial}{\partial \boldsymbol{\theta}} \log f(Y; \boldsymbol{\theta})\right)^2 \middle| |Y| > \epsilon\right] < \infty.$$

Theorem 3.2.2. *Let f satisfies the assumptions (A1) to (A3) and let $\hat{\mu}_n$ be the maximiser of $L^{\mathrm{LOO}}(\mu; \boldsymbol{y})$. Then $\hat{\mu}_n$ is consistent estimator of μ and for any $\beta < 1/(1+\alpha)$,*

$$n^{\beta}(\hat{\mu}_n - \mu) \xrightarrow{p} 0$$

where α is defined in (A1).

Proof. See Podgórski and Wallin [89]. □

This theorem states the lower bound for the rate of convergence $n^{-\beta}$ for the maximum LOO likelihood location estimator. For univariate VG distribution, $\alpha = 2\nu - 1$ from (1.39). Hence setting $\beta = 1/(1+\alpha) = 1/(2\nu)$ possibly gives us the index for the optimal rate of convergence (or the *proposed optimal rate*) for $\nu < 1/2$. Additionally, $n^{\beta}(\hat{\mu}_n - \mu)$ will converge to some asymptotic distribution for some suitable choice of β. We investigate these asymptotic properties in Section 3.4 using simulations from univariate symmetric VG distribution. We remark that currently, there is no multivariate extension of Theorem 3.2.2 and further research is needed to investigate such extension.

3.3 AECM algorithm using LOO likelihood

Directly finding the maximum LOO likelihood estimator $\hat{\boldsymbol{\theta}}_n$ of VG distribution can be difficult as the observed LOO likelihood function has many non-differentiable points when $\nu \leq d/2$, and the LOO index $k(\boldsymbol{\mu})$ makes derivatives tedious to work with since the summation and the differential with respect to $\boldsymbol{\mu}$ can not be interchanged due to the dependency of the summation index on $\boldsymbol{\mu}$ in (3.1). Alternatively, we can implement the AECM algorithm to not only maximise the conditional expectation of the complete data LOO likelihood, but also improve convergence and computational time.

Given the complete data $(\boldsymbol{y}, \boldsymbol{u})$, we can use the NMVM representation in (1.37) to represent the complete data LOO log-likelihood as

$$\ell^{\mathrm{LOO}}(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{u}) = \ell_N^{\mathrm{LOO}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}; \boldsymbol{y}, \boldsymbol{u}) + \ell_G^{\mathrm{LOO}}(\nu; \boldsymbol{u}) \tag{3.5}$$

where the LOO log-likelihood of the conditional normal distribution ignoring additive constants is given by

$$\ell_N^{\mathrm{LOO}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}; \boldsymbol{y}, \boldsymbol{u}) = -\frac{n-1}{2} \log |\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{i \neq k(\boldsymbol{\mu})} \frac{1}{u_i} (\boldsymbol{y}_i - \boldsymbol{\mu} - u_i \boldsymbol{\gamma})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu} - u_i \boldsymbol{\gamma}) \tag{3.6}$$

and the LOO log-likelihood of the gamma distribution is given by

$$\ell_G^{\mathrm{LOO}}(\nu; \boldsymbol{u}) = (n-1)(\nu \log \nu - \log \Gamma(\nu)) + (\nu - 1) \sum_{i \neq k(\boldsymbol{\mu})} \log u_i - \nu \sum_{i \neq k(\boldsymbol{\mu})} u_i. \tag{3.7}$$

We have proposed the AECM algorithm for the VG distribution using the full likelihood in Section 2.2. However, modifications to the algorithm are necessary when using the LOO likelihood. We discuss new techniques to maximise the LOO likelihood while avoiding some numerical issues. We remark that the E-step using the LOO likelihood is the same as with the full likelihood in Section 2.1.1.

3.3.1 E-step

Refer to the E-step in Section 2.1.1 for the conditional expectations. Recall from Section 2.3 that for the unbounded density case, $\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{1}{u_i} \middle| \boldsymbol{y}_i\right]$ diverges to infinity at a hyperbolic rate as $\boldsymbol{\mu} \to \boldsymbol{y}_i$. This leads to numerical problem when the maximum of the likelihood becomes unbounded at the data points. The LOO likelihood avoids this by preventing the location estimate to converge towards the data points as it was shown that the maximum of LOO likelihood tends to be between data points from Figure 3.1(b) and Proposition 3.2.1.

3.3.2 CM-step

We encounter two types of difficulties in calculating the derivative of ℓ_N with respect to $\boldsymbol{\mu}$ for the CM-step.

Firstly, even when the LOO likelihood removes the unbounded points from the full likelihood, there still exist non-differentiable points in the LOO likelihood function. Consequently, we cannot completely rely on derivative based methods to find the maximum of the LOO likelihood with respect to the location parameter $\boldsymbol{\mu}$.

Secondly, given the unobserved data $\boldsymbol{u}$, the first order derivative of the complete data LOO log-likelihood in (3.6) with respect to $\boldsymbol{\mu}$ is

$$\frac{\partial}{\partial \boldsymbol{\mu}} \ell_N^{\mathrm{LOO}} = -\frac{1}{2}\left(\frac{\partial}{\partial \boldsymbol{\mu}} \sum_{i \neq k(\boldsymbol{\mu})} \frac{1}{u_i} (\boldsymbol{y}_i - \boldsymbol{\mu} - u_i \boldsymbol{\gamma})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu} - u_i \boldsymbol{\gamma}) \right). \tag{3.8}$$

Since the summation index depends on $\boldsymbol{\mu}$, the differential and the summation cannot simply be interchanged. Thus the CM-step for $\boldsymbol{\mu}$ does not have a closed-form solution.

To solve these two problems, we propose the *local midpoint search* and *local point search* algorithms for the first problem, and the *approximate derivative* of the complete data LOO log-likelihood for the second problem.

3.3.2.1 Local midpoint search (for one-dimensional case)

As seen in Figure 3.1(b) and Proposition 3.2.1, the maximum of the LOO log-likelihood tends to occur at the non-differentiable points which are located between data points for the one-dimensional case. So ideally we want to search along these midpoints to maximise the LOO log-likelihood with respect to μ. This leads to the local midpoint search. The idea is to search for midpoints around the current iterate $\mu^{(t)}$ and choose the one that maximises the LOO log-likelihood.

Local midpoint search algorithm: Let $(\mu^{(t)}, \Sigma^{(t)}, \gamma^{(t)}, \nu^{(t)})$ be our current estimates, and $y_{(i)}$ be the ordered data. The procedures are:

Step 1: Calculate Euclidean distances $\left|x_i - \mu^{(t)}\right|$ for $i = 1, \dots, n-1$ where $x_i := (y_{(i)} + y_{(i+1)})/2$, choose the least m Euclidean distances with corresponding midpoints $x_{i_1}, \dots, x_{i_m}$ and let $x_{i_0} = \mu^{(t)}$.

Step 2: Update the location estimate by choosing μ out of $\{x_{i_0}, \dots, x_{i_m}\}$ such that it maximises the LOO likelihood in (3.4). That is,

$$\hat{\boldsymbol{\mu}} = \underset{\mu \in \{x_{i_0}, \dots, x_{i_m}\}}{\operatorname{argmax}} \ell^{\mathrm{LOO}}\big(\mu, \Sigma^{(t)}, \gamma^{(t)}, \nu^{(t)}; \boldsymbol{y}\big).$$

Step 3: Repeat steps 1 and 2 until the location estimate converges.

In practice, we search over data points with the least m Euclidean distances from the midpoints and set $m = \max\{20, n/100\}$ in this simulation study given that $n \geq 20$. This

choice of m was defined to balance the computational time and accuracy since small m results in the algorithm being incapable of escaping the local maximum, whereas large m results in slower computational time. Finding out the optimal choice of m requires further research and is not considered in this thesis. Hence we simply take $m = \max\{20, n/100\}$ as an ad hoc choice.

3.3.2.2 Local point search (for higher dimensional case)

In general, finding the maximum in higher dimensions is more computationally demanding. For two-dimensional data, the maximum occurs at the non-differentiable lines which is demonstrated later in Figure 3.3. For d-dimensional data, the maximum occur on the $(d-1)$ dimensional non-differentiable manifolds.

So for simplicity, we propose to search for data points around the current iterate $\hat{\boldsymbol{\mu}}^{(t)}$ and choose the one that increases the LOO log-likelihood.

Local point search algorithm:

The algorithm is similar to the local midpoint search algorithm in Section 3.3.2.1 except we search over the data points instead of the midpoints, and replace the Euclidean distance with the Mahalanobis distance

$$(\boldsymbol{y}_i - \boldsymbol{\mu}^{(t)})'(\boldsymbol{\Sigma}^{(t)})^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}^{(t)})$$

for $i = 1, \ldots, n$. For the rest of this thesis, we simply refer to these two algorithms as the *local point search* (LPS) algorithms.

3.3.2.3 Approximated derivative of the complete data LOO log-likelihood

To evaluate the first order derivative in (3.8), we propose to approximate the derivative by considering the LOO index in (3.2) to be fixed at the current estimate $\boldsymbol{\mu}^{(t)}$ so that we leave out the data point closest to $\boldsymbol{\mu}^{(t)}$ instead of $\boldsymbol{\mu}$. This gives us an approximation to the derivative of the LOO log-likelihood for the conditional normal distribution with

respect to $\boldsymbol{\mu}$ from (3.8) given the mixing variables $\boldsymbol{u}$,

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\mu}}\ell_N^{\text{LOO}} &\approx -\frac{1}{2}\left(\frac{\partial}{\partial \boldsymbol{\mu}}\sum_{i\neq k(\boldsymbol{\mu}^{(t)})}\frac{1}{u_i}(\boldsymbol{y}_i-\boldsymbol{\mu}-u_i\boldsymbol{\gamma})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_i-\boldsymbol{\mu}-u_i\boldsymbol{\gamma})\right)\\ &= \boldsymbol{\Sigma}^{-1}\sum_{i\neq k(\boldsymbol{\mu}^{(t)})}\frac{1}{u_i}(\boldsymbol{y}_i-\boldsymbol{\mu}-u_i\boldsymbol{\gamma}).\end{aligned}$$

Similarly, applying the approximate partial derivative to ℓ_N^{LOO} and ℓ_G^{LOO} with respect to other parameters and solving the approximate derivatives at zero gives us the following CM-steps.

CM-step for $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\gamma}$:

Suppose that the current iterate is $\boldsymbol{\theta}^{(t)}$ and $\boldsymbol{u}$ is given. After equating each component of the approximate partial derivatives of $\ell_N^{\text{LOO}}(\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\gamma};\boldsymbol{y},\boldsymbol{u})$ to zero, we obtain the following estimates:

$$\hat{\boldsymbol{\mu}} = \frac{S_{\boldsymbol{y}/u}S_u-(n-1)S_{\boldsymbol{y}}}{S_{1/u}S_u-(n-1)^2}, \tag{3.9}$$

$$\hat{\boldsymbol{\gamma}} = \frac{S_{\boldsymbol{y}}-(n-1)\hat{\boldsymbol{\mu}}}{S_u}, \tag{3.10}$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\sum_{i\neq k(\boldsymbol{\mu}^{(t)})}\frac{1}{u_i}(\boldsymbol{y}_i-\hat{\boldsymbol{\mu}})(\boldsymbol{y}_i-\hat{\boldsymbol{\mu}})' - \frac{1}{n-1}\hat{\boldsymbol{\gamma}}\hat{\boldsymbol{\gamma}}'S_u \tag{3.11}$$

where the sufficient statistics to the approximate LOO log-likelihood are:

$$S_{\boldsymbol{y}} = \sum_{i\neq k(\boldsymbol{\mu}^{(t)})}\boldsymbol{y}_i,\quad S_{\boldsymbol{y}/u} = \sum_{i\neq k(\boldsymbol{\mu}^{(t)})}\frac{1}{u_i}\boldsymbol{y}_i,\quad S_u = \sum_{i\neq k(\boldsymbol{\mu}^{(t)})}u_i,\quad S_{1/u} = \sum_{i\neq k(\boldsymbol{\mu}^{(t)})}\frac{1}{u_i}. \tag{3.12}$$

For the AECM algorithm, the CM-step for κ and the CM-step for ν using the LOO likelihood are similar to Section 2.2 and 2.1.3 respectively.

3.3.2.4 Line search

The estimates in (3.9) to (3.11) using approximate derivatives will not guarantee the LOO likelihood to increase. In this regard, we propose to apply a line search to ensure the LOO likelihood increase after each CM-step. This line search is part of a class of adaptive over-relaxed methods which can also improve the efficiency of EM algorithm [97]. Let $\boldsymbol{\theta}^{(t)}$ be the current estimate and $\boldsymbol{\theta}^{(t+1)}$ be the updated estimate after the

CM-step in Section 3.3.2. We propose to construct a direct line search by defining

$$\boldsymbol{\theta}^* = \boldsymbol{\theta}^{(t)} + \xi\left(\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}\right)$$

where $\xi \in I \subset \mathbb{R}$ and the interval I is chosen so that $\boldsymbol{\theta}^*$ remains in the parameter space.

Using the `optimise` function in `R`, ξ is estimated to be ξ^* such that it maximises the LOO log-likelihood

$$\xi^* = \underset{\xi \in I}{\operatorname{argmax}}\ \ell^{\text{LOO}}(\boldsymbol{\theta}^*).$$

Since finding the maximum of a non-differentiable function is difficult, we can alternatively choose $\boldsymbol{\theta}^*$ such that it improves the LOO likelihood over the previous estimate such that

$$\ell^{\text{LOO}}(\boldsymbol{\theta}^*; \boldsymbol{y}) \geq \ell^{\text{LOO}}(\boldsymbol{\theta}^{(t)}; \boldsymbol{y}).$$

3.3.3 AECM algorithm

Combining the steps we introduced earlier gives us the ACME algorithm for the VG distribution using the LOO likelihood:

Initialisation step: Choose suitable starting values $(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0, \boldsymbol{\gamma}_0, \nu_0)$. It is recommended to choose starting values $(\bar{\boldsymbol{y}}, \text{cov}(\boldsymbol{y}), \mathbf{0}, d+3)$.

At the t^{th} iteration with current estimates $(\boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}, \boldsymbol{\gamma}^{(t)}, \nu^{(t)})$:

Local Point Search: Update the parameter $\boldsymbol{\mu}$ using local midpoint or point search in Sections 3.3.2.1 and 3.3.2.2 respectively.

E-step 1: Calculate $\widehat{u}_i$ and $\widehat{1/u_i}$ for $i = 1, \ldots, n$ in (2.7) and (2.8) respectively using parameters from the local point search. Also calculate the sufficient statistics $S_{\boldsymbol{y}/u}$, S_u and $S_{1/u}$ in (3.12).

CM-step 1: Update the parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\gamma})$ in (3.9) to (3.11) using the sufficient statistics in E-step 1. Then apply the line search in Section 3.3.2.4 to ensure monotonic convergence.

CM-step 2: Estimate κ to update the parameters $(\boldsymbol{\Sigma}, \boldsymbol{\gamma})$ using the data augmentation scheme similar to Section 2.2.

CM-step 3: Update the parameter ν by maximising the observed LOO log-likelihood with respect to ν while keeping the other parameters fixed.

Stopping rule: Repeat the procedures until the relative increment of LOO log-likelihood function is sufficiently small as in (1.8).

We remark that the LPS algorithm ensure the location estimate does not get stuck around the local maximas whereas the line search in Section 3.3.2.4 is applied after each CM-step that maximise the Q-function to ensure monotonic convergence of the AECM algorithm.

We numerically verify the accuracy of this algorithm in Section 3.4.1 using Monte Carlo simulations.

Algorithm 10: AECM algorithm for VG using LOO likelihood

Input: Initial value $\boldsymbol{\theta}^{(0)}$

while $\ell^{\text{LOO}}(\boldsymbol{\theta}^{(t+1)};\boldsymbol{y}) - \ell^{\text{LOO}}(\boldsymbol{\theta}^{(t)};\boldsymbol{y}) > \delta$ **do**

$$\boldsymbol{\theta}^{(t+1/5)} \leftarrow \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}_1}{\operatorname{argmax}} \left\{ \ell^{\text{LOO}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{(t)}, \boldsymbol{\gamma}^{(t)}, \nu^{(t)}; \boldsymbol{y}) : \boldsymbol{\mu} \in \{\boldsymbol{y}_{i_0}, \dots, \boldsymbol{y}_{i_m}\} \right\} ;$$

$$Q^{\text{LOO}}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t+1/5)}) \leftarrow \mathbb{E}_{\boldsymbol{\theta}^{(t+1/5)}} \left[\ell^{\text{LOO}}(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{u}) | \boldsymbol{y} \right] ;$$

$$\boldsymbol{\theta}^{(t+2/5)} \leftarrow \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}_2}{\operatorname{argmax}} Q^{\text{LOO}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}, \nu^{(t+1/5)}; \boldsymbol{\theta}^{(t+1/5)}) ;$$

$$\boldsymbol{\theta}^{(t+3/5)} \leftarrow \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}_3}{\operatorname{argmax}} \left\{ \ell^{\text{LOO}}(\boldsymbol{\theta}; \boldsymbol{y}) : \boldsymbol{\theta} = \boldsymbol{\theta}^{(t+1/5)} + \xi(\boldsymbol{\theta}^{(t+2/5)} - \boldsymbol{\theta}^{(t+1/5)}) \right\} ;$$

$$\boldsymbol{\theta}^{(t+4/5)} \leftarrow \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}_4}{\operatorname{argmax}} \ell^{\text{LOO}}(\boldsymbol{\mu}^{(t+3/5)}, \tfrac{1}{\kappa}\boldsymbol{\Sigma}^{(t+3/5)}, \tfrac{1}{\kappa}\boldsymbol{\gamma}^{(t+3/5)}, \nu^{(t+3/5)}; \boldsymbol{y}) ;$$

$$\boldsymbol{\theta}^{(t+1)} \leftarrow \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}_5}{\operatorname{argmax}} \ell^{\text{LOO}}(\boldsymbol{\mu}^{(t+4/5)}, \boldsymbol{\Sigma}^{(t+4/5)}, \boldsymbol{\gamma}^{(t+4/5)}, \nu; \boldsymbol{y}) ;$$

end

3.3.4 Convergence of AECM algorithm using LOO likelihood

The AECM algorithm described in Section 3.3.3 can be thought of as an ECME algorithm with additional CM-step for κ. So for this case, it is sufficient to prove the monotonic convergence of the ECME algorithm using the LOO likelihood.

Let the approximate LOO log-likelihood be defined as

$$\widetilde{\ell^{\text{LOO}}}(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i \neq k(\boldsymbol{\mu}^{(t)})} \log f(\boldsymbol{y}_i; \boldsymbol{\theta}) \tag{3.13}$$

with the LOO index fixed at $k(\boldsymbol{\mu}^{(t)})$. To show the convergence of the ECME algorithm using approximate LOO log-likelihood, we first prove the convergence of the ECME algorithm with one CM-step for the approximate LOO log-likelihood and then extend the proof for multiple CM-steps. For the case with one CM-step, we apply the idea in Section 1.3 to the LOO log-likelihood and state two fundamental results below:

$$\widetilde{\ell^{\mathrm{LOO}}}(\boldsymbol{\theta};\boldsymbol{y}) = \widetilde{Q^{\mathrm{LOO}}}(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}) - \widetilde{H^{\mathrm{LOO}}}(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)})$$

and

$$\widetilde{H^{\mathrm{LOO}}}(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}) \leq \widetilde{H^{\mathrm{LOO}}}(\boldsymbol{\theta}^{(t)};\boldsymbol{\theta}^{(t)})$$

where we let

$$\widetilde{Q^{\mathrm{LOO}}}(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}) = \int \widetilde{\ell^{\mathrm{LOO}}}(\boldsymbol{\theta};\boldsymbol{y},\boldsymbol{u}) f(\boldsymbol{u}|\boldsymbol{y};\boldsymbol{\theta}^{(t)})\,\mathrm{d}\boldsymbol{u}$$

with $f(\boldsymbol{u}|\boldsymbol{y};\boldsymbol{\theta}^{(t)}) = \prod_{i=1}^{n} f(u_i|\boldsymbol{y}_i;\boldsymbol{\theta}^{(t)})$, $\widetilde{\ell^{\mathrm{LOO}}}(\boldsymbol{\theta};\boldsymbol{y},\boldsymbol{u}) = \sum_{i\neq k(\boldsymbol{\mu}^{(t)})} \log f(\boldsymbol{y}_i,u_i;\boldsymbol{\theta})$, and

$$\widetilde{H^{\mathrm{LOO}}}(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}) = \int \widetilde{\ell^{\mathrm{LOO}}}(\boldsymbol{\theta};\boldsymbol{u}|\boldsymbol{y}) f(\boldsymbol{u}|\boldsymbol{y};\boldsymbol{\theta}^{(t)})\,\mathrm{d}\boldsymbol{u}$$

with $\widetilde{\ell^{\mathrm{LOO}}}(\boldsymbol{\theta};\boldsymbol{u}|\boldsymbol{y}) = \sum_{i\neq k(\boldsymbol{\mu}^{(t)})} \log f(u_i|\boldsymbol{y}_i;\boldsymbol{\theta})$. The idea of the proof are exactly the same as in Lemma 1.3.1 and 1.3.2 by replacing the full likelihood with the LOO likelihood.

However, choosing $\boldsymbol{\theta}^{(t+1)}$ such that

$$\widetilde{Q^{\mathrm{LOO}}}\big(\boldsymbol{\theta}^{(t+1)},\boldsymbol{\theta}^{(t)}\big) \geq \widetilde{Q^{\mathrm{LOO}}}\big(\boldsymbol{\theta}^{(t)},\boldsymbol{\theta}^{(t)}\big)$$

guarantee that $\widetilde{\ell^{\mathrm{LOO}}}\big(\boldsymbol{\theta}^{(t+1)};\boldsymbol{y}\big) \geq \widetilde{\ell^{\mathrm{LOO}}}\big(\boldsymbol{\theta}^{(t)};\boldsymbol{y}\big)$ but not $\ell^{\mathrm{LOO}}\big(\boldsymbol{\theta}^{(t+1)};\boldsymbol{y}\big) \geq \ell^{\mathrm{LOO}}\big(\boldsymbol{\theta}^{(t)};\boldsymbol{y}\big)$. For this reason we perform a line search in Section 3.3.2.4 so that the LOO log-likelihood improves and thus guarantee the monotonic convergence of the LOO log-likelihood.

For the case with multiple CM-steps, the monotonic convergence of the ECME algorithm only applies if all the CM-steps applied to Q-functions are performed before the CM-step applied to the observed LOO log-likelihood (see Section 1.4.3). Thus for this case, we apply the line search to the CM-steps involving the Q-function to ensure that the observed LOO log-likelihood increase after each CM-step. Thus this guarantees the monotonic convergence of the LOO log-likelihood in Section 3.3.3.

3.4 Simulation studies

3.4.1 Accuracy of estimates for AECM algorithm

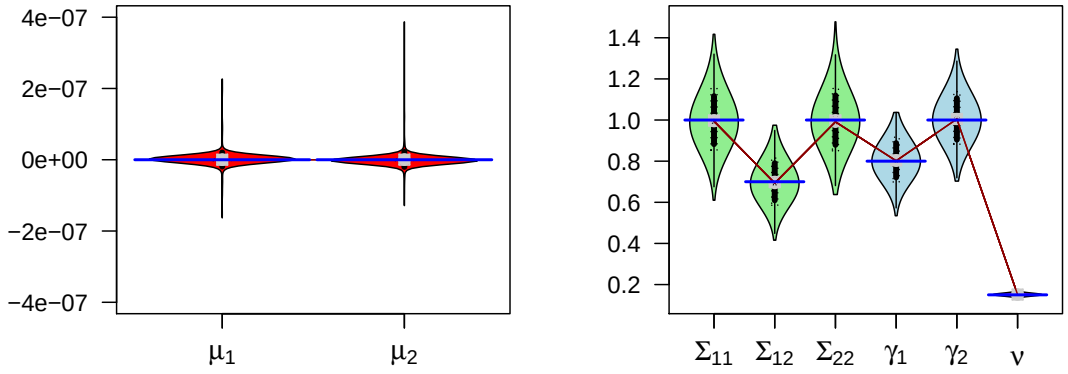
To demonstrate the accuracy of the proposed AECM algorithm, we simulate $n = 1000$ bivariate skewed VG samples with parameter values

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}, \quad \boldsymbol{\gamma} = \begin{pmatrix} 0.8 \\ 1 \end{pmatrix}, \text{ and } \quad \nu = 0.15 \tag{3.14}$$

and estimate the parameters using the AECM algorithm in Section 3.3.3. We repeat this experiment 1000 times and present the results in Figures 3.2 and 3.3.

Figure 3.2 shows the violin plots implemented using the `caroline` package [99] in `R` which presents the density estimate of the parameter estimates using a Gaussian kernel. The medians of the estimates are very close to the true parameters of the distribution implying that the algorithm gives consistent estimates for these parameters, even when $\nu < \frac{d}{2}$ leads to unbounded likelihood. Moreover, the distribution of the parameters $\hat{\boldsymbol{\Sigma}}$, $\hat{\boldsymbol{\gamma}}$, and $\hat{\nu}$ appears to approximately follow a normal distribution. On the other hand, the distribution of $\hat{\boldsymbol{\mu}}$ is non-Gaussian with high density around $\mathbf{0}$ and extreme heavy-tails.

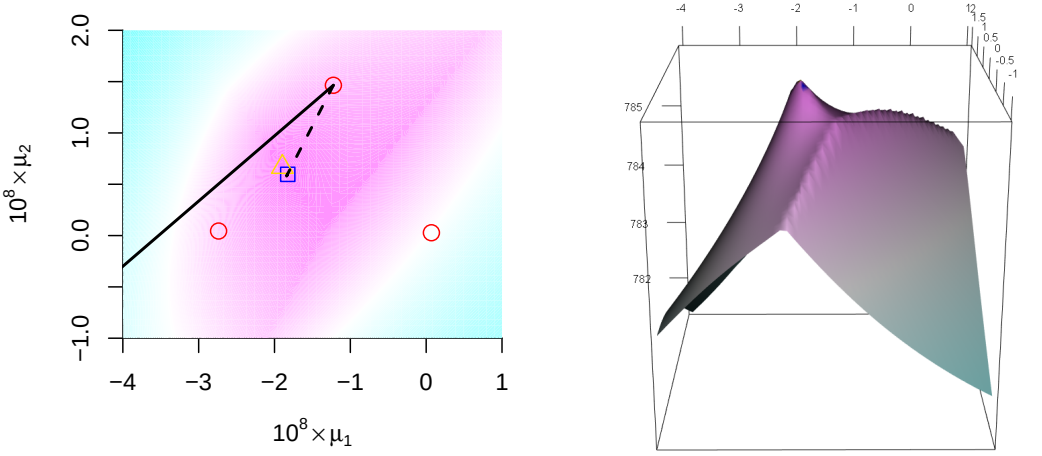
Figure 3.3(a) gives a contour plot of the LOO log-likelihood for one set of simulated data while tracking the path of the location parameter for each iterate from the LPS algorithm, CM-step for $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\gamma})$ and line search. The estimate converges to the final estimate which is close to the local maximum, lying roughly between the data points and along the non-differentiable lines as discussed in Section 3.3.2.2. Furthermore, in Figure 3.3(b), we provide a three-dimensional plot of the LOO log-likelihood which is viewed from the bottom side of the contour plot. The maximum lying along the non-differentiable lines makes the computation more demanding as we cannot purely rely on derivative based methods. The LPS algorithm along with the line search serve as efficient iterative methods to obtain parameter estimates. The idea behind these search methods is that the estimate from the LPS jumps to the point broadly close to the maximum, while the CM-step and line search improves the estimates so that they converge closer towards the maximum which lie on the non-differentiable lines.



(a) Vioplot of $\boldsymbol{\mu}$

(b) Vioplot of $\boldsymbol{\Sigma}$, $\boldsymbol{\gamma}$, ν

Figure 3.2. Vioplots of the parameter estimates. The median is displayed as a grey box which is connected by a crimson line. Also the true parameter values represented by the blue lines is drawn for comparison.



(a) Contours of the LOO log-likelihood

(b) 3D plot of the LOO log-likelihood

Figure 3.3. Contour and 3D plot of the LOO log-likelihood for one set of simulated data: (a) Contour plot with the path of the algorithm's iterated values (solid black from local point search, and dashed black from CM-step and line search) converging towards the final estimate (blue square). This estimate is close to the local maximum (gold triangle) obtained by fine grid search, and is roughly between the data points (red open circles). (b) 3D plot viewed from roughly the bottom side of the contour plot. It can be observed that the local maximum is visible at the peak, and that it lies on the cusp lines which is generated from two closest points. For both plots, the subdued blue-pink palette is used to represent lower (blue) and higher (pink) values.

3.4.2 Asymptotic properties for the location estimates of VG distribution

Podgórski and Wallin [89] proved the consistency and super-efficiency of the location estimator using the LOO likelihood as stated in Theorem 3.2.2. They also stated the upper bound for the index of the rate of convergence $\beta < 1/(1+\alpha)$ where $\alpha = 2\nu - 1$ for the univariate VG distribution with unbounded density. The aim of this section is to determine these optimal rates through Monte Carlo simulations, and to analyse the asymptotic distribution of the location parameter estimator for the cases of cusp and unbound densities.

We present the set-up of the simulation study below:

Step 1: Set the true shape parameters ν to be one of the 50 shape parameters $\{0.02, 0.04, \dots, 0.98, 1\}$.

Step 2: For each shape parameter, set the sample size n to be one of the 41 sample sizes
$\{500, 1000, \dots, 19500, 20000\} \cup \{100000\}$.

Step 3: For each (ν, n), generate 20000 different sets of samples, each set from standardised univariate symmetric VG distribution with shape parameter ν and sample size n.

Step 4: For each set of samples, estimate $\hat{\mu}_n$ by searching through the midpoints and choosing the one that maximises the LOO log-likelihood where the other parameters $(\sigma^2 = 1, \gamma = 0, \nu)$ are fixed.

This gives us 20000 $\hat{\mu}_n$'s for each (ν, n).

3.4.2.1 Optimal Convergence rate of $\hat{\mu}$

Since the scale of asymptotic distribution of $\hat{\mu}_n$ increases according to a power law with respect to n, we fit a power curve to estimate the optimal rate. To measure the spread of $\hat{\mu}_n$ centred from the true parameter value $\mu = 0$, we choose a robust measure of spread called the median absolute deviation from 0 (MAD_0) defined by

$$\text{MAD}_0(\boldsymbol{x}) = \text{median}(|\boldsymbol{x}|)$$

for some univariate data set $\boldsymbol{x} = (x_1, \dots, x_n)$.

For each (ν, n), we calculate the MAD_0 of the 20000 $\hat{\mu}_n$'s. Then for each ν, we fit a power curve to the MAD_0 against n. In other words, we find parameters a and b such that $\text{MAD}_0 = an^b$. This is equivalent to fitting a simple linear regression model $\log \text{MAD}_0 = \log a + b \log n$ to obtain the estimates $(\widehat{\log a}, \hat{b})$. Then an estimate of the optimal rate for a given ν is obtained by setting $\hat{\beta} = -\hat{b}$. We repeat this process for the other choices of ν.

Figure 3.4 plots the relative error of $\hat{\beta}$ against ν along with its confidence intervals. From this figure, $\hat{\beta}$ appears to follow the proposed optimal rate of $\frac{1}{2\nu}$ when $0 < \nu \leq 0.4$. However, when $0.4 < \nu < 1$, $\hat{\beta}$ appears to be different from $\frac{1}{2\nu}$ and the relative error follows a wave-like pattern. In fact, for $0.4 < \nu < 0.76$, $\hat{\beta}$ appears to be greater than the proposed optimal convergence rate index whereas for $0.76 < \nu < 1$, $\hat{\beta}$ appears to be less than the proposed optimal convergence rate index. As ν approaches to 1, $\hat{\beta}$ approaches the convergence rate for asymptotic normality. Overall, the estimated optimal rate is consistent with Theorem 3.2.2 in the range $\nu < 0.5$ for unbounded density. As for $0.5 \leq \nu \leq 1$, more theoretical studies is needed to understand the behaviour of the location estimate $\hat{\mu}_n$ of distribution with cusp density. To investigate this peculiar behaviour of $\hat{\mu}_n$, we further examine the asymptotic distribution using our simulated results.

3.4.2.2 Asymptotic distribution of $\hat{\mu}$

We begin by plotting a Gaussian kernel density estimate in Figure 3.5 of the simulated estimates $\hat{\mu}_n$ with its scale standardised using MAD_0 when $n = 10000$. We note that the estimated density exhibits heavier tails and sharper peaks at the expense of intermediate tails as ν decreases. We transform the $\hat{\mu}_n$'s by considering $\log |\hat{\mu}_n|$ in order to observe the behaviour on a more appropriate scale, and plot the kernel density estimates in Figure 3.6 for various (ν, n). Generally, as the sample size increases, the location of the distribution shifts to the left. The scale and shape roughly stay the same with $\nu = 0.8$ as an exception since the scale gets slightly larger while the shape becomes more left skewed.

Comparing these plots with Figure C.1, we see that the density of $\log |\hat{\mu}_n|$ resembles that of a generalised Gumbel (GG) distribution which is discussed in more detail in Appendix C3. To investigate this further, we first fit the 20000 $\log |\hat{\mu}_n|$'s to a GG distribution

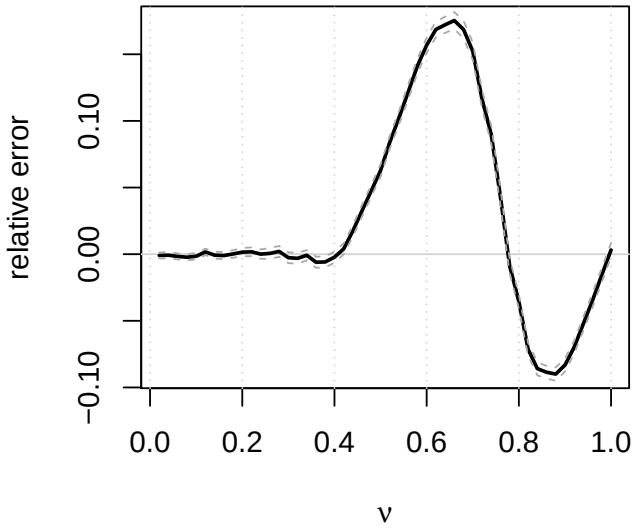


Figure 3.4. Plots of the relative error of $\frac{\hat{\beta}-\beta}{\beta}$ (solid black) against ν. The horizontal solid grey line indicates agreement of $\hat{\beta}$ with the proposed optimal rate $\beta = \frac{1}{2\nu}$. The vertical grey dotted lines represents grid lines for $\nu = \{0, 0.2, 0.4, 0.6, 0.8, 1\}$. We also include the 95% confidence interval for the relative error (dashed grey).

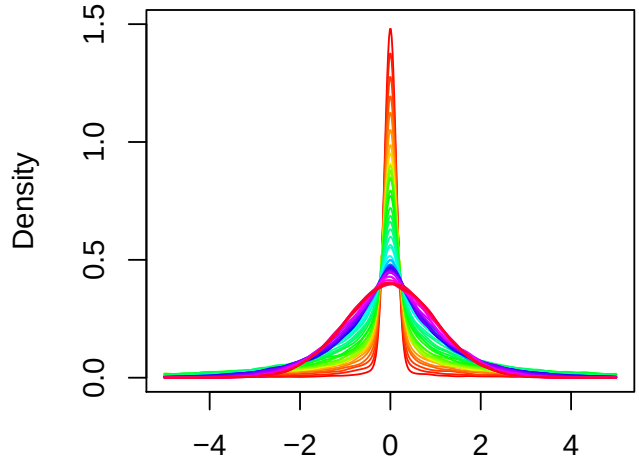
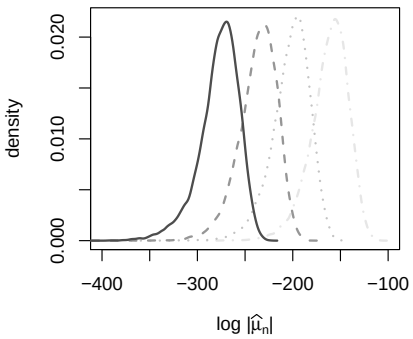
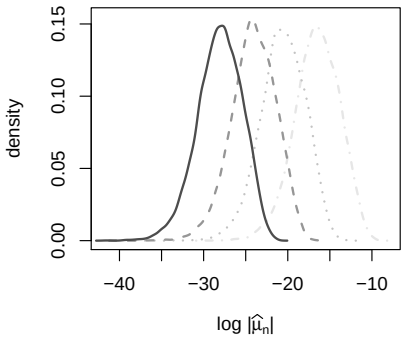


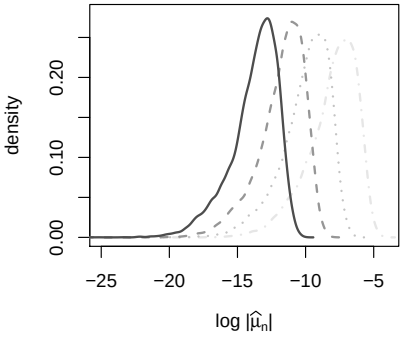
Figure 3.5. Density plots of the simulated $\hat{\mu}_n$ with its scale standardised using MAD_0 for each ν where $n = 100000$. We use a rainbow colour scheme ranging from red ($\nu = 0.02$) to magenta ($\nu = 1$).



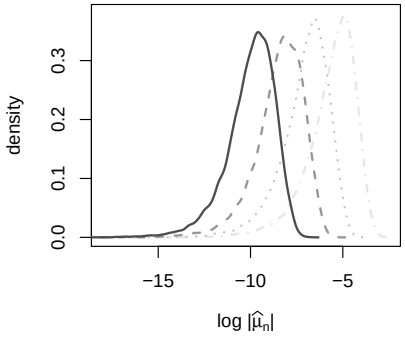
(a) density plots for $\nu = 0.02$



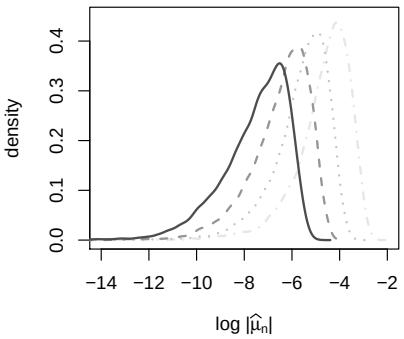
(b) density plots for $\nu = 0.2$



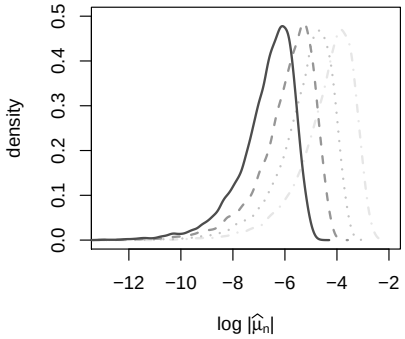
(c) density plots for $\nu = 0.4$



(d) density plots for $\nu = 0.6$



(e) density plots for $\nu = 0.8$



(f) density plots for $\nu = 1$

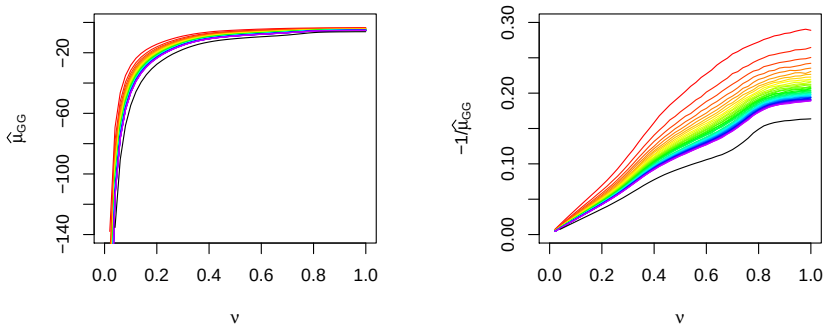
Figure 3.6. Kernel density estimates of $\log|\hat{\mu}_n|$'s for $\nu = 0.02, 0.2, 0.4, 0.6, 0.8, 1$ and $n = 1000$ (dash-dotted light grey), 5000 (dotted grey), 20000 (dashed dark grey), 100000 (solid black) with each n being combined into a single plot for comparison.

for each (ν, n). The parameter estimates of a GG distribution are represented by $(\hat{\mu}_{\mathrm{GG}}, \hat{\sigma}_{\mathrm{GG}}, \hat{m}_{\mathrm{GG}})$.

We plot the parameter estimates against ν in Figure 3.7, while also combining the plots for different n for comparison. We also plot the transformed parameter estimates to identify the behaviour across ν. In Figures 3.7(a) and 3.7(b), as ν decreases, the $\hat{\mu}_{\mathrm{GG}}$ appears to decrease roughly at a hyperbolic rate curve with some minor curvature for larger values of ν. In Figures 3.7(c) and 3.7(d), as ν decreases, $\hat{\sigma}_{\mathrm{GG}}$ increases at a hyperbolic rate with two bumps. One major bump occurs around $\nu = 0.2$ and a minor bump around $\nu = 0.7$. For the major bump, there is no clear distinction between each n due to the fluctuation with $\hat{\sigma}_{\mathrm{GG}}$. The source of the fluctuation is possibly due to sampling error. For the minor bump, the distinction between each n is more clear especially when the estimates for $n = 100000$ are distinct from the other n. This seems to suggest that the asymptotic distribution for $\hat{\mu}_n$ has yet to converge. How large should n be so that the asymptotic distribution converges is unclear for $0.4 \leq \nu \leq 0.9$. Moreover, the minor bump falls into the range $0.4 < \nu < 0.76$ in which the estimated convergence rate index $\hat{\beta}$ is larger than $\frac{1}{2\nu}$, as shown in Figure 3.4. In Figures 3.7(e) and 3.7(f), $\hat{m}_{\mathrm{GG}}$ also has a major and minor bumps similar to $\hat{\sigma}_{\mathrm{GG}}$. Unlike $\hat{\mu}_{\mathrm{GG}}$ and $\hat{\sigma}_{\mathrm{GG}}$, $\hat{m}_{\mathrm{GG}}$ tends to some constant value as ν approaches to 0.

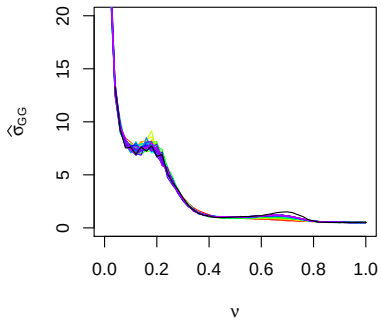
Lastly we provide the P-P plots in Figures 3.8 and 3.9 to check the goodness-of-fit for the GG distribution. The P-P plots are generated by applying the cumulative distribution function (CDF) of the GG distribution in (C.3) fitted to the 20000 $\log|\hat{\mu}_n|$ against the ordered sequence $\{i/(20001)\}$ for $i = 1, \ldots, 20000$. For comparison, we also combine all the sample sizes analysed into one plot for each $\nu = 0.02, \ldots, 1$. From the P-P plots, it appears that the GG distribution fit the simulated $\log|\hat{\mu}_n|$ really well since the plots roughly follow a straight line, although there are some small deviation from the straight line for $0.22 \leq \nu \leq 0.34$. Note that fitting the GG distribution to $\log|\hat{\mu}_n|$ corresponded to fitting a double generalised gamma (DGGamma) distribution to $\hat{\mu}_n$ from Theorem C4.1.

Thus we can perform statistical inference on $\hat{\boldsymbol{\mu}}_n$ using the DGGamma distribution as an approximation. To briefly demonstrate this, we apply the approximate distribution to estimate the variability of $\hat{\boldsymbol{\mu}}$ in Section 3.4.1. Using the true parameter $\nu = 0.15$ with $n = 1000$ to extrapolate values $(-1/\hat{\mu}_{\mathrm{GG}} = 0.0466, -1/\hat{\sigma}_{\mathrm{GG}} = -0.1291, \log \hat{m}_{\mathrm{GG}} = 1.8733)$ by applying the `spline` function in `R` to Figures 3.7(b), (d), and (f), this gives

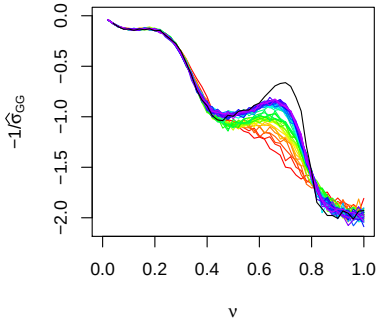


(a) $\hat{\mu}_{\mathrm{GG}}$ vs ν

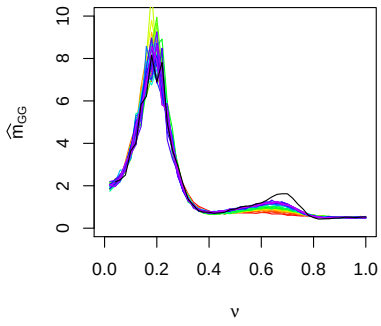
(b) $-1/\hat{\mu}_{\mathrm{GG}}$ vs ν



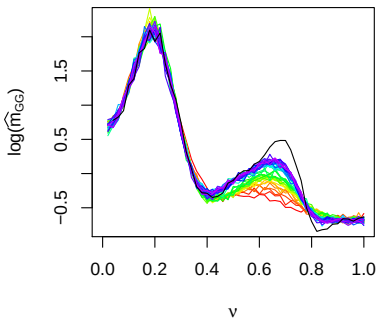
(c) $\hat{\sigma}_{\mathrm{GG}}$ vs ν



(d) $-1/\hat{\sigma}_{\mathrm{GG}}$ vs ν



(e) $\hat{m}_{\mathrm{GG}}$ vs ν



(f) $\log \hat{m}_{\mathrm{GG}}$ vs ν

Figure 3.7. Plot of estimates of GG distribution fitted to the distribution of $\log|\hat{\mu}_n|$ against ν. On the left column, we plot the $(\hat{\mu}_{\mathrm{GG}}, \hat{\sigma}_{\mathrm{GG}}, \hat{m}_{\mathrm{GG}})$ against ν respectively while on the right column, we plot the transformation $(-1/\hat{\mu}_{\mathrm{GG}}, -1/\hat{\sigma}_{\mathrm{GG}}, \log \hat{m}_{\mathrm{GG}})$ against ν respectively to enlarge certain portion of the plots. A rainbow colour scheme ranging from red ($n = 500$) to magenta ($n = 20000$) is used to denote sample size. In addition, the black line represents $n = 100000$.

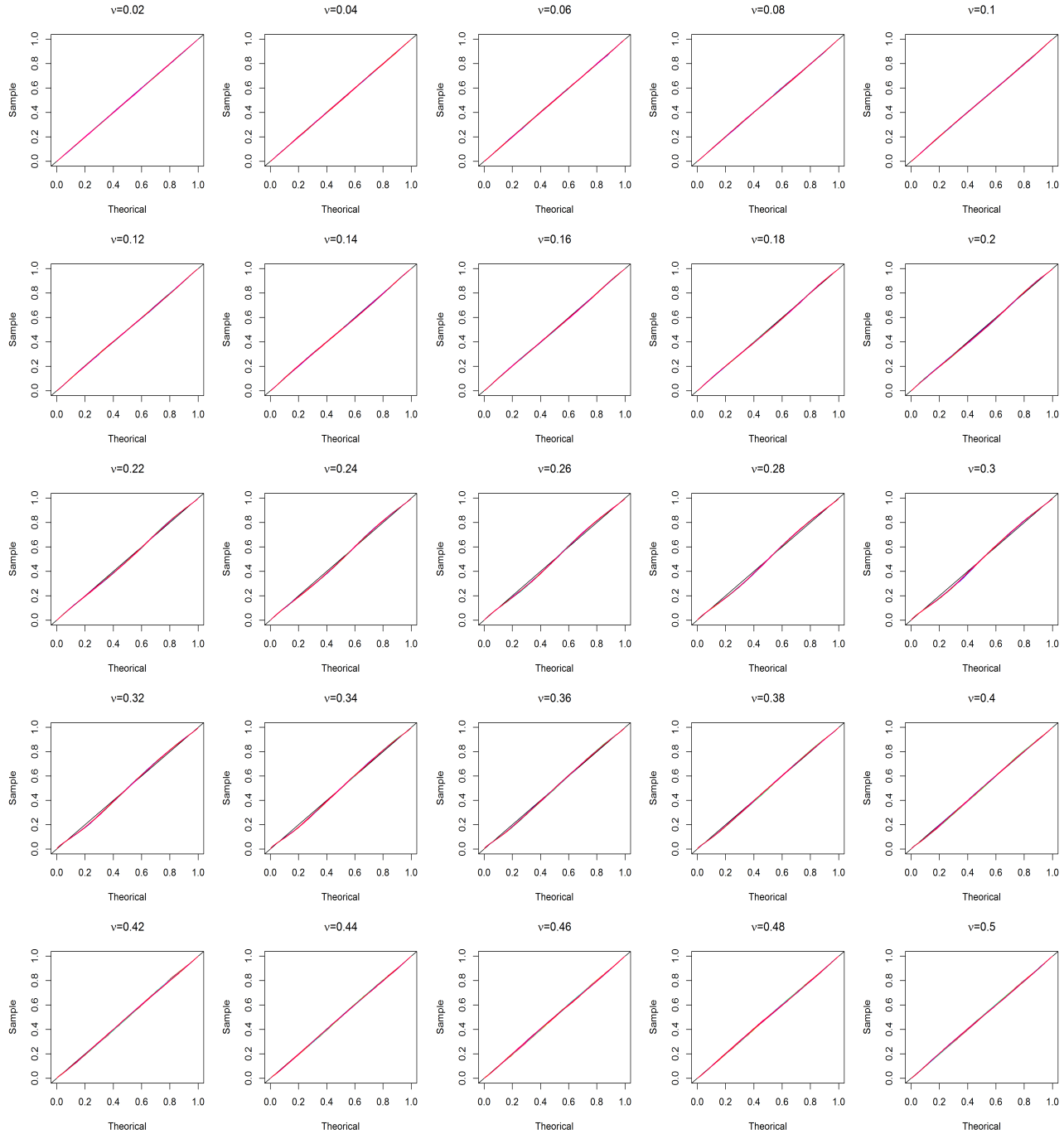


Figure 3.8. P-P plots for $0.02 \leq \nu \leq 0.5$ where the x-axis represent the empirical CDF $i/(20001)$, $i = 1, \ldots, 20000$, and y-axis represents the ordered $F_{\mathrm{GG}}(\log|\hat{\mu}_n|)$ where F_{GG} is the CDF of GG distribution based on the fitted parameters $(\hat{\mu}_{\mathrm{GG}}, \hat{\sigma}_{\mathrm{GG}}, \hat{m}_{\mathrm{GG}})$. For comparison, we also combine the sample sizes $n = 500, 1000, \ldots, 19500, 20000$ and $n = 100000$ into one plot for each $\nu = 0.02, \ldots, 1$. The same rainbow colour scheme is used from Figure 3.7.

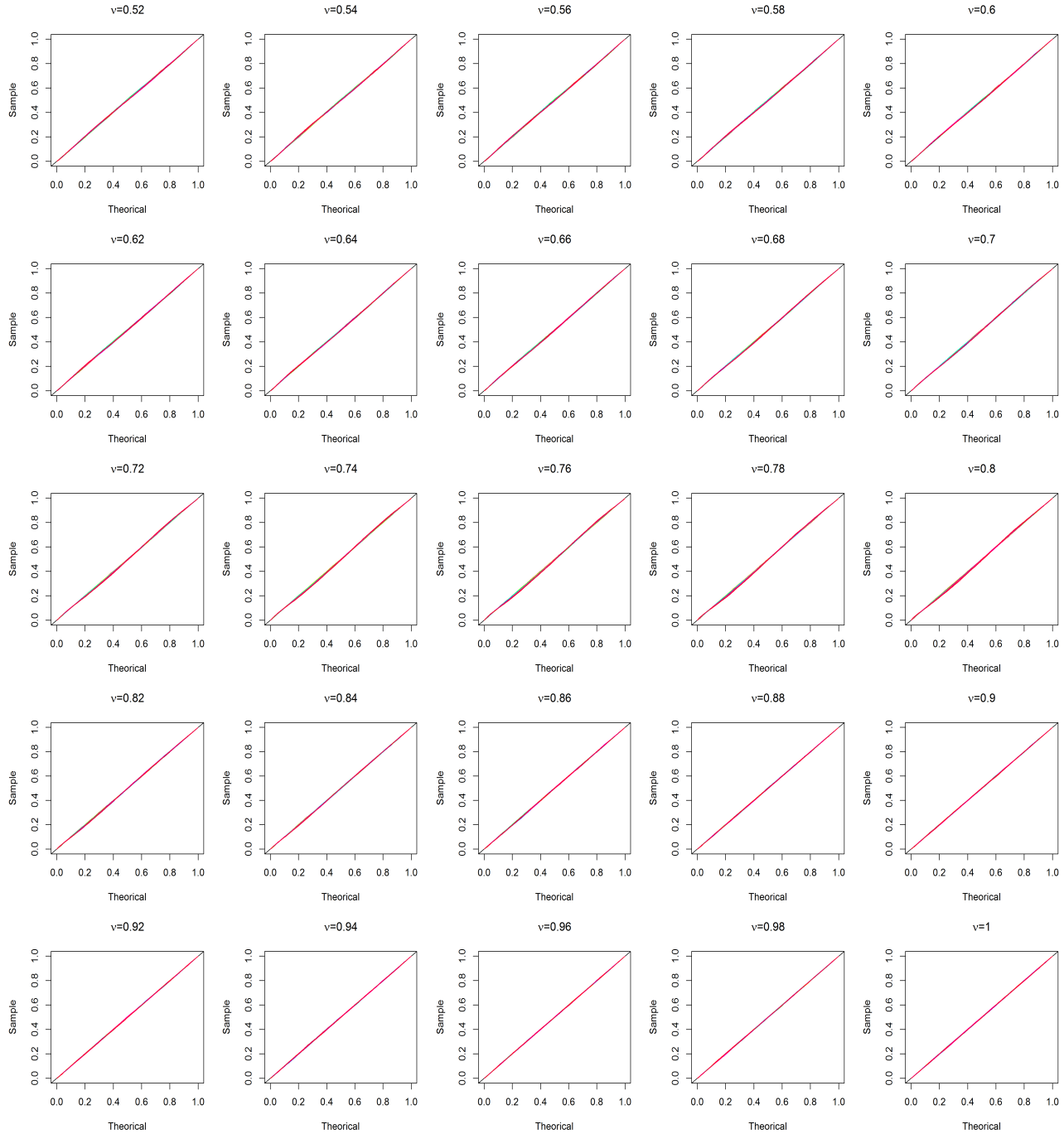


Figure 3.9. P-P plots for $0.52 \leq \nu \leq 1$.

the values for the GG parameter estimates ($\hat{\mu}_{\text{GG}} = -21.4431, \hat{\sigma}_{\text{GG}} = 7.7431, \hat{m}_{\text{GG}} = 6.5098$). Applying these estimates to equation (C.6) with $\Sigma_{11} = 1$ and $\Sigma_{22} = 1$ gives us the approximation based on the DGGamma distribution

$$\text{MAD}_{\text{DGGamma}}(\hat{\boldsymbol{\mu}}) = \begin{pmatrix} \text{MAD}_{\text{DGGamma}}(\hat{\mu}_1) \\ \text{MAD}_{\text{DGGamma}}(\hat{\mu}_2) \end{pmatrix} \approx \begin{pmatrix} 3.25 \times 10^{-10} \\ 3.25 \times 10^{-10} \end{pmatrix},$$

whereas the sample median absolute deviation (MAD) applied to each element of the 1000 replicates of the location estimate $\hat{\boldsymbol{\mu}}$ gives us

$$\text{MAD}_{\text{sam}}(\hat{\boldsymbol{\mu}}) = \begin{pmatrix} 3.16 \times 10^{-10} \\ 2.85 \times 10^{-10} \end{pmatrix}$$

where the sample MAD is defined by $\text{MAD}_{\text{sam}}(\boldsymbol{x}) = \text{median}(|\boldsymbol{x} - \text{median}(\boldsymbol{x})|)$ for some univariate data $\boldsymbol{x}$. Since the MAD using the DGGamma distribution is similar to the MAD from the simulated location estimates, we conclude that the DGG distribution can provide reasonably accurate estimates for the SE of the location parameter. In conclusion, we can construct confidence intervals and approximate the SE for the location parameter, especially when the shape parameter falls into the range that gives rise to the unbounded or cusp density ($\nu \leq \frac{d+1}{2}$).

3.5 Conclusion

We propose an AECM algorithm to estimate parameters of the VG distribution using the LOO likelihood when the density is unbounded. Our first simulation study shows that all parameters for the VG distribution are estimated to a high level of accuracy. Looking at the first simulated data, we also demonstrate how the AECM algorithm estimates the location parameter which lies along the non-differentiable lines of the LOO likelihood.

We conduct our second simulation study to empirically explore the optimal convergence rate and asymptotic distribution for the location parameter estimator using the maximum LOO likelihood method. Results show that the index for the optimal rate of convergence follows $\frac{1}{2\nu}$ when $0 < \nu \leq 0.4$. However, when $0.4 < \nu < 1$, the index appears to be slightly different from $\frac{1}{2\nu}$ with a wave-like pattern for the relative error. As ν approaches to 1, the optimal rate approaches the convergence rate for asymptotic

normality. Furthermore, we demonstrate how the asymptotic distribution for $\hat{\mu}_n$ can be approximated using the DGGamma distribution for all ν. Hence we can approximate the SE for $\hat{\mu}$ and construct confidence intervals based on the DGGamma distribution.

However, we see some limitations in the simulation study such as the assumption of univariate symmetric VG distribution and the ignorance of dependency of location parameter with other parameters. This is discussed in Chapter 6. In terms of model applicability, there are two issues. The first issue is that the LOO likelihood method fails where there are repeated data points. The second is that the VG distribution fails to capture the high persistence and time series data structures often present in financial return series. The next two chapters deal with these two issues.

CHAPTER 4

Weighted Leave-one-out Likelihood for data multiplicity

4.1 Introduction

The LOO likelihood in Chapter 3 performs well when there are no repeated data points since the contribution of the unboundedness for each data point occurs once. Thus leaving out a single data point removes the unboundedness in the LOO likelihood. When there exist repeated data points, then even if we leave out one of the data points, the LOO likelihood would still blow up to infinity. Data multiplicity is common when the measurements have limited level of accuracy.

One method to circumvent the problem is to leave out multiple data points depending on the multiplicity of the data point in the likelihood function. This modified likelihood is called the leave-multiple-out (LMO) likelihood. However, the number of data points to leave out is not fixed but instead varies depending on the data multiplicity. For the LOO likelihood, there are always $(n-1)$ data contribution from a sample of size n whereas the LMO likelihood have varied data contribution across the parameter space if there are varied data multiplicities. Moreover, there are discontinuities between data points for the LMO likelihood which is described in more detail in Section 4.3.

The aim of this section is to modify the LOO likelihood by adding weights so that the likelihood not only prevent the unboundedness of the LOO likelihood in the case of data multiplicity, but also have the number of data contribution consistent with the LOO likelihood as well as no discontinuities for the symmetric case.

For the rest of the chapter, we begin with defining the LMO and WLOO likelihoods in Section 4.2. Then Section 4.3 gives three simple examples to illustrate the data multiplicity problem and find the weights to ensure consistent data contribution and continuity between data points for the WLOO likelihood. A simulation study is conducted in Section 4.4 to compare the performance of the WLOO likelihood method with other likelihood methods, and lastly, a conclusion is drawn in Section 4.5.

4.2 Leave-multiple-out and weighted LOO likelihoods

When there are data multiplicity, one way is to leave out data points with data multiplicity to avoid the unbounded density. The leave-multiple-out (LMO) likelihood is defined as

$$L^{\text{LMO}}(\boldsymbol{\theta}; \boldsymbol{y}) = \prod_{i \neq K(\boldsymbol{\mu})} f(\boldsymbol{y}_i; \boldsymbol{\theta})$$

where

$$K(\boldsymbol{\mu}) = \left\{ i \in \{1, \dots, n\} | \boldsymbol{y}_i = \boldsymbol{y}_{k(\boldsymbol{\mu})} \right\} \tag{4.1}$$

represents the LMO indices which corresponds to the data points identical to $\boldsymbol{y}_{k(\boldsymbol{\mu})}$, and $k(\boldsymbol{\mu})$ represents the LOO index defined in (3.2) in Section 3.2.1.

When there are no data multiplicity in the data set, the LMO likelihood reduces to the LOO likelihood function. However, when there are data multiplicity, the LMO likelihood leaves out all those data points which contribute to the unbounded likelihood whereas the LOO likelihood leave out just one data point which is not enough to remove the unbounded likelihood.

A major drawback of the LMO likelihood is that the number of data contribution is not consistent throughout the parameter space when there is data multiplicity resulting in a discontinuous LMO likelihood. To remedy this, we consider the weighted LOO (WLOO) likelihood defined as

$$L^{\text{WLOO}}(\boldsymbol{\theta}; \boldsymbol{y}) = \prod_{i=1}^{n} f(\boldsymbol{y}_i; \boldsymbol{\theta})^{w_i} \tag{4.2}$$

for $i = 1, \dots, n$ where we choose the weights to be

$$w_i = \begin{cases} 0 & , \text{ if } i \in K(\boldsymbol{\mu}) \\ \frac{|K(\boldsymbol{\mu})|+|J(\boldsymbol{\mu})|-1}{|J(\boldsymbol{\mu})|} & , \text{ if } i \in J(\boldsymbol{\mu}) \\ 1 & , \text{ otherwise} \end{cases} \tag{4.3}$$

such that $|K(\boldsymbol{\mu})|$ represents the cardinality of the set $K(\boldsymbol{\mu})$,

$$J(\boldsymbol{\mu}) = \left\{ i \in \{1, \dots, n\} \backslash K(\boldsymbol{\mu}) : \boldsymbol{y}_i = \boldsymbol{y}_{j(\boldsymbol{\mu})} \right\} \tag{4.4}$$

where $K(\boldsymbol{\mu})$ is defined in (4.1), and

$$j(\boldsymbol{\mu}) = \underset{i \in I \backslash K(\boldsymbol{\mu})}{\operatorname{argmin}} (\boldsymbol{y}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu})$$

represents the secondary LOO index. Similarly, the WLOO log-likelihood is defined as

$$\ell^{\text{WLOO}}(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{n} w_i f(\boldsymbol{y}_i; \boldsymbol{\theta}). \tag{4.5}$$

It is clear that for the case with no data multiplicity, the WLOO likelihood is equivalent to the LOO likelihood whereas for the classical likelihood, the weights are chosen to be $w_i = 1$.

The following section demonstrate with three examples on how the weights in (4.3) are derived based on the criteria that the WLOO likelihood removes the unbounded point with data multiplicity, the likelihood is continuous (for the symmetric case) and the data contribution is consistent with the LOO likelihood, or in other words, $\sum_{i=1}^{n} w_i = n - 1$.

When the density function is skewed, both the LOO and WLOO likelihoods are not continuous between data points. Nevertheless, the density function for the VG distribution in (1.39) as $\boldsymbol{\mu}$ approaches any data point is approximately symmetric. So by having more data points, the effect of the discontinuities due to skewness is negligible. Moreover, the alternate LOO index in (3.3) can also be adopted to make the WLOO likelihood continuous, even for the skewed case, though this alternate LOO index is not considered in this thesis.

4.3 Examples

Three examples are considered in this section. The first example considers the case with data multiplicity at a single location whereas the second example considers the case with data multiplicity at two different locations. The last example considers a general case with multiple data multiplicities at two different locations and verifies the formula for the weights in (4.3) which satisfy the three previously mentioned conditions that the WLOO likelihood prevents the unbounded likelihood from data multiplicity, data contribution is consistent with the LOO likelihood such that $\sum_{i=1}^{n} w_i = n-1$, and has no discontinuities at the midpoints for the symmetric case. For the figures in each example, we consider the symmetric VG distribution with shape parameter $\nu = 0.4$ which is in the region that causes the unbounded likelihood.

4.3.1 Example 1: data multiplicity at a single location

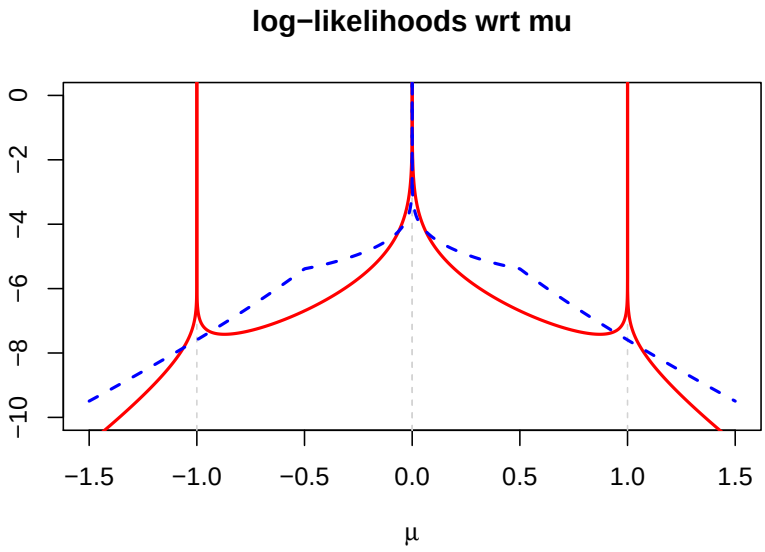
In this first example, the data set $\{-1, 0, 1, 0\}$ of size 4 contains a data multiplicity at 0. Figure 4.1 plots the LOO log-likelihood across location parameter μ. In plot (a), we observe that the LOO log-likelihood is unbounded at 0 even after leaving out one of the problematic data point. The LMO log-likelihood in plot (b) is bounded after leaving out multiple data points at 0 that cause the unboundedness but it also produces some discontinuities at the midpoints of -0.5 and 0.5. This behaviour of the log-likelihood is undesirable.

To this end, we consider the WLOO log-likelihood given by

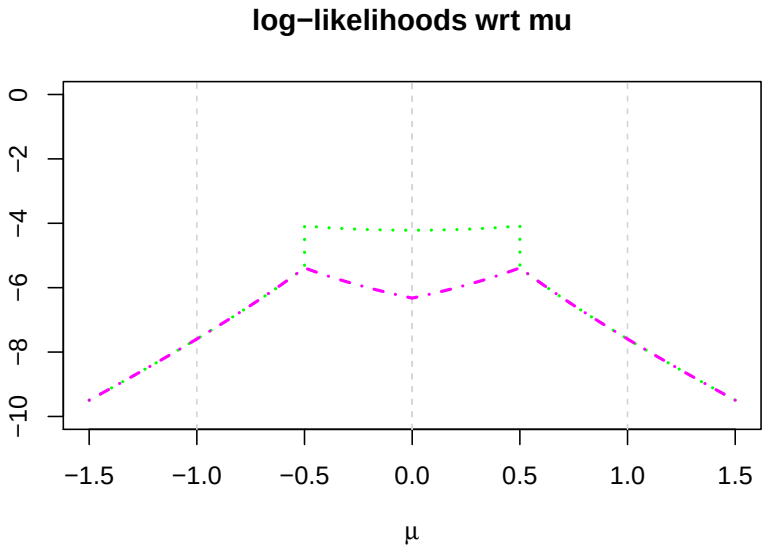
$$\ell^{\text{WLOO}}(\mu) = \sum_{i=1}^{4} w_i \log f(x_i; \mu)$$

with some suitably chosen weights such that the WLOO log-likelihood is bounded, continuous and data contribution consistent with the LOO log-likelihood. This means the weights need to satisfy the condition $\sum_{i=1}^{n} w_i = n-1$. For this example, the WLOO log-likelihood becomes

$$\ell^{\text{WLOO}}(\mu) = w_1 \log f(-1) + 2w_2 \log f(0) + w_3 \log f(1)$$



(a) Plot of full (solid red) and LOO (striped blue) log-likelihoods.



(b) Plot of LMO (dotted green) and WLOO (dot and striped magenta) log-likelihoods.

Figure 4.1. Plots of full, LOO, LMO, and WLOO log-likelihoods of univariate symmetric VG distribution with $\nu = 0.4$ for data set $\{-1, 0, 1, 0\}$ represented by light grey vertical strips.

where we let $w_2 = w_4$ due to the data multiplicity at 0. Since the density function is assumed to be symmetric, we define

$$g_i(|x_i - \mu|) = f(x_i; \mu).$$

To find the appropriate weights, we analyse the WLOO log-likelihood around the midpoints where the discontinuities occur. In particular, we look at the small neighbourhood around the midpoints located at -0.5 and 0.5. It is sufficient to look at one of the midpoints by the symmetry of the data.

Midpoint of 0 and 1: Let $\epsilon > 0$ be a small constant. On the right hand side of 0.5, the data point $x_3 = 1$ is closest to μ. So we leave out that data point in the WLOO likelihood by setting $w_3 = 0$. This gives us

$$\ell^{\text{WLOO}}(0.5 + \epsilon) = w_1 \log g_1(1.5 + \epsilon) + 2w_2 \log g_2(0.5 + \epsilon) + 0 \log g_3(0.5 - \epsilon).$$

Since the data point $x_1 = -1$ has a single contribution to the likelihood, we set $w_1 = 1$. This leaves the other weights $w_2 = w_4 = 1$ so that $\sum_{i=1}^{4} w_i = 3$.

On the left hand side of 0.5, the data points $x_2 = x_4 = 0$ is closest to μ. so we set $w_2 = w_4 = 0$ which gives us

$$\ell^{\text{WLOO}}(0.5 - \epsilon) = w_1 \log g_1(1.5 - \epsilon) + 0 \log g_2(0.5 - \epsilon) + w_3 \log g_3(0.5 + \epsilon). \tag{4.6}$$

Also since the data point $x_1 = -1$ has a single contribution to the likelihood, we set $w_1 = 1$. This leaves $w_3 = 2$ so that $\sum_{i=1}^{4} w_i = 3$. Choosing these weights gives us a continuous likelihood at the midpoint 0.5 with WLOO likelihood at $\mu = 0.5$

$$\ell^{\text{WLOO}}(0.5) = \log g_1(1.5) + 2 \log g_2(0.5) \tag{4.7}$$

where $g_2(|0 - 0.5|) = g_3(|1 - 0.5|)$.

One way to think about the chosen weights for the WLOO likelihood is that after leaving out data points with data multiplicity, extra weight is added to the neighbouring data points to compensate for the missing weights. The next example considers a data set with data multiplicities at two different data points.

4.3.2 Example 2: data multiplicities at two locations

In this second example, the data set $\{-1, 0, 1, 0, 0, 1\}$ have size 6 with data multiplicities at 0 and 1. As before we want to choose the weights so that the WLOO log-likelihood

$$\ell^{\text{WLOO}}(\mu) = w_1 \log f(-1) + 3w_2 \log f(0) + 2w_3 \log f(1)$$

is bounded, continuous at the midpoints and has data contribution such that $\sum_{i=1}^{n} w_i = n - 1$.

Midpoint of -1 and 0: Using the same argument as in the first example, the choice of weights should satisfy the bounded, continuity and $\sum_{i=1}^{n} w_i = n-1$ conditions. This gives us the WLOO log-likelihood

$$\ell^{\text{WLOO}}(-0.5) = 3 \log g_1(0.5) + 2 \log g_3(1.5).$$

Midpoint of 0 and 1: Unlike the previous midpoint, this midpoint is between two data points with different data multiplicities. On the right hand side of 0.5, the data point $x_3 = x_6 = 1$ is closer to μ, so we leave out these data points by setting $w_3 = w_6 = 0$. This gives us the WLOO log-likelihood

$$\ell^{\text{WLOO}}(0.5 + \epsilon) = w_1 \log g_1(1.5 + \epsilon) + 3w_2 \log g_2(0.5 + \epsilon) + 0 \log g_3(0.5 - \epsilon).$$

Based on the single contribution of $x_1 = -1$, we set $w_1 = 1$, and $w_2 = w_4 = w_5 = 4/3$ for the remaining weights so that $\sum_{i=1}^{6} w_i = 5$.

On the left hand side of 0.5, we leave out the data points $x_2 = x_4 = x_5 = 0$ by setting $w_2 = w_4 = w_5 = 0$. This gives us the WLOO log-likelihood

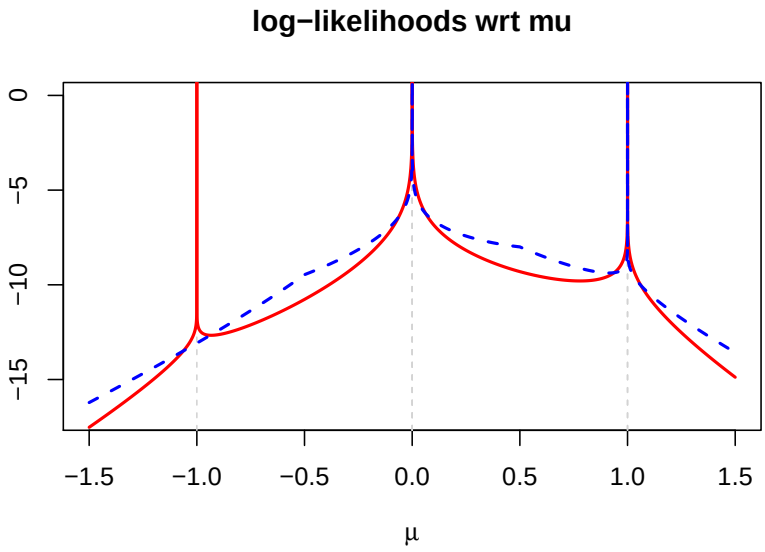
$$\ell^{\text{WLOO}}(0.5 - \epsilon) = w_1 \log g_1(1.5 - \epsilon) + 0w_2 \log g_2(0.5 - \epsilon) + 2w_3 \log g_3(0.5 + \epsilon).$$

We set $w_1 = 1$ for the single contribution, and $w_3 = w_6 = 4/2$ for the remaining weights.

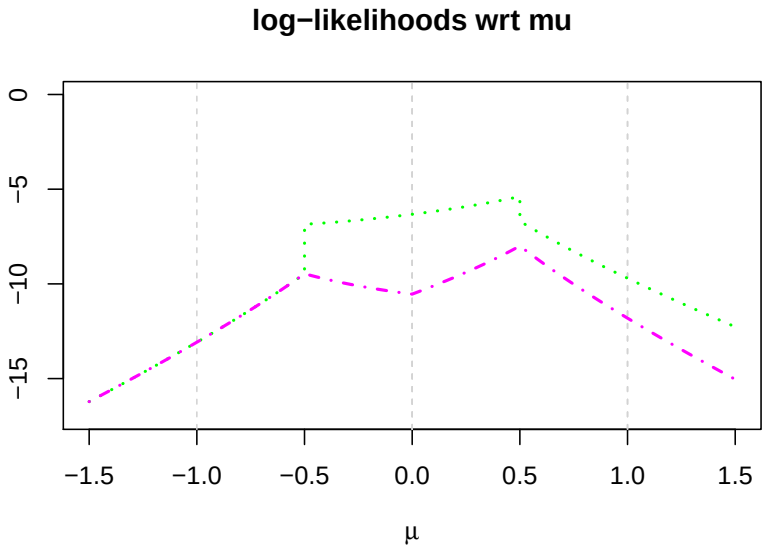
In the end, this gives us the WLOO log-likelihood of

$$\ell^{\text{WLOO}}(0.5) = \log g_1(1.5) + 4 \log g_2(0.5)$$

where $g_2(0.5) = g_3(0.5)$.



(a) Plot of full (solid red) and LOO (striped blue) log-likelihoods.



(b) Plot of LMO (dotted green) and WLOO (dot and striped magenta) log-likelihoods.

Figure 4.2. Plot of the full, LOO, LMO, and WLOO log-likelihoods of symmetric VG distribution with $\nu = 0.4$ for data set $\{-1, 0, 1, 0, 0, 1\}$ represented by light grey vertical strips.

4.3.3 Example 3: general data multiplicities at two locations

For the general case, we consider without loss of generality, the neighbourhood around the midpoint which is between two data points x_i, x_j with data multiplicities m_i, m_j.

Data points that are not around the neighbourhood of the midpoint have the same contribution to the WLOO likelihood regardless of whether the right or left side of the midpoint is considered, so these data points not around the neighbourhood of the midpoint always have weight of 1. Next, we consider the side closer to x_j. The data points with the same value as x_j would be left out. This is done by setting the weights for m_j data points of x_j to be 0. The weights corresponding to x_i is then set to be

$$w_i = (m_i + m_j - 1)/m_i$$

so that $\sum_{i=1}^{n} w_i = n - 1$. This gives us the formula for the weights in (4.3).

4.4 Simulation study

To assess the performance of the WLOO likelihood for data sets with data multiplicity, we conduct a simulation study to compare the WLOO likelihood method with several other likelihood methods proposed or discussed in Chapters 2 and 3, including the `R` package called `ghyp`. Some of these likelihood methods depend on different chosen cap regions called Δ as defined in (2.19) in Section 2.3.

The following likelihood methods are considered in this study:

(i) `ghyp` package: MCECM algorithm in Section 2.1 using the full likelihood with Δ set to `.Machine$double.eps^0.25` $\approx$ `1.2e-4` that is a tolerance level many `R` functions use.

(ii) Full likelihood: AECM algorithm in Section 2.2 using the full likelihood with the smallest positive Δ defined in Section 2.5.1 for some numerical stability. This method resembles the classical likelihood.

(iii) Adaptive Δ likelihood: AECM algorithm using the full likelihood with adaptive Δ as described in Section 2.5.2.

(iv) LOO likelihood: AECM algorithm using the LOO likelihood along with the local point search (LPS) 3.3.2.2 and line search 3.3.2.4 with the smallest positive Δ. The Δ is relevant when there is data multiplicity.

(v) WLOO likelihood: AECM algorithm using the WLOO likelihood, LPS and line search with smallest positive Δ. This Δ is irrelevant since the WLOO takes care of the unbounded likelihood caused by data multiplicity.

Details of these five likelihood methods are summarised in Table 4.1.

Table 4.1. Summary of `ghyp`, full, adaptive Δ, LOO and WLOO likelihood methods.

Names	ECM algorithm	likelihood	Δ
`ghyp` package	MCECM	full	`1.2e-4`
Full	AECM	full	`1.5e-154`
Adaptive Δ	AECM	full	adaptive
LOO	AECM with LPS & line search	LOO	`1.5e-154`
WLOO	AECM with LPS & line search	WLOO	`1.5e-154`

To conduct the simulation study, we create data multiplicity in two ways: replicate each data point R times or round each data point to D decimal places. The procedure for the simulation study are summarised below:

Step 1: Choose $\nu = 0.05, 0.1, \dots, 1.45, 1.5$ and $R = 1, \dots, 5$ or $D = \infty, 8, 6, 5, 4$ for Sections 4.4.1 and 4.4.2 respectively. We remark that $\nu \leq 1$ gives rise to unbounded density and $1 \leq \nu \leq 1.5$ is cusped for dimension $d = 2$.

Step 2: Simulate $n = 1000$ data from the bivariate VG distribution with true parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\gamma}$ given in (3.14) and the chosen ν. Then repeat each data point R times, or round each data point to D decimal places.

Step 3: Apply the five different likelihood methods to the data sets to obtain five different sets of estimates.

Step 4: Repeat these steps until we have 1000 replicates for each method, each level of ν, and each level of R or D.

Likelihood methods are compared based on some measures of accuracy for parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, $\boldsymbol{\gamma}$ and ν. For vectors on $\mathbb{R}^d$ such as $\boldsymbol{\mu}$, the accuracy is measured by $|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}|$. For positive scalars such as ν, the accuracy is measured on the logarithmic scale by $|\log \hat{\nu} -$

$\log \nu|$. For positive definite matrices such as $\boldsymbol{\Sigma}$, we can consider the determinant $|\cdot|$ which becomes a positive scalar, and so the accuracy is measured by $\big|\log|\hat{\boldsymbol{\Sigma}}| - \log|\boldsymbol{\Sigma}|\big|$. The median of these measure of accuracy are reported in Table 4.2 only for the regular case with no replication or rounding.

To visualise the differences in performance across different likelihood methods, we consider some transformations of the accuracy measures reported in Table 4.2 and are given by:

$$\begin{aligned}
\hat{\boldsymbol{\mu}}&: -\log\Big[\log\big(-\log|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}|\big)\Big],\\
\hat{\boldsymbol{\Sigma}}&: \log\big|\log|\hat{\boldsymbol{\Sigma}}| - \log|\boldsymbol{\Sigma}|\big|,\\
\hat{\boldsymbol{\gamma}}&: \log|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}|,\\
\hat{\nu}&: |\log\hat{\nu} - \log\nu|.
\end{aligned}$$

We remark that since the results for $\hat{\boldsymbol{\mu}}$ vary substantially in Table 4.2, we adopt the transformation $f(x) = -\log[\log(-\log x)]$ which is a monotonically increasing function defined on $f : (0, 1/e) \to \mathbb{R}$. These transformed accuracy measures are graphed in Figures 4.3 and 4.4 for replicated and rounded data points respectively. For some ν and parameter estimate, smaller transformed accuracy measures indicate better accuracies. So the likelihood method that comparatively have smaller transformed accuracy measures for a wide range of ν and parameter estimates are preferred.

4.4.1 Results for data multiplicity due to repetition

Figure 4.3 presents these accuracy measures onto a transformed scale as previously mentioned to facilitate comparison. These figures show that the WLOO likelihood generally performs better than other likelihoods for different levels of ν. The adaptive Δ and `ghyp` seem to provide reasonable accuracy when $\nu > 0.2$. When ν is small, the accuracy of `ghyp` becomes very poor as expected since the Δ is fixed to a relatively higher level of 1.2e-4 as reported in Table 4.1. The full and LOO likelihoods has the worst performance showing that they are sensitive to data multiplicity.

Comparing across likelihood methods, there is not much variation in the performance of $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\gamma}}$ except when ν is very small. For $\hat{\boldsymbol{\Sigma}}$ and $\hat{\nu}$, the variations of the median estimates between each likelihood method are much greater. As ν approaches 1.5, $\hat{\boldsymbol{\Sigma}}$ and $\hat{\nu}$ roughly approach the same value for each likelihood method. The accuracy for

Table 4.2. Median of 500 accuracy measures of parameter estimates across five likelihood methods with no data multiplicity ($R = 1$).

Measures	Likelihoods	$\nu = 0.05$	$\nu = 0.1$	$\nu = 0.2$	$\nu = 0.5$	$\nu = 0.75$	$\nu = 1$	$\nu = 1.5$
$\lvert\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\rvert$	`ghyp`	1.2e-10	9.3e-10	7.7e-7	0.0067	0.020	0.031	0.050
	Full	3.5e-28	7.1e-14	6.8e-7	0.0054	0.017	0.030	0.049
	Adaptive Δ	3.8e-28	7.7e-14	6.8e-7	0.0036	0.012	0.026	0.049
	LOO	1.6e-28	1.5e-14	1.1e-7	0.0014	0.012	0.027	0.050
	WLOO	1.6e-28	1.5e-14	1.1e-7	0.0014	0.012	0.027	0.050
$\log\lvert\hat{\boldsymbol{\Sigma}}\rvert - \log\lvert\boldsymbol{\Sigma}\rvert$	`ghyp`	-3.943	-0.601	-0.038	-0.016	0.009	0.023	-0.013
	Full	-0.103	0.010	0.117	0.339	0.472	0.549	-0.012
	Adaptive Δ	-0.143	-0.059	-0.048	-0.027	-0.010	-0.007	-0.012
	LOO	-0.231	-0.126	-0.074	-0.026	-0.015	-0.009	-0.011
	WLOO	-0.232	-0.126	-0.074	-0.026	-0.015	-0.009	-0.011
$\lvert\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\rvert$	`ghyp`	0.243	0.060	0.038	0.038	0.043	0.048	0.059
	Full	0.048	0.043	0.038	0.037	0.039	0.048	0.059
	Adaptive Δ	0.048	0.042	0.039	0.037	0.038	0.045	0.059
	LOO	0.061	0.050	0.041	0.036	0.038	0.045	0.059
	WLOO	0.061	0.050	0.041	0.036	0.038	0.045	0.059
$\log\hat{\nu} - \log\nu$	`ghyp`	0.1376	0.0368	0.0067	-0.0047	-0.0418	-0.0809	0.0033
	Full	-0.0259	-0.0667	-0.1571	-0.4636	-0.7097	-0.9199	0.0022
	Adaptive Δ	0.0052	0.0055	0.0091	0.0103	0.0024	0.0010	0.0033
	LOO	0.0092	0.0093	0.0110	0.0123	0.0142	0.0090	0.0167
	WLOO	0.0092	0.0093	0.0110	0.0123	0.0142	0.0090	0.0167

the full and LOO likelihoods are generally quite poor when there are repeated data points since they both adopt a fixed capping level. The only difference is that the LOO likelihood leaves out a data point while the full likelihood does not. However, as the repetition of data points increases, leaving a data point out becomes insignificant and so the results of the LOO likelihood converge to the full likelihood.

Comparing across different R, the graphs look similar. In fact, repeating data points only changes the scale of the classical log-likelihood, that is,

$$\sum_{i=1}^{n} R \log f(x_i) = R \sum_{i=1}^{n} \log f(x_i).$$

So the estimates using the classical likelihoods such as the full and `ghyp` should not change for different R. Similarly, the WLOO likelihood should not change for different R since it leaves out data points based on its data multiplicity. However, small changes in the figure for each R are possibly due to sampling errors.

In practice, it is unclear the range ν will fall into in real application. Hence, it is better to use the WLOO likelihood which provides the best overall performance.

4.4.2 Results for data multiplicity due to rounding

By rounding the data, we have effectively changed the data distribution. Nevertheless, the accuracy measures still serves as a guide to assess the performance of the likelihood methods. The trends in Figure 4.4 are somewhat similar to those in Figure 4.3 in general but there are also some clear differences. The LOO likelihood is now much better than the full likelihood when $\nu > 0.3$. While the WLOO likelihood is generally still better for most cases, it is generally less accurate than the `ghyp` package when $\nu < 0.5$ and is also less accurate than the adaptive Δ likelihood when $\nu > 0.5$.

We remark that for larger ν, the data rounding is insignificant because of the lower peak and hence less data multiplicity around $\hat{\boldsymbol{\mu}}$. For $\hat{\boldsymbol{\mu}}$, the behaviour is similar across likelihoods for each rounding when $\nu > 0.3$. However when $\nu < 0.3$, the accuracy of WLOO likelihood stabilises while other likelihoods improve substantially as rounding increases.

For $\hat{\boldsymbol{\Sigma}}$, all the likelihood methods apart from the full likelihood appear to be similar and perform reasonably well whereas the full likelihood performs worse. With rounding, the `ghyp` package and adaptive Δ likelihood performs similarly and the WLOO likelihood performs slightly better for $\nu < 0.2$.

For each parameter estimate, the LOO and WLOO likelihoods are similar for larger ν but vary for smaller ν. The LOO and full likelihood also performs similarly for smaller ν. The full likelihood performs well at around $\nu = 1.4$, whereas the adaptive

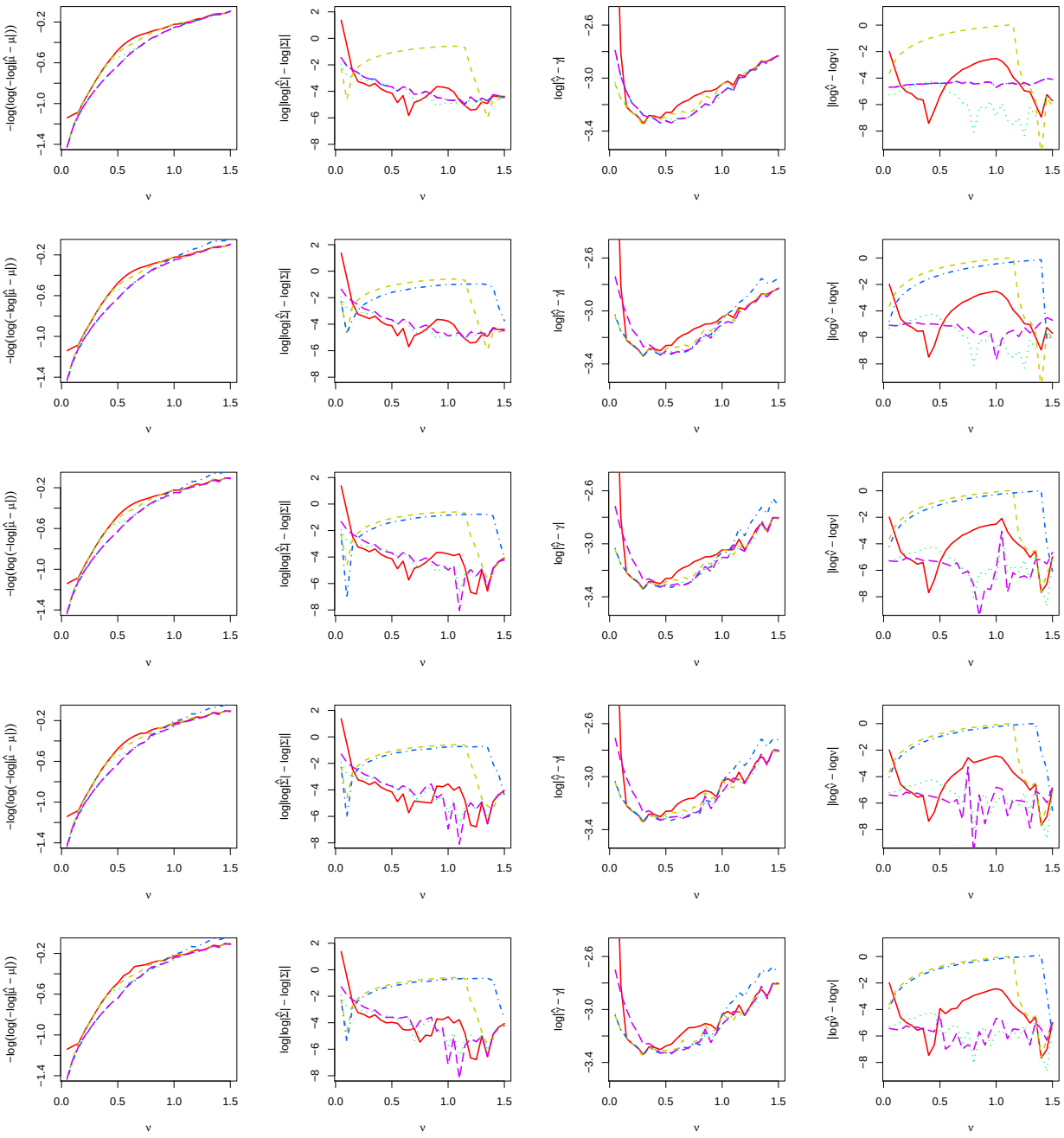


Figure 4.3. Plot of transformed accuracy measures for `ghyp` package (red solid line), full and smallest Δ (light green striped line), full and adaptive Δ (dotted green line), LOO (blue dot & striped line), and WLOO (dash magenta line). The columns from left to right represents the median parameter accuracy measures for $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\gamma}}, \hat{\nu})$ respectively. The rows from top to bottom represents $R = 1, \ldots, 5$ respectively where R is the number of times each data point is repeated.

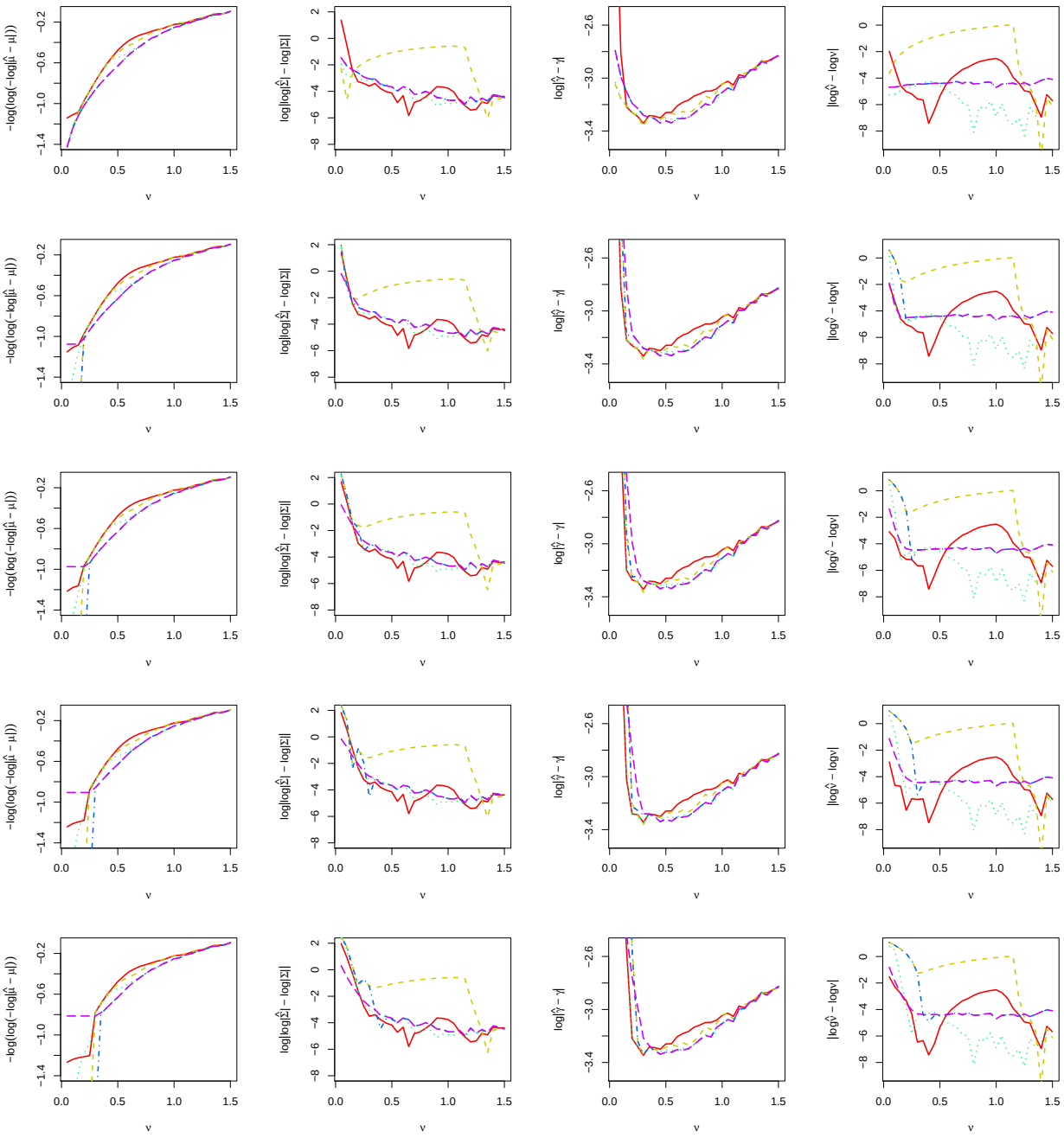


Figure 4.4. Plot of transformed accuracy measures for `ghyp` package (red solid line), full and smallest Δ (light green striped line), full and adaptive Δ (dotted green line), LOO (blue dot & striped line), and WLOO (dash magenta line). The columns from left to right represents the median parameter accuracy measures for $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\gamma}}, \hat{\nu})$ respectively. The rows from top to bottom represents $D = \infty, 8, 6, 5, 4$ respectively where D is the number of decimal places to be rounded off.

Δ likelihood performs well in the mid-range at around $0.6 < \nu < 1.2$ and the `ghyp` package performs well at around $0.2 < \nu < 0.5$.

Overall, for both simulation studies with data multiplicity due to repeated values or rounding, we have demonstrated in Section 2.5.1 that capping the density for the cusp and unbounded density cases improves the performance of the AECM algorithm. The fact that the full likelihood generally performs worse in this simulation study suggests that the full likelihood with a fixed capping level only provides a temporary solution to deal with the unbounded likelihood with data multiplicity. This full likelihood method can be improved by providing an adaptive capping level in Section 2.5.2. On the other hand, the LOO likelihood method provides a better alternative for dealing with cusp and unbounded density. Moreover, the LOO likelihood method can be extended to the WLOO likelihood to prevent the unbounded likelihood due to data multiplicity.

4.5 Conclusion

We propose the WLOO likelihood to estimate the parameters of the VG distribution when the likelihood function is unbounded and the data set has repeated data points. Without data multiplicity, the WLOO likelihood is equivalent to the LOO likelihood. When there are repeated data points, then leaving out a single data point in the LOO likelihood is not enough to remove the unbounded likelihood as there are still other data points that contribute to the unbounded likelihood. We illustrate through three examples the way to choose suitable weights for the WLOO likelihood so that it does not only remove the unbounded likelihood, but also preserves continuity at the midpoints as well as data contribution consistent with the LOO likelihood. In the simulation study, we compare the WLOO likelihood with other likelihoods using the AECM algorithm for data sets where we artificially create data multiplicity by either repeating or rounding each data point. Overall, the WLOO likelihood gives the most stable and accurate results.

In summary, it is important to address the data multiplicity issue when applying the LOO likelihood as the issue is likely to occur for three reasons. Firstly, data rounding is common in practice simply for convenience or for reducing data storage. To fit this kind of discretised data, one may consider the order probit or logit models but for

a continuous model like the VG distribution, there are non-ignorable chance of data multiplicity. Secondly, as high frequency data are more prevalence in these recent years, the larger data size also increases the chance of data multiplicity. Lastly, as the measurement is made more instantaneously for high frequency data, the price changes will be minimal and hence return will be extremely small or even zero again giving rise to data multiplicity. We demonstrate the applicability of WLOO likelihood through two real applications in Chapter 5 when the model parameters with time series structure are estimated using the WLOO likelihood. Although the chance of data multiplicity is actually lower when using returns for time series models, when the means from the time series model change over time and when using multi-dimensional models, one will never be sure in practice if data multiplicity exists and so it is always advisable to consider using the WLOO likelihood to provide numerically stable estimates.

CHAPTER 5

Applications to Financial Time Series

5.1 Introduction

After proposing various methodologies for the estimation of the VG distribution, this chapter focus on applications to solve real financial problems. As financial time series often display autocorrelation and high kurtosis, we address these issues by extending the VG distribution with constant mean in Chapter 2 to adopt some time series structures.

From a modelling perspective, the vector autoregressive moving average (VARMA) model have been widely considered in many fields such as econometrics, dynamical systems, and finance (see [32, 87] for other examples) as it allows for a parsimonious description of stationary stochastic processes while also modelling the dependence structure between different components. The subclasses of the VARMA model such as the vector autoregressive (VAR) and the univariate autoregressive moving average (ARMA) model are also popular due to its simplicity and easy interpretability along with many other desirable properties. There are a rich literature analysing properties of the VARMA model, including identifiability, causality and invertibility. We provide a review of these properties in Section 5.3.2. See also [44, 69, 103, 104] for a brief overview of these models.

A common assumption with these models is that the innovations follow a Gaussian distribution. Estimation methods of the VAR model with normal innovations include the generalised least squares (GLS) [110] and Bayesian methods [103]. However, in financial markets, the distributions of asset prices tend to have kurtosis much higher than the Gaussian distribution [25, 33, 73]. This feature is especially prominent when

looking at high frequency returns data [57]. Not capturing the extra leptokurtosis can seriously affect the prediction of risk in asset forecasts. To account for the extra leptokurtosis, we propose the VARMA model with VG innovations called the VARMA-VG model. We also discuss the challenges faced by implementing the VARMA-VG model, in ways similar to the VARMA model. In addition, we also consider VARMA model with Student's t innovations as the Student's t distribution is very popular in financial time series modelling. This allows comparison of performance between the two models in Section 5.6.

There are many problems when implementing the MLE for the VARMA model as there is no closed-form solution making the maximisation more complicated. This has led many researchers to find various approximation techniques to remedy this problem. One popular technique is the two-stage approximation method to first estimate the error terms by fitting the series with a high order VAR model and then use these fitted errors to estimate the model parameters using the GLS method. See [26, 45, 59, 92, 107] for the approximate MLE based on the GLS method and its extensions. Other methods include the EM algorithm using a state-space representation [80], and the structured matrix norm optimisation method for the stochastic multivariate ARMA model to approximate the VARMA model [106].

Apart from the various ML approaches, the Bayesian paradigm is getting popular in recent years. Particularly, for some complicated models such as the VARMA-VG and VARMA-t, it can avoid the problem of maximising the log-likelihood function and replace it with posterior sampling. However, it also has several disadvantages in estimating the VARMA-VG and VARMA-t models. Firstly, the running of MCMC can be very computationally demanding with slow convergence for some complicated models. Secondly, the specification of the prior distributions is not straight forward and maybe subject to debate. Lastly, there is no guarantee that the parameters sampled from MCMC satisfy the causality and invertibility conditions.

In this chapter, we choose to adopt the EM algorithm and extend it to estimate the additional parameters involving the ARMA mean structure. One challenge in this extension is the non-existence of a closed-form solution. To handle this problem, we follow the idea of [104] to adopt the two-stage approximation method, based on the ML method instead of the GLS method. This method works well for the case when the

true parameters are far from the non-causal, and non-invertible region. Further details are given in Section 5.3.3.

We illustrate the applicability of the VARMA-VG and VARMA-t models by analysing returns of high frequency market indices as well as returns of cryptocurrency exchanges. Due to the advances in computer capacity and storage, price movements in stock market are captured nearly instantaneously. Cryptocurrency market recently received a lot of attention and thus it has very limited market share in the currency exchange market. Currently, studies into the characteristics of cryptocurrency are very limited. In particular, factors such as high observed frequency and small market share may give rise to volatile returns. We investigate how the VG innovations can describe the features of the volatile returns by lowering the shape parameter to capture high kurtosis. We also compare the performance of the VARMA-VG model to the VARMA-t model and highlight the advantages of the VARMA-VG model.

In summary, this chapter provides a useful illustration of the applicability of our proposed VARMA-VG model and its implementation using the AECM algorithm. Sections 5.2 and 5.3 report these details for the VAR-VG and VARMA-VG models and how the ECM algorithm developed in the previous chapters can be extended to estimate the additional parameters in the ARMA mean function. Section 5.4 extends the methodologies for Student's t innovations. Section 5.5 assess the performance of the AECM algorithm for VARMA-VG though simulation studies with various choices of skewness and shape parameters including cases when the density is unbounded. We also study the identifiability issue with the AR and MA parameters and provide the SE calculation for all parameters. Section 5.6 demonstrates the application of the VARMA-VG model by empirically studying the stock indices and cryptocurrency returns and compare the performance with the VARMA-t model. We conclude our contribution and discuss further extensions in Section 5.7.

5.2 Estimation of VAR-VG model

In this section, we introduce the VAR-VG model and develop the AECM algorithm to estimate parameters of the model.

5.2.1 VAR-VG model

Suppose the d-dimensional time series $\{\boldsymbol{y}_t\}$ follow a VAR-VG model of order p denoted by VAR(p)-VG. Then we can represent the series as

$$\boldsymbol{y}_t = \boldsymbol{c} + \boldsymbol{A}_1\boldsymbol{y}_{t-1} + \ldots + \boldsymbol{A}_p\boldsymbol{y}_{t-p} + \boldsymbol{\varepsilon}_t \tag{5.1}$$

for $t = 1, \ldots, n$ where the previous observations $\{\boldsymbol{y}_{1-p}, \ldots, \boldsymbol{y}_0\}$ are assumed to be observed, $\boldsymbol{\varepsilon}_t \sim \mathcal{VG}_d(-\boldsymbol{\gamma}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}, \nu)$ so that $\mathbb{E}(\boldsymbol{\varepsilon}_t) = \boldsymbol{0}$, $\boldsymbol{c} \in R^d$, and $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_p$ are $d \times d$ matrix coefficients for the AR terms. Alternatively, equation (5.1) can be rewritten as

$$\boldsymbol{y}_t' = \boldsymbol{x}_t'\boldsymbol{\beta} + (\boldsymbol{\varepsilon}_t + \boldsymbol{\gamma})' \tag{5.2}$$

where

$$\begin{aligned}
\boldsymbol{x}_t' &= \begin{pmatrix} 1 & \boldsymbol{y}_{t-1}' & \cdots & \boldsymbol{y}_{t-p}' \end{pmatrix} \text{ is a } (dp+1) \text{ vector,} \\
\boldsymbol{\beta}' &= \begin{pmatrix} \boldsymbol{\mu} & \boldsymbol{A}_1 & \cdots & \boldsymbol{A}_p \end{pmatrix} \text{ is a } d \times (dp+1) \text{ matrix, and} \\
\boldsymbol{\mu} &= \boldsymbol{c} - \boldsymbol{\gamma}\,.
\end{aligned}$$

Equivalently, we can also write model (5.1) as

$$\boldsymbol{y}_t | \mathscr{F}_{t-1} \sim \mathcal{VG}_d(\boldsymbol{\beta}'\boldsymbol{x}_t, \boldsymbol{\Sigma}, \boldsymbol{\gamma}, \nu) \tag{5.3}$$

where $\mathscr{F}_t = \{\boldsymbol{y}_s : s \leq t\}$ represents the filtration (or information) up to time t.

To represent the model (5.1) using matrices, we first define the following matrices

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1' \\ \vdots \\ \boldsymbol{x}_n' \end{pmatrix}, \; \boldsymbol{Y} = \begin{pmatrix} \boldsymbol{y}_1' \\ \vdots \\ \boldsymbol{y}_n' \end{pmatrix}, \; \boldsymbol{\mathcal{E}} = \begin{pmatrix} \boldsymbol{\varepsilon}_1' \\ \vdots \\ \boldsymbol{\varepsilon}_n' \end{pmatrix} \tag{5.4}$$

where $\boldsymbol{X}$ has dimensions $n \times (dp+1)$, whereas $\boldsymbol{Y}$ and $\boldsymbol{\mathcal{E}}$ has dimensions $n \times d$. Then with these matrices, we can write the whole model as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \mathbb{1}_n\boldsymbol{\gamma}' + \boldsymbol{\mathcal{E}} \tag{5.5}$$

where $\mathbb{1}_n$ is a n-dimensional column vector of ones.

5.2.2 Likelihood functions for VAR-VG model

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}, \nu)$ and $\boldsymbol{y} = (\boldsymbol{y}_1, \dots, \boldsymbol{y}_n)$, then the (conditional) observed log-likelihood function

$$\ell(\boldsymbol{\theta}; \boldsymbol{y} | \mathscr{F}_0) = \sum_{t=1}^{n} \log f_{VG}(\boldsymbol{y}_t | \mathscr{F}_{t-1}; \boldsymbol{\theta}) \tag{5.6}$$

where $\mathscr{F}_0 = \{\boldsymbol{y}_s : 1 - p \leq s \leq 0\}$ and $f_{VG}(\cdot)$ denotes the pdf of the VG distribution in (1.36). Using the NMVM representation of the VG distribution in (1.37), we can decompose the (conditional) complete data log-likelihood (ignoring additive constants) as follows

$$\ell(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{u} | \mathscr{F}_0) = \ell_N(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}; \boldsymbol{y}, \boldsymbol{u} | \mathscr{F}_0) + \ell_G(\nu; \boldsymbol{u}) \tag{5.7}$$

where $\boldsymbol{u} = \{u_1, \dots, u_n\}$ and the (conditional) log-likelihood for the conditional normal distribution is given by

$$\ell_N(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}; \boldsymbol{y}, \boldsymbol{u} | \mathscr{F}_0) = -\frac{n}{2} \log |\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{t=1}^{n} \frac{1}{u_t} (\boldsymbol{y}_t - \boldsymbol{\beta}' \boldsymbol{x}_t - u_t \boldsymbol{\gamma})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}_t - \boldsymbol{\beta}' \boldsymbol{x}_t - u_t \boldsymbol{\gamma}) \tag{5.8}$$

and the log-likelihood for the gamma distribution is given by

$$\ell_G(\nu; \boldsymbol{u}) = n\nu \log \nu - n \log \Gamma(\nu) + (\nu - 1) \sum_{t=1}^{n} \log u_t - \nu \sum_{t=1}^{n} u_t \,. \tag{5.9}$$

This decomposition allows for the implementation of the EM algorithm that is discussed in the later sections.

5.2.3 E-step

The conditional distribution of u_t given $\mathscr{F}_t$ has pdf

$$f(u_t | \mathscr{F}_t) \propto u_t^{\nu - d/2 - 1} \exp\left(-\frac{1}{2u_t} z_t^2 - \frac{u_t}{2} \left(2\nu + \boldsymbol{\gamma}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\gamma} \right) \right) \tag{5.10}$$

for $t = 1, \dots, n$ which corresponds to the $\mathcal{GIG}(\nu - d/2, z_t^2, 2\nu + \boldsymbol{\gamma}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\gamma})$ distribution in Section 1.5.1 with

$$z_t^2 = (\boldsymbol{y}_t - \boldsymbol{\beta}' \boldsymbol{x}_t)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}_t - \boldsymbol{\beta}' \boldsymbol{x}_t) \,. \tag{5.11}$$

Refer to the E-step in Section 2.1.1 for the relevant conditional expectations.

5.2.4 CM-step for $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\gamma}$

In order to obtain the parameter estimates that maximise the complete data log-likelihood function in (5.8) we differentiate the conditional normal log-likelihood using the matrix derivative results in Section A9. This gives us the following first order derivatives for the conditional normal log-likelihood

$$\frac{\partial \ell_N}{\partial \boldsymbol{\beta}'} = \boldsymbol{\Sigma}^{-1} \sum_{t=1}^{n} \frac{1}{u_t} (\boldsymbol{y}_t - \boldsymbol{\beta}' \boldsymbol{x}_t - u_t \boldsymbol{\gamma}) \boldsymbol{x}_t' \,, \tag{5.12}$$

$$\frac{\partial \ell_N}{\partial \boldsymbol{\gamma}} = \boldsymbol{\Sigma}^{-1} \sum_{t=1}^{n} (\boldsymbol{y}_t - \boldsymbol{\beta}' \boldsymbol{x}_t - u_t \boldsymbol{\gamma}) \,, \tag{5.13}$$

$$\frac{\partial \ell_N}{\partial \boldsymbol{\Sigma}} = \boldsymbol{D}_d^\top \left(-\frac{n}{2} \boldsymbol{\Sigma}^{-1} + \frac{1}{2} \boldsymbol{\Sigma}^{-1} S_{\tilde{\boldsymbol{y}}\tilde{\boldsymbol{y}}/u} \boldsymbol{\Sigma}^{-1} \right) , \tag{5.14}$$

where $\boldsymbol{D}_d$ represents the duplication matrix (A.7) and

$$S_{\tilde{\boldsymbol{y}}\tilde{\boldsymbol{y}}/u} = \sum_{t=1}^{n} \frac{1}{u_t} (\boldsymbol{y}_t - \boldsymbol{\beta}' \boldsymbol{x}_t - u_t \boldsymbol{\gamma})(\boldsymbol{y}_t - \boldsymbol{\beta}' \boldsymbol{x}_t - u_t \boldsymbol{\gamma})' . \tag{5.15}$$

Setting $\dfrac{\partial \ell_N}{\partial \boldsymbol{\beta}} = \boldsymbol{0}$, and $\dfrac{\partial \ell_N}{\partial \boldsymbol{\gamma}'} = \boldsymbol{0}$ gives us

$$\sum_{t=1}^{n} \frac{1}{u_t} \boldsymbol{x}_t \boldsymbol{x}_t' \boldsymbol{\beta} + \sum_{t=1}^{n} \boldsymbol{x}_t \boldsymbol{\gamma}' = \sum_{t=1}^{n} \frac{1}{u_t} \boldsymbol{x}_t \boldsymbol{y}_t' \,, \tag{5.16}$$

and

$$\sum_{t=1}^{n} \boldsymbol{x}_t' \boldsymbol{\beta} + \sum_{t=1}^{n} u_t \boldsymbol{\gamma}' = \sum_{t=1}^{n} \boldsymbol{y}_t' \,. \tag{5.17}$$

Alternatively, representing (5.16) and (5.17) as matrices gives us

$$\begin{pmatrix} \boldsymbol{X}' \boldsymbol{U}^{-1} \boldsymbol{X} & \boldsymbol{X}' \mathbb{1}_n \\ \mathbb{1}_n' \boldsymbol{X} & \mathbb{1}_n' \boldsymbol{u} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma}' \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}' \boldsymbol{U}^{-1} \boldsymbol{Y} \\ \mathbb{1}_n' \boldsymbol{Y} \end{pmatrix} \tag{5.18}$$

where $\boldsymbol{u} = (u_1, \dots, u_n)$ and $\boldsymbol{U}^{-1} = \begin{pmatrix} \frac{1}{u_1} & & \boldsymbol{0} \\ & \ddots & \\ \boldsymbol{0} & & \frac{1}{u_n} \end{pmatrix}$.

Thus, solving for the first derivative gives us

$$\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\gamma}}' \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}'\boldsymbol{U}^{-1}\boldsymbol{X} & \boldsymbol{X}'\mathbb{1}_n \\ \mathbb{1}_n'\boldsymbol{X} & \mathbb{1}_n'\boldsymbol{u} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{X}'\boldsymbol{U}^{-1}\boldsymbol{Y} \\ \mathbb{1}_n'\boldsymbol{Y} \end{pmatrix}. \tag{5.19}$$

Next, solving for $\dfrac{\partial \ell_N}{\partial \boldsymbol{\Sigma}} = \mathbf{0}$ gives us

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{t=1}^{n} \frac{1}{u_t}(\boldsymbol{y}_t - \boldsymbol{\beta}'\boldsymbol{x}_t)(\boldsymbol{y}_t - \boldsymbol{\beta}'\boldsymbol{x}_t)' - \frac{1}{n}\boldsymbol{\gamma}\boldsymbol{\gamma}'\sum_{t=1}^{n} u_t\,. \tag{5.20}$$

5.2.5 CM-step for ν

Estimation for ν can be obtained in the same way using the log-likelihood of the gamma distribution in (5.9) or the observed log-likelihood in (5.6). For further details, see Section 2.1.

5.2.6 Summary of ECME algorithm

We present here the ECME algorithm to estimate parameters of the VAR-VG model using the classical likelihood. Other extensions such as the AECM algorithm in Section 2.2 or the LOO log-likelihood in Section 3.3 can be adopted into the algorithm.

Initialisation step: We first initialise the algorithm by choosing suitable initial parameter estimates

$$\left(\boldsymbol{c}^{(0)}, \boldsymbol{A}_1^{(0)}, \dots, \boldsymbol{A}_p^{(0)}, \boldsymbol{\Sigma}^{(0)}, \boldsymbol{\gamma}^{(0)}, \nu^{(0)}\right) = (\bar{\boldsymbol{y}}, \mathbf{0}, \dots, \mathbf{0}, \operatorname{cov}(\boldsymbol{y}), \mathbf{0}, d+3). \tag{5.21}$$

At the k^{th} iteration with current estimates $(\boldsymbol{\beta}^{(k)}, \boldsymbol{\Sigma}^{(k)}, \boldsymbol{\gamma}^{(k)}, \nu^{(k)})$:

E-step: Calculate $\mathbb{E}(u_t|\mathscr{F}_t)$ and $\mathbb{E}(\frac{1}{u_t}|\mathscr{F}_t)$ for $t = 1, \dots, n$ using the conditional distribution in Section 5.2.3.

CM-step 1: Update the parameters $(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\gamma})$ using (5.19) and (5.20).

CM-step 2: Update the parameter ν by maximising (5.6).

Stopping rule: Repeat the procedure until the algorithm converges.

Algorithm 11: ECME algorithm for VAR-VG model

Input: Initial value $\boldsymbol{\theta}^{(0)}$
while $\ell(\boldsymbol{\theta}^{(k+1)}; \boldsymbol{y}|\mathscr{F}_0) - \ell(\boldsymbol{\theta}^{(k)}; \boldsymbol{y}|\mathscr{F}_0) > \delta$ **do**
  $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}) \leftarrow \mathbb{E}_{\boldsymbol{\theta}^{(k)}}[\ell(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{u})|\mathscr{F}_n]$;
  $\boldsymbol{\theta}^{(k+1/2)} \leftarrow \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}_1}{\operatorname{argmax}}\, Q(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}, \nu^{(k)}; \boldsymbol{\theta}^{(k)})$;
  $\boldsymbol{\theta}^{(k+1)} \leftarrow \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}_2}{\operatorname{argmax}}\, \ell(\boldsymbol{\beta}^{(k+1/2)}, \boldsymbol{\Sigma}^{(k+1/2)}, \boldsymbol{\gamma}^{(k+1/2)}, \nu; \boldsymbol{y}|\mathscr{F}_0)$;
end

5.3 Estimation of VARMA-VG model

In this section, we generalise the VAR-VG model by including moving average (MA) components into the mean function. Unlike the VAR-VG model, the CM-step for $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ in general does not have closed-form solution, and model non-identifiability can occur which may cause problems in the estimation procedure, statistical inference and interpretability of the model. Here we provide a brief overview of the properties of VARMA model and the ways to handle these problems.

5.3.1 VARMA-VG model

Let $\{\boldsymbol{y}_t\}$ be a d-dimensional time series. It follows a VARMA(p, q)-VG process if it has the following structure,

$$\boldsymbol{y}_t = \boldsymbol{c} + \boldsymbol{A}_1\boldsymbol{y}_{t-1} + \cdots + \boldsymbol{A}_p\boldsymbol{y}_{t-p} - \boldsymbol{B}_1\boldsymbol{\varepsilon}_t - \cdots - \boldsymbol{B}_q\boldsymbol{\varepsilon}_{t-q} + \boldsymbol{\varepsilon}_t \tag{5.22}$$

for $t = 1, \ldots, n$ where the previous observations $\{\boldsymbol{y}_{1-\max(p,q)}, \ldots, \boldsymbol{y}_0\}$ are assumed to be observed, $\boldsymbol{\varepsilon}_t \sim \mathcal{VG}_d(-\boldsymbol{\gamma}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}, \nu)$, $\boldsymbol{c} \in \mathbb{R}^d$, $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_p$ are the AR coefficient matrices with dimensions $d \times d$, and $\boldsymbol{B}_1, \ldots, \boldsymbol{B}_q$ are the MA coefficient matrices with dimensions $d \times d$. We can summarise the model in (5.22) as

$$\boldsymbol{\mathcal{A}}(L)\boldsymbol{y}_t = \boldsymbol{\mathcal{B}}(L)\boldsymbol{\varepsilon}_t \tag{5.23}$$

where L represents the lag operator, and $\boldsymbol{\mathcal{A}}(z)$ and $\boldsymbol{\mathcal{B}}(z)$ are matrix polynomial functions defined by

$$\boldsymbol{\mathcal{A}}(z) = \boldsymbol{I}_d - \boldsymbol{A}_1 z - \ldots - \boldsymbol{A}_p z^p, \tag{5.24}$$

$$\boldsymbol{\mathcal{B}}(z) = \boldsymbol{I}_d - \boldsymbol{B}_1 z - \ldots - \boldsymbol{B}_q z^q, \tag{5.25}$$

respectively, for any complex number z with no common factors.

Similar to the VAR-VG model, the VARMA-VG model can be alternatively expressed as

$$\boldsymbol{y}_t' = \boldsymbol{x}_t'\boldsymbol{\beta} + (\boldsymbol{\varepsilon}_t + \boldsymbol{\gamma})' \tag{5.26}$$

for $t = 1, \dots, n$ where

$$\boldsymbol{\beta}' = \begin{pmatrix} \boldsymbol{c} & \boldsymbol{A}_1 & \cdots & \boldsymbol{A}_p & \boldsymbol{B}_1 & \cdots & \boldsymbol{B}_q \end{pmatrix} \text{ is a } d \times (d(p+q)+1) \text{ matrix,} \tag{5.27}$$

$$\boldsymbol{x}_t' = \begin{pmatrix} 1 & \boldsymbol{y}_{t-1}' & \cdots & \boldsymbol{y}_{t-p}' & -\boldsymbol{\varepsilon}_{t-1}' & \cdots & -\boldsymbol{\varepsilon}_{t-q}' \end{pmatrix} \text{ is a } (d(p+q)+1) \text{ vector,} \tag{5.28}$$

and so (5.22) admits a matrix representation

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \mathbb{1}_n\boldsymbol{\gamma}' + \mathcal{E} \tag{5.29}$$

similar to (5.5), except that $\boldsymbol{\beta}$ is defined in (5.27) and $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_n \end{pmatrix}'$ where $\boldsymbol{x}_t$ is defined in (5.28). Although the error terms in $\boldsymbol{X}$ depends on $\boldsymbol{\beta}$, the linear regression form in (5.27) facilitates the use of the linear approximation by estimating the error terms using a high order VAR-VG model as a proxy. More specifically, we use the Akaike information criterion (AIC) to select a suitable order p for the VAR(p)-VG model which is described in Section 5.3.3.3.

5.3.2 Properties

We give as brief summary of some important properties of the VARMA model where the innovations do not necessarily follow a Gaussian distribution. Hence, these properties also apply to the VG innovations. See Gouriéroux and Zakoïan [40] and Tsay [104].

5.3.2.1 Causality and Invertibility

Definition 5.3.1 (Causality)**.** *The process* $\{\boldsymbol{y}_t\}$ *in (5.23) is said to be a* causal *if it can be expressed in the form of*

$$\boldsymbol{y}_t = \sum_{k=0}^{\infty} \boldsymbol{\Psi}_k \boldsymbol{\varepsilon}_{t-k} \tag{5.30}$$

for a sequence of coefficient matrices $\{\boldsymbol{\Psi}_k\}$ *such that* $\sum_{k=0}^{\infty} \boldsymbol{\Psi}_k < \infty$.

In terms of the polynomial function $\boldsymbol{\mathcal{A}}(z)$, the process is causal if and only if $\det(\boldsymbol{\mathcal{A}}(z)) \neq 0$ for all $z \in \mathbb{C}$ such that $|z| \leq 1$.

This is a desirable property as it has a natural interpretation that the process is independent of the future values, thus allowing the process to be forecasted using current and past values. Another desirable property is invertibility.

Definition 5.3.2 (Invertibility)**.** *The process* $\{\boldsymbol{y}_t\}$ *is said to be* invertible *if the error terms can be expressed in the form of*

$$\boldsymbol{\varepsilon}_t = \sum_{k=0}^{\infty} \boldsymbol{\Pi}_k \boldsymbol{y}_k$$

for a sequence of coefficient matrices $\{\boldsymbol{\Pi}_k\}$ *such that* $\sum_{k=0}^{\infty} \boldsymbol{\Pi}_k < \infty$.

In terms of the polynomial function $\boldsymbol{\mathcal{B}}(z)$, the process is invertible if and only if $\det(\boldsymbol{\mathcal{B}}(z)) \neq 0$ for all $z \in \mathbb{C}$ such that $|z| \leq 1$.

5.3.2.2 Identifiability Issue

Unlike the VAR-VG model, the VARMA-VG model faces model identification problem which can lead to wrong interpretation of the model. So additional assumptions needs to be imposed to the model to avoid identification problem. We refer to the assumptions by Gouriéroux and Monfort [38].

Assumptions:

(i) $\boldsymbol{\varepsilon}_t$ are independent and identically distributed such that $\mathbb{E}(\|\boldsymbol{\varepsilon}_t\|^s) < \infty$ for some $s > 0$, and there exist matrix $\boldsymbol{C}$ such that the components of $\boldsymbol{C}\boldsymbol{\varepsilon}_t$ are mutually independent.

(ii) If $\boldsymbol{\mathcal{A}}(L)$ and $\boldsymbol{\mathcal{B}}(L)$ have left common factor $\boldsymbol{\mathcal{C}}(L)$ such that $\boldsymbol{\mathcal{A}}(L) = \boldsymbol{\mathcal{C}}(L)\widetilde{\boldsymbol{\mathcal{A}}}(L)$ and $\boldsymbol{\mathcal{B}}(L) = \boldsymbol{\mathcal{C}}(L)\widetilde{\boldsymbol{\mathcal{B}}}(L)$ for some polynomial functions $\widetilde{\boldsymbol{\mathcal{A}}}(L)$ and $\widetilde{\boldsymbol{\mathcal{B}}}(L)$, then $|\boldsymbol{\mathcal{C}}(L)|$ is independent of L.

(iii) The process $\{\boldsymbol{y}_t\}$ is causal and invertible.

The first assumption is based on the choice of error distributions. By choosing the distribution to have a NMVM representation, this allows for the first assumption to be

satisfied. Thus, choosing the error terms to follow the VG or Student's t distribution takes care of this assumption.

The second assumption is based on a property called *left coprimeness*. This assumption states that the polynomials $\boldsymbol{\mathcal{A}}(z)$ and $\boldsymbol{\mathcal{B}}(z)$ in equations (5.24) and (5.25) respectively have no common factors. Essentially, this assumption ensures the representation of the VARMA model is minimal in the sense that all possible simplifications have been performed. Checking this assumption is equivalent to checking if the two polynomials have at least one common eigenvalue [58]. These eigenvalues can be found by solving the scalar polynomials $|\boldsymbol{\mathcal{A}}(L)| = 0$ and $|\boldsymbol{\mathcal{B}}(L)| = 0$. Directly calculating these eigenvalues can be complicated especially for higher orders of p and q. See Gourieroux et al. [39] for a hypothesis test of common root for the univariate ARMA process. Also see Gouriéroux and Monfort [38] and Hannan and Deistler [45] for a treatment of the identification issue using the structural VARMA model under a non-Gaussian framework.

Gohberg and Lerer [37] proved that the Fisher information matrix becomes singular if and only if the matrix polynomials $\boldsymbol{\mathcal{A}}(L)$ and $\boldsymbol{\mathcal{B}}(L)$ have at least one common eigenvalue. Similarly, Klein et al. [58] proved that this is equivalent to the singularity of the tensor Sylvester matrix defined by

$$S^{\otimes}(-\boldsymbol{B}, \boldsymbol{A}) = \begin{pmatrix} (-\boldsymbol{I}_d)\otimes\boldsymbol{I}_d & (-\boldsymbol{B}_1)\otimes\boldsymbol{I}_d & \cdots & (-\boldsymbol{B}_q)\otimes\boldsymbol{I}_d & \boldsymbol{0}_{d^2\times d^2} & \cdots & \boldsymbol{0}_{d^2\times d^2} \\ \boldsymbol{0}_{d^2\times d^2} & \ddots & \ddots & & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \ddots & \boldsymbol{0}_{d^2\times d^2} \\ \boldsymbol{0}_{d^2\times d^2} & \cdots & \boldsymbol{0}_{d^2\times d^2} & (-\boldsymbol{I}_d)\otimes\boldsymbol{I}_d & (-\boldsymbol{B}_1)\otimes\boldsymbol{I}_d & \cdots & (-\boldsymbol{B}_q)\otimes\boldsymbol{I}_d \\ \boldsymbol{I}_d\otimes\boldsymbol{I}_d & \boldsymbol{I}_d\otimes\boldsymbol{A}_1 & \cdots & \boldsymbol{I}_d\otimes\boldsymbol{A}_p & \boldsymbol{0}_{d^2\times d^2} & \cdots & \boldsymbol{0}_{d^2\times d^2} \\ \boldsymbol{0}_{d^2\times d^2} & \ddots & \ddots & & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \ddots & \boldsymbol{0}_{d^2\times d^2} \\ \boldsymbol{0}_{d^2\times d^2} & \cdots & \boldsymbol{0}_{d^2\times d^2} & \boldsymbol{I}_d\otimes\boldsymbol{I}_d & \boldsymbol{I}_d\otimes\boldsymbol{A}_1 & \cdots & \boldsymbol{I}_d\otimes\boldsymbol{A}_p \end{pmatrix} \tag{5.31}$$

which is a $d^2(p+q) \times d^2(p+q)$ matrix and '$\otimes$' is the Kronecker product defined in (A.2). From a statistical point of view, testing for common roots between the AR and MA polynomials is equivalent to testing the determinant of the tensor Sylvester matrix (which is called the *resultant*) is equal to zero. However, formulating such a test requires further research.

5.3.3 ECM algorithm for VARMA-VG model

The structure of the ECM algorithm is similar to the VAR-VG model in Section 5.2.6 except some modifications are needed for the CM-step since the solution for $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ generally does not have closed-form. Rather than maximising the conditional normal log-likelihood function directly, some approximate techniques consider fitting the residuals $\boldsymbol{\varepsilon}_t$ using some higher order VAR-VG models as a proxy model. Then parameter estimates are obtained using the ML method with the fitted residuals. Thus, for the ECM algorithm of the VARMA-VG model, we only need to adjust the CM-step for $(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\gamma})$ which is described later in this section.

5.3.3.1 Likelihood functions

Since the initial residuals $\boldsymbol{\varepsilon}_t$ are not observed in (5.22), we make additional assumptions to these residuals before initialising the estimation procedure. This gives rise to the conditional likelihood and exact likelihood methods.

For the *conditional likelihood* method, the initial observations are considered to be observed. More formally, we observe $\{\boldsymbol{y}_s : 1 - \max(p, q) \leq s \leq 0\}$ but only observations $\{\boldsymbol{y}_s : s > 0\}$ contributes to the likelihood function. On the other hand, the initial observations and residuals under the *exact likelihood* method are considered to be random variables. As a result, these initial observations and residuals need to be estimated as additional parameters.

Since estimation of parameters via exact likelihood is more computationally intensive, we focus on the conditional likelihood method. Note that for large sample size n, the two likelihood methods provide similar results, especially when parameters lie further away from the non-invertible region. On the other hand, if the parameters lie close to the non-invertible region, then the exact likelihood method is preferred [48].

5.3.3.2 CM-step for $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\gamma}$

Unlike the VAR-VG model, the CM-step for parameters $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ in the VARMA-VG models does not have closed-form solution since the residuals $\boldsymbol{\varepsilon}_t$ depends on these parameters. One way to address this issue is to use numerical optimisation techniques

such as the NR algorithm to maximise the conditional normal log-likelihood function directly. However, due to the large number of parameters involved with the optimisation, the algorithm is computationally intensive and may not always lead to convergent estimates. Moreover, it relies heavily on the starting values.

Alternately, we use a higher order VAR-VG model given $\boldsymbol{u}$ and ν as a proxy model to approximate $\boldsymbol{\varepsilon}_t$. Specifically, we first choose an appropriate order for the VAR-VG model based on the AIC to obtain the fitted residuals $\hat{\boldsymbol{\varepsilon}}_t$. Then estimate parameters of the new approximate model

$$\boldsymbol{y}_t = \boldsymbol{\phi}_0 + \boldsymbol{A}_1\boldsymbol{y}_{t-1} + \cdots + \boldsymbol{A}_p\boldsymbol{y}_{t-p} - \boldsymbol{B}_1\hat{\boldsymbol{\varepsilon}}_{t-1} - \cdots - \boldsymbol{B}_q\hat{\boldsymbol{\varepsilon}}_{t-q} + \boldsymbol{\varepsilon}_t \tag{5.32}$$

for $t = 1, \ldots, n$ using the estimation procedures in Section 5.2.4. This is equivalent to fitting $\boldsymbol{\beta}$ from the linear equation in (5.29).

Roy et al. [95] mentioned that unlike the MLE of VAR model which is causal and invertible, there is no guarantee the MLE of VARMA model is causal and invertible. So for our proposed ECM algorithms, the estimate of $\boldsymbol{\beta}$ after the CM-step is not guaranteed to be in the causal and invertible region for the VARMA-VG model. Thus, we apply line search in Section 3.3.2.4 to ensure the parameter estimates stays within the causal and invertible region of the parameter space.

5.3.3.3 Order Selection of VAR-VG models

The information criteria method is used to select the order for the VAR-VG model given $\boldsymbol{u}$ and ν in order to approximate $\boldsymbol{\varepsilon}_t$ in the CM-step for $(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\gamma})$. The maximised conditional normal likelihood for the VAR(p)-VG model is essentially equivalent to the determinant of the covariance matrix of the innovations since

$$\begin{aligned}
L(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Sigma}}_p, \hat{\boldsymbol{\gamma}}; \boldsymbol{y}, \boldsymbol{u}|\mathscr{F}_0) &= \prod_{t=1}^{n} f_N(\boldsymbol{y}_t|\boldsymbol{u}_t, \mathscr{F}_{t-1}; \hat{\boldsymbol{\beta}}'\boldsymbol{x}_t + u_t\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\Sigma}}_p) \\
&= |\hat{\boldsymbol{\Sigma}}_p|^{-n/2}(2\pi)^{-nd/2}\exp\left(-\frac{1}{2}\mathrm{tr}\left(\hat{\boldsymbol{\Sigma}}_p^{-1}\widehat{S}_{\tilde{\boldsymbol{y}}\tilde{\boldsymbol{y}}/u}\right)\right)\prod_{t=1}^{n} u_t^{-d/2} \\
&= |\hat{\boldsymbol{\Sigma}}_p|^{-n/2}(2\pi)^{-nd/2}\exp\left(-\frac{n}{2}\underbrace{\mathrm{tr}(\boldsymbol{I}_d)}_{d}\right)\prod_{t=1}^{n} u_t^{-d/2} \\
&\propto |\hat{\boldsymbol{\Sigma}}_p|^{-n/2}
\end{aligned}$$

where $\mathscr{F}_0 = \{\boldsymbol{y}_s : 1 - P \leq s \leq 0\}$, P is the maximum AR order considered for the approximation, f_N represents the density function of the multivariate normal distribution, $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Sigma}}_p, \hat{\boldsymbol{\gamma}})$ are the estimates of $(\boldsymbol{\beta}, \boldsymbol{\Sigma}_p, \boldsymbol{\gamma})$ for the VAR(p)-VG model which are obtained using (5.19) and (5.20), and $\widehat{S}_{\tilde{\boldsymbol{\mathbf{y}}}\tilde{\boldsymbol{\mathbf{y}}}/u}$ is defined as

$$\widehat{S}_{\tilde{\boldsymbol{\mathbf{y}}}\tilde{\boldsymbol{\mathbf{y}}}/u} = \sum_{t=1}^{n} \frac{1}{u_t} \Big(\boldsymbol{y}_t - \hat{\boldsymbol{\beta}}' \boldsymbol{x}_t - u_t \hat{\boldsymbol{\gamma}}\Big) \Big(\boldsymbol{y}_t - \hat{\boldsymbol{\beta}}' \boldsymbol{x}_t - u_t \hat{\boldsymbol{\gamma}}\Big)'.$$

Thus, the AIC can be constructed as

$$\mathrm{AIC}(p) = \log \left|\hat{\boldsymbol{\Sigma}}_p\right| + \frac{2}{n} p d^2.$$

When implementing the order selection procedure to the CM-step of $(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\gamma})$ for the VARMA(p, q)-VG model with filtration $\{\boldsymbol{y}_s : 1 - \max(p, q) \leq s \leq 0\}$, the initial summation indices needs to be adjusted so that $\boldsymbol{y}_{1-P}$ in this section aligns with $\boldsymbol{y}_{1-\max(p,q)}$ in (5.22). This alignment leads to summation indices $t = R+1, \ldots, n$ where $R = P - \max(p, q)$ with filtration $\{\boldsymbol{y}_s : 1 - \max(p, q) \leq s \leq R\}$ which can then be used to implement the order selection in the CM-step.

In this thesis, We choose $P = 13$ when implementing the order selection in the ECM algorithm for VARMA-VG model.

5.3.4 Forecasting using VARMA-VG model

One of the main purpose of financial time series modelling is to provide forecasts for trading strategy formulation.

5.3.4.1 l-step ahead forecast

Let $\mathscr{F}_t = \{\boldsymbol{y}_s : s \leq t\}$, suppose that parameter estimates $\boldsymbol{\theta}_t$ for the VARMA-VG model are obtained through some in-sample model fittings using $\mathscr{F}_t$ and it is used to forecast l time points ahead. Then the one-step ahead forecast is

$$\boldsymbol{y}_t(1) = \mathbb{E}[\boldsymbol{y}_{t+1} | \mathscr{F}_t] = \boldsymbol{c} + \sum_{i=1}^{p} \boldsymbol{A}_i \boldsymbol{y}_{t+1-i} - \sum_{j=1}^{q} \boldsymbol{B}_j \boldsymbol{\varepsilon}_{t+1-j}$$

where $\mathbb{E}[\boldsymbol{\varepsilon}_{t+i}|\mathscr{F}_t] = \mathbf{0}$ for $i > 0$, the associated forecast error is

$$\boldsymbol{e}_t(1) = \boldsymbol{y}_{t+1} - \boldsymbol{y}_t(1) = \boldsymbol{\varepsilon}_{t+1}$$

and the covariance of the forecast is

$$\text{cov}[\boldsymbol{e}_t(1)] = \text{cov}(\boldsymbol{\varepsilon}_{t+1}) = \boldsymbol{\Sigma}.$$

For the two-step ahead forecast, we have that

$$\boldsymbol{y}_t(2) = \mathbb{E}[\boldsymbol{y}_{t+2}|\mathscr{F}_t] = \boldsymbol{c} + \boldsymbol{A}_1\boldsymbol{y}_t(1) + \sum_{i=2}^{p} \boldsymbol{A}_i\boldsymbol{y}_{t+2-i} - \sum_{j=2}^{q} \boldsymbol{B}_j\boldsymbol{\varepsilon}_{t+2-j}.$$

In general, the l-step ahead forecast is given by

$$\boldsymbol{y}_t(l) = \boldsymbol{c} + \sum_{i=1}^{p} \boldsymbol{A}_i\mathbb{E}[\boldsymbol{y}_{t+l-i}|\mathscr{F}_t] - \sum_{j=1}^{q} \boldsymbol{B}_j\mathbb{E}[\boldsymbol{\varepsilon}_{t+l-j}|\mathscr{F}_t]$$

where

$$\mathbb{E}[\boldsymbol{y}_{t+i}|\mathscr{F}_t] = \begin{cases} \boldsymbol{y}_{t+i} & \text{if } i \leq 0, \\ \boldsymbol{y}_t(i) & \text{if } i > 0, \end{cases} \quad \text{and} \quad \mathbb{E}[\boldsymbol{\varepsilon}_{t+i}|\mathscr{F}_t] = \begin{cases} \boldsymbol{\varepsilon}_{t+i} & \text{if } i \leq 0, \\ \mathbf{0} & \text{if } i > 0. \end{cases}$$

The l-step ahead forecast error is

$$\boldsymbol{e}_t(l) = \boldsymbol{\varepsilon}_{t+l} + \boldsymbol{\Psi}_1\boldsymbol{\varepsilon}_{t+l-1} + ... + \boldsymbol{\Psi}_{l-1}\boldsymbol{\varepsilon}_{t+1}$$

where $\{\boldsymbol{\Psi}_k\}$ represents the coefficients from the MA representation. Thus the covariance of the l-step ahead forecast error is

$$\text{cov}[\boldsymbol{e}_t(l)] = \boldsymbol{\Sigma} + \boldsymbol{\Psi}_1\boldsymbol{\Sigma}\boldsymbol{\Psi}_1' + ... + \boldsymbol{\Psi}_{l-1}\boldsymbol{\Sigma}\boldsymbol{\Psi}_{l-1}'.$$

This means that the stationary VARMA-VG process is mean-reverting. That is, $\boldsymbol{y}_t(l) \to \mathbb{E}[\boldsymbol{y}_t]$ as $l \to \infty$. Additionally $\text{cov}[\boldsymbol{e}_t(l)] \to \text{cov}[\boldsymbol{y}_t]$ as $l \to \infty$.

5.3.4.2 Updating the forecast

When new information arises, the previous VARMA-VG forecast can be easily updated in the following way. Using the MA representation, the l-step ahead forecast at t is

$$\begin{aligned}\boldsymbol{y}_t(l) &= \sum_{k=1}^{\infty} \boldsymbol{\Psi}_k \mathbb{E}[\boldsymbol{\varepsilon}_{t+l-k}|\mathscr{F}_t] \\ &= \boldsymbol{\Psi}_l \boldsymbol{\varepsilon}_t + \boldsymbol{\Psi}_{l+1}\boldsymbol{\varepsilon}_{t-1} + ... \end{aligned} \tag{5.33}$$

and the $(l-1)$-step ahead forecast at $t+1$ is

$$\boldsymbol{y}_{t+1}(l-1) = \boldsymbol{\Psi}_{l-1}\boldsymbol{\varepsilon}_{t+1} + \boldsymbol{\Psi}_l \boldsymbol{\varepsilon}_t + \boldsymbol{\Psi}_{l+1}\boldsymbol{\varepsilon}_{t-1} + \dots . \tag{5.34}$$

Subtracting (5.33) from (5.34) gives us the formula for updating the VARMA-VG forecast

$$\boldsymbol{y}_{t+1}(l-1) = \boldsymbol{y}_t(l) + \boldsymbol{\Psi}_{l-1}\boldsymbol{\varepsilon}_{t+1}$$

where $\boldsymbol{\varepsilon}_{t+1}$ is observed since the new information is available at time $t+1$.

5.4 Estimation of VARMA-t model

Another important extension worth considering is the VARMA-t model where instead the innovations follow the Student's t distribution.

5.4.1 VARMA-t model

The d-dimensional time series $\{\boldsymbol{y}_t\}$ follows a VARMA(p,q)-t process if it has the VARMA structure as in (5.22) where instead $\boldsymbol{\varepsilon}_t \sim t_d(-\boldsymbol{\gamma}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}, \upsilon)$ for $\upsilon > 2$ which has density function in (C.1). This parametrisation is chosen so that the mean is well-defined, and the mixing variable which follows $\mathcal{IG}(\frac{\upsilon}{2}, \frac{\upsilon}{2} - 1)$ for $\upsilon > 2$ such that it has expectation of one.

The estimation procedure is similar to the VARMA-VG model in Section 5.3, except the E-step and CM-step for υ needs to be modified. Details of these steps is presented in the following sections.

5.4.2 E-step

The conditional distribution of u_t given $\mathscr{F}_t$ has pdf

$$f(u_t|\mathscr{F}_t) \propto u_t^{-(\upsilon+d)/2-1} \exp\left(-\frac{1}{2u_t}(\upsilon - 2 + z_t^2) - \frac{u_t}{2}\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}\right) \tag{5.35}$$

for $t = 1, \dots, n$ which corresponds to the $\mathcal{GIG}(-(\upsilon+d)/2, \upsilon - 2 + z_t^2, \boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma})$ distribution with z_t^2 defined in (5.11). Let $\boldsymbol{\theta}^{(k)} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}, \upsilon)$, then we have the following conditional expectations:

$$\widehat{u_t} = \mathbb{E}_{\boldsymbol{\theta}^{(k)}}[u_t|\mathscr{F}_t] = \frac{\eta_t \, K_{\frac{\upsilon+d}{2}-1}(\omega_t)}{K_{\frac{\upsilon+d}{2}}(\omega_t)}, \tag{5.36}$$

$$\widehat{1/u_t} = \mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[\frac{1}{u_t}\middle|\mathscr{F}_t\right] = \frac{K_{\frac{\upsilon+d}{2}+1}(\omega_t)}{\eta_t \, K_{\frac{\upsilon+d}{2}}(\omega_t)}, \tag{5.37}$$

$$\widehat{\log u_t} = \mathbb{E}_{\boldsymbol{\theta}^{(k)}}[\log u_t|\mathscr{F}_t] = \log \eta_t - \frac{K^{(1,0)}_{\frac{\upsilon+d}{2}}(\omega_t)}{K_{\frac{\upsilon+d}{2}}(\omega_t)} \tag{5.38}$$

where $\eta_t = \sqrt{\upsilon - 2 + z_t^2}/\sqrt{\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}}$, $\omega_t = \sqrt{(\upsilon - 2 + z_t^2)\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}}$.

Symmetric Student's t distribution case:

For the case when $\boldsymbol{\gamma} \to \mathbf{0}$,

$$\mathbb{E}_{\boldsymbol{\theta}^{(k)}}[u_t|\mathscr{F}_t] \sim \begin{cases} \frac{\Gamma\left(1-\frac{\upsilon+d}{2}\right)}{\Gamma\left(\frac{\upsilon+d}{2}\right)} 2^{1-d-\upsilon} \left(\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}\right)^{\frac{\upsilon+d}{2}-1} \left(\upsilon - 2 + z_t^2\right)^{\frac{\upsilon+d}{2}} & \text{if } \upsilon < 2 - d, \\ -\log(\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma})\left(\upsilon - 2 + z_t^2\right) & \text{if } \upsilon = 2 - d, \\ \frac{\upsilon - 2 + z_t^2}{\upsilon - 2 + d} & \text{if } \upsilon > 2 - d, \end{cases}$$

$$\mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[\frac{1}{u_t}\middle|\mathscr{F}_t\right] \sim \frac{\upsilon + d}{\upsilon - 2 + z_t^2},$$

$$\mathbb{E}_{\boldsymbol{\theta}^{(k)}}[\log u_t|\mathscr{F}_t] \sim \log\left(\frac{\upsilon - 2 + z_t^2}{2}\right) - \psi\left(\frac{\upsilon + d}{2}\right).$$

5.4.3 CM-step for υ

Given the mixing variables $\boldsymbol{u}$, the MLE of υ can be obtained by maximising the log-likelihood of the inverse-gamma distribution

$$\ell_{IG}(\upsilon; \boldsymbol{u}) = \frac{n\upsilon}{2}\log\Big(\frac{\upsilon}{2}-1\Big) - n\log\Gamma\Big(\frac{\upsilon}{2}\Big) - \Big(\frac{\upsilon}{2}+1\Big)\sum_{t=1}^{n}\log u_t - \Big(\frac{\upsilon}{2}-1\Big)\sum_{t=1}^{n}\frac{1}{u_t}$$

and it has derivatives

$$\frac{\partial \ell_{IG}}{\partial \upsilon} = \frac{n}{2}\bigg(1 + \frac{2}{\nu-2} + \log\Big(\frac{\upsilon}{2}-1\Big) - \psi\Big(\frac{\upsilon}{2}\Big)\bigg) - \frac{1}{2}\sum_{t=1}^{n}\log u_t - \frac{1}{2}\sum_{t=1}^{n}\frac{1}{u_t} \tag{5.39}$$

$$\frac{\partial^2 \ell_{IG}}{\partial \upsilon^2} = \frac{\nu-4}{2(\nu-2)^2} - \frac{1}{4}\psi'\Big(\frac{\upsilon}{2}\Big) \tag{5.40}$$

for derivative-based optimisation methods. For the ECME algorithm, the MLE of υ can be obtained by maximising the observed log-likelihood which involves the log of the Student's t density in (C.1).

5.5 Simulation study

In order to evaluate the performance of the AECM algorithm proposed in Section 5.3.3, we conduct two simulation studies where the first one considers the two-dimensional VARMA(1,1)-VG model when the AR and MA parameters fall into the identifiability region with different sets of values for the skewness and shape parameters. The second one considers the case when the true AR and MA parameters fall outside the identifiability region and describes ways of dealing with the non-identifiability issue.

5.5.1 Identifiable VARMA-VG model

We consider the two-dimensional VARMA(1,1)-VG model with the following true parameters

$$\boldsymbol{c} = \begin{pmatrix}0\\0\end{pmatrix},\ \boldsymbol{A}_1 = \begin{pmatrix}0.5 & -0.25\\0.3 & 0.4\end{pmatrix},\ \boldsymbol{B}_1 = \begin{pmatrix}0.2 & -0.1\\0.05 & 0.3\end{pmatrix},\ \boldsymbol{\Sigma} = \begin{pmatrix}1 & 0.7\\0.7 & 1\end{pmatrix}. \tag{5.41}$$

As for the skewness and shape parameters, we consider four different sets of parameters.

	$\boldsymbol{\gamma} = (0.2, 0.3)$	$\boldsymbol{\gamma} = (1, 2)$
$\nu = 3$	Case 1	Case 3
$\nu = 0.7$	Case 2	Case 4

We remark that when $\nu < 1$, the density function of the VG distribution becomes unbounded. This corresponds to cases 2 and 4.

We simulate data sets each consisting of a time series of 4000 data points, then discard the first 2000 data points to ensure convergence of the process. Then we use the remaining 2000 data points to implement the AECM algorithm with the WLOO likelihood extended from the ECM algorithm in Section 5.3.3. This procedure is replicated 1000 times to obtain 1000 parameter estimates.

5.5.1.1 Parameter estimates

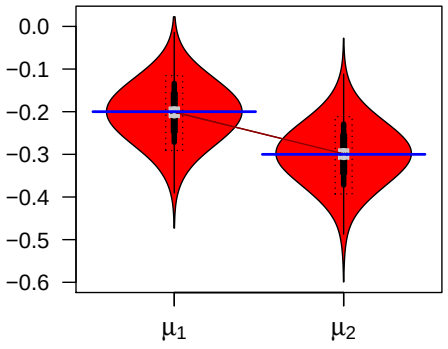
In Figures 5.1 to 5.4, we plot the 1000 parameter estimates using violin plots from the `R` package called `Caroline`. The true parameter value represented by a blue horizontal line are also included to facilitate comparison. For all four cases, the algorithm seems to perform reasonably well as the medians are very close to the true values. For case 1 when $\nu = 3$ and $\boldsymbol{\gamma} = (0.2, 0.3)$, each of the parameters seems to be normally distributed which suggests that the algorithm is numerically stable in this region of the parameter space. However, for case 2 when $\nu = 0.7$, most parameters still seem to follow normal distributions but μ_2 and γ_2 have heavier tails to one side. For case 3 when the skewness increases to $\boldsymbol{\gamma} = (1, 2)$ but $\nu = 3$, the distributions for μ_2, Σ_{22}, γ_2 and ν have even heavier one-side tail and this phenomenon gets even more severe for case 4 when $\nu = 0.7$. Nevertheless, since medians of all distributions match closely with the true parameter values, we conclude that the ECM algorithm performs well even when the shape parameter falls into the unbounded range. Not only is the algorithm able to estimate parameters of the VARMA(1,1)-VG model to a high level of accuracy but it is also able to demonstrate a high level of computational efficiency. Specifically, the median time it takes to run the algorithm for case 1 to 4 is 63, 75, 142 and 82 seconds respectively.

5.5.1.2 Standard error calculation

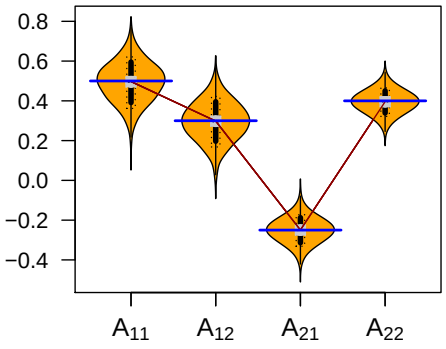
We test the performance of SE calculation for the four cases. For each parameter, we calculate the standard deviation (SD) of 1000 parameter estimates and report the SD under the column called "Simulated" in Tables 5.1 and 5.2. The SD of the simulated estimates are compared with the SE calculated using the Louis' method described in Section 1.3.4 and equation (1.10). Results show that they are highly consistent for $\nu = 3$ and with slight discrepancy for $\nu = 0.7$ since there are a few outliers with the simulated estimates. We remark that for the unbounded cases (case 2 and 4 with $\nu = 0.7$), we use the double generalised gamma approximation to calculate the SE for $\boldsymbol{\mu}$. See Section 3.4.2.2 and appendix C4 for details.

Table 5.1. SEs based on simulated estimates and calculation using Louis' method for cases 1 and 3.

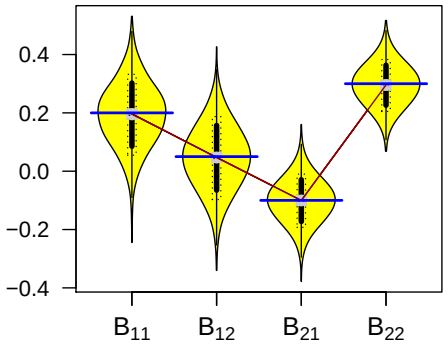
case	parameter	Simulated	Louis
Case 1: $\boldsymbol{\gamma} = (0.2, 0.3)$ $\nu = 3$	$\boldsymbol{\mu}'$	$\begin{pmatrix} 0.067 & 0.070 \end{pmatrix}$	$\begin{pmatrix} 0.069 & 0.071 \end{pmatrix}$
	$\boldsymbol{A}_1$	$\begin{pmatrix} 0.098 & 0.061 \\ 0.090 & 0.053 \end{pmatrix}$	$\begin{pmatrix} 0.102 & 0.063 \\ 0.099 & 0.056 \end{pmatrix}$
	$\boldsymbol{B}_1$	$\begin{pmatrix} 0.104 & 0.068 \\ 0.101 & 0.063 \end{pmatrix}$	$\begin{pmatrix} 0.108 & 0.072 \\ 0.111 & 0.069 \end{pmatrix}$
	$\boldsymbol{\Sigma}$	$\begin{pmatrix} 0.0.38 & 0.033 \\ & 0.039 \end{pmatrix}$	$\begin{pmatrix} 0.038 & 0.032 \\ & 0.039 \end{pmatrix}$
	$\boldsymbol{\gamma}'$	$\begin{pmatrix} 0.071 & 0.073 \end{pmatrix}$	$\begin{pmatrix} 0.073 & 0.075 \end{pmatrix}$
	ν	0.368	0.375
Case 3: $\boldsymbol{\gamma} = (1, 2)$ $\nu = 3$	$\boldsymbol{\mu}'$	$\begin{pmatrix} 0.082 & 0.127 \end{pmatrix}$	$\begin{pmatrix} 0.106 & 0.206 \end{pmatrix}$
	$\boldsymbol{A}_1$	$\begin{pmatrix} 0.083 & 0.046 \\ 0.088 & 0.047 \end{pmatrix}$	$\begin{pmatrix} 0.085 & 0.048 \\ 0.094 & 0.051 \end{pmatrix}$
	$\boldsymbol{B}_1$	$\begin{pmatrix} 0.090 & 0.051 \\ 0.101 & 0.055 \end{pmatrix}$	$\begin{pmatrix} 0.092 & 0.054 \\ 0.108 & 0.061 \end{pmatrix}$
	$\boldsymbol{\Sigma}$	$\begin{pmatrix} 0.052 & 0.064 \\ & 0.099 \end{pmatrix}$	$\begin{pmatrix} 0.059 & 0.086 \\ & 0.158 \end{pmatrix}$
	$\boldsymbol{\gamma}'$	$\begin{pmatrix} 0.086 & 0.131 \end{pmatrix}$	$\begin{pmatrix} 0.108 & 0.206 \end{pmatrix}$
	ν	0.278	0.283



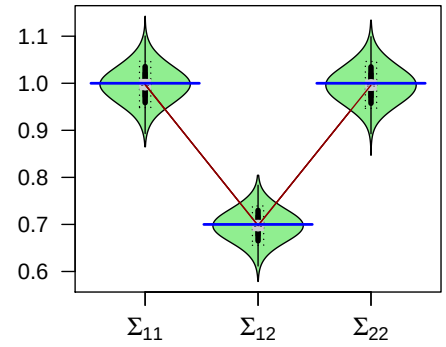
(a) true $\boldsymbol{\mu} = (-0.2, -0.3)$



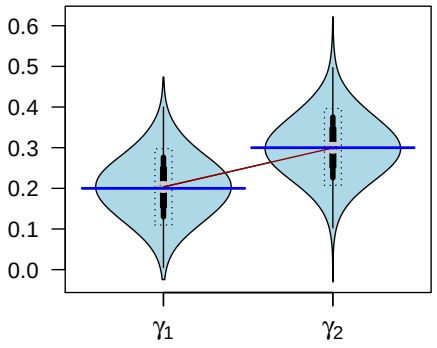
(b) true $\boldsymbol{A}_1 = \begin{pmatrix} 0.5 & 0.3 \\ -0.25 & 0.4 \end{pmatrix}$



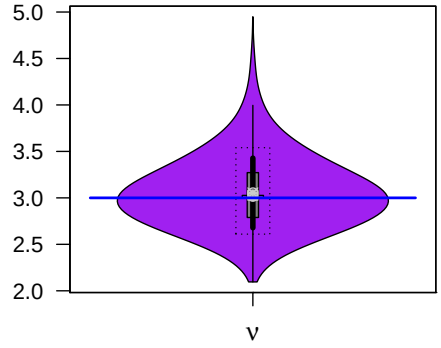
(c) true $\boldsymbol{B}_1 = \begin{pmatrix} 0.2 & 0.05 \\ -0.1 & 0.3 \end{pmatrix}$



(d) true $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}$

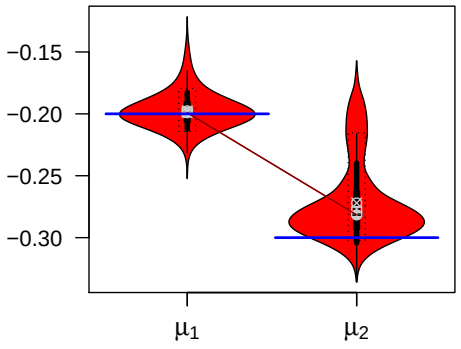


(e) true $\boldsymbol{\gamma} = (0.2, 0.3)$

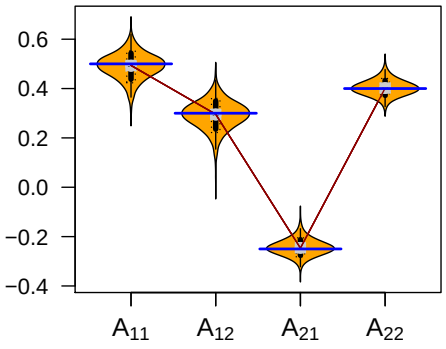


(f) true $\nu = 3$

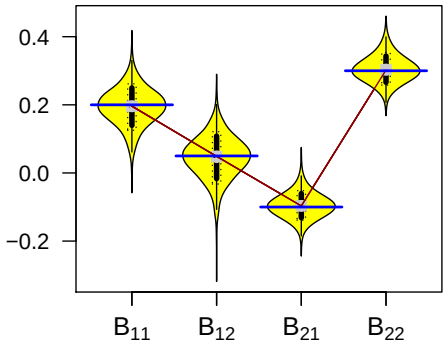
Figure 5.1. Case 1: Vioplots for parameter estimates of VARMA(1,1)-VG model with $\boldsymbol{\gamma} = (0.2, 0.3)$ and $\nu = 3$.



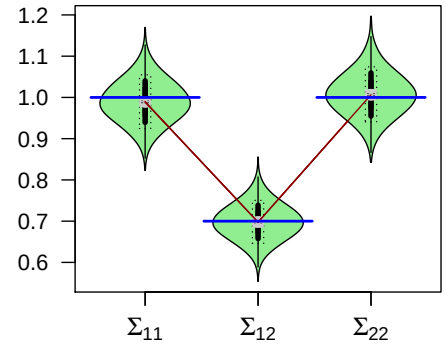
(a) true $\boldsymbol{\mu} = (-0.2, -0.3)$



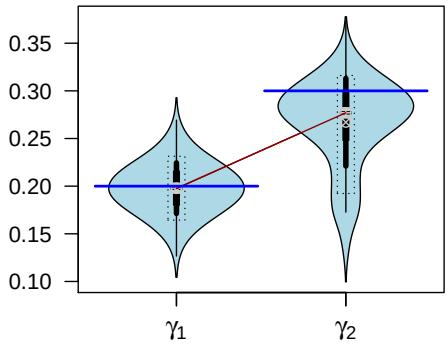
(b) true $\boldsymbol{A}_1 = \begin{pmatrix} 0.5 & 0.3 \\ -0.25 & 0.4 \end{pmatrix}$



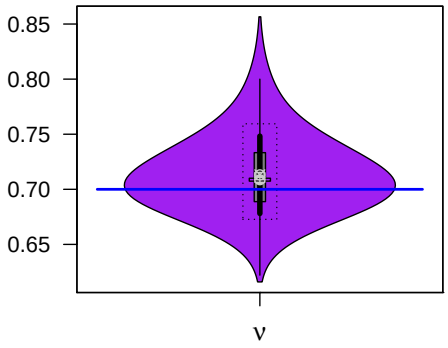
(c) true $\boldsymbol{B}_1 = \begin{pmatrix} 0.2 & 0.05 \\ -0.1 & 0.3 \end{pmatrix}$



(d) true $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}$

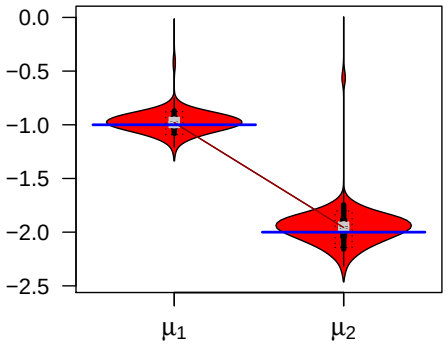


(e) true $\boldsymbol{\gamma} = (0.2, 0.3)$

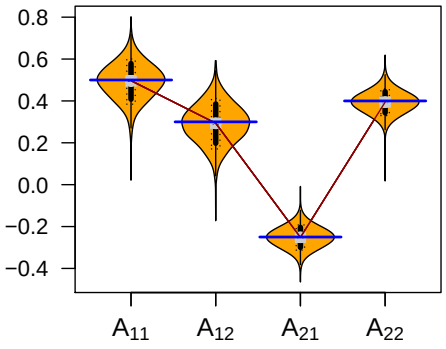


(f) true $\nu = 0.7$

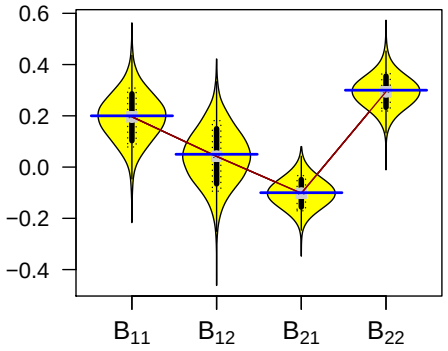
Figure 5.2. Case 2: Vioplots for parameter estimates of VARMA(1,1)-VG model with $\boldsymbol{\gamma} = (0.2, 0.3)$ and $\nu = 0.7$.



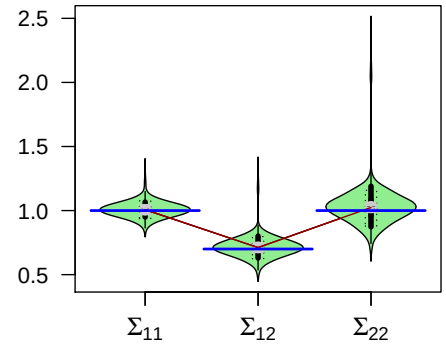
(a) true $\boldsymbol{\mu} = (-1, -2)$



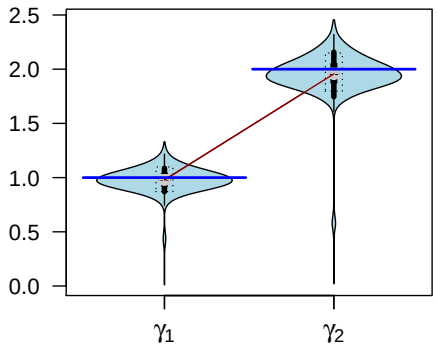
(b) true $\boldsymbol{A}_1 = \begin{pmatrix} 0.5 & 0.3 \\ -0.25 & 0.4 \end{pmatrix}$



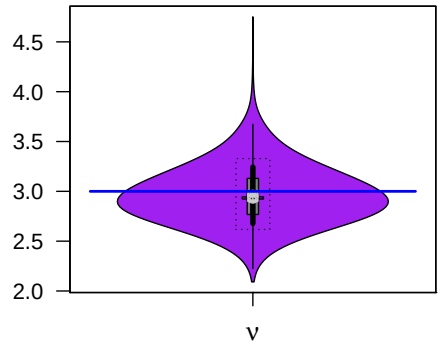
(c) true $\boldsymbol{B}_1 = \begin{pmatrix} 0.2 & 0.05 \\ -0.1 & 0.3 \end{pmatrix}$



(d) true $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}$

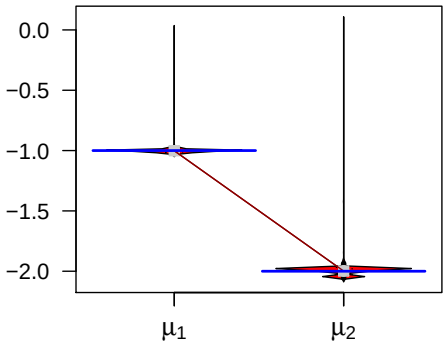


(e) true $\boldsymbol{\gamma} = (0.2, 0.3)$

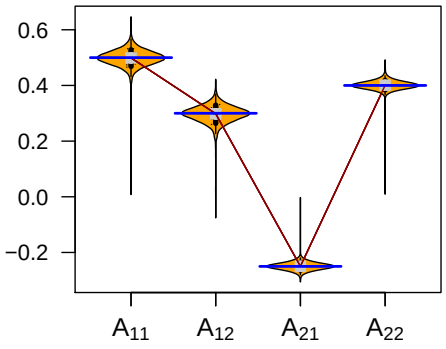


(f) true $\nu = 3$

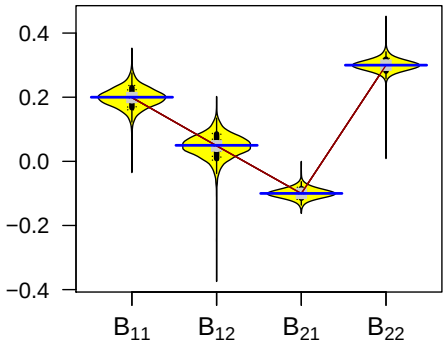
Figure 5.3. Case 3: Vioplots for parameter estimates of VARMA(1,1)-VG model with $\boldsymbol{\gamma} = (1, 2)$ and $\nu = 3$.



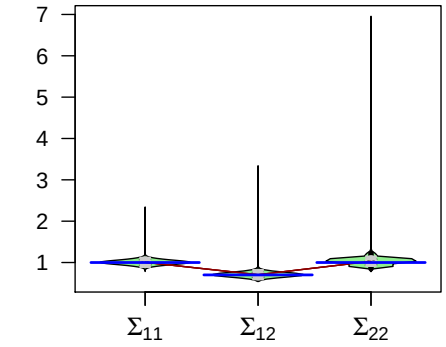
(a) true $\boldsymbol{\mu} = (-1, -2)$



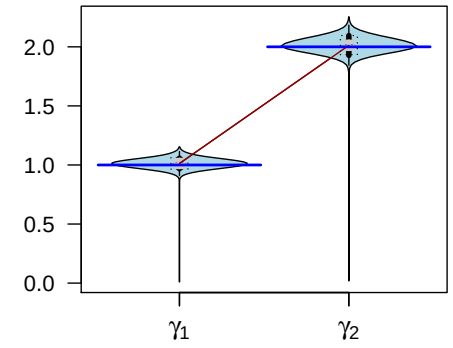
(b) true $\boldsymbol{A}_1 = \left(\begin{smallmatrix} 0.5 & 0.3 \\ -0.25 & 0.4 \end{smallmatrix}\right)$



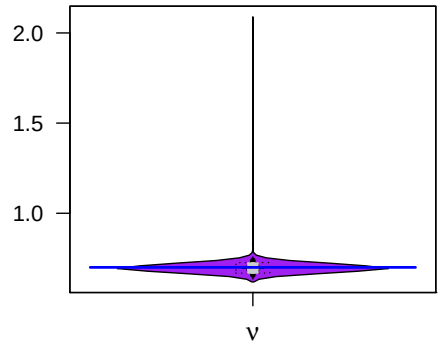
(c) true $\boldsymbol{B}_1 = \left(\begin{smallmatrix} 0.2 & 0.05 \\ -0.1 & 0.3 \end{smallmatrix}\right)$



(d) true $\boldsymbol{\Sigma} = \left(\begin{smallmatrix} 1 & 0.7 \\ 0.7 & 1 \end{smallmatrix}\right)$



(e) true $\boldsymbol{\gamma} = (1, 2)$



(f) true $\nu = 0.7$

Figure 5.4. Case 4: Vioplots for parameter estimates of VARMA(1,1)-VG model with $\boldsymbol{\gamma} = (1, 2)$ and $\nu = 0.7$.

Table 5.2. SEs based on simulated estimates and calculation using Louis' method for $(\hat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\gamma}}, \hat{\nu})$ and the double generalised gamma approximation in Section 3.4.2.2 for $\hat{\boldsymbol{\mu}}$.

case	parameter	Simulated	Louis
Case 2: $\boldsymbol{\gamma} = (0.2, 0.3)$ $\nu = 0.7$	$\boldsymbol{\mu}'$	$\begin{pmatrix} 0.006 & 0.009 \end{pmatrix}$	$\begin{pmatrix} 0.015 & 0.032 \end{pmatrix}$
	$\boldsymbol{A}_1$	$\begin{pmatrix} 0.028 & 0.017 \\ 0.028 & 0.017 \end{pmatrix}$	$\begin{pmatrix} 0.053 & 0.032 \\ 0.057 & 0.030 \end{pmatrix}$
	$\boldsymbol{B}_1$	$\begin{pmatrix} 0.030 & 0.019 \\ 0.031 & 0.020 \end{pmatrix}$	$\begin{pmatrix} 0.055 & 0.035 \\ 0.062 & 0.038 \end{pmatrix}$
	$\boldsymbol{\Sigma}$	$\begin{pmatrix} 0.049 & 0.040 \\ & 0.050 \end{pmatrix}$	$\begin{pmatrix} 0.050 & 0.041 \\ & 0.052 \end{pmatrix}$
	$\boldsymbol{\gamma}'$	$\begin{pmatrix} 0.023 & 0.024 \end{pmatrix}$	$\begin{pmatrix} 0.027 & 0.046 \end{pmatrix}$
	ν	0.028	0.035
Case 4: $\boldsymbol{\gamma} = (1, 2)$ $\nu = 0.7$	$\boldsymbol{\mu}'$	$\begin{pmatrix} 0.005 & 0.020 \end{pmatrix}$	$\begin{pmatrix} 0.027 & 0.061 \end{pmatrix}$
	$\boldsymbol{A}_1$	$\begin{pmatrix} 0.019 & 0.009 \\ 0.023 & 0.011 \end{pmatrix}$	$\begin{pmatrix} 0.030 & 0.014 \\ 0.030 & 0.014 \end{pmatrix}$
	$\boldsymbol{B}_1$	$\begin{pmatrix} 0.021 & 0.010 \\ 0.026 & 0.013 \end{pmatrix}$	$\begin{pmatrix} 0.032 & 0.015 \\ 0.039 & 0.019 \end{pmatrix}$
	$\boldsymbol{\Sigma}$	$\begin{pmatrix} 0.051 & 0.049 \\ & 0.065 \end{pmatrix}$	$\begin{pmatrix} 0.062 & 0.088 \\ & 0.189 \end{pmatrix}$
	$\boldsymbol{\gamma}'$	$\begin{pmatrix} 0.036 & 0.059 \end{pmatrix}$	$\begin{pmatrix} 0.045 & 0.084 \end{pmatrix}$
	ν	0.022	0.041

5.5.2 Non-identifiable VARMA-VG model

We now consider the case when the true parameters of the VARMA(1,1)-VG model fall into the non-identifiable region of the parameter space. It is important to address the effect of non-identifiability with parameter estimation.

We adopt similar true parameters as in (5.41) except that we modify the parameters in $\boldsymbol{A}_1$ and $\boldsymbol{B}_1$ so that they together fall into the non-identifiable region of the parameter space. In particular, we choose parameter values

$$\boldsymbol{A}_1 = \begin{pmatrix} 0.8 & 2 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad \boldsymbol{B}_1 = \begin{pmatrix} 0.3 & 0 \\ 0 & 0 \end{pmatrix} \tag{5.42}$$

which is the set of parameters taken from example 3.5 in [104].

5.5.2.1 Model identification

It can be shown that the model parameters are identical to

$$\boldsymbol{A}_1 = \begin{pmatrix} 0.8 & 2+\alpha_1 \\ 0 & \alpha_2 \end{pmatrix} \quad \text{and} \quad \boldsymbol{B}_1 = \begin{pmatrix} 0.3 & \alpha_1 \\ 0 & \alpha_2 \end{pmatrix}. \tag{5.43}$$

Looking at the two polynomial matrices $\boldsymbol{\mathcal{A}}(L)$ and $\boldsymbol{\mathcal{B}}(L)$ for equation (5.23), we can show that there exist a left common factor that is a non-zero constant such that for matrix polynomial function $\boldsymbol{\mathcal{A}}(L)$

$$\begin{pmatrix} 1-0.8L & -(2+\alpha_1)L \\ 0 & 1-\alpha_2 L \end{pmatrix} = \begin{pmatrix} 1 & -\alpha_1 L \\ 0 & 1-\alpha_2 L \end{pmatrix} \begin{pmatrix} 1-0.8L & -2\alpha_2 \\ 0 & 1 \end{pmatrix},$$

and for matrix polynomial function $\boldsymbol{\mathcal{B}}(L)$

$$\begin{pmatrix} 1-0.3L & -\alpha_1 L \\ 0 & 1-\alpha_2 L \end{pmatrix} = \begin{pmatrix} 1 & -\alpha_1 L \\ 0 & 1-\alpha_2 L \end{pmatrix} \begin{pmatrix} 1-0.3L & 0 \\ 0 & 1 \end{pmatrix}.$$

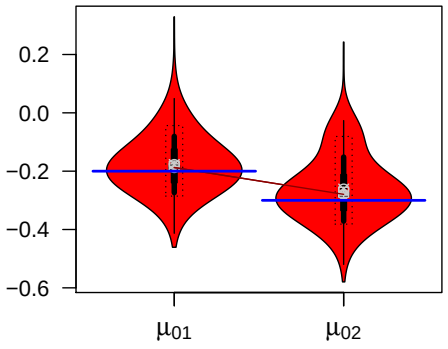
Since the determinant of the left common factor depends on L, then the second assumption in Section 5.3.2.2 is clearly violated for the model parameters in (5.43). In the light of this, the model parameters in (5.42) can be thought of as the parametrisation in its most simplest form. Additionally, $\boldsymbol{B}_1$ in (5.42) corresponds to a pure white noise process in the model (5.23).

Furthermore, we only consider the following parameter values

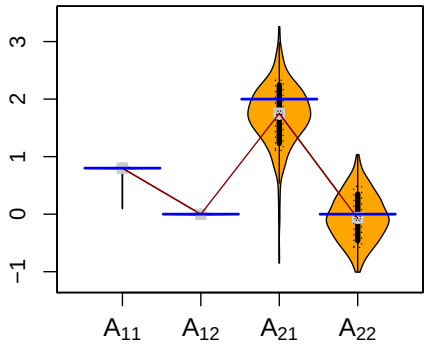
$$\boldsymbol{\gamma} = \begin{pmatrix} 0.2 \\ 0.3 \end{pmatrix}, \ \nu = 3,$$

for the skewness and shape parameters. Similar to the previous simulation study, we repeat the experiment 1000 times with a sample size of 2000 for each experiment fixing $\alpha_1 = \alpha_2 = 0$. This allows for a more parsimonious description of the process. Then we apply the ECM algorithm in Section 5.3.3 to fit the data in each experiment.

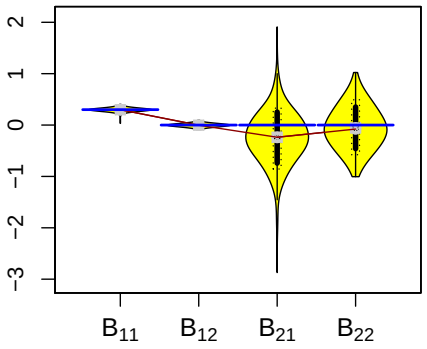
Again we use violin plots to present the distributions of the parameter estimates in Figure 5.5. A key feature to point out is that the variabilities for some of the AR and MA parameters are extremely large in comparison to the results in Figures 5.1 to 5.4. These large variabilities are due to their non-identifiable nature when α_1 and α_2 can take any arbitrary value. However, due to the sampling error in the simulations, these parameter estimates are distributed around zero.



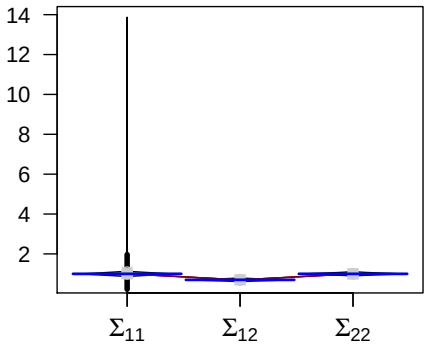
(a) true $\boldsymbol{\mu} = (-0.2, -0.3)$



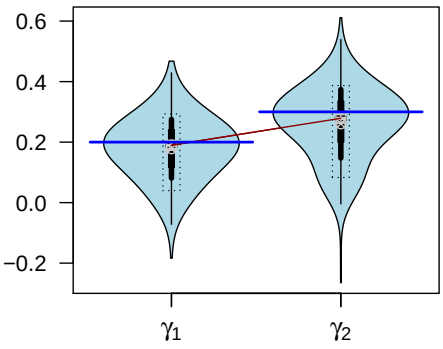
(b) true $\boldsymbol{A}_1 = \left(\begin{smallmatrix} 0.8 & 2 \\ 0 & 0 \end{smallmatrix}\right)$



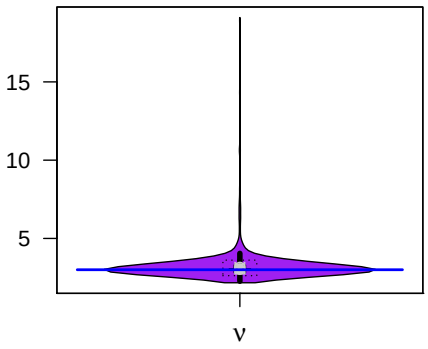
(c) true $\boldsymbol{B}_1 = \left(\begin{smallmatrix} 0.3 & 0 \\ 0 & 0 \end{smallmatrix}\right)$



(d) true $\boldsymbol{\Sigma} = \left(\begin{smallmatrix} 1 & 0.7 \\ 0.7 & 1 \end{smallmatrix}\right)$



(e) true $\boldsymbol{\gamma} = (0.2, 0.3)$



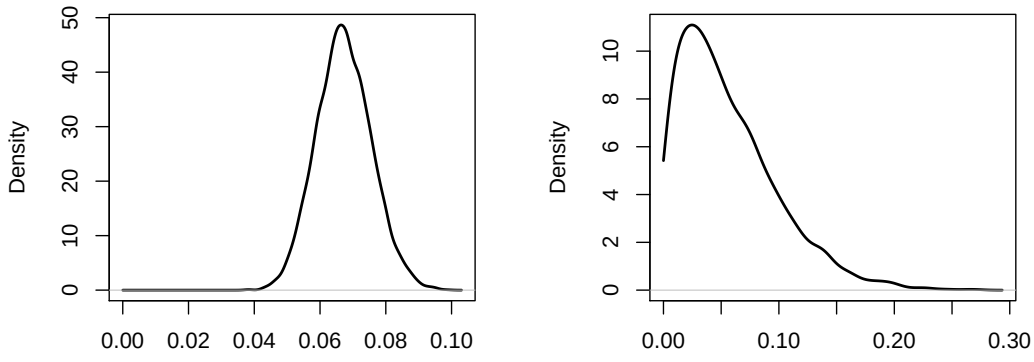
(f) true $\nu = 3$

Figure 5.5. Non-identifiable case: Vioplots for parameter estimates of non-identifiable VARMA(1,1)-VG model with $\boldsymbol{\gamma} = (0.2, 0.3)$ and $\nu = 3$.

5.5.2.2 Test for identifiability

As discussed in Section 5.5.1, one way to detect non-identifiability problem is to test if there is a common eigenvalue for the two matrix polynomials using the resultant which is the determinant of the tensor Sylvester matrix in (5.31). Theoretically, the resultant is zero when there is a common eigenvalue. Taking into account the random nature of the sampling error, we instead get values of the determinant following a certain distribution.

To investigate the characteristic of the resultant, we plot the kernel density estimate of the square root of the resultant in Figure 5.6. For the identifiable cases in Section 5.5.1, the square root of the resultant seems to follow some symmetric distribution with non-zero mean. On the other hand, for the non-identifiable case in this section, it seems to follow roughly a chi-squared distribution. In summary, the large variabilities in the AR and MA parameter estimates as well as in the root of the resultant give some indication whether there is common eigenvalue in the two matrix polynomials. These characteristics motivate one to construct a hypothesis test for the existence of a common eigenvalue. However, more research is required to develop such test for VARMA related models.



(a) Identifiable case

(b) Non-identifiable case

Figure 5.6. Density plots of the square root of the resultant applied to VARMA-VG model with parameters from Case 1 in Section 5.5.1 (identifiable) and in Section 5.5.2 (non-identifiable).

5.5.2.3 Results and remarks

The proposed AECM algorithm described in Section 5.3.3 not only provides accurate results but also computationally efficient procedure to estimate parameters of the VARMA-VG model. The main extension of the ECM algorithm in this chapter is the application of the approximation technique discussed in Section 5.3.3 to estimate the MA parameters. This approximation technique is useful especially for lower order VARMA-VG process. For higher order process, one needs to take some precautions in applying this approximation technique due to the extreme non-linear behaviour of the log-likelihood function, thus resulting in estimation procedures which are highly reliant on the starting values. Nevertheless, experience shows that only low orders of AR and MA terms are necessary in most real applications.

5.6 Applications

We illustrate the practical application of the VARMA-VG model through real data analyses for two return series, cryptocurrency exchange rates and stock market indices. The first analysis investigates the features of the newly emerged cryptocurrencies. The second analysis aims to study the effect of increasing sampling frequency on the characteristics of returns and hence the choice of model. For our statistical analysis, the returns of the time series is defined in (2.28).

5.6.1 Application to cryptocurrency

Bitcoin is the first decentralised cryptocurrency which is a digital payment system using blockchain technology allowing direct peer-to-peer transactions without a central repository or single administrator. The idea of this direct peer-to-peer transaction system without an intermediary was first proposed by Nakamoto [82]. These transactions are verified by network nodes and recorded in a public distributed ledger. Thus, Bitcoin as well as other cryptocurrencies display features distinct from ordinary fiat currencies traded in the market.

Since cryptocurrencies emerged only very recently, there are very limited studies analysing their market behaviour. Due to their short life period, many people do not fully understand their features of wild volatility and hence doubt their roles as currencies rather than merely some speculative investment assets. As cryptocurrencies are still on their early stages of development, the volatilities of their returns are higher than ordinary currencies. This poses great challenges when it comes to modelling and forecasting cryptocurrency as typically the innovations are assumed to follow a normal distribution. Instead, we see great opportunities for applying our proposed VARMA-VG model to study cryptocurrencies while also comparing the performance with the VARMA-t model in Section 5.4.1.

From the Brave New Coin (BNC) Digital Currency indices database, we obtained the closing price from 21st May 2014 to 17th July 2017 for the global weighted average of Bitcoin, Ripple, Litecoin, and Dash which are some of the most popular cryptocurrencies to date. Extracting the closing prices for each day gives us a sample size of 1154 for each component and returns are calculated using (2.28) and are plotted in Figure 5.7.

5.6.1.1 Numerical summary and statistical tests

Numerical summaries of the returns are presented in Table 5.3. The SD for the returns of Bitcoin is slightly larger than other cryptocurrencies. Moreover, all cryptocurrencies have kurtosis much larger than the normal distribution and have some positive skewness. Table 5.4 reports the correlation coefficient between each pair of cryptocurrencies. It is clear that the returns of cryptocurrencies are not strongly correlated though Dash and Litecoin exhibit some level of positive correlation.

All these features indicate that the distribution of the returns should have heavier tails and sharper peak than the normal distribution. Thus the multivariate skewed VG and Student's t distributions are suitable for this data set as they can capture the positive skewness and the large leptokurtosis of the data. These model choices do not pose any problem to our proposed ECM algorithms as our simulation studies have confirmed the accuracy of the VG parameter estimates even when the shape parameter falls into the unbounded density region and the Student's t distribution does not have any unbounded density issue. In this case, we use WLOO likelihood methodologies developed in Section

4.2 to overcome this unbounded density problem from the VG distribution while also allowing comparison with other models adopting the Student's t distribution.

Table 5.3. Numerical summaries of the daily returns of cryptocurrencies along with p-values of Box-Pierce test for serial correlation.

Cryptocurrency	median	mean	SD	skewness	kurtosis	Box-Pierce
Bitcoin	−0.0018	0.0013	0.133	1.25	37.6	3.6e-9
Ripple	−0.0022	0.0029	0.076	1.68	41.7	0.0015
Litecoin	−0.0003	0.0012	0.055	0.40	23.7	0.1992
Dash	−0.0021	0.0027	0.070	1.19	17.9	0.6779

Table 5.4. Correlation matrix of the daily returns of cryptocurrencies.

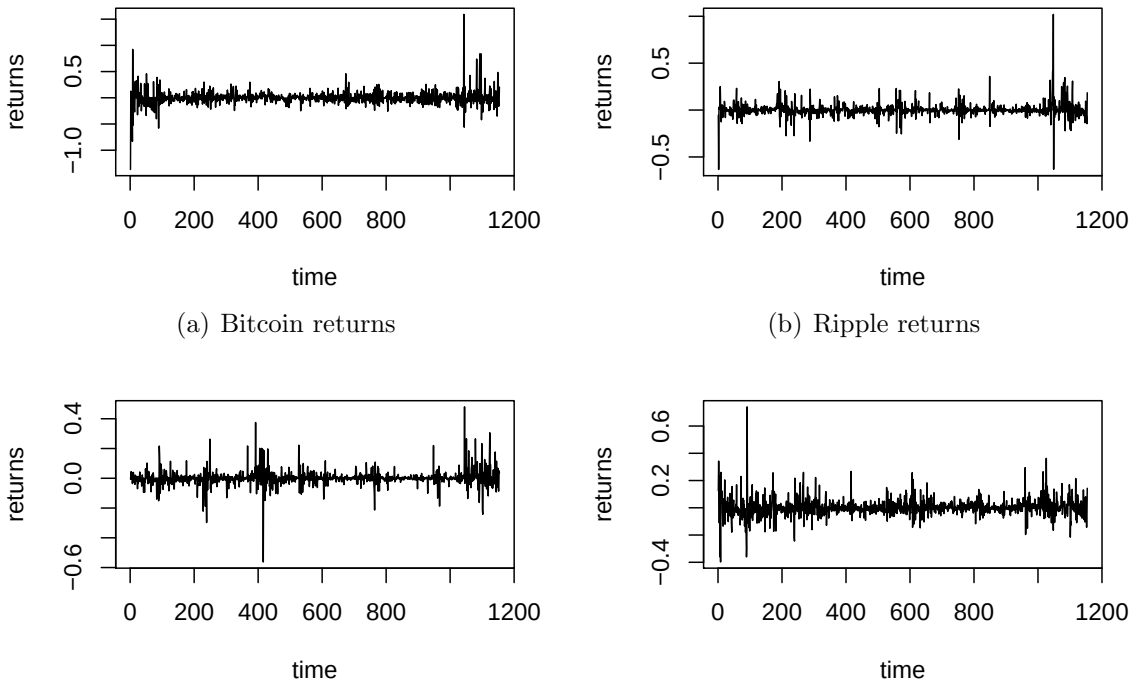
Correlation	Bitcoin	Ripple	Litecoin	Dash
Bitcoin	1.000	0.088	0.131	0.010
Ripple		1.000	0.146	0.029
Litecoin			1.000	0.279
Dash				1.000

The Box-Pierce test [15] is used to test whether there is serial correlation in a return series. The p-values of the test reported in Table 5.3 show significant serial correlations for both Bitcoin and Ripple and suggest the suitability of fitting the returns with time series models. Moreover, the Box-Pierce test in Table 5.3 also shows that Litecoin and Dash have no significant serial correlation. However, the autocorrelation functions (ACF) of the returns in Figures 5.8 show short memory feature of the four return series. In conclusion, it is suitable to use VARMA-VG and VARMA-t model with low orders of p and q to describe the short memory feature.

The Box-Pierce test can also be used to test if there is serial conditional heteroscedasticity and serial conditional correlation between the return series by considering the

Table 5.5. P-values for the Box-Pierce test of serial conditional heteroscedasticity (diagonal entries) and serial conditional correlation (off-diagonal entries) for the series $\{y_{ti}y_{tj}\}$ for $i, j = 1, \dots, 4$.

	Bitcoin	Ripple	Litecoin	Dash
Bitcoin	1.4e-7	0.4756	0.0240	0.0001
Ripple		<2.2e-16	0.0016	0.0276
Litecoin			0.0003	1.7e-15
Dash				6.7e-16



(a) Bitcoin returns

(b) Ripple returns

(c) Litecoin returns

(d) Dash returns

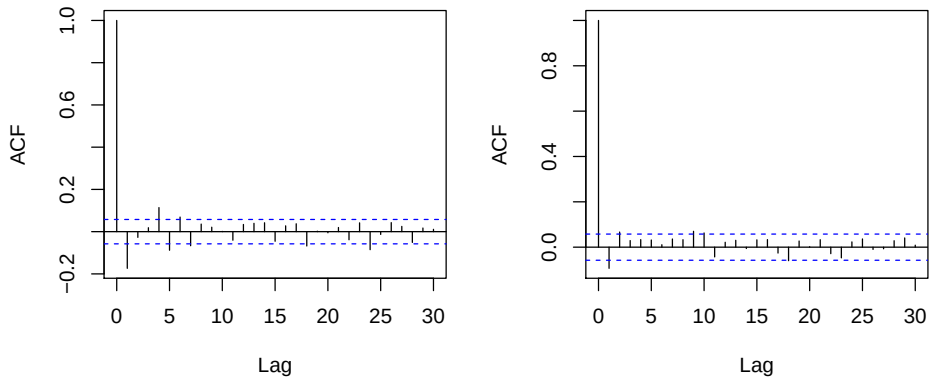
Figure 5.7. Time series plots for the returns of cryptocurrencies.

series $\{y_{t,i}y_{t,j}\}$ for $i, j = 1, \ldots, 4$. The p-values displayed in Table 5.5 show that each return series have serial conditional heteroscedasticity, and most of them have serial conditional correlation. However, current models do not consider these features as they assume that these volatilities and correlations do not depend on time. Though they can be extended to adopt a GARCH-type volatility and dynamic correlation model for future research.

5.6.1.2 Model fitting

To determine the order of the VARMA-VG (or VARMA-t) model, we use the corrected AIC (AICc) based on the WLOO log-likelihood to choose the appropriate orders for the models where the AICc is defined by

$$\text{AICc} = \text{AIC} + \frac{2K(K+1)}{n-K-1}, \tag{5.44}$$



(a) ACF of Bitcoin returns (b) ACF of Ripple returns

(c) ACF of Litecoin returns (d) ACF of Dash returns

Figure 5.8. ACF plots of returns of cryptocurrency

K represents the number of parameters and the AIC is defined as

$$\text{AIC} = -2\ell(\hat{\boldsymbol{\theta}}; \boldsymbol{y}) + 2K. \tag{5.45}$$

The estimation procedures for determining the order of p and q are given as follows:

Step 1: Remove the first few data points so that each estimation procedure contains the same number of data points. This enables model comparison using AICc. In our analysis, we choose to remove the first five data points.

Step 2: Start the estimation procedure choosing $p = 0$ and $q = 0$, then fit the VARMA(0,0)-VG model to the return series and calculate the AICc.

Step 3: For $k \mapsto k+1$ iteration, fit VARMA(p,q)-VG model to the return series for all p and q such that $p+q=k$. Then calculate the AICc.

Step 4: Repeat step 3 until the AICc no longer decreases.

Table 5.6 and 5.7 reports the AICc for the VARMA-VG and VARMA-t models respectively where the AICc are based on the WLOO log-likelihood to facilitate comparison between the VG and Student's t distributions for the innovations when using the ECM algorithm with CM-step for κ in Section 2.2 and the observed log-likelihood in the CM-step for ν for the VG distribution (or υ for the Student's t distribution). Results show that the best p and q are 2 and 0 respectively for both VARMA-VG and VARMA-t models, that is, the best models are VARMA(2,0)-VG and VARMA(2,0)-t respectively as they give the smallest AICc which are bolded in these tables.

Comparing the AICc for these two models, one finds that overall VARMA(2,0)-t model gives a better model fit than VARMA(2,0)-VG model. Although the VG distribution captures the behaviour around the peak better than the Student's t distribution as demonstrated by comparing Figures 5.9 and 5.10, the Student's t distribution captures the heavy-tailed behaviour better than VG distribution which is the dominating factor in the AICc for the cryptocurrency data.

Apart from comparing model fit, Table 5.8 compares computational time and number of iterations using various estimation methods for the VARMA(2,0) model using the two distributions. Using the first method in the table, the computational efficiency for VG versus Student's t innovations differ by a factor of 77 times in terms of both computation time (15 vs 1152) and number of iterations (12 vs 925). These trends roughly apply to other methods involving other combination of steps in the ECM algorithm such as the full log-likelihood with adaptive Δ for the maximisation and the Q-function in the CM-step for ν (or υ). One possible reason for the faster convergence and shorter computational time is due to the super-efficiency property for the location estimate of VG distribution using the LOO or WLOO log-likelihood which assists the convergence of the ECM algorithm.

Comparing with the full likelihood methods, the WLOO likelihood converges faster and is computationally more efficient than the full likelihood with adaptive Δ. The improvement in computational efficiency can range up to a factor of seven for the VG distribution and a factor of three for the Student's t distribution. Moreover, when

comparing the accuracy of these likelihood methods, Section 4.4 shows that WLOO likelihood method overall performs better than full likelihood method. Thus, in this case, the WLOO likelihood is more preferable than the full likelihood method with adaptive Δ.

The parameter estimates and SE estimates using the first method in Table 5.8 are reported in Tables 5.9 and 5.10 for the VARMA(2,0)-VG and VARMA(2,0)-t model respectively. Looking at the scale and skewness parameter estimates, the fact that $\hat{\Sigma}_{1,1}$ is higher than other diagonal entries of $\hat{\boldsymbol{\Sigma}}$ and all elements of $\hat{\boldsymbol{\gamma}}$ are positive for both models are consistent with the observations that Bitcoin has higher SD and each cryptocurrencies is slightly positively skewed as shown in Table 5.3. Moreover $\hat{\Sigma}_{3,4}$ is most significant whereas $\hat{\Sigma}_{1,4}$ is least significant for VARMA-VG model. These results also agree with the observation in Table 5.4. In fact, estimates in $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\gamma}}$ are all significant for VARMA-VG model and mostly significant for VARMA-t model which again testify the suitability of the VARMA(2,0) models.

Looking at the persistence parameters, the diagonal entries of $\hat{\boldsymbol{A}}_1$ and $\hat{\boldsymbol{A}}_2$ show that the persistence is generally negative and it is much weaker for Litecoin and Dash compared with Bitcoin and Ripple. These results are consistent with the Box-Pierce test in Table 5.3. For the off-diagonal entities, most of the cross-persistence are not significant. Since the shape parameter estimate for VG innovations falls inside the unbounded density region ($\hat{\nu} \leq \frac{d}{2}$), it suggests the high density of points about the centre of the distribution. So we apply the WLOO likelihood methodology in Section 4.2 to remove the data point that gives rise to unbounded density so that the ML estimation is well-defined. The small Student's t shape parameter lying very close to the boundary of the parameter space also suggest the extreme heavy-tailness of the cryptocurrency returns.

The SE estimates for VARMA-VG model can be calculated using Louis' method in (2.20) except the SE for $\hat{\boldsymbol{\mu}}$ is obtained using the double generalised gamma approximation for WLOO likelihood method as described in Section 3.4.2.2 and the derivatives are given in Appendix A9. Louis' method also applies to the VARMA-t model with adjustments made to the conditional expectations using the conditional distribution in (5.35) and the inverse-gamma log-likelihood using the derivatives in (5.39) and (5.40).

Table 5.6. AICc of VARMA-VG model for different p's and q's.

AICc	$q=0$	$q=1$	$q=2$	$q=3$
$p=0$	-13287	-13325	-13374	-13357
$p=1$	-13314	-13328	-13188	
$p=2$	$\mathbf{-13384}$	-13187		
$p=3$	-13356			

Table 5.7. AICc of VARMA-t model for different p's and q's.

AICc	$q=0$	$q=1$	$q=2$	$q=3$
$p=0$	-13344	-13411	-13432	-13412
$p=1$	-13400	-13387	-13331	
$p=2$	$\mathbf{-13432}$	-13205		
$p=3$	-13420			

Table 5.8. Computational time and number of iterations until convergence using different ECM algorithms for VARMA(2,0)-VG and VARMA(2,0)-t models.

			VG		Student's t	
Log-likelihood	CM-step for κ	CM-step for ν (or υ)	Time (s)	iter	Time (s)	iter
WLOO	✓	$\ell_{\text{obs}}^{\text{WLOO}}$	15	12	1152	925
WLOO	✓	Q^{WLOO}	11	17	402	721
WLOO		$\ell_{\text{obs}}^{\text{WLOO}}$	40	27	3608	2579
WLOO		Q^{WLOO}	38	38	1989	2513
adaptive Δ	✓	ℓ_{obs}	93	67	1679	2072
adaptive Δ	✓	Q	85	88	1166	1299
adaptive Δ		ℓ_{obs}	66	53	5922	4679
adaptive Δ		Q	108	98	3514	4745

5.6.1.3 Model fit assessment

To assess the performance of the VARMA(2,0)-VG and VARMA(2,0)-t models, we plot the density of the errors using Gaussian kernel density estimation in Figure 5.9 and 5.10 respectively. For comparison, we also include the marginal pdfs for VG and Student's t distributions respectively and each plot also includes a univariate normal pdf to show the high level of kurtosis for each component of the errors. These plots show that these models capture the overall shape of the density especially the high peak

Table 5.9. Parameter estimates and SEs for the VARMA(2,0)-VG model using the first method in Table 5.8.

parameter	estimate	SE
$\boldsymbol{\mu}^\top$	$\begin{pmatrix} -0.004 & -0.003 & -0.001 & -0.003 \end{pmatrix}$	$\begin{pmatrix} 0.002 & 0.001 & 0.001 & 0.001 \end{pmatrix}$
$\boldsymbol{A}_1$	$\begin{pmatrix} -0.191 & 0.063 & 0.054 & 0.079 \\ -0.003 & -0.131 & 0.013 & 0.004 \\ -0.007 & 0.003 & -0.072 & 0.007 \\ -0.010 & -0.039 & -0.048 & -0.013 \end{pmatrix}$	$\begin{pmatrix} 0.031 & 0.043 & 0.059 & 0.083 \\ 0.010 & 0.026 & 0.036 & 0.036 \\ 0.011 & 0.014 & 0.026 & 0.011 \\ 0.014 & 0.024 & 0.026 & 0.018 \end{pmatrix}$
$\boldsymbol{A}_2$	$\begin{pmatrix} -0.069 & 0.075 & -0.037 & 0.016 \\ -0.026 & -0.028 & 0.003 & -0.038 \\ 0.003 & 0.043 & -0.129 & 0.022 \\ -0.049 & 0.077 & -0.030 & -0.084 \end{pmatrix}$	$\begin{pmatrix} 0.006 & 0.034 & 0.048 & 0.037 \\ 0.002 & 0.022 & 0.025 & 0.014 \\ 0.002 & 0.012 & 0.021 & 0.010 \\ 0.002 & 0.021 & 0.025 & 0.010 \end{pmatrix}$
$\boldsymbol{\Sigma}$	$\begin{pmatrix} 0.0138 & 0.0010 & 0.0008 & 0.0004 \\ & 0.0038 & 0.0005 & 0.0006 \\ & & 0.0021 & 0.0008 \\ & & & 0.0051 \end{pmatrix}$	$\begin{pmatrix} 0.0009 & 0.0002 & 0.0002 & 0.0003 \\ & 0.0002 & 0.0001 & 0.0001 \\ & & 0.0001 & 0.0001 \\ & & & 0.0003 \end{pmatrix}$
$\boldsymbol{\gamma}^\top$	$\begin{pmatrix} 0.007 & 0.006 & 0.002 & 0.006 \end{pmatrix}$	$\begin{pmatrix} 0.003 & 0.002 & 0.001 & 0.002 \end{pmatrix}$
ν	0.7447	0.0333

Table 5.10. Parameter estimates and SEs for the VARMA(2,0)-t model using the first method in Table 5.8.

parameter	estimate	SE
$\boldsymbol{\mu}^\top$	$\begin{pmatrix} -0.004 & -0.003 & -0.001 & -0.002 \end{pmatrix}$	$\begin{pmatrix} 0.002 & 0.001 & 0.001 & 0.001 \end{pmatrix}$
$\boldsymbol{A}_1$	$\begin{pmatrix} -0.219 & 0.033 & 0.055 & 0.060 \\ -0.008 & -0.131 & 0.025 & -0.013 \\ -0.003 & 0.006 & -0.064 & 0.014 \\ 0.011 & -0.046 & -0.030 & -0.035 \end{pmatrix}$	$\begin{pmatrix} 0.027 & 0.042 & 0.054 & 0.042 \\ 0.012 & 0.026 & 0.026 & 0.020 \\ 0.009 & 0.016 & 0.020 & 0.015 \\ 0.015 & 0.026 & 0.032 & 0.029 \end{pmatrix}$
$\boldsymbol{A}_2$	$\begin{pmatrix} -0.056 & 0.080 & -0.076 & -0.021 \\ -0.017 & -0.049 & 0.022 & -0.049 \\ 0.005 & 0.028 & -0.104 & -0.002 \\ -0.032 & 0.030 & -0.024 & -0.076 \end{pmatrix}$	$\begin{pmatrix} 0.025 & 0.036 & 0.052 & 0.041 \\ 0.012 & 0.022 & 0.027 & 0.020 \\ 0.009 & 0.013 & 0.022 & 0.016 \\ 0.015 & 0.022 & 0.031 & 0.028 \end{pmatrix}$
$\boldsymbol{\Sigma}$	$\begin{pmatrix} 0.0929 & 0.0065 & 0.0056 & 0.0032 \\ & 0.0242 & 0.0031 & 0.0043 \\ & & 0.0135 & 0.0056 \\ & & & 0.0355 \end{pmatrix}$	$\begin{pmatrix} 0.0126 & 0.0017 & 0.0024 & 0.0070 \\ & 0.0030 & 0.0185 & 0.0334 \\ & & 0.0200 & 0.0109 \\ & & & 0.0201 \end{pmatrix}$
$\boldsymbol{\gamma}^\top$	$\begin{pmatrix} 0.040 & 0.033 & 0.012 & 0.025 \end{pmatrix}$	$\begin{pmatrix} 0.048 & 0.004 & 0.003 & 0.003 \end{pmatrix}$
υ	2.0805	0.0441

for the errors of Bitcoin and Dash. For Ripple and Litecoin, the peaks using the VG and Student's t distribution is lower than the kernel density estimate suggesting that the shape parameter for the VG and Student's t distributions is not small enough for these two cryptocurrencies. This is expected as the model assumes a common shape parameter for all four cryptocurrencies. In all cases, the normal distribution can not capture the behaviour of the peak, intermediate tails and extreme tails.

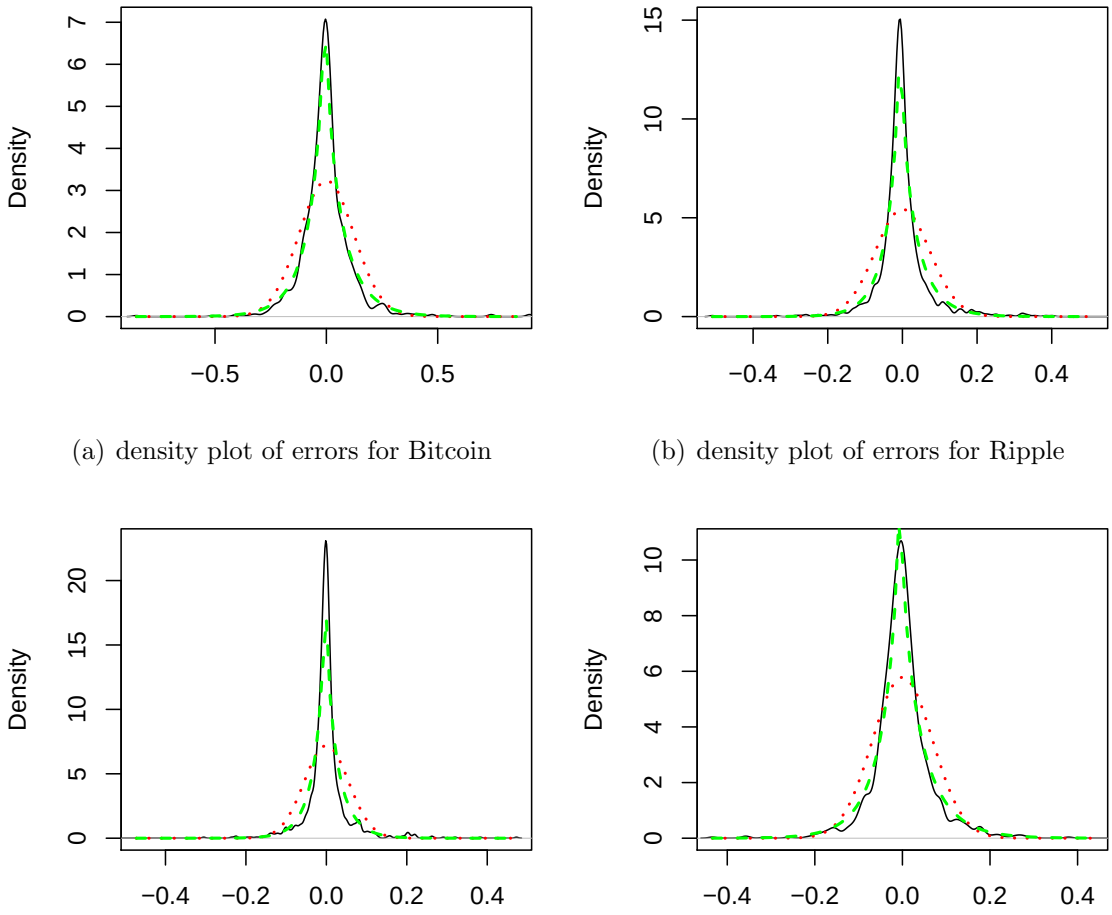
In summary, the VARMA(2,0) model captures the short memory feature with weaker persistence, cross-correlation, positive skewness as well as high kurtosis. For the features of high kurtosis, the VG distribution describes the high concentration of data points around the centre better whereas the Student's t distribution models the heavy tails better. Although Student's t model is preferred in terms of AICc which is often dominated by fitting better the few outliers at the cost of fitting the peak behaviour, VG model demonstrates high computational efficiency due to its super-efficiency property for its location estimate. This is particularly an advantage when analysing high frequency financial time series data which is the focus for the next application. The next section aims to study the effects of increasing sampling frequency on the kurtosis, in particular, the peak of the data distribution and the ability for the shape parameter of the VG distribution to capture the varying levels of peakness. In this regard, the VARMA-t model is not considered.

5.6.2 Stock market indices

This analysis is divided into two parts. The first analysis considers the VAR(1)-VG model fitted to the same daily return data in Section 2.6. This analysis demonstrates the improvement in model performance of VAR-VG model over the VG model in Section 2.6. The second analysis studies how the sampling frequency affects the characteristics of the return series.

5.6.2.1 Daily stock indices

The description of the data can be found in Section 2.6, the time series plots are displayed in Figure 2.2, the summary of the data are given in Table 2.4, and the plots



(a) density plot of errors for Bitcoin

(b) density plot of errors for Ripple

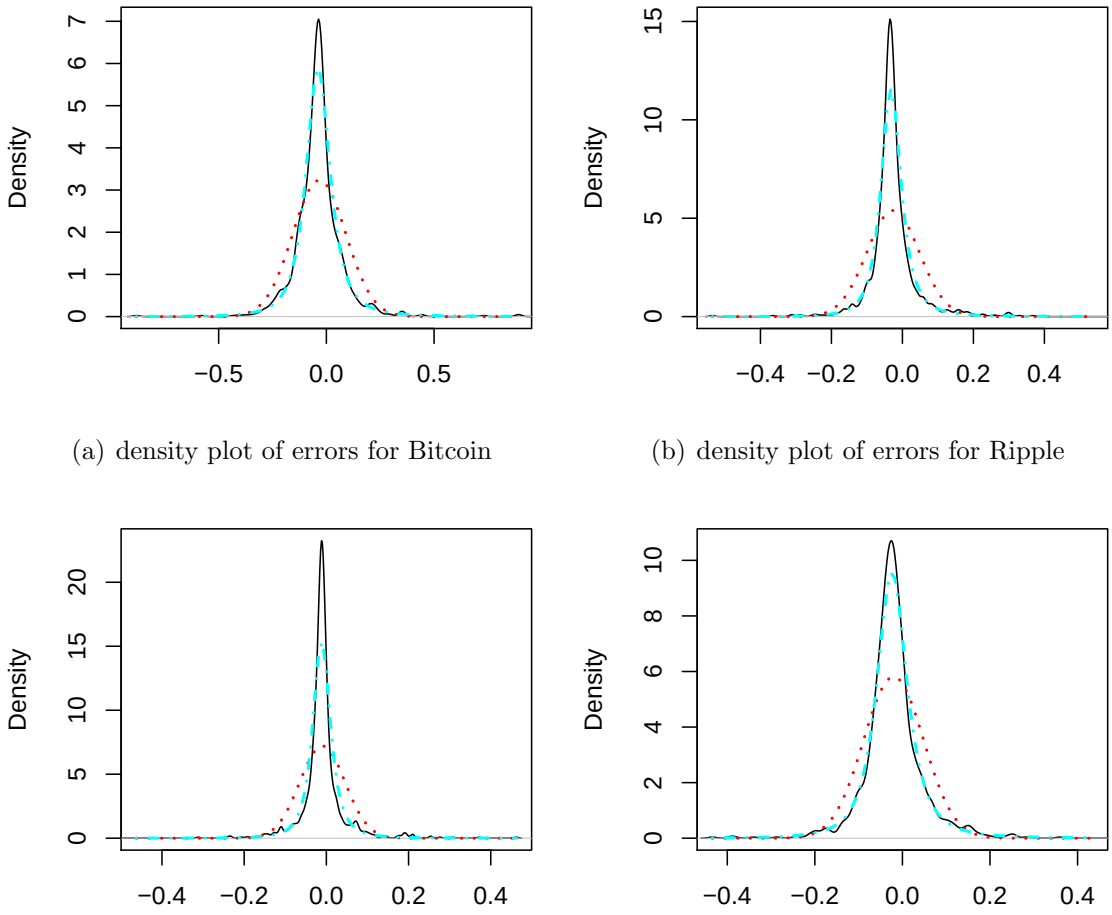
(c) density plot of errors for Litecoin

(d) density plot of errors for Dash

Figure 5.9. Density plots (black solid line) of the errors of the VARMA(2,0)-VG model for Bitcoin, Ripple, Litecoin and Dash returns. The density for VG (green dashed line) and univariate normal (red dotted line) are included for comparison.

of the autocorrelation function (ACF) are given in Figure 5.11. Results for fitting VAR(1)-VG model is reported in Table 5.11 based on the AECM algorithm using the full likelihood with adaptive Δ to facilitate comparison with the VAR(0)-VG results in Section 2.6.

The VAR(1)-VG model provides good performance as illustrated in the density plots of the residuals in Figure 5.12 after filtering out the AR(1) term. Fitted marginal



(a) density plot of errors for Bitcoin

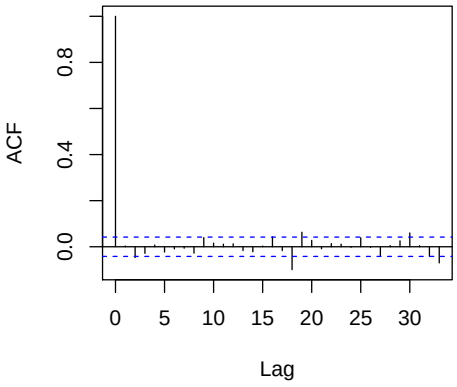
(b) density plot of errors for Ripple

(c) density plot of errors for Litecoin

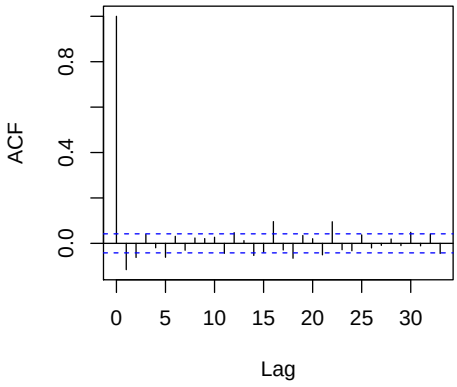
(d) density plot of errors for Dash

Figure 5.10. Density plots (black solid line) of the errors of the VARMA(2,0)-t model for Bitcoin, Ripple, Litecoin and Dash returns. The density for Student's t (cyan dashed-dotted line) and univariate normal (red dotted line) are included for comparison.

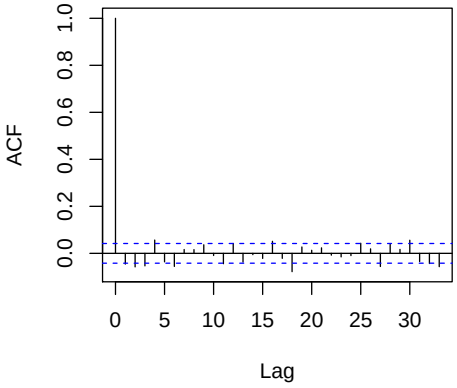
pdfs of the VG distribution are added to the figure to facilitate comparison. However, occasionally the peaks of the density estimates and fitted pdfs does not match, for example, the peaks of S&P 500 and CAC 40 are underestimated whereas the peak for AORD is overestimated. This is due to the rather strong assumption of a common shape parameter ν across all components. We also note that AICc is not adopted for model fit comparison because the capping level differ between the two models making the two capped likelihoods incomparable.



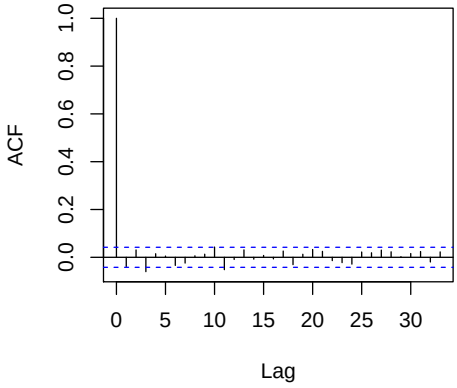
(a) DAX



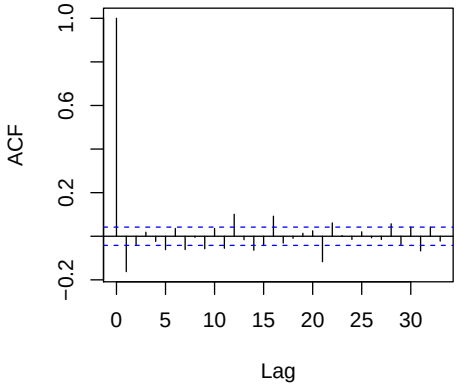
(b) S&P 500



(c) FTSE 100



(d) AORD



(e) CAC 40

Figure 5.11. Plot of ACF for DAX, S&P 500, FTSE 100, AORD and CAC 40 daily returns.

Table 5.11. Estimates, SEs and correlation matrix $\boldsymbol{\rho}$ for the VAR(1)-VG model using DAX, S&P 500, FTSE 100, AORD and CAC 40 daily return series

	Estimates	Standard errors
$\boldsymbol{\mu}'$	$10^{-4}\begin{pmatrix}16.0 & 13.0 & 9.6 & 13.7 & 2.1\end{pmatrix}$	$10^{-4}\begin{pmatrix}3.4 & 2.8 & 2.7 & 2.3 & 6.3\end{pmatrix}$
$\boldsymbol{A}_1$	$10^{-2}\begin{pmatrix}-8.2 & 34.7 & -18.0 & -1.4 & 2.1\\ 7.5 & -10.7 & -3.6 & -6.8 & -0.6\\ -13.9 & 36.8 & -12.9 & -1.2 & -0.4\\ -3.9 & 43.1 & 23.7 & -20.8 & -2.0\\ 10.3 & -27.3 & 0.9 & -9.3 & -4.6\end{pmatrix}$	$10^{-2}\begin{pmatrix}3.7 & 3.3 & 4.6 & 2.6 & 1.1\\ 3.0 & 2.8 & 3.9 & 2.3 & 1.0\\ 2.5 & 2.7 & 3.5 & 2.1 & 1.0\\ 2.4 & 2.3 & 3.0 & 1.8 & 0.8\\ 5.7 & 6.3 & 8.0 & 4.9 & 2.3\end{pmatrix}$
$\boldsymbol{\Sigma}$	$10^{-5}\begin{pmatrix}17.3 & 9.9 & 12.1 & 3.2 & 12.6\\ & 13.5 & 8.1 & 2.1 & 16.5\\ & & 11.9 & 3.1 & 9.6\\ & & & 7.4 & 2.0\\ & & & & 64.1\end{pmatrix}$	$10^{-6}\begin{pmatrix}6.8 & 5.4 & 5.3 & 3.7 & 10.9\\ & 5.2 & 4.4 & 3.2 & 9.9\\ & & 4.6 & 3.1 & 9.0\\ & & & 2.9 & 6.9\\ & & & & 24.9\end{pmatrix}$
$\boldsymbol{\gamma}'$	$10^{-4}\begin{pmatrix}-13.0 & -11.9 & -8.2 & -12.4 & -0.2\end{pmatrix}$	$10^{-4}\begin{pmatrix}5.2 & 3.7 & 4.3 & 2.3 & 8.5\end{pmatrix}$
ν	1.45	0.057
$\boldsymbol{\rho}$	$\begin{pmatrix}1 & 0.65 & 0.85 & 0.29 & 0.38\\ & 1 & 0.64 & 0.22 & 0.56\\ & & 1 & 0.33 & 0.35\\ & & & 1 & 0.09\\ & & & & 1\end{pmatrix}$	

Comparing the estimates of the VAR(0)-VG model in Table 2.5 with the VAR(1)-VG model in Table 5.11, the estimate $\hat{\boldsymbol{\Sigma}}$ are very similar while $\hat{\boldsymbol{\gamma}}$ are also quite similar. Since the second column of the $\hat{\boldsymbol{A}}_1$ matrix has relatively larger values, all five stocks are strongly cross-correlated (of lag one) with S&P 500. It is not surprising to know that the returns in S&P 500 has the most impact on each of the five returns the next day because S&P 500 has been shown to be a strong predictor for a number of market indices. This is due to its large market share and its minimal real time difference with lag-one return on the other markets.

In addition, by observing that the first and third rows of $\hat{\boldsymbol{A}}_1$ matrix in Table 5.11 are similar, both market indices DAX and FTSE 100 have similar cross-persistence with the other stocks. Moreover, the correlation matrices of the two models in Tables 2.5 and 5.11 respectively show that DAX and FTSE 100 are highly correlated ($\hat{\rho}_{13} \simeq 0.85$). These strong cross persistence and correlation are possibly due to the strong competitiveness of the German and UK markets as they are the major stock markets in Europe.

To visualise the strong correlation, Figure 5.13 gives the scatter plot of the residuals of DAX and FTSE 100 returns along with the fitted contour plot using estimates from the VAR(1)-VG model. It shows high density of points in the central region as well as strong linear dependence between the two market indices. In summary, the VAR(1)-VG model can capture the strong cross-persistence between the four stocks with S&P 500 as well as between DAX and FTSE 100.

5.6.2.2 Forecast

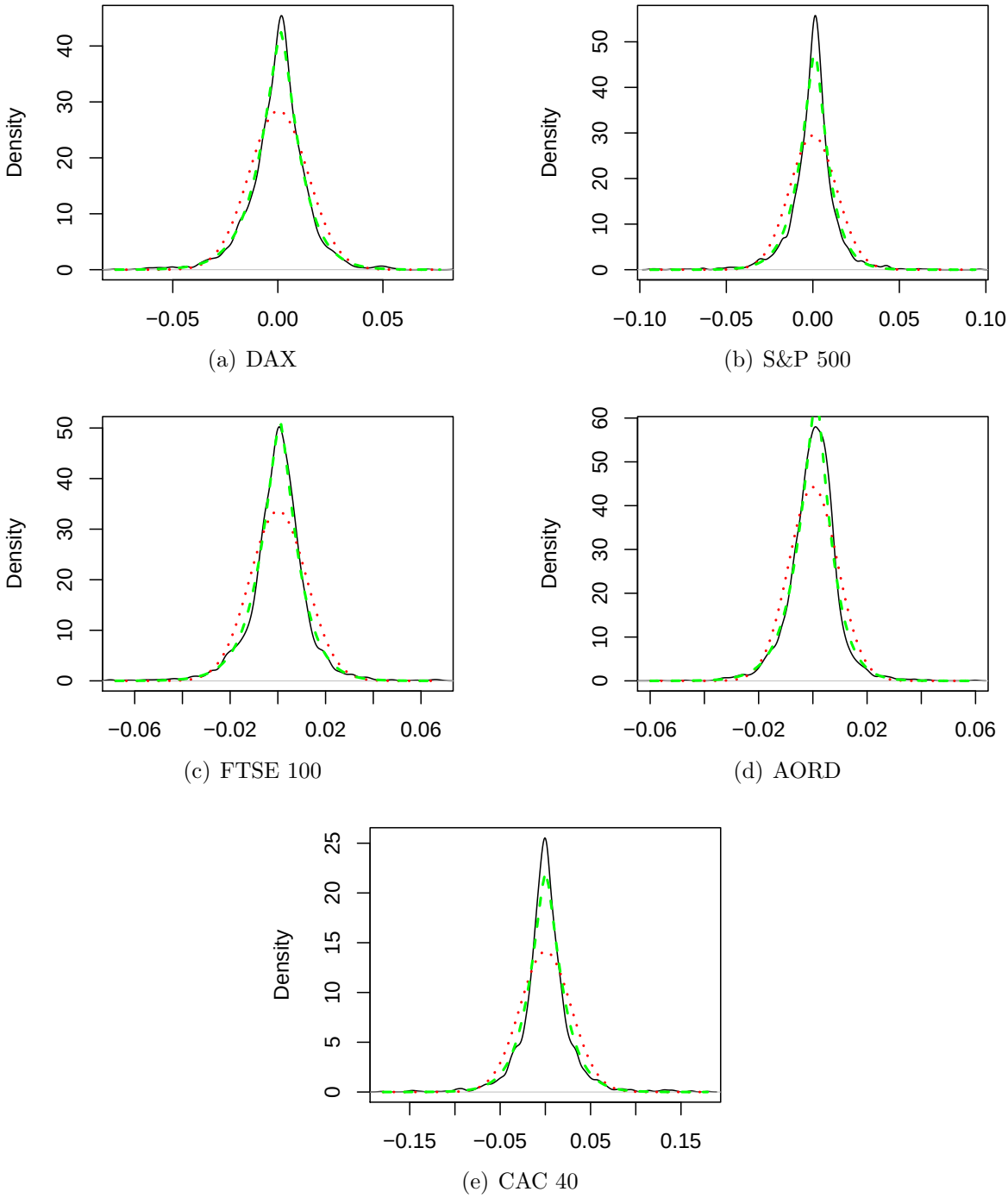
To demonstrate the forecast ability of the VAR-VG model, we consider the VAR(1)-VG model and forecast $\boldsymbol{y}_{T+s}$ for $T = 2000$ and $s = 1, \ldots, 188$ via a sequence of 1-step ahead forecasts by fitting repeatedly to the sliding window $\mathscr{F}_{s:T+s-1} = (\boldsymbol{y}_s, \ldots, \boldsymbol{y}_{T+s-1})$ to obtain parameter estimates $\hat{\boldsymbol{\theta}}_s = (\hat{\boldsymbol{c}}_s, \hat{\boldsymbol{A}}_{s,1}, \widehat{\boldsymbol{\Sigma}}_s, \hat{\boldsymbol{\gamma}}_s, \hat{\nu}_s)$. The forecasts $\hat{\boldsymbol{y}}_{T+s}$ are obtained by

$$\hat{\boldsymbol{y}}_{T+s} = \mathbb{E}(\boldsymbol{y}_{T+s}) = \hat{\boldsymbol{c}}_s + \hat{\boldsymbol{A}}_{s,1}\boldsymbol{y}_{T+s-1} + \hat{\boldsymbol{\gamma}}_s \tag{5.46}$$

using (5.46) and (5.26) such that $\mathbb{E}(\boldsymbol{\varepsilon}_{s,t}) = \mathbf{0}$. We note that $\mathbb{E}(\boldsymbol{y}_{T+s})$ aims to capture the general trend of the process excluding noises. The forecasts as plotted in Figure 5.14 can capture the general trends of all return series, especially for AORD and FTSE 100 daily stock indices. Results agree with the in-sample filtered residuals in Figure 5.12 that the VAR(1)-VG model can capture the kurtosis and dynamics of AORD and FTSE 100 indices better than to other indices. After demonstrating the improvement of model performance with the AR time series structure, we show in the second analysis how the more general VARMA model can accommodate increasing kurtosis level due to increasing sampling frequency for these market indices.

5.6.3 High frequency stock indices

Again, we consider the returns of four different stock market indices, namely, Australian Securities Exchange 200 index (ASX 200), Cotation Assistée en Continu 40 index (CAC 40), Financial Times Stock Exchange 100 index (FTSE 100) and Standard and Poor 500 index (S&P 500) which are based on the Australian, French, London, and U.S. stock exchange respectively. We take the closing prices from 29th August 2016 to 29th August 2017 with sampling frequencies of 1 hour, 15 minutes, 5 minutes and 1 minute.



(a) DAX

(b) S&P 500

(c) FTSE 100

(d) AORD

(e) CAC 40

Figure 5.12. Density plots of VAR(1)-VG residuals (solid black line), pdf of the VG distribution after filtering the mean function (green dash line) and pdf of fitted univariate normal (red dotted line) for the daily returns of DAX, S&P 500, FTSE 100, AORD and CAC 40.

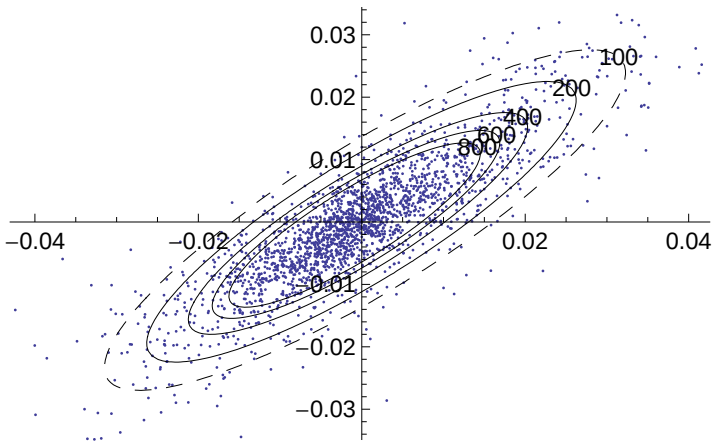
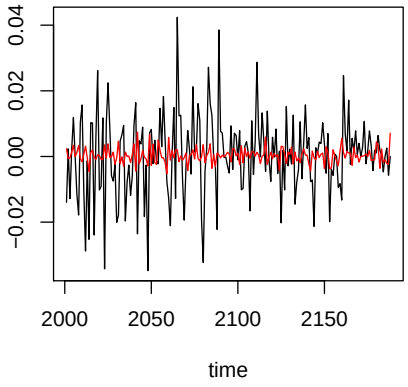


Figure 5.13. Fitted contour plot of VAR(1)-VG for DAX and FTSE 100 data sets after filtering the mean function.

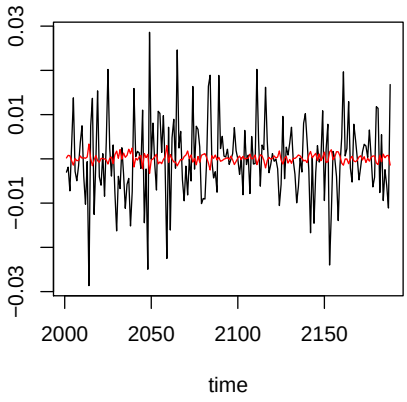
The series are adjusted so that the prices are aligned based on time from the stock market opening times for each day. The sample sizes are 1606, 5604, 16523 and 79814 for the sampling frequencies of 1 hour, 15 minutes, 5 minutes and 1 minute respectively.

From the numeric summaries in Table 5.12, as the sampling frequency increases, the mean, SD and correlation decrease while the skewness and kurtosis increase. The large increase in skewness and kurtosis are due to the large number of outliers along with the increasing sampling frequency. To mitigate the effect of the outliers, robust estimation of moments is included in Table 5.13 using the sample median, median absolute deviation (MAD), Bowley's skewness and Moors' kurtosis [13]. Based on these new robust estimates, as the sampling frequency increases, the overall magnitude of the skewness decreases, the kurtosis remains roughly constant but the serial correlation as well as heteroscedasticity increases. So we expect estimates for the skewness and shape parameter to exhibit behaviours in agreement with the robust estimates. The only exception to the decreasing trend of correlation is the correlation between CAC 40 and FTSE 100 indices which stays strong across increasing sampling frequencies.

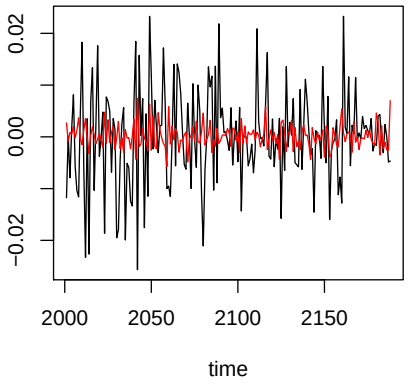
The Box-Pierce test in Table 5.13 is applied to test if there is autocorrelation in the high frequency returns for each index and sampling frequency. From this test, there is no autocorrelation for 1 hour returns, whereas there is autocorrelation for all except FTSE 100 indices for 1 minute returns. In general, there is a trend of increasing autocorrelation with sampling frequency. From the autocorrelation plots in Figure



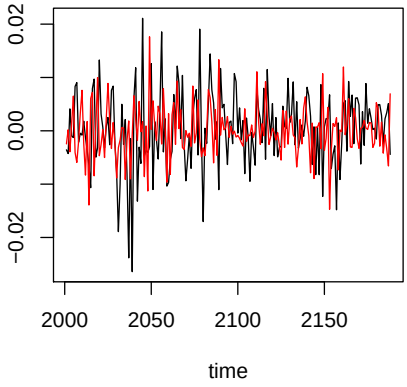
(a) forecasting DAX



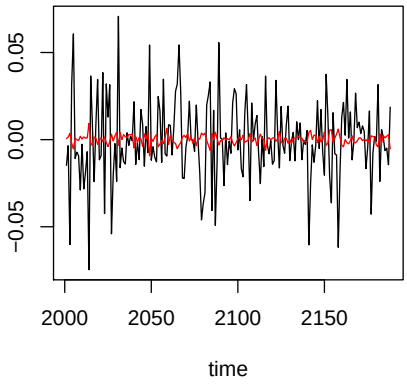
(b) forecasting S&P 500



(c) forecasting FTSE 100



(d) forecasting AORD



(e) forecasting CAC 40

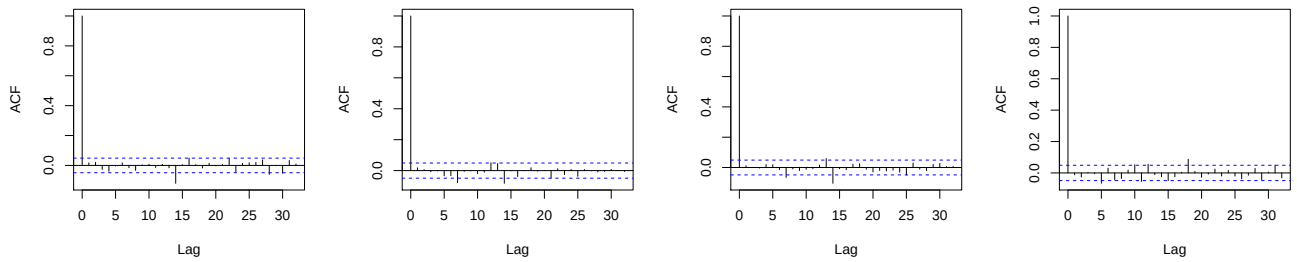
Figure 5.14. Observed (black line) and predicted (red line) returns for Bitcoin, Ripple, Litecoin and Dash using VAR(1)-VG model.

Table 5.12. Numerical summaries of 1 hour, 15 min, 5 min and 1 min sampling frequencies of ASX 200, CAC 40, FTSE 100 and S&P 500 returns.

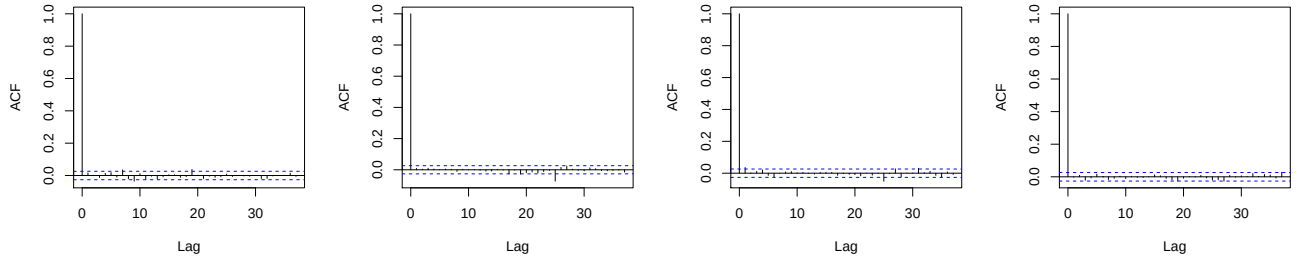
Freq	Index	Mean	SD	Skewness	Kurtosis	Correlation
1 hr	ASX 200	2.8e-5	2.7e-3	0.11	20.0	$\begin{pmatrix} 1 & 0.31 & 0.30 & 0.23 \\ & 1 & 0.76 & 0.25 \\ & & 1 & 0.22 \\ & & & 1 \end{pmatrix}$
	CAC 40	8.4e-5	3.3e-3	0.66	20.3	
	FTSE 100	5.0e-5	2.6e-3	−0.16	11.4	
	S&P 500	6.9e-5	1.9e-3	0.63	12.8	
15 min	ASX 200	0.8e-5	1.4e-3	0.67	57.7	$\begin{pmatrix} 1 & 0.32 & 0.30 & 0.22 \\ & 1 & 0.79 & 0.20 \\ & & 1 & 0.22 \\ & & & 1 \end{pmatrix}$
	CAC 40	2.5e-5	1.7e-3	1.36	64.7	
	FTSE 100	1.4e-5	1.3e-3	0.28	38.0	
	S&P 500	2.0e-5	1.0e-3	1.28	38.2	
5 min	ASX 200	2.8e-6	7.3e-4	0.87	68.5	$\begin{pmatrix} 1 & 0.18 & 0.17 & 0.13 \\ & 1 & 0.79 & 0.20 \\ & & 1 & 0.21 \\ & & & 1 \end{pmatrix}$
	CAC 40	8.7e-6	10.6e-4	3.10	181.6	
	FTSE 100	4.9e-6	7.9e-4	0.94	101.6	
	S&P 500	6.7e-6	5.9e-4	2.42	98.9	
1 min	ASX 200	0.6e-6	2.9e-4	5.16	343.3	$\begin{pmatrix} 1 & 0.11 & 0.12 & 0.08 \\ & 1 & 0.76 & 0.20 \\ & & 1 & 0.20 \\ & & & 1 \end{pmatrix}$
	CAC 40	1.8e-6	4.7e-4	8.14	1107.8	
	FTSE 100	1.0e-6	3.5e-4	2.51	620.3	
	S&P 500	1.4e-6	2.6e-4	4.69	460.2	

Table 5.13. Robust numerical summaries for different sampling frequencies and indices. P-values of the Box-Pierce test on the returns and returns squared are also included.

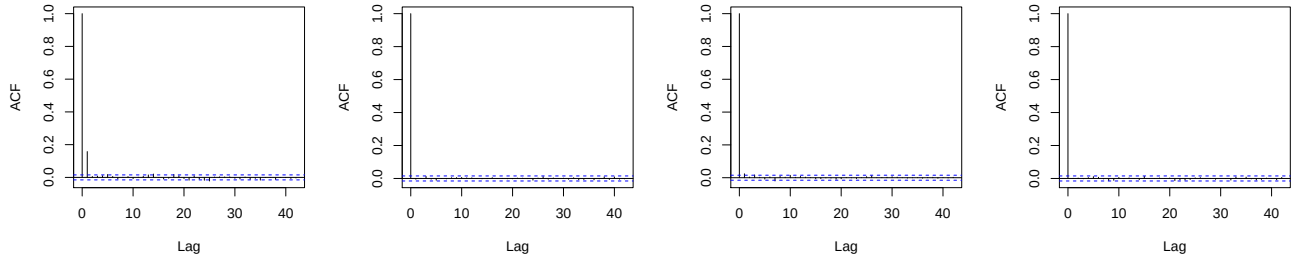
Freq	Index	Median	MAD	Bowley's skewness	Moors' kurtosis	BP $\{y_{t,j}\}$	BP $\{y_{t,j}^2\}$
1 hr	ASX 200	8.7e-5	1.0e-3	−1.2e-2	0.24	0.49	0.00
	CAC 40	3.9e-5	1.2e-3	4.4e-2	0.12	0.47	0.00
	FTSE 100	3.1e-5	1.0e-3	5.9e-2	0.19	0.64	0.00
	S&P 500	0.0e-5	6.6e-4	7.6e-2	0.95	0.62	0.00
15 min	ASX 200	1.7e-5	5.0e-4	−0.2e-2	0.13	0.23	0.00
	CAC 40	0.7e-5	5.6e-4	−1.1e-2	0.25	0.53	0.57
	FTSE 100	1.7e-5	4.6e-4	−4.4e-2	0.25	0.01	0.14
	S&P 500	0.0e-5	2.7e-4	7.9e-2	0.99	0.84	0.36
5 min	ASX 200	0	2.6e-4	2.0e-2	0.20	0.00	0.00
	CAC 40	0	3.0e-4	−1.3e-2	0.22	0.55	0.00
	FTSE 100	0	2.5e-4	−0.4e-2	0.18	0.00	0.00
	S&P 500	0	1.5e-4	5.7e-2	0.99	0.89	0.13
1 min	ASX 200	0	1.0e-4	4.1e-3	0.19	0.00	0.00
	CAC 40	0	1.1e-4	4.4e-3	0.34	0.03	0.00
	FTSE 100	0	1.0e-4	−0.4e-3	0.22	0.36	0.00
	S&P 500	0	0.6e-4	31.5e-3	1.04	0.00	0.00



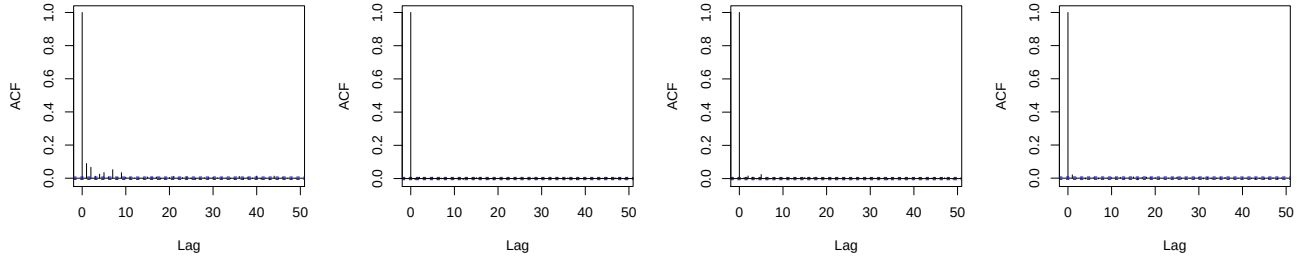
(a) 1 hr, ASX 200 (b) 1 hr, CAC 40 (c) 1 hr, FTSE 100 (d) 1 hr, S&P 500



(e) 15 min, ASX 200 (f) 15 min, CAC 40 (g) 15 min, FTSE 100 (h) 15 min, S&P 500



(i) 5 min, ASX 200 (j) 5 min, CAC 40 (k) 5 min, FTSE 100 (l) 5 min, S&P 500



(m) 1 min, ASX 200 (n) 1 min, CAC 40 (o) 1 min, FTSE 100 (p) 1 min, S&P 500

Figure 5.15. ACFs of the returns for the 1 hr, 15 min, 5 min and 1 min (rows) of ASX 200, CAC 40, FTSE 100 and S&P 500 returns (columns).

5.15, all the autocorrelations except for 1 minute ASX 200 decay very quickly to 0 suggesting a low order VARMA-VG model is suitable for this data set.

Using the model fitting procedure in Section 5.3.3 to estimate the orders of VARMA(p, q)-VG model, we obtained the results in Tables 5.14, 5.15 and 5.16. As $\hat{\nu} < d/2$ for all sampling frequencies, the VG distribution has unbounded density indicating the high level of kurtosis in the data.

Table 5.14. AICc of VARMA-VG model for different p and q's and different sampling frequencies.

1 hour	$\begin{array}{c} \\ p=0 \\ p=1 \end{array}\begin{array}{c} \begin{array}{cc} q=0 & q=1 \end{array} \\ \begin{bmatrix} \mathbf{-62166} & -62157 \\ -62157 & \end{bmatrix} \end{array}$
15 min	$\begin{array}{c} \\ p=0 \\ p=1 \end{array}\begin{array}{c} \begin{array}{cc} q=0 & q=1 \end{array} \\ \begin{bmatrix} \mathbf{-252346} & -252341 \\ -252342 & \end{bmatrix} \end{array}$
5 min	$\begin{array}{c} \\ p=0 \\ p=1 \\ p=2 \\ p=3 \end{array}\begin{array}{c} \begin{array}{cccc} q=0 & q=1 & q=2 & q=3 \end{array} \\ \begin{bmatrix} -825114 & -825231 & \mathbf{-825238} & -825234 \\ -825232 & -825192 & -825025 & \\ -825237 & -825031 & & \\ -825231 & & & \end{bmatrix} \end{array}$
1 min	$\begin{array}{c} \\ p=0 \\ p=1 \\ p=2 \\ p=3 \\ p=4 \\ p=5 \end{array}\begin{array}{c} \begin{array}{cccccc} q=0 & q=1 & q=2 & q=3 & q=4 & q=5 \end{array} \\ \begin{bmatrix} -4559145 & -4560916 & -4561001 & -4561028 & -4561042 & -4561067 \\ -4560918 & -4560962 & -4561022 & -4561034 & -4561031 & \\ -4561005 & -4561022 & -4561018 & -4560246 & & \\ -4561031 & -4561028 & -4561001 & & & \\ -4561045 & -4561006 & & & & \\ \mathbf{-4561071} & & & & & \end{bmatrix} \end{array}$

For the order of VARMA model, results show that indices with both 1 hour and 15 minute sampling frequencies have no time series structure while VMA-VG(2) and VAR(5)-VG provide the best fit for 5 minute and 1 minute respectively which is consistent with the autocorrelation plots in Figure 5.15 and the p-values from Box Pierce test in Table 5.13. As the sampling frequency increases, the means decrease, diagonals of the scale matrix decrease, the correlations in $\hat{\boldsymbol{\rho}}$ drop slightly, and the magnitude of

Table 5.15. Parameter estimates and correlation matrix $\boldsymbol{\rho}$ of VARMA(p,q)-VG model for 1hr and 15min high frequency returns.

estimates	1 hour	15 min
(p,q)	(0,0)	(0,0)
$\boldsymbol{\mu}'$	$10^{-4}\begin{pmatrix}1.2 & 1.0 & 1.6 & 0.2\end{pmatrix}$	$10^{-5}\begin{pmatrix}1.3 & 0.8 & 2.7 & -0.2\end{pmatrix}$
$\boldsymbol{\Sigma}$	$10^{-6}\begin{pmatrix}5.6 & 0.8 & 0.7 & 0.5\\ & 8.5 & 4.6 & 0.6\\ & & 5.7 & 0.3\\ & & & 2.9\end{pmatrix}$	$10^{-7}\begin{pmatrix}14.0 & 1.1 & 0.9 & 0.6\\ & 18.1 & 9.8 & 0.3\\ & & 11.8 & 0.4\\ & & & 6.4\end{pmatrix}$
$\boldsymbol{\gamma}'$	$10^{-4}\begin{pmatrix}-1.0 & 0.1 & -1.0 & 0.5\end{pmatrix}$	$10^{-5}\begin{pmatrix}-0.4 & 1.6 & -1.3 & 2.2\end{pmatrix}$
ν	1.0254	0.9196
$\boldsymbol{\rho}$	$\begin{pmatrix}1 & 0.11 & 0.12 & 0.12\\ & 1 & 0.66 & 0.11\\ & & 1 & 0.08\\ & & & 1\end{pmatrix}$	$\begin{pmatrix}1 & 0.07 & 0.07 & 0.07\\ & 1 & 0.67 & 0.02\\ & & 1 & 0.04\\ & & & 1\end{pmatrix}$

$\hat{\boldsymbol{\gamma}}$ decreases. All these phenomena agree with the numerical summaries in Tables 5.12 and 5.13.

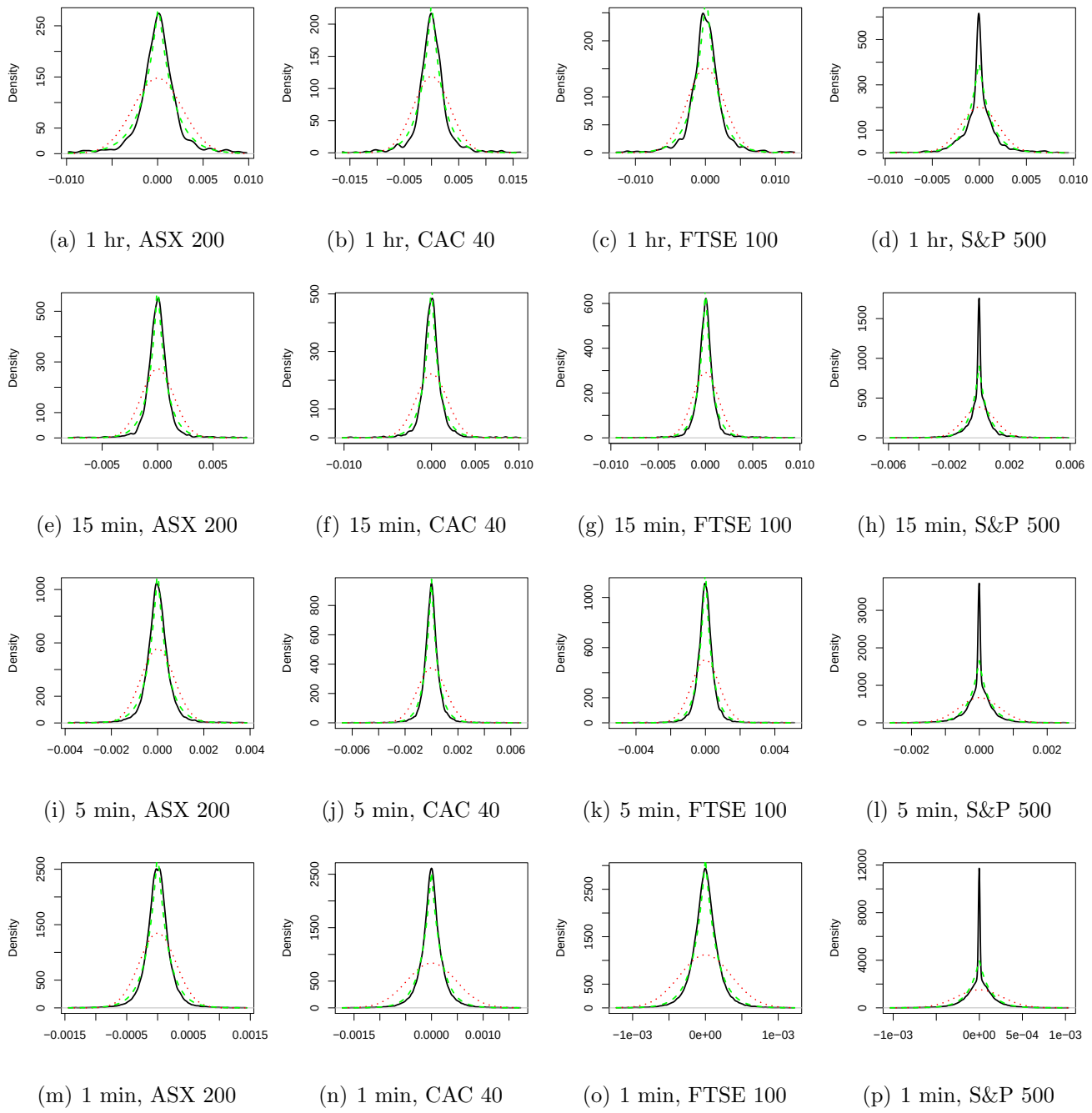
Comparing the shape parameter estimate, it drops from $\hat{\nu} = 1.40$ for the daily returns to $\hat{\nu} = 1.0254$ for 1 hour returns (over different sampling period). However, as the sampling frequency continues to increase, the shape parameter varies between 0.9 to 1. This can be explained from the fact that although the robust kurtosis of CAC 40 and S&P 500 show increasing trends, the other two indices do not show such trend. As a result, the shape parameter does not drop with sampling frequency to indicate a consistent trend for all four indices.

To check the model performance, we display in Figure 5.16 density plots of errors given in (5.26). For comparison, we also include the pdfs of marginal VARMA(p,q)-VG model and univariate normal distribution to each component of the errors. The plots demonstrate good fit of VARMA(p,q)-VG model except S&P 500 for all sampling frequencies. It shows clearly that S&P 500 has higher kurtosis than the shape parameter can allow for as the peak of the pdf is lower than the peak of the residual density due to again the assumption of consistent shape parameter across all components.

In summary, the VG distribution fits the four indices well for all sampling frequencies as it can capture adequately the peak of each distribution. Due to the constraint of a common shape parameter, it was incapable to capture the extreme leptokurtosis of

Table 5.16. Parameter estimates and correlation matrix $\boldsymbol{\rho}$ of VARMA(p,q)-VG model for 5 min and 1 min high frequency returns.

estimates	5 min	1 min
(p,q)	(0,2)	(5,0)
$\boldsymbol{\mu}'$	$10^{-6}\begin{pmatrix}3.3 & -7.3 & -5.8 & 1.4\end{pmatrix}$	$10^{-6}\begin{pmatrix}0.8 & -2.0 & -0.7 & 0.1\end{pmatrix}$
	$\boldsymbol{B}_1 = 10^{-2}\begin{pmatrix}-3.3 & -3.0 & -0.6 & -5.3\\ 1.0 & 4.7 & -6.4 & 0.7\\ 0.9 & 2.5 & -2.6 & 0.3\\ -0.3 & 0.4 & -0.4 & 1.7\end{pmatrix}$ $\boldsymbol{B}_2 = 10^{-2}\begin{pmatrix}-0.5 & 0.2 & -1.1 & -0.1\\ -1.6 & 0.5 & 1.2 & 2.8\\ -0.9 & 0.0 & 1.3 & 0.0\\ -0.3 & 0.2 & -0.3 & 0.9\end{pmatrix}$	$\boldsymbol{A}_1 = 10^{-2}\begin{pmatrix}2.3 & -0.2 & 0.5 & 1.3\\ -0.8 & -5.4 & 18.0 & -1.0\\ -0.2 & -1.8 & 5.8 & -0.6\\ 0.3 & 0.0 & 0.2 & 0.4\end{pmatrix}$ $\boldsymbol{A}_2 = 10^{-2}\begin{pmatrix}1.5 & 0.6 & 0.4 & -0.2\\ -0.2 & 0.2 & -1.0 & 1.2\\ -0.1 & 0.9 & -0.2 & 0.8\\ 0.1 & 0.0 & 0.1 & -1.2\end{pmatrix}$ $\boldsymbol{A}_3 = 10^{-2}\begin{pmatrix}1.0 & -0.2 & 0.0 & 0.3\\ -0.2 & -0.2 & -0.5 & 1.0\\ 0.4 & 0.1 & -0.3 & 0.4\\ 0.3 & 0.0 & 0.0 & -0.8\end{pmatrix}$ $\boldsymbol{A}_4 = 10^{-2}\begin{pmatrix}-0.2 & 0.4 & -0.2 & -0.1\\ 0.2 & 0.1 & -0.4 & -0.9\\ -0.2 & -0.1 & -0.2 & -0.2\\ 0.1 & -0.2 & 0.2 & -1.1\end{pmatrix}$ $\boldsymbol{A}_5 = 10^{-2}\begin{pmatrix}0.6 & 1.0 & -0.5 & 0.6\\ 0.4 & -0.5 & 1.2 & -0.2\\ 0.4 & -0.9 & 1.6 & -0.3\\ 0.0 & 0.0 & -0.1 & -0.3\end{pmatrix}$
$\boldsymbol{\Sigma}$	$10^{-8}\begin{pmatrix}38.5 & 1.2 & 0.9 & 0.6\\ & 52.2 & 26.8 & 1.1\\ & & 33.3 & 0.8\\ & & & 18.8\end{pmatrix}$	$10^{-9}\begin{pmatrix}61.1 & 0.6 & 0.7 & 0.1\\ & 75.4 & 31.8 & 0.8\\ & & 49.9 & 0.7\\ & & & 31.2\end{pmatrix}$
$\boldsymbol{\gamma}'$	$10^{-6}\begin{pmatrix}-0.7 & 15.7 & 10.7 & 5.4\end{pmatrix}$	$10^{-6}\begin{pmatrix}-0.4 & 3.7 & 1.7 & 1.3\end{pmatrix}$
ν	0.9591	0.9962
$\boldsymbol{\rho}$	$\begin{pmatrix}1 & 0.03 & 0.02 & 0.02\\ & 1 & 0.64 & 0.03\\ & & 1 & 0.03\\ & & & 1\end{pmatrix}$	$\begin{pmatrix}1 & 0.01 & 0.01 & 0.00\\ & 1 & 0.52 & 0.02\\ & & 1 & 0.02\\ & & & 1\end{pmatrix}$



(a) 1 hr, ASX 200 (b) 1 hr, CAC 40 (c) 1 hr, FTSE 100 (d) 1 hr, S&P 500

(e) 15 min, ASX 200 (f) 15 min, CAC 40 (g) 15 min, FTSE 100 (h) 15 min, S&P 500

(i) 5 min, ASX 200 (j) 5 min, CAC 40 (k) 5 min, FTSE 100 (l) 5 min, S&P 500

(m) 1 min, ASX 200 (n) 1 min, CAC 40 (o) 1 min, FTSE 100 (p) 1 min, S&P 500

Figure 5.16. Density plots of the residuals (solid black line), pdf of VARMA-VG after filtering the mean function (green dash line) and fitted univariate normal (red dotted line) for the 1 hr, 15 min, 5 min and 1 min (rows) of ASX 200, CAC 40, FTSE 100 and S&P 500 returns (columns).

S&P 500. Nevertheless, the VARMA-VG model has demonstrated its applicability to capture the important features of the high frequency returns.

5.7 Conclusion

In this chapter, we extend the VG distribution in Chapter 1.5.3 to VAR-VG and VARMA-VG models with additional AR and MA parameters to model a wide range of persistence structures in financial time series. We also extend the AECM algorithm developed in Chapters 2 to 4 to estimate the additional AR and MA parameters. We note that there is no closed-form expression for the AR, MA and skewness parameters for the VARMA-VG model, and the optimisation techniques such as NR method may be problematic due to the large number of parameters. Instead, we consider the approximation method to first fit a suitable order VAR-VG model and estimate the parameters based on the fitted residuals.

We test the performance of the AECM algorithm in the first simulation study for four cases: low or high level of skewness as well as small or large shape parameter. Results show that the AECM algorithm provides sufficiently accurate parameter estimates even when the density is unbounded in which case the WLOO likelihood method of Chapter 4 is applied. We also demonstrate that the SE estimates using Louis' method give reasonably accurate results and the SE calculation is computationally efficient.

Apart from the levels of skewness and kurtosis, other factors that may affect the performance of AECM algorithm include the model identifiability issue. We discuss the properties of the VARMA-VG model and the conditions when the model identification problem arises. The second simulation study is designed so that the AR and MA parameters are subject to such problem. We demonstrate that parameter estimates are subject to extra variabilities due to the redundant variables in the AR and MA parameters. Nevertheless, parameter estimates under such condition are still reasonable.

After checking the model performance, we demonstrate the applicability of VARMA-VG model through analysing two return series both displaying distinct characteristics. We find that the AECM algorithm provides reasonable estimates which agree closely with the observed characteristics. The first type of return series we consider is cryptocurrency returns which is less stable than the common stock indices due to its short life period.

We find that the cryptocurrency daily returns have kurtosis even higher than the high frequency stock index returns. Nevertheless, these behaviour of the cryptocurrency daily returns can be captured with the VARMA(2,0)-VG and VARMA(2,0)-t models where the latter model performs better in model-fit in terms of AICc but it converges much slower than the former model using the AECM algorithm. For the second type of stock index returns, we consider daily returns as well as high frequency returns at 1 hr, 15 min, 5 min and 1 min sampling frequencies. We find the shape parameters for both types of return series are lower than $d/2$ which signifies unboundedness for the VG density. This confirms the usefulness of VARMA-VG model to capture the high level of kurtosis for various financial time series. Moreover SEs are also calculated using Louis' method to allow us to assess the significance of each parameter estimate. For the location parameters, SEs are approximated using the double generalised gamma distribution.

CHAPTER 6

Conclusion

This thesis is the first of its kind to develop ECM algorithms for VG distribution with unbounded density. Various improvements have been made to the ECM algorithm to improve numerical stability, computational efficiency and accuracy to estimate the parameters, particularly, the location parameter of the VG distribution. Model performance is verified using simulation studies. This chapter concludes the contributions in each chapter and proposes future research directions.

In Chapter 2, the AECM algorithm is derived to provide efficient estimators for the VG distribution and the capped likelihood method is proposed to deal with the numerical issues for the location parameter estimate when the density is unbounded. Simulation studies confirm the good performance of the AECM algorithm, determine the optimal choice of capping level and test the three methods for calculating SE. A comprehensive set of formulas are derived for calculating the observed and Fisher information matrix in the appendices. These formulas also apply to other distributions with NMVM representations such as the Student's t and GH distributions. Application to stock index returns shows that the VG model can capture high kurtosis and some skewness in the stock index returns.

Chapter 3 considers the LOO likelihood methods and extends the AECM algorithm to accommodate the LOO likelihood. Simulation studies again confirm the good performance and further investigate the optimal convergence rate and asymptotic distribution for the location parameter estimator. We further demonstrate how the double generalised gamma distribution can describe the distribution of the location estimator for cusp and unbounded densities and hence the SE can be calculated, even when the information for the location parameter is not well-defined.

In case there are data multiplicity, the LOO likelihood method for location estimate may fail. Chapter 4 proposes the WLOO likelihood method and compares it with four other estimators proposed in Chapters 2 and 3. Results show that the AECM algorithm with WLOO likelihood provides good estimates for all parameters. Nevertheless, the adaptive capped likelihood method also provides reasonably good accuracy.

With an aim to provide real financial applications, the constant mean VG model in Chapter 2 is found to be inadequate to describe the persistence in most financial return series. Chapter 5 extends the model to VARMA-VG model and modifies the AECM algorithm using WLOO likelihood to estimate additional AR and MA parameters. Simulation studies confirm the performance of the AECM algorithm with accurate parameter and SE estimates. In case of non-identifiable model, it still provides reasonable estimates with enlarged variability for some AR and MA parameters. Two extensive real data analyses are performed with returns from cryptocurrency daily exchange and high frequency stock indices. Characteristics of these two markets are extracted showing that smaller market share like cryptocurrency and higher sampling frequency may increase the kurtosis of a data to a level that most distributions fail to capture. The VARMA-VG and VARMA-t models can capture extreme kurtosis and so are very favourable to model these types of data. Their ability to capture the cross-correlation between assets is also useful in portfolio setting and trading strategy formulation for financial institutions. These analyses are very new in the modelling and finance literature and provide great contribution to the understanding of the properties of high frequency index and cryptocurrency returns.

There are many promising area to pursue in the future.

Firstly, the proposed AECM algorithm in Chapters 2 and 3 can be applied to estimate parameters for other distributions with cusp or unbounded densities. When the NMVM representation is unknown for some distributions such as exponential power and double generalised gamma distributions, numerical optimisation techniques such as NR method with local point search and line search can be applied to maximise directly the WLOO likelihood. The estimation of these distributions with extreme leptokurtosis is very significant in the methodological development for models to describe the increasingly prevalent high frequency data.

Secondly, the simulation study for the optimal convergence rate and asymptotic distribution of the location estimate in Chapter 3 should be extended. Currently, we consider only the univariate symmetric VG distribution. For future work, it is worth studying these properties for the multivariate skew case while allowing for dependence between the location and other parameters from VG distribution. In addition, we suspect that the range $0.4 < \nu < 1$ corresponds to the transition range when the optimal rate of convergence changes from $1/2$ (when $\nu \geq 1$) to $1/(2\nu)$ (when $\nu < 0.4$), and the asymptotic distribution converges more slowly in this transition range. More numerical studies as well as theoretical developments should be directed to extract more distinct behaviours when ν lies within this transition range. Although proving these asymptotic results analytically is still an open question, our numerical results provide useful insight for the theoretical development of properties of the location estimate when the density is cusped or unbounded.

Thirdly, statistical test should be developed to see if the VARMA type models such as the VARMA-VG have a common root. This is equivalent to testing if the resultant defined by the determinant of the tensor Sylvester matrix in (5.31) is equal to zero which indicates that the model is non-identifiable. This is important as it causes the information matrix to be singular and gives unstable parameter estimates. Once the model is classified to be non-identifiable, the model can then be further simplified by setting some of the elements in the AR and MA matrices to be zero or incorporate structural specification into the VARMA model [104].

Fourthly, the VARMA-VG model should be further extended to adopt more advanced time series features. This includes extending the mean function to adopt the autoregressive fractional integrated moving average (ARFIMA) [43] and its generalisation called Gegenbauer autoregressive moving-average (GARMA) [49] structures to describe the strong and possibly periodic persistence that are present in some financial time series. Moreover, as Figure 2.2 displays some volatility clustering, the variance-covariance matrix $\mathbf{\Sigma}$ can be assigned a dynamic rather than static structure such as the popular GARCH-type volatility structure [12] and the covariance regression structure [111].

Lastly, the VG distribution should be extended to allow multiple shape parameters. As demonstrated in all applications, the VG distribution with one consistent shape parameter for all components is very restrictive. To improve the model flexibility, one may consider a modified VG distribution, similar to the modified multivariate Student's t

distribution in Choy et al. [23] or the multiple scaled distribution in Wraith and Forbes [108] so that each marginal distribution has a separate shape parameter. More generally, we can group sets of marginal distributions to have group-specific shape parameters. The ACME algorithm can be extended to estimate these additional shape parameters. We expect the conditional expectations for the mixing variables of VG distribution with different ν to have a more complicated functional form and so further research is required to develop estimation methods to implement these extended distributions.

CHAPTER A

Matrix Differentiation

A1 Introduction

Evaluating derivatives of the log-likelihood function is important for maximisation in the ML approach and calculation of the the information matrix to obtain standard errors. As the log-likelihood function commonly involves multiple parameters, the maximisation requires differentiation with respect to each of these individual parameters. Moreover, the second order derivatives required in the calculation of the information matrix is even more tedious to evaluate as it requires differentiation with respect to numerous parameter pairs. Hence, the element-wise differentiation becomes inefficient particularly for multivariate models like the VAR-VG or VARMA-VG. In this section, we introduce the theory of matrix differentiation where the log-likelihood function can be differentiated with respect to matrix arguments. This approach can greatly simplify the amount of computation when designing a computer program to evaluate derivatives. See Boik [11], Petersen and Pedersen [88], Lütkepohl [69], Schonemann [98], Minka [81] for more information on matrix differentiation.

A2 Matrix operators

Let $\boldsymbol{X}$ be a $m \times n$ matrix. We define the following functions to obtain dimensions of a matrix $\boldsymbol{X}$:

$$\dim\boldsymbol{X} = (m, n), \qquad \text{row}\boldsymbol{X} = m, \qquad \text{col}\boldsymbol{X} = n.$$

We then define the following constant vectors and matrices:

(i) $\boldsymbol{I}_{\text{col}\boldsymbol{X}}$ represents an $n \times n$ identity matrix,
(ii) $\mathbb{1}_{\text{row}\boldsymbol{X}}$ represents a m-dimensional vector of ones, and
(iii) $\mathbb{1}_{\text{dim}\boldsymbol{X}}$ represents a $m \times n$ matrix of ones.

If $\boldsymbol{Y}$ has the same dimension as $\boldsymbol{X}$, then the *Hadamard product* $\boldsymbol{X} \circ \boldsymbol{Y}$ is a matrix of same dimension with elements given by

$$(\boldsymbol{X} \circ \boldsymbol{Y})_{ij} = X_{ij} \times Y_{ij} \,. \tag{A.1}$$

If $\boldsymbol{Y}$ is a $p \times q$ matrix, then the *Kronecker product* is defined as

$$\boldsymbol{X} \otimes \boldsymbol{Y} = \begin{pmatrix} X_{11}\boldsymbol{Y} & \cdots & X_{1n}\boldsymbol{Y} \\ \vdots & & \vdots \\ X_{m1}\boldsymbol{Y} & \cdots & X_{mn}\boldsymbol{Y} \end{pmatrix} \tag{A.2}$$

which is a $mp \times nq$ matrix.

If $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_n \end{pmatrix}$, then the *vectorisation* of $\boldsymbol{X}$ stacks the columns into a vector. That is,

$$\text{vec}(\boldsymbol{X}) = \begin{pmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_n \end{pmatrix} \tag{A.3}$$

is a mn-dimensional vector.

If $\boldsymbol{\Sigma}$ is a $d \times d$ symmetric matrix, then the *half-vectorisation* of $\boldsymbol{\Sigma}$ stacks the columns of the lower triangular matrix of $\boldsymbol{\Sigma}$ into a vector. That is,

$$\text{vech}(\boldsymbol{\Sigma}) = \begin{pmatrix} \Sigma_{11} & \cdots & \Sigma_{d1} & \Sigma_{22} & \cdots & \Sigma_{d2} & \cdots & \Sigma_{dd} \end{pmatrix}^\top \tag{A.4}$$

is a $d(d+1)/2$-dimensional vector.

A *commutation matrix* denoted by $\boldsymbol{K}^{(m,n)}$ is a $mn \times mn$ permutation matrix such that

$$\boldsymbol{K}^{(m,n)}\text{vec}(\boldsymbol{X}) = \text{vec}(\boldsymbol{X}^\top). \tag{A.5}$$

See [72] for properties of the commutation matrix.

An *elimination matrix* denoted by $\boldsymbol{L}_d$ is a $d(d+1)/2 \times d^2$ matrix such that

$$\boldsymbol{L}_d\text{vec}(\boldsymbol{\Sigma}) = \text{vech}(\boldsymbol{\Sigma}) \,. \tag{A.6}$$

A *duplication matrix* denoted by $\boldsymbol{D}_d$ is a $d^2 \times d(d+1)/2$ matrix such that

$$\boldsymbol{D}_d \text{vech}(\boldsymbol{\Sigma}) = \text{vec}(\boldsymbol{\Sigma})\,. \tag{A.7}$$

A3 Definitions and basic rules

Let the $m \times n$ matrix $\boldsymbol{X}$ be a general matrix with no particular structure (such as symmetric, Toeplitz, etc.) so that the elements of the matrix are mutually independent, we write

$$\boldsymbol{X} = \begin{pmatrix} X_{11} & \cdots & X_{1n} \\ \vdots & & \vdots \\ X_{m1} & \cdots & X_{mn} \end{pmatrix},$$

and $f(\boldsymbol{X})$ is a scalar differentiable function with respect to $\boldsymbol{X}$. Then the first order partial derivative of f with respect to $\boldsymbol{X}$ is defined as

$$\frac{\partial f}{\partial \boldsymbol{X}} = \begin{pmatrix} \frac{\partial f}{\partial X_{11}} & \cdots & \frac{\partial f}{\partial X_{1n}} \\ \vdots & & \vdots \\ \frac{\partial f}{\partial X_{m1}} & \cdots & \frac{\partial f}{\partial X_{mn}} \end{pmatrix}$$

which is a $m \times n$ matrix.

Let $\boldsymbol{Y}(x)$ be a $p \times q$ matrix function with each element $Y_{ij}(x)$ a function of a scalar variable x. Then

$$\frac{\partial \boldsymbol{Y}}{\partial x} = \begin{pmatrix} \frac{\partial \boldsymbol{Y}_{11}}{\partial x} & \cdots & \frac{\partial \boldsymbol{Y}_{1q}}{\partial x} \\ \vdots & & \vdots \\ \frac{\partial \boldsymbol{Y}_{p1}}{\partial x} & \cdots & \frac{\partial \boldsymbol{Y}_{pq}}{\partial x} \end{pmatrix}$$

is a $p \times q$ matrix.

If $\boldsymbol{f}(\boldsymbol{x}) = (f_1(\boldsymbol{x}), \ldots, f_p(\boldsymbol{x}))'$ is a p-dimensional vector function with respect to m-dimensional vector $\boldsymbol{x}$, then

$$\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}'} = \begin{pmatrix} \frac{\partial \boldsymbol{f}}{\partial x_1} & \cdots & \frac{\partial \boldsymbol{f}}{\partial x_m} \end{pmatrix}$$

which is a $p \times m$ matrix.

More generally, if $\boldsymbol{Y}(\boldsymbol{X})$ is a matrix function depending on matrix $\boldsymbol{X}$, then

$$\frac{\partial \boldsymbol{Y}}{\partial \boldsymbol{X}} = \frac{\partial}{\partial \boldsymbol{X}} \otimes \boldsymbol{Y} = \begin{pmatrix} \frac{\partial \boldsymbol{Y}}{\partial X_{11}} & \cdots & \frac{\partial \boldsymbol{Y}}{\partial X_{1n}} \\ \vdots & & \vdots \\ \frac{\partial \boldsymbol{Y}}{\partial X_{m1}} & \cdots & \frac{\partial \boldsymbol{Y}}{\partial X_{mn}} \end{pmatrix}$$

which is a $pm \times qn$ matrix.

Using this notation, we can represent the Hessian matrix as

$$\frac{\partial^2 f}{\partial \boldsymbol{x} \partial \boldsymbol{x}'} = \frac{\partial^2 f}{\partial \boldsymbol{x}' \partial \boldsymbol{x}} = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_m \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_m \partial x_m} \end{pmatrix}$$

where $\boldsymbol{x}$ is a m-dimensional vector.

The following rules are useful to derive many differential results [81]:

$$\partial \boldsymbol{A} = \boldsymbol{0} \quad \text{for some constant matrix } \boldsymbol{A}, \tag{A.8}$$

$$\partial(\alpha \boldsymbol{X}) = \alpha\, \partial \boldsymbol{X} \quad \text{for some scalar constant } \alpha, \tag{A.9}$$

$$\partial(\boldsymbol{X} + \boldsymbol{Y}) = \partial \boldsymbol{X} + \partial \boldsymbol{Y}, \tag{A.10}$$

$$\partial(\mathrm{tr}(\boldsymbol{X})) = \mathrm{tr}(\partial \boldsymbol{X}), \tag{A.11}$$

$$\partial(\boldsymbol{X}\boldsymbol{Y}) = (\partial \boldsymbol{X})\boldsymbol{Y} + \boldsymbol{X}(\partial \boldsymbol{Y}), \tag{A.12}$$

$$\partial(\boldsymbol{X} \circ \boldsymbol{Y}) = (\partial \boldsymbol{X}) \circ \boldsymbol{Y} + \boldsymbol{X} \circ (\partial \boldsymbol{Y}), \tag{A.13}$$

$$\partial(\boldsymbol{X} \otimes \boldsymbol{Y}) = (\partial \boldsymbol{X}) \otimes \boldsymbol{Y} + \boldsymbol{X} \otimes (\partial \boldsymbol{Y}), \tag{A.14}$$

$$\partial\big(\boldsymbol{X}^{-1}\big) = -\boldsymbol{X}^{-1}(\partial \boldsymbol{X})\boldsymbol{X}^{-1}, \tag{A.15}$$

$$\partial(\det \boldsymbol{X}) = \det(\boldsymbol{X})\, \mathrm{tr}\big(\boldsymbol{X}^{-1}\partial \boldsymbol{X}\big), \tag{A.16}$$

$$\partial(\log \det \boldsymbol{X}) = \mathrm{tr}\big(\boldsymbol{X}^{-1}\partial \boldsymbol{X}\big), \tag{A.17}$$

$$\partial \boldsymbol{X}' = (\partial \boldsymbol{X})'\,. \tag{A.18}$$

The following theorem is a direct application of (A.18).

Theorem A3.1 (Transpose of derivative)**.** *Let $\boldsymbol{Y}$ be a matrix function with respect to matrix $\boldsymbol{X}$, then*

$$\left(\frac{\partial \boldsymbol{Y}}{\partial \boldsymbol{X}}\right)' = \frac{\partial \boldsymbol{Y}'}{\partial \boldsymbol{X}'}. \tag{A.19}$$

Other differential properties can be obtained from the rules above,

$$\frac{\partial \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}'} = \boldsymbol{A}, \tag{A.20}$$

$$\frac{\partial \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = \mathrm{vec}(\boldsymbol{A}), \tag{A.21}$$

$$\frac{\partial \boldsymbol{x}'\boldsymbol{A}}{\partial \boldsymbol{x}} = \boldsymbol{A}, \tag{A.22}$$

$$\frac{\partial \boldsymbol{x}'\boldsymbol{A}}{\partial \boldsymbol{x}'} = \mathrm{vec}(\boldsymbol{A}')', \tag{A.23}$$

$$\frac{\partial \log|\boldsymbol{X}|}{\partial \boldsymbol{X}} = \boldsymbol{X}^{-\top}, \tag{A.24}$$

$$\frac{\partial \mathrm{tr}(\boldsymbol{Y})}{\partial \boldsymbol{X}} = \frac{\partial \mathrm{tr}(\boldsymbol{Y}')}{\partial \boldsymbol{X}}, \tag{A.25}$$

$$\frac{\partial \mathrm{tr}(\boldsymbol{A}\boldsymbol{X})}{\partial \boldsymbol{X}} = \frac{\partial \mathrm{tr}(\boldsymbol{X}'\boldsymbol{A}')}{\partial \boldsymbol{X}} = \boldsymbol{A}'. \tag{A.26}$$

The aim of this appendix is to present some matrix derivative results useful to differentiate the log-likelihood of many different statistical models for the maximisation likelihood approach.

A4 Derivative of products

In this section, the following product rules are considered: Matrix product rule, Hadamard product rule and Kronecker product rule. Other product rules regarding the trace and vectorisation is given in the later sections.

Theorem A4.1 (Matrix product rule)**.** *Let $\boldsymbol{U}$ and $\boldsymbol{V}$ be matrix functions with respect to $\boldsymbol{X}$ such that $\boldsymbol{U}\boldsymbol{V}$ is conformable, then*

$$\frac{\partial(\boldsymbol{U}\boldsymbol{V})}{\partial \boldsymbol{X}} = \frac{\partial \boldsymbol{U}}{\partial \boldsymbol{X}}(\boldsymbol{I}_{\mathrm{col}\boldsymbol{X}} \otimes \boldsymbol{V}) + (\boldsymbol{I}_{\mathrm{row}\boldsymbol{X}} \otimes \boldsymbol{U})\frac{\partial \boldsymbol{V}}{\partial \boldsymbol{X}} \tag{A.27}$$

where $\boldsymbol{I}_d$ represents an $d \times d$ identity matrix.

Proof. Consider the derivative with respect to the $(i,j)^{\text{th}}$ component of $\boldsymbol{X}$, using the differential property $\partial(\boldsymbol{U}\boldsymbol{V}) = (\partial \boldsymbol{U})\boldsymbol{V} + \boldsymbol{U}(\partial \boldsymbol{V})$ from equation (A.12),

$$\frac{\partial(\boldsymbol{U}\boldsymbol{V})}{\partial X_{ij}} = \frac{\partial \boldsymbol{U}}{\partial X_{ij}}\boldsymbol{V} + \boldsymbol{U}\frac{\partial \boldsymbol{V}}{\partial X_{ij}}.$$

Extending this to the whole matrix,

$$\frac{\partial(\boldsymbol{UV})}{\partial \boldsymbol{X}} = \begin{pmatrix} \frac{\partial \boldsymbol{U}}{\partial X_{11}}\boldsymbol{V} & \cdots & \frac{\partial \boldsymbol{U}}{\partial X_{1n}}\boldsymbol{V} \\ \vdots & & \vdots \\ \frac{\partial \boldsymbol{U}}{\partial X_{m1}}\boldsymbol{V} & \cdots & \frac{\partial \boldsymbol{U}}{\partial X_{mn}}\boldsymbol{V} \end{pmatrix} + \begin{pmatrix} \boldsymbol{U}\frac{\partial \boldsymbol{V}}{\partial X_{11}} & \cdots & \boldsymbol{U}\frac{\partial \boldsymbol{V}}{\partial X_{1n}} \\ \vdots & & \vdots \\ \boldsymbol{U}\frac{\partial \boldsymbol{V}}{\partial X_{m1}} & \cdots & \boldsymbol{U}\frac{\partial \boldsymbol{V}}{\partial X_{mn}} \end{pmatrix}.$$

Using block matrix multiplication, we can write it as

$$\begin{aligned} &= \begin{pmatrix} \frac{\partial \boldsymbol{U}}{\partial X_{11}} & \cdots & \frac{\partial \boldsymbol{U}}{\partial X_{1n}} \\ \vdots & & \vdots \\ \frac{\partial \boldsymbol{U}}{\partial X_{m1}} & \cdots & \frac{\partial \boldsymbol{U}}{\partial X_{mn}} \end{pmatrix} \begin{pmatrix} \boldsymbol{V} & & \boldsymbol{0} \\ & \ddots & \\ \boldsymbol{0} & & \boldsymbol{V} \end{pmatrix} + \begin{pmatrix} \boldsymbol{U} & & \boldsymbol{0} \\ & \ddots & \\ \boldsymbol{0} & & \boldsymbol{U} \end{pmatrix} \begin{pmatrix} \frac{\partial \boldsymbol{V}}{\partial X_{11}} & \cdots & \frac{\partial \boldsymbol{V}}{\partial X_{1n}} \\ \vdots & & \vdots \\ \frac{\partial \boldsymbol{V}}{\partial X_{m1}} & \cdots & \frac{\partial \boldsymbol{V}}{\partial X_{mn}} \end{pmatrix} \\ &= \frac{\partial \boldsymbol{U}}{\partial \boldsymbol{X}}(\boldsymbol{I}_{\mathrm{col}\boldsymbol{X}} \otimes \boldsymbol{V}) + (\boldsymbol{I}_{\mathrm{row}\boldsymbol{X}} \otimes \boldsymbol{U})\frac{\partial \boldsymbol{V}}{\partial \boldsymbol{X}}. \end{aligned}$$

□

Theorem A4.2 (Hadamard product rule)**.** *Let $\boldsymbol{U}$ and $\boldsymbol{V}$ be matrix functions with respect to $\boldsymbol{X}$ such that $\boldsymbol{U} \circ \boldsymbol{V}$ is conformable, then*

$$\frac{\partial(\boldsymbol{U} \circ \boldsymbol{V})}{\partial \boldsymbol{X}} = \frac{\partial \boldsymbol{U}}{\partial \boldsymbol{X}} \circ (\mathbb{1}_{\dim \boldsymbol{X}} \otimes \boldsymbol{V}) + (\mathbb{1}_{\dim \boldsymbol{X}} \otimes \boldsymbol{U}) \circ \frac{\partial \boldsymbol{V}}{\partial \boldsymbol{X}} \tag{A.28}$$

where $\mathbb{1}_{\dim \boldsymbol{X}}$ represents a matrix of ones with same dimensions as $\boldsymbol{X}$.

Proof. Consider the derivative with respect to the $(i,j)^{\text{th}}$ component of $\boldsymbol{X}$, using the differential property $\partial(\boldsymbol{U} \circ \boldsymbol{V}) = (\partial \boldsymbol{U}) \circ \boldsymbol{V} + \boldsymbol{U} \circ (\partial \boldsymbol{V})$ from equation (A.13),

$$\frac{\partial(\boldsymbol{U} \circ \boldsymbol{V})}{\partial X_{ij}} = \frac{\partial \boldsymbol{U}}{\partial X_{ij}} \circ \boldsymbol{V} + \boldsymbol{U} \circ \frac{\partial \boldsymbol{V}}{\partial X_{ij}}.$$

Extending this to the whole matrix,

$$\begin{aligned} &\frac{\partial(\boldsymbol{U} \circ \boldsymbol{V})}{\partial \boldsymbol{X}} \\ &= \begin{pmatrix} \frac{\partial \boldsymbol{U}}{\partial X_{11}} \circ \boldsymbol{V} & \cdots & \frac{\partial \boldsymbol{U}}{\partial X_{1n}} \circ \boldsymbol{V} \\ \vdots & & \vdots \\ \frac{\partial \boldsymbol{U}}{\partial X_{m1}} \circ \boldsymbol{V} & \cdots & \frac{\partial \boldsymbol{U}}{\partial X_{mn}} \circ \boldsymbol{V} \end{pmatrix} + \begin{pmatrix} \boldsymbol{U} \circ \frac{\partial \boldsymbol{V}}{\partial X_{11}} & \cdots & \boldsymbol{U} \circ \frac{\partial \boldsymbol{V}}{\partial X_{1n}} \\ \vdots & & \vdots \\ \boldsymbol{U} \circ \frac{\partial \boldsymbol{V}}{\partial X_{m1}} & \cdots & \boldsymbol{U} \circ \frac{\partial \boldsymbol{V}}{\partial X_{mn}} \end{pmatrix} \\ &= \begin{pmatrix} \frac{\partial \boldsymbol{U}}{\partial X_{11}} & \cdots & \frac{\partial \boldsymbol{U}}{\partial X_{1n}} \\ \vdots & & \vdots \\ \frac{\partial \boldsymbol{U}}{\partial X_{m1}} & \cdots & \frac{\partial \boldsymbol{U}}{\partial X_{mn}} \end{pmatrix} \circ \begin{pmatrix} \boldsymbol{V} & \cdots & \boldsymbol{V} \\ \vdots & & \vdots \\ \boldsymbol{V} & \cdots & \boldsymbol{V} \end{pmatrix} + \begin{pmatrix} \boldsymbol{U} & \cdots & \boldsymbol{U} \\ \vdots & & \vdots \\ \boldsymbol{U} & \cdots & \boldsymbol{U} \end{pmatrix} \circ \begin{pmatrix} \frac{\partial \boldsymbol{V}}{\partial X_{11}} & \cdots & \frac{\partial \boldsymbol{V}}{\partial X_{1n}} \\ \vdots & & \vdots \\ \frac{\partial \boldsymbol{V}}{\partial X_{m1}} & \cdots & \frac{\partial \boldsymbol{V}}{\partial X_{mn}} \end{pmatrix} \\ &= \frac{\partial \boldsymbol{U}}{\partial \boldsymbol{X}} \circ (\mathbb{1}_{\dim \boldsymbol{X}} \otimes \boldsymbol{V}) + (\mathbb{1}_{\dim \boldsymbol{X}} \otimes \boldsymbol{U}) \circ \frac{\partial \boldsymbol{V}}{\partial \boldsymbol{X}}. \end{aligned}$$

□

Theorem A4.3 (Kronecker product rule)**.** *Let* $\boldsymbol{U}$ *and* $\boldsymbol{V}$ *be matrix functions with respect to* $\boldsymbol{X}$*, then*

$$\frac{\partial(\boldsymbol{U}\otimes\boldsymbol{V})}{\partial\boldsymbol{X}} = \frac{\partial\boldsymbol{U}}{\partial\boldsymbol{X}}\otimes\boldsymbol{V} + \left(\boldsymbol{I}_{\mathrm{row}\boldsymbol{X}}\otimes\boldsymbol{K}^{(\mathrm{row}\boldsymbol{V},\mathrm{row}\boldsymbol{U})}\right)\left(\frac{\partial\boldsymbol{V}}{\partial\boldsymbol{X}}\otimes\boldsymbol{U}\right)\left(\boldsymbol{I}_{\mathrm{col}\boldsymbol{X}}\otimes\boldsymbol{K}^{(\mathrm{col}\boldsymbol{U},\mathrm{col}\boldsymbol{V})}\right) \tag{A.29}$$

where $\boldsymbol{K}^{(m,n)}$ *is defined in* (A.5)*.*

Proof. Consider the derivative with respect to the $(i,j)^{\text{th}}$ component of $\boldsymbol{X}$, using the differential property $\partial(\boldsymbol{U}\otimes\boldsymbol{V}) = (\partial\boldsymbol{U})\otimes\boldsymbol{V} + \boldsymbol{U}\otimes(\partial\boldsymbol{V})$ from equation (A.14),

$$\begin{aligned}\frac{\partial(\boldsymbol{U}\otimes\boldsymbol{V})}{\partial X_{ij}} &= \frac{\partial\boldsymbol{U}}{\partial X_{ij}}\otimes\boldsymbol{V} + \boldsymbol{U}\otimes\frac{\partial\boldsymbol{V}}{\partial X_{ij}} \\ &= \frac{\partial\boldsymbol{U}}{\partial X_{ij}}\otimes\boldsymbol{V} + \boldsymbol{K}^{(\mathrm{row}\boldsymbol{U},\mathrm{row}\boldsymbol{V})}\left(\frac{\partial\boldsymbol{V}}{\partial X_{ij}}\otimes\boldsymbol{U}\right)\boldsymbol{K}^{(\mathrm{col}\boldsymbol{V},\mathrm{col}\boldsymbol{U})}\end{aligned}$$

where the last equality holds using the commutation property. Extending this to the whole matrix,

$$\begin{aligned}&\frac{\partial(\boldsymbol{U}\otimes\boldsymbol{V})}{\partial\boldsymbol{X}} \\ &= \begin{pmatrix} \frac{\partial\boldsymbol{U}}{\partial X_{11}}\otimes\boldsymbol{V} & \cdots & \frac{\partial\boldsymbol{U}}{\partial X_{1n}}\otimes\boldsymbol{V} \\ \vdots & & \vdots \\ \frac{\partial\boldsymbol{U}}{\partial X_{m1}}\otimes\boldsymbol{V} & \cdots & \frac{\partial\boldsymbol{U}}{\partial X_{mn}}\otimes\boldsymbol{V} \end{pmatrix} + \begin{pmatrix} \boldsymbol{K}^{(\mathrm{row}\boldsymbol{U},\mathrm{row}\boldsymbol{V})} & & \boldsymbol{0} \\ & \ddots & \\ \boldsymbol{0} & & \boldsymbol{K}^{(\mathrm{row}\boldsymbol{U},\mathrm{row}\boldsymbol{V})} \end{pmatrix} \\ &\qquad \begin{pmatrix} \frac{\partial\boldsymbol{V}}{\partial X_{11}}\otimes\boldsymbol{U} & \cdots & \frac{\partial\boldsymbol{V}}{\partial X_{1n}}\otimes\boldsymbol{U} \\ \vdots & & \vdots \\ \frac{\partial\boldsymbol{V}}{\partial X_{m1}}\otimes\boldsymbol{U} & \cdots & \frac{\partial\boldsymbol{V}}{\partial X_{mn}}\otimes\boldsymbol{U} \end{pmatrix} \begin{pmatrix} \boldsymbol{K}^{(\mathrm{col}\boldsymbol{V},\mathrm{col}\boldsymbol{U})} & & \boldsymbol{0} \\ & \ddots & \\ \boldsymbol{0} & & \boldsymbol{K}^{(\mathrm{col}\boldsymbol{V},\mathrm{col}\boldsymbol{U})} \end{pmatrix} \\ &= \frac{\partial\boldsymbol{U}}{\partial\boldsymbol{X}}\otimes\boldsymbol{V} + \left(\boldsymbol{I}_{\mathrm{row}\boldsymbol{X}}\otimes\boldsymbol{K}^{(\mathrm{row}\boldsymbol{U},\mathrm{row}\boldsymbol{V})}\right)\left(\frac{\partial\boldsymbol{V}}{\partial\boldsymbol{X}}\otimes\boldsymbol{U}\right)\left(\boldsymbol{I}_{\mathrm{col}\boldsymbol{X}}\otimes\boldsymbol{K}^{(\mathrm{col}\boldsymbol{V},\mathrm{col}\boldsymbol{U})}\right).\end{aligned}$$

□

Corollary A4.4. *Let* f *be a scalar function and* $\boldsymbol{V}$ *be a matrix function with respect to* $\boldsymbol{X}$*, then*

$$\frac{\partial(f\,\boldsymbol{V})}{\partial\boldsymbol{X}} = \frac{\partial f}{\partial\boldsymbol{X}}\otimes\boldsymbol{V} + f\,\frac{\partial\boldsymbol{V}}{\partial\boldsymbol{X}}\,. \tag{A.30}$$

Using these product rules, we have the following differentiation formulas:

$$\frac{\partial \boldsymbol{x}\boldsymbol{x}'}{\partial \boldsymbol{x}} = \boldsymbol{I}_d \otimes \boldsymbol{x} + \mathrm{vec}(\boldsymbol{I}_d)\boldsymbol{x}', \tag{A.31}$$

$$\frac{\partial \boldsymbol{x} \otimes \boldsymbol{x}\boldsymbol{x}'}{\partial \boldsymbol{x}'} = \boldsymbol{I}_d \otimes \boldsymbol{x}\boldsymbol{x}' + \boldsymbol{x} \otimes \boldsymbol{I}_d \otimes \boldsymbol{x}' + (\boldsymbol{x} \otimes \boldsymbol{x})\mathrm{vec}(\boldsymbol{I}_d)'. \tag{A.32}$$

where x is a d-dimensional vector.

A5 Derivative of trace

Theorem A5.1 (Product rule with trace)**.** *Let $\boldsymbol{U}$ and $\boldsymbol{V}$ be matrix functions depending on matrix $\boldsymbol{X}$ such that $\boldsymbol{U}\boldsymbol{V}$ and $\boldsymbol{V}\boldsymbol{U}$ are conformable, then*

$$\frac{\partial \mathrm{tr}(\boldsymbol{U}\boldsymbol{V})}{\partial \boldsymbol{X}} = \frac{\partial}{\partial \boldsymbol{X}}\mathrm{tr}(\boldsymbol{U}_c\boldsymbol{V}) + \frac{\partial}{\partial \boldsymbol{X}}\mathrm{tr}(\boldsymbol{U}\boldsymbol{V}_c) \tag{A.33}$$

where a matrix with subscript "c" is treated as a constant matrix inside the differential operator.

Proof. See equation 14, page 122 of Schonemann [98]. □

Theorem A5.2 (Derivative of Inverse with Trace)**.**

$$\frac{\partial}{\partial \boldsymbol{X}} tr\big(\boldsymbol{Y}^{-1}\big) = -\frac{\partial}{\partial \boldsymbol{X}} tr\big(\boldsymbol{Y}_c^{-2}\boldsymbol{Y}\big). \tag{A.34}$$

Or more generally

$$\frac{\partial}{\partial \boldsymbol{X}} tr\big(\boldsymbol{Y}^{-1}\boldsymbol{A}\big) = -\frac{\partial}{\partial \boldsymbol{X}} tr\big(\boldsymbol{Y}_c^{-1}\boldsymbol{Y}\boldsymbol{Y}_c^{-1}\boldsymbol{A}\big). \tag{A.35}$$

Proof. See equation 15, page 123 of Schonemann [98] □

Using these differential results, we have the following differentiation formulas:

$$\frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{a}-\boldsymbol{x})'\boldsymbol{W}(\boldsymbol{a}-\boldsymbol{x}) = -2\boldsymbol{W}(\boldsymbol{a}-\boldsymbol{x}), \tag{A.36}$$

$$\frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{a}-\boldsymbol{B}\boldsymbol{x})'\boldsymbol{W}(\boldsymbol{a}-\boldsymbol{B}\boldsymbol{x}) = -2\boldsymbol{B}'\boldsymbol{W}(\boldsymbol{a}-\boldsymbol{B}\boldsymbol{x}), \tag{A.37}$$

$$\frac{\partial}{\partial \boldsymbol{X}}\mathrm{tr}\big(\boldsymbol{A}\boldsymbol{X}^{-1}\boldsymbol{B}\big) = -\big(\boldsymbol{X}^{-1}\boldsymbol{B}\boldsymbol{A}\boldsymbol{X}^{-1}\big)', \tag{A.38}$$

$$\frac{\partial \|\boldsymbol{x}\|}{\partial \boldsymbol{x}} = \frac{\partial \sqrt{\boldsymbol{x}'\boldsymbol{x}}}{\partial \boldsymbol{x}} = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} \tag{A.39}$$

where $\boldsymbol{a}$ is a constant vector, $\boldsymbol{W}$ is a constant symmetric matrix, and $\boldsymbol{A}$ and $\boldsymbol{B}$ are constant matrices.

A6 Derivative of vectorisation

Theorem A6.1 (Product Rule with vectorisation)**.** *Let $\boldsymbol{U}$ and $\boldsymbol{V}$ be matrix functions depending on vector $\boldsymbol{x}$ such that $\boldsymbol{U}\boldsymbol{V}$ are conformable, then*

$$\frac{\partial \mathrm{vec}(\boldsymbol{U}\boldsymbol{V})}{\partial \boldsymbol{x}'} = (\boldsymbol{I}_{\mathrm{col}\boldsymbol{V}} \otimes \boldsymbol{U})\frac{\partial \mathrm{vec}\boldsymbol{V}}{\partial \boldsymbol{x}'} + (\boldsymbol{V} \otimes \boldsymbol{I}_{\mathrm{row}\boldsymbol{U}})\frac{\partial \mathrm{vec}\boldsymbol{U}}{\partial \boldsymbol{x}'}. \tag{A.40}$$

Proof. See equation 7, page 668 of Lütkepohl [69]. □

Corollary A6.2. *Let $\boldsymbol{Y}$ be a matrix function depending on vector $\boldsymbol{x}$, and $\boldsymbol{A}$ and $\boldsymbol{B}$ are constant matrices such that $\boldsymbol{A}\boldsymbol{Y}\boldsymbol{B}$ are conformable. Then*

$$\frac{\partial \mathrm{vec}(\boldsymbol{A}\boldsymbol{Y}\boldsymbol{B})}{\partial \boldsymbol{x}'} = (\boldsymbol{B}' \otimes \boldsymbol{A})\frac{\partial \mathrm{vec}\boldsymbol{Y}}{\partial \boldsymbol{x}'}. \tag{A.41}$$

Proof. See equation 7, page 668 of Lütkepohl [69]. □

Theorem A6.3 (Derivative of Inverse with vectorisation)**.**

$$\frac{\partial \mathrm{vec}\boldsymbol{Y}^{-1}}{\partial \boldsymbol{x}'} = -\big(\boldsymbol{Y}^{-\top} \otimes \boldsymbol{Y}^{-1}\big)\frac{\partial \mathrm{vec}\boldsymbol{Y}}{\partial \boldsymbol{x}'}. \tag{A.42}$$

Proof. Suppose $\boldsymbol{Y}$ is a $d \times d$ matrix that depends on vector $\boldsymbol{x}$. We can represent the derivative as $\frac{\partial \mathrm{vec}\left(\boldsymbol{Y}^{-1}\boldsymbol{Y}\boldsymbol{Y}^{-1}\right)}{\boldsymbol{x}'}$. Using the product rule in (A.27) on this representation

gives us

$$\frac{\partial \text{vec}(\boldsymbol{Y}^{-1}\boldsymbol{Y}\boldsymbol{Y}^{-1})}{\partial \boldsymbol{x}'} = \big(\boldsymbol{I}_d \otimes \boldsymbol{Y}^{-1}\big)\frac{\partial \text{vec}(\boldsymbol{Y}\boldsymbol{Y}^{-1})}{\partial \boldsymbol{x}'} + \underbrace{\big(\boldsymbol{Y}^{-\top}\boldsymbol{Y}^{\top} \otimes \boldsymbol{I}_d\big)}_{\boldsymbol{I}_{d^2}}\frac{\partial \text{vec}(\boldsymbol{Y}^{-1})}{\partial \boldsymbol{x}'}. \quad \text{(A.43)}$$

For the first term, we can use the product rule again

$$\frac{\partial \text{vec}(\boldsymbol{Y}\boldsymbol{Y}^{-1})}{\partial \boldsymbol{x}'} = (\boldsymbol{I}_d \otimes \boldsymbol{Y})\frac{\partial \text{vec}(\boldsymbol{Y}^{-1})}{\partial \boldsymbol{x}'} + \big(\boldsymbol{Y}^{-\top} \otimes \boldsymbol{I}_d\big)\frac{\partial \text{vec}(\boldsymbol{Y})}{\partial \boldsymbol{x}'}.$$

Applying this to equation (A.43) and expanding gives us

$$\frac{\partial \text{vec}(\boldsymbol{Y}^{-1}\boldsymbol{Y}\boldsymbol{Y}^{-1})}{\partial \boldsymbol{x}'} = \underbrace{(\boldsymbol{I}_d \otimes \boldsymbol{I}_d)}_{\boldsymbol{I}_{d^2}}\frac{\partial \text{vec}(\boldsymbol{Y}^{-1})}{\partial \boldsymbol{x}'} + \big(\boldsymbol{Y}^{-\top} \otimes \boldsymbol{Y}^{-1}\big)\frac{\partial \text{vec}(\boldsymbol{Y})}{\partial \boldsymbol{x}'} + \frac{\partial \text{vec}(\boldsymbol{Y}^{-1})}{\partial \boldsymbol{x}'}.$$

Note that the left hand side was originally $\frac{\partial \text{vec}(\boldsymbol{Y}^{-1})}{\partial \boldsymbol{x}'}$, rearranging the terms gives us the required result. □

Using these vectorisation differential results, we have the following differentiation formulas:

$$\frac{\partial \text{vec}\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}}{\partial \text{vec}(\boldsymbol{X})'} = \boldsymbol{B}' \otimes \boldsymbol{A}, \quad \text{(A.44)}$$

$$\frac{\partial \text{vec}\boldsymbol{X}^{-1}}{\partial \text{vec}(\boldsymbol{X})'} = -\big(\boldsymbol{X}^{-\top} \otimes \boldsymbol{X}^{-1}\big), \quad \text{(A.45)}$$

$$\frac{\partial \text{vec}\boldsymbol{X}^{-1}\boldsymbol{A}\boldsymbol{X}^{-1}}{\partial \text{vec}(\boldsymbol{X})'} = -\big(\boldsymbol{X}^{-\top} \otimes \boldsymbol{X}^{-1}\boldsymbol{A}\boldsymbol{X}^{-1}\big) - \big(\boldsymbol{X}^{-\top}\boldsymbol{A}^{\top}\boldsymbol{X}^{-\top} \otimes \boldsymbol{X}^{-1}\big), \quad \text{(A.46)}$$

$$\frac{\partial \text{vec}\boldsymbol{X}^{-1}\boldsymbol{B}}{\partial \text{vec}(\boldsymbol{X})'} = -\big(\boldsymbol{B}^{\top}\boldsymbol{X}^{-\top} \otimes \boldsymbol{X}^{-1}\big), \quad \text{(A.47)}$$

$$\frac{\partial \text{vec}\boldsymbol{A}\boldsymbol{X}^{-1}\boldsymbol{B}}{\partial \text{vec}(\boldsymbol{X})'} = -\big(\boldsymbol{B}^{\top}\boldsymbol{X}^{-\top} \otimes \boldsymbol{A}\boldsymbol{X}^{-1}\big), \quad \text{(A.48)}$$

$$\quad \text{(A.49)}$$

where we let $\boldsymbol{A}$ and $\boldsymbol{B}$ constant matrices, and $\boldsymbol{Y}$ be a function of matrix $\boldsymbol{X}$. See Lütkepohl [69] and Petersen and Pedersen [88] for more results on derivatives with vectorisation.

A7 Derivative with respect to structured matrix

In the previous section, we assumed the matrix to have no particular structure. However, for the case when the matrix $\boldsymbol{X}$ has some structure (eg. symmetric, Toeplitz etc.), then the results presented in the previous section does not apply in general. Modifications is required if a structured matrix is considered for differentiation. Here we focus on derivatives with respect to a symmetric matrix. Our aim is to obtain results on derivatives with respect to symmetric matrices based on unstructured matrices. We first introduce the chain rule from multivariable calculus.

Theorem A7.1 (Chain rule)**.** *Suppose $f(\boldsymbol{X})$ is a scalar function that depends on matrix $\boldsymbol{X}$, then*

$$\left[\frac{\partial f}{\partial \boldsymbol{X}}\right]_{ij} = \operatorname{tr}\left(\frac{\partial f}{\partial \boldsymbol{X}^\top}\frac{\partial \boldsymbol{X}}{\partial X_{ij}}\right). \tag{A.50}$$

Note that the term $\frac{\partial \boldsymbol{X}}{\partial X_{ij}}$ is referred to as the *structure matrix of* $\boldsymbol{X}$. For the case when $\boldsymbol{X}$ has no structure, then the structure matrix of $\boldsymbol{X}$ is simply given by $\frac{\partial \boldsymbol{X}}{\partial X_{ij}} = \boldsymbol{J}^{ij}$, where $\boldsymbol{J}^{ij}$ represents a single-entry matrix with one in the $(i,j)^{\text{th}}$ entry and zeroes everywhere else.

Now suppose $\boldsymbol{\Sigma}$ is a $d \times d$ symmetric matrix, then the structure matrix of $\boldsymbol{\Sigma}$ is given by

$$\frac{\partial \boldsymbol{\Sigma}}{\partial \Sigma_{ij}} = \boldsymbol{J}^{ij} + \boldsymbol{J}^{ji} - \boldsymbol{J}^{ij}\boldsymbol{J}^{ji}. \tag{A.51}$$

A7.1 Derivatives with respect to symmetric matrix

Applying the structure matrix of $\boldsymbol{\Sigma}$ in (A.51) to the chain rule in (A.50), this gives us the derivative formula with respect to symmetric matrix $\boldsymbol{\Sigma}$

$$\frac{\partial f}{\partial \boldsymbol{\Sigma}} = \frac{\partial f}{\partial \boldsymbol{\Sigma}_u} + \frac{\partial f}{\partial \boldsymbol{\Sigma}_u^\top} - \operatorname{diag}\left(\frac{\partial f}{\partial \boldsymbol{\Sigma}_u}\right) \tag{A.52}$$

where $\boldsymbol{\Sigma}_u$ represents an unstructured matrix version of $\boldsymbol{\Sigma}$, and $\operatorname{diag}(\boldsymbol{X})$ is diagonal matrix with diagonal entries from square matrix $\boldsymbol{X}$ and zeroes everywhere else.

If $\frac{\partial f}{\partial \boldsymbol{\Sigma}_u} = \frac{\partial f}{\partial \boldsymbol{\Sigma}_u^\top}$ or equivalently $f(\boldsymbol{\Sigma}_u) = f(\boldsymbol{\Sigma}_u^\top)$, then the derivative of f with respect to symmetric matrix $\boldsymbol{\Sigma}$ can be represented as

$$\frac{\partial f}{\partial \boldsymbol{\Sigma}} = \boldsymbol{C} \circ \frac{\partial f}{\partial \boldsymbol{\Sigma}_u} \tag{A.53}$$

where $\boldsymbol{C} = 2(\mathbb{1}_{\dim \boldsymbol{\Sigma}}) - \boldsymbol{I}_d$. Applying the vectorisation and half-vectorisation operation

$$\begin{aligned}
\frac{\partial f}{\partial \mathrm{vec}\boldsymbol{\Sigma}} &= \mathrm{vec}(\boldsymbol{C}) \circ \frac{\partial f}{\partial \mathrm{vec}\boldsymbol{\Sigma}_u}, \\
\frac{\partial f}{\partial \mathrm{vech}\boldsymbol{\Sigma}} &= \boldsymbol{L}_d \left(\mathrm{vec}(\boldsymbol{C}) \circ \frac{\partial f}{\partial \mathrm{vec}\boldsymbol{\Sigma}_u} \right)
\end{aligned}$$

where $\boldsymbol{L}_d$ represents an elimination matrix defined in (A.6). Alternatively, we can write

$$\frac{\partial f}{\partial \mathrm{vech}\boldsymbol{\Sigma}} = \boldsymbol{D}_d^\top \frac{\partial f}{\partial \mathrm{vec}\boldsymbol{\Sigma}_u} \tag{A.54}$$

where $\boldsymbol{D}_d$ represents a duplication matrix defined in (A.7).

A7.2 Second order derivatives with respect to symmetric matrix

Here we represent the second order derivatives with respect to symmetric matrix which includes $\frac{\partial^2}{\partial \boldsymbol{\Sigma} \otimes \partial \boldsymbol{\Sigma}}$, $\frac{\partial^2}{\partial \mathrm{vec}\boldsymbol{\Sigma}\, \partial \mathrm{vec}(\boldsymbol{\Sigma})'}$, $\frac{\partial^2}{\partial \mathrm{vech}\boldsymbol{\Sigma}\, \partial \mathrm{vech}(\boldsymbol{\Sigma})'}$, and the cross derivatives $\frac{\partial^2}{\partial \mathrm{vec}\boldsymbol{\Sigma}\, \partial \boldsymbol{x}'}$, $\frac{\partial^2}{\partial \mathrm{vech}\boldsymbol{\Sigma}\, \partial \boldsymbol{x}'}$:

$$\begin{aligned}
\frac{\partial^2 f}{\partial \boldsymbol{\Sigma}\, \partial \boldsymbol{\Sigma}} &= \left(\frac{\partial}{\partial \boldsymbol{\Sigma}} \right) \otimes \left(\frac{\partial}{\partial \boldsymbol{\Sigma}} \right) f \\
&= \left(\boldsymbol{C} \circ \frac{\partial}{\partial \boldsymbol{\Sigma}_u} \right) \otimes \left(\boldsymbol{C} \circ \frac{\partial}{\partial \boldsymbol{\Sigma}_u} \right) f \\
&= (\boldsymbol{C} \otimes \boldsymbol{C}) \circ \left(\frac{\partial^2 f}{\partial \boldsymbol{\Sigma}_u \partial \boldsymbol{\Sigma}_u} \right)
\end{aligned}$$

using (B.1) for the last equality and

$$\begin{aligned}
\frac{\partial^2 f}{\partial \mathrm{vec}\boldsymbol{\Sigma}\, \partial \mathrm{vec}(\boldsymbol{\Sigma})'} &= \mathrm{vec}\left(\frac{\partial}{\partial \boldsymbol{\Sigma}} \right) \mathrm{vec}\left(\frac{\partial}{\partial \boldsymbol{\Sigma}} \right)' f \\
&= \mathrm{vec}\left(\boldsymbol{C} \circ \frac{\partial}{\partial \boldsymbol{\Sigma}_u} \right) \mathrm{vec}\left(\boldsymbol{C} \circ \frac{\partial}{\partial \boldsymbol{\Sigma}_u} \right)' f \\
&= \left(\mathrm{vec}\boldsymbol{C} \circ \frac{\partial}{\partial \mathrm{vec}\boldsymbol{\Sigma}_u} \right) \left(\mathrm{vec}(\boldsymbol{C})' \circ \frac{\partial}{\partial \mathrm{vec}(\boldsymbol{\Sigma}_u)'} \right) f \\
&= (\mathrm{vec}\boldsymbol{C}\, \mathrm{vec}(\boldsymbol{C})') \circ \left(\frac{\partial^2 f}{\partial \mathrm{vec}\boldsymbol{\Sigma}_u\, \partial \mathrm{vec}(\boldsymbol{\Sigma}_u)'} \right)
\end{aligned}$$

using the property $\mathrm{vec}(\boldsymbol{A} \circ \boldsymbol{B}) = \mathrm{vec}(\boldsymbol{A}) \circ \mathrm{vec}(\boldsymbol{B})$ for the third equality, and (B.2) for the last equality.

Similarly, we have

$$\frac{\partial^2 f}{\partial \mathrm{vech}\boldsymbol{\Sigma}\, \partial \mathrm{vech}(\boldsymbol{\Sigma})'} = \boldsymbol{D}_d^\top \frac{\partial^2 f}{\partial \mathrm{vec}\boldsymbol{\Sigma}_u\, \partial \mathrm{vec}(\boldsymbol{\Sigma}_u)'} \boldsymbol{D}_d \tag{A.55}$$

and

$$\begin{aligned} \frac{\partial^2 f}{\partial \mathrm{vec}\boldsymbol{\Sigma}\, \partial \boldsymbol{x}'} &= (\mathrm{vec}\boldsymbol{C}\, \mathbb{1}_d') \circ \left(\frac{\partial^2 f}{\partial \mathrm{vec}\boldsymbol{\Sigma}_u\, \partial \boldsymbol{x}'} \right), \\ \frac{\partial^2 f}{\partial \mathrm{vech}\boldsymbol{\Sigma}\, \partial \boldsymbol{x}'} &= \boldsymbol{D}_d^\top \frac{\partial^2 f}{\partial \mathrm{vec}\boldsymbol{\Sigma}_u\, \partial \boldsymbol{x}'}. \end{aligned} \tag{A.56}$$

A8 Derivatives of complete data log-likelihood for VG distribution

In this section, we present the derivatives of the log-likelihood for the VG distribution. These techniques can be applied to other models with NMVM representation such as the GH distribution.

Recall from (2.3) the complete data log-likelihood function of the d-dimensional VG distribution is given by

$$\ell(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{u}) = \ell_N(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}; \boldsymbol{y}, \boldsymbol{u}) + \ell_G(\nu; \boldsymbol{u})$$

where

$$\ell_N(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}; \boldsymbol{y}, \boldsymbol{u}) = -\frac{n}{2} \log|\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{i=1}^n \frac{1}{u_i} (\boldsymbol{y}_i - \boldsymbol{\mu} - u_i \boldsymbol{\gamma})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu} - u_i \boldsymbol{\gamma})$$

and

$$\ell_G(\nu; \boldsymbol{u}) = n\nu \log \nu - n \log \Gamma(\nu) + (\nu - 1) \sum_{i=1}^n \log u_i - \nu \sum_{i=1}^n u_i \,.$$

A8.1 First order derivatives

The first order derivatives of the complete data log-likelihood for the VG distribution is given by

$$\frac{\partial \ell_N}{\partial \boldsymbol{\mu}} = \boldsymbol{\Sigma}^{-1} \sum_{i=1}^{n} \frac{1}{u_i} (\boldsymbol{y}_i - \boldsymbol{\mu} - u_i \boldsymbol{\gamma}), \tag{A.57}$$

$$\frac{\partial \ell_N}{\partial \boldsymbol{\gamma}} = \boldsymbol{\Sigma}^{-1} \sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\mu} - u_i \boldsymbol{\gamma}), \tag{A.58}$$

$$\frac{\partial \ell_N}{\partial \text{vech}\boldsymbol{\Sigma}} = \boldsymbol{D}_d^{\top} \frac{\partial \ell_N}{\partial \text{vec}\boldsymbol{\Sigma}_u}, \tag{A.59}$$

$$\frac{\partial \ell_G}{\partial \nu} = n + n \log \nu - n\psi(\nu) + \sum_{i=1}^{n} \log(u_i) - \sum_{i=1}^{n} u_i \tag{A.60}$$

where (A.36) is used for the first equation, (A.37) for the second equation, and (A.54) for the third equation with

$$\frac{\partial \ell_N}{\partial \text{vec}\boldsymbol{\Sigma}_u} = \text{vec}\left(\frac{\partial \ell_N}{\partial \boldsymbol{\Sigma}_u} \right) = \text{vec}\left(-\frac{n}{2} \boldsymbol{\Sigma}^{-1} + \frac{1}{2} \boldsymbol{\Sigma}^{-1} S_{\tilde{\boldsymbol{y}}\tilde{\boldsymbol{y}}/u} \boldsymbol{\Sigma}^{-1} \right) \tag{A.61}$$

which follows from (A.24) and (A.38), and

$$S_{\tilde{\boldsymbol{y}}\tilde{\boldsymbol{y}}/u} = \sum_{i=1}^{n} \frac{1}{u_i} (\boldsymbol{y}_i - \boldsymbol{\mu}_i - u_i \boldsymbol{\gamma})(\boldsymbol{y}_i - \boldsymbol{\mu}_i - u_i \boldsymbol{\gamma})'.$$

A8.2 Second order derivatives

The second order derivatives of the complete data log-likelihood for the VG distribution is given by

$$\frac{\partial^2 \ell_N}{\partial \boldsymbol{\mu} \partial \boldsymbol{\mu}'} = -\boldsymbol{\Sigma}^{-1} \sum_{i=1}^{n} \frac{1}{u_i}$$

$$\frac{\partial^2 \ell_N}{\partial \boldsymbol{\gamma} \, \partial \boldsymbol{\gamma}'} = -\boldsymbol{\Sigma}^{-1} \sum_{i=1}^{n} u_i$$

$$\frac{\partial^2 \ell_N}{\partial \text{vech}\boldsymbol{\Sigma} \, \partial \text{vech}(\boldsymbol{\Sigma})'} = \boldsymbol{D}_d^{\top} \frac{\partial^2 \ell_N}{\partial \text{vec}\boldsymbol{\Sigma}_u \, \partial \text{vec}(\boldsymbol{\Sigma}_u)'} \boldsymbol{D}_d$$

$$\frac{\partial^2 \ell_N}{\partial \nu^2} = \frac{n}{\nu} - n\psi'(\nu)$$

where (A.20) is used for the first and second equation, (A.55) for the third equation with

$$\begin{aligned}&\frac{\partial^2 \ell_N}{\partial \mathrm{vec}\boldsymbol{\Sigma}_u \, \partial \mathrm{vec}(\boldsymbol{\Sigma}_u)'} \\ &= \frac{n}{2}\big(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1}\big) - \frac{1}{2}\big(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1} S_{\tilde{\boldsymbol{y}}\tilde{\boldsymbol{y}}/u} \boldsymbol{\Sigma}^{-1}\big) - \frac{1}{2}\big(\boldsymbol{\Sigma}^{-1} S_{\tilde{\boldsymbol{y}}\tilde{\boldsymbol{y}}/u} \boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1}\big)\end{aligned}$$

which follows from (A.45) and (A.46).

A8.3 Cross derivatives

The cross derivatives of the complete data log-likelihood for the VG distribution is given by

$$\begin{aligned}\frac{\partial^2 \ell_N}{\partial \mathrm{vech}\boldsymbol{\Sigma} \, \partial \boldsymbol{\mu}'} &= \boldsymbol{D}_d^\top \frac{\partial^2 \ell_N}{\partial \mathrm{vec}\boldsymbol{\Sigma}_u \, \partial \boldsymbol{\mu}'}, \\ \frac{\partial^2 \ell_N}{\partial \mathrm{vech}\boldsymbol{\Sigma} \, \partial \boldsymbol{\gamma}'} &= \boldsymbol{D}_d^\top \frac{\partial^2 \ell_N}{\partial \mathrm{vec}\boldsymbol{\Sigma}_u \, \partial \boldsymbol{\gamma}'}, \\ \frac{\partial^2 \ell_N}{\partial \boldsymbol{\mu} \, \partial \boldsymbol{\gamma}'} &= -n\boldsymbol{\Sigma}^{-1}\end{aligned}$$

where (A.56) is used for the first and second equation, and (A.20) for the third equation with

$$\begin{aligned}\frac{\partial^2 \ell_N}{\partial \mathrm{vec}\boldsymbol{\Sigma} \, \partial \boldsymbol{\mu}'} &= -\boldsymbol{\Sigma}^{-1} \sum_{i=1}^{n} \frac{1}{u_i}(\boldsymbol{y}_i - \boldsymbol{\mu} - u_i\boldsymbol{\gamma}) \otimes \boldsymbol{\Sigma}^{-1}, \\ \frac{\partial^2 \ell_N}{\partial \mathrm{vec}\boldsymbol{\Sigma} \, \partial \boldsymbol{\gamma}'} &= -\boldsymbol{\Sigma}^{-1} \sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\mu} - u_i\boldsymbol{\gamma}) \otimes \boldsymbol{\Sigma}^{-1}\end{aligned}$$

which follows from (A.47). Note that the other cross derivatives are just a zero matrix.

A9 Derivative of complete data log-likelihood for VARMA-VG model

The derivatives for $\ell_G(\nu; \boldsymbol{u})$ is the same as in Section A8. So it is sufficient to focus on the the (conditional) log-likelihood of the conditional normal distribution for the

d-dimensional VARMA-VG model in Section 5.3 which is given by

$$\begin{aligned}\ell_N(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\Sigma};\boldsymbol{y},\boldsymbol{u}|\mathscr{F}_0) &= -\frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{t=1}^{n}\frac{1}{u_t}\tilde{\boldsymbol{\varepsilon}}_t'\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{\varepsilon}}_t \\ &= -\frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{t=1}^{n}\frac{1}{u_t}(\boldsymbol{\varepsilon}_t+\boldsymbol{\gamma})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\varepsilon}_t+\boldsymbol{\gamma}) \\ &\quad - \frac{1}{2}\sum_{t=1}^{n}u_t\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma} + \sum_{t=1}^{n}(\boldsymbol{\varepsilon}_t+\boldsymbol{\gamma})'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}\end{aligned} \tag{A.62}$$

where $\mathscr{F}_0$ represents the filtration up to time t and

$$\begin{aligned}\tilde{\boldsymbol{\varepsilon}}_t &= \boldsymbol{\varepsilon}_t + \boldsymbol{\gamma} - u_t\boldsymbol{\gamma} \\ &= \boldsymbol{y}_t - \boldsymbol{\beta}'\boldsymbol{x}_t - u_t\boldsymbol{\gamma}\end{aligned} \tag{A.63}$$

for $t = 1, \dots, n$.

A9.1 First order derivatives

Since the method to obtain the first order derivatives with respect to vech$\boldsymbol{\Sigma}$ and ν are similar to the previous section, we only need to focus on derivatives with respect to vec$(\boldsymbol{\beta}')$ and $\boldsymbol{\gamma}$.

Derivative with respect to vec$(\boldsymbol{\beta}')$:

The first derivative of ℓ_N with respect to vec$(\boldsymbol{\beta}')$ is given by

$$\frac{\partial\ell_N}{\partial\text{vec}(\boldsymbol{\beta}')} = -\sum_{t=1}^{n}\frac{1}{u_t}\frac{\partial\tilde{\boldsymbol{\varepsilon}}_t'}{\partial\text{vec}(\boldsymbol{\beta}')}\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{\varepsilon}}_t \tag{A.64}$$

where differentiating the transpose of $\tilde{\boldsymbol{\varepsilon}}_t$ in (A.63) with respect to vec$(\boldsymbol{\beta}')$ gives us

$$\frac{\partial\tilde{\boldsymbol{\varepsilon}}_t'}{\partial\text{vec}(\boldsymbol{\beta}')} = \frac{\partial\boldsymbol{\varepsilon}_t'}{\partial\text{vec}(\boldsymbol{\beta}')} = -\frac{\partial(\boldsymbol{x}_t'\boldsymbol{\beta})}{\partial\text{vec}(\boldsymbol{\beta}')} = -\frac{\partial\text{vec}(\boldsymbol{\beta}'\boldsymbol{x}_t)'}{\partial\text{vec}(\boldsymbol{\beta}')} = -\left(\frac{\partial\boldsymbol{x}_t'}{\partial\text{vec}(\boldsymbol{\beta}')}\right)\boldsymbol{\beta} - \underbrace{\frac{\partial\text{vec}(\boldsymbol{\beta}')'}{\partial\text{vec}(\boldsymbol{\beta}')}}_{\boldsymbol{I}}(\boldsymbol{x}_t\otimes\boldsymbol{I}_d) \tag{A.65}$$

with Theorem A6.1 applied to the last equality.

$$\frac{\partial \tilde{\boldsymbol{\varepsilon}}_t'}{\partial \text{vec}(\boldsymbol{\beta}')} = -\left(\frac{\partial \boldsymbol{x}_t'}{\partial \text{vec}(\boldsymbol{\beta}')}\right)\boldsymbol{\beta} - \underbrace{\frac{\partial \text{vec}(\boldsymbol{\beta}')'}{\partial \text{vec}(\boldsymbol{\beta}')}}_{\boldsymbol{I}}(\boldsymbol{x}_t \otimes \boldsymbol{I}_d)$$

For $\frac{\partial \boldsymbol{x}_t'}{\partial \text{vec}(\boldsymbol{\beta}')}$, recall that

$$\boldsymbol{x}_t' = \begin{pmatrix} 1 & \boldsymbol{y}_{t-1}' & \cdots & \boldsymbol{y}_{t-p}' & -\boldsymbol{\varepsilon}_{t-1}' & \cdots & -\boldsymbol{\varepsilon}_{t-q}' \end{pmatrix}$$

defined in (5.28). So differentiating with respect to $\text{vec}(\boldsymbol{\beta}')$ gives us

$$\begin{aligned} \frac{\partial \boldsymbol{x}_t'}{\partial \text{vec}(\boldsymbol{\beta}')} &= \begin{pmatrix} 0 & \mathbf{0} & \cdots & \mathbf{0} & -\frac{\partial \boldsymbol{\varepsilon}_{t-1}'}{\partial \text{vec}(\boldsymbol{\beta}')} & \cdots & -\frac{\partial \boldsymbol{\varepsilon}_{t-q}'}{\partial \text{vec}(\boldsymbol{\beta}')} \end{pmatrix} \\ &= \begin{pmatrix} 0 & \mathbf{0} & \cdots & \mathbf{0} & -\frac{\partial \tilde{\boldsymbol{\varepsilon}}_{t-1}'}{\partial \text{vec}(\boldsymbol{\beta}')} & \cdots & -\frac{\partial \tilde{\boldsymbol{\varepsilon}}_{t-q}'}{\partial \text{vec}(\boldsymbol{\beta}')} \end{pmatrix} \end{aligned}$$

where and $\frac{\partial \boldsymbol{\varepsilon}_t'}{\partial \text{vec}(\boldsymbol{\beta}')} = \frac{\partial (\tilde{\boldsymbol{\varepsilon}}_t - \boldsymbol{\gamma} + u\boldsymbol{\gamma})'}{\partial \text{vec}(\boldsymbol{\beta}')} = \frac{\partial \tilde{\boldsymbol{\varepsilon}}_t'}{\partial \text{vec}(\boldsymbol{\beta}')}$.

Note that $\frac{\partial \tilde{\boldsymbol{\varepsilon}}_t'}{\partial \text{vec}(\boldsymbol{\beta}')}$ and $\frac{\partial \boldsymbol{x}_t'}{\partial \text{vec}(\boldsymbol{\beta}')}$ can be computed iteratively by calculating, $\frac{\partial \boldsymbol{x}_t'}{\partial \text{vec}(\boldsymbol{\beta}')}$, then $\frac{\partial \tilde{\boldsymbol{\varepsilon}}_t'}{\partial \text{vec}(\boldsymbol{\beta}')}$ for $t \mapsto t+1$ iteration where we assume $\boldsymbol{\varepsilon}_t = \mathbf{0}$ for $t \leq 0$.

Algorithm 12: Computing $\frac{\partial \boldsymbol{x}_t'}{\partial \text{vec}(\boldsymbol{\beta}')}$ and $\frac{\partial \tilde{\boldsymbol{\varepsilon}}_t'}{\partial \text{vec}(\boldsymbol{\beta}')}$ for $t = 1, \ldots, n$

Input: Initial value $\boldsymbol{\varepsilon}_t = \mathbf{0}$ for $t \leq 0$
for $t = 1, \ldots, n$ **do**
  $\frac{\partial \boldsymbol{x}_t'}{\partial \text{vec}(\boldsymbol{\beta}')} \leftarrow \begin{pmatrix} 0 & \mathbf{0} & \cdots & \mathbf{0} & -\frac{\partial \tilde{\boldsymbol{\varepsilon}}_{t-1}'}{\partial \text{vec}(\boldsymbol{\beta}')} & \cdots & -\frac{\partial \tilde{\boldsymbol{\varepsilon}}_{t-q}'}{\partial \text{vec}(\boldsymbol{\beta}')} \end{pmatrix}$;
  $\frac{\partial \tilde{\boldsymbol{\varepsilon}}_t'}{\partial \text{vec}(\boldsymbol{\beta}')} \leftarrow -\left(\frac{\partial \boldsymbol{x}_t'}{\partial \text{vec}(\boldsymbol{\beta}')}\right)\boldsymbol{\beta} - (\boldsymbol{x}_t \otimes \boldsymbol{I}_d)$;
end

Derivative with respect to $\boldsymbol{\gamma}$:

The first derivative of ℓ_N with respect to $\boldsymbol{\gamma}$ is given by

$$\begin{aligned} \frac{\partial \ell_N}{\partial \boldsymbol{\gamma}} = &-\sum_{t=1}^{n} \frac{1}{u_t}\left(\frac{\partial \boldsymbol{\varepsilon}_t'}{\partial \boldsymbol{\gamma}} + \boldsymbol{I}_d\right)\boldsymbol{\Sigma}^{-1}(\boldsymbol{\varepsilon}_t + \boldsymbol{\gamma}) + \sum_{t=1}^{n}\left(\frac{\partial \boldsymbol{\varepsilon}_t'}{\partial \boldsymbol{\gamma}} + \boldsymbol{I}_d\right)\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma} \\ &- \sum_{t=1}^{n} u_t \boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma} + \sum_{t=1}^{n} \boldsymbol{\Sigma}^{-1}(\boldsymbol{\varepsilon}_t + \boldsymbol{\gamma}) \end{aligned} \tag{A.66}$$

where

$$\frac{\partial \boldsymbol{\varepsilon}_t'}{\partial \boldsymbol{\gamma}} = -\frac{\partial(\boldsymbol{x}_t'\boldsymbol{\beta} - \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} = -\left(\frac{\partial \boldsymbol{x}_t'}{\partial \boldsymbol{\gamma}}\right)\boldsymbol{\beta} - \boldsymbol{I}_d$$

and

$$\frac{\partial \boldsymbol{x}_t'}{\partial \boldsymbol{\gamma}} = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & -\frac{\partial \boldsymbol{\varepsilon}_{t-1}'}{\partial \boldsymbol{\gamma}} & \cdots & -\frac{\partial \boldsymbol{\varepsilon}_{t-q}'}{\partial \boldsymbol{\gamma}} \end{pmatrix}.$$

The first and second term of (A.66) in probability goes to zero as $n \to \infty$ since $\mathbb{E}[\frac{1}{u_t}(\boldsymbol{\varepsilon}_t + \boldsymbol{\gamma})] = \boldsymbol{\gamma}$ which follows from (1.33) and $\frac{\partial \boldsymbol{\varepsilon}_t'}{\partial \boldsymbol{\gamma}}$ does not depend on $\boldsymbol{\varepsilon}_t$. So the derivative in (A.66) can be approximated by

$$\begin{aligned} \frac{\partial \ell_N}{\partial \boldsymbol{\gamma}} &\approx -\sum_{t=1}^{n} u_t \boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma} + \sum_{t=1}^{n} \boldsymbol{\Sigma}^{-1}(\boldsymbol{\varepsilon}_t + \boldsymbol{\gamma}) \\ &= \boldsymbol{\Sigma}^{-1} \sum_{t=1}^{n} \tilde{\boldsymbol{\varepsilon}}_t \end{aligned} \tag{A.67}$$

A9.2 Second order derivatives

In this section, we focus on the second order derivatives that involves derivatives with respect to $\text{vec}(\boldsymbol{\beta}')$ and $\boldsymbol{\gamma}$.

Second order derivative with respect to $\text{vec}(\boldsymbol{\beta}')$:

Differentiating (A.64) with respect to $\text{vec}(\boldsymbol{\beta}')'$ gives us

$$\begin{aligned} &\frac{\partial^2 \ell_N}{\partial \text{vec}(\boldsymbol{\beta}') \partial \text{vec}(\boldsymbol{\beta}')'} \\ &= -\sum_{t=1}^{n} \frac{1}{u_t} \frac{\partial}{\partial \text{vec}(\boldsymbol{\beta}')'} \left[\frac{\partial \tilde{\boldsymbol{\varepsilon}}_t'}{\partial \text{vec}(\boldsymbol{\beta}')} \boldsymbol{\Sigma}^{-1} \tilde{\boldsymbol{\varepsilon}}_t \right] \\ &= -\sum_{t=1}^{n} \frac{1}{u_t} \left[\frac{\partial \tilde{\boldsymbol{\varepsilon}}_t'}{\partial \text{vec}(\boldsymbol{\beta}')} \boldsymbol{\Sigma}^{-1} \frac{\partial \tilde{\boldsymbol{\varepsilon}}_t}{\partial \text{vec}(\boldsymbol{\beta}')} + \left(\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{\varepsilon}}_t \otimes \boldsymbol{I}\right) \frac{\partial \text{vec}\left(\frac{\partial \tilde{\boldsymbol{\varepsilon}}_t'}{\partial \text{vec}(\boldsymbol{\beta}')}\right)}{\partial \text{vec}(\boldsymbol{\beta}')'} \right] \end{aligned}$$

where (A.40) was used for the last equality. Note that the second term in probability goes to zero as $n \to \infty$ since $\mathbb{E}[\tilde{\boldsymbol{\varepsilon}}_t] = \mathbf{0}$ and $\frac{\partial \tilde{\boldsymbol{\varepsilon}}_t'}{\partial \text{vec}(\boldsymbol{\beta}')}$ is independent of $\tilde{\boldsymbol{\varepsilon}}_t$ where the idea of the proof is similar to [69, Lemma 12.1]. Then

$$\frac{\partial^2 \ell_N}{\partial \text{vec}(\boldsymbol{\beta}')\partial \text{vec}(\boldsymbol{\beta}')'} \approx -\sum_{t=1}^{n} \frac{1}{u_t} \frac{\partial \tilde{\boldsymbol{\varepsilon}}_t'}{\partial \text{vec}(\boldsymbol{\beta}')} \boldsymbol{\Sigma}^{-1} \frac{\partial \tilde{\boldsymbol{\varepsilon}}_t}{\partial \text{vec}(\boldsymbol{\beta}')}.$$

Second order derivative with respect to $\boldsymbol{\gamma}$:

Differentiating (A.66) with respect to $\boldsymbol{\gamma}$ gives us

$$\begin{aligned}\frac{\partial^2 \ell_N}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'} = &-\sum_{t=1}^{n} \frac{1}{u_t} \left(\frac{\partial \boldsymbol{\varepsilon}_t'}{\partial \boldsymbol{\gamma}} + \boldsymbol{I}_d\right) \boldsymbol{\Sigma}^{-1} \left(\frac{\partial \boldsymbol{\varepsilon}_t}{\partial \boldsymbol{\gamma}'} + \boldsymbol{I}_d\right) - \sum_{t=1}^{n} \frac{1}{u_t} \left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{\varepsilon}_t + \boldsymbol{\gamma}) \otimes \boldsymbol{I}_d\right] \frac{\partial \text{vec}\left(\frac{\partial \boldsymbol{\varepsilon}_t'}{\partial \boldsymbol{\gamma}} + \boldsymbol{I}_d\right)}{\partial \boldsymbol{\gamma}'} \\ &+ \sum_{t=1}^{n} \left(\frac{\partial \boldsymbol{\varepsilon}_t'}{\partial \boldsymbol{\gamma}} + \boldsymbol{I}_d\right) \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\gamma}}{\partial \boldsymbol{\gamma}'} + \sum_{t=1} (\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma} \otimes \boldsymbol{I}_d) \frac{\partial \text{vec}\left(\frac{\partial \boldsymbol{\varepsilon}_t'}{\partial \boldsymbol{\gamma}} + \boldsymbol{I}_d\right)}{\partial \boldsymbol{\gamma}'} \\ &- \boldsymbol{\Sigma}^{-1} \sum_{t=1}^{n} u_t + \sum_{t=1}^{n} \boldsymbol{\Sigma}^{-1} \left(\frac{\partial \boldsymbol{\varepsilon}_t'}{\partial \boldsymbol{\gamma}} + \boldsymbol{I}_d\right).\end{aligned}$$

Since $\mathbb{E}[\frac{1}{u_t}(\boldsymbol{\varepsilon}_t + \boldsymbol{\gamma})] = \boldsymbol{\gamma}$ and $\frac{\partial \boldsymbol{\varepsilon}_t'}{\partial \boldsymbol{\gamma}}$ does not depend on $\boldsymbol{\varepsilon}_t$, this gives us

$$\begin{aligned}\frac{\partial^2 \ell_N}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'} = &-\sum_{t=1}^{n} \frac{1}{u_t} \left(\frac{\partial \boldsymbol{\varepsilon}_t'}{\partial \boldsymbol{\gamma}} + \boldsymbol{I}_d\right) \boldsymbol{\Sigma}^{-1} \left(\frac{\partial \boldsymbol{\varepsilon}_t}{\partial \boldsymbol{\gamma}'} + \boldsymbol{I}_d\right) - \boldsymbol{\Sigma}^{-1} \sum_{t=1}^{n} u_t \\ &+ \sum_{t=1}^{n} \left(\frac{\partial \boldsymbol{\varepsilon}_t'}{\partial \boldsymbol{\gamma}} + \boldsymbol{I}_d\right) \boldsymbol{\Sigma}^{-1} + \sum_{t=1}^{n} \boldsymbol{\Sigma}^{-1} \left(\frac{\partial \boldsymbol{\varepsilon}_t'}{\partial \boldsymbol{\gamma}} + \boldsymbol{I}_d\right).\end{aligned}$$

A9.3 Cross derivatives

In this section, we focus on the cross derivatives that involves derivatives with respect to $\text{vec}(\boldsymbol{\beta}')$, $\boldsymbol{\gamma}$ and $\text{vech}(\boldsymbol{\Sigma})$

Cross derivative with respect to $(\text{vec}(\boldsymbol{\beta}'), \boldsymbol{\gamma})$:

To obtain the cross derivative with respect to $(\text{vec}(\boldsymbol{\beta}'), \boldsymbol{\gamma})$, note that (A.64) can be represented as

$$\frac{\partial \ell_N}{\partial \text{vec}(\boldsymbol{\beta}')} = -\sum_{t=1}^{n} \frac{1}{u_t} \frac{\partial \boldsymbol{\varepsilon}_t'}{\partial \text{vec}(\boldsymbol{\beta}')} \boldsymbol{\Sigma}^{-1}(\boldsymbol{\varepsilon}_t + \boldsymbol{\gamma}) + \sum_{t=1}^{n} \frac{\partial \boldsymbol{\varepsilon}_t'}{\partial \text{vec}(\boldsymbol{\beta}')} \boldsymbol{\Sigma}^{-1} \boldsymbol{\gamma}. \tag{A.68}$$

Taking derivative with respect to $\boldsymbol{\gamma}'$ gives us

$$\frac{\partial^2 \ell_N}{\partial \mathrm{vec}(\boldsymbol{\beta}')\,\partial \boldsymbol{\gamma}'} = -\sum_{t=1}^{n} \frac{1}{u_t} \frac{\partial \boldsymbol{\varepsilon}_t'}{\partial \mathrm{vec}(\boldsymbol{\beta}')} \boldsymbol{\Sigma}^{-1} \left(\frac{\partial \boldsymbol{\varepsilon}_t'}{\partial \boldsymbol{\gamma}} + \boldsymbol{I}_d \right) + \sum_{t=1}^{n} \frac{\partial \boldsymbol{\varepsilon}_t'}{\partial \mathrm{vec}(\boldsymbol{\beta}')} \boldsymbol{\Sigma}^{-1}$$

where $\frac{\partial \boldsymbol{\varepsilon}_t'}{\partial \mathrm{vec}(\boldsymbol{\beta}')}$ in (A.65) does not depend on $\boldsymbol{\gamma}$.

Cross derivative with respect to $(\mathbf{vec}(\boldsymbol{\beta}'), \mathbf{vech}\boldsymbol{\Sigma})$:

To obtain the cross derivative with respect to $(\mathrm{vec}(\boldsymbol{\beta}'), \mathrm{vech}\boldsymbol{\Sigma})$, differentiate (A.68) with respect to $\mathrm{vec}(\boldsymbol{\Sigma}_u)'$ by using formula (A.48) gives us

$$\begin{aligned}
&\frac{\partial^2 \ell_N}{\partial \mathrm{vec}(\boldsymbol{\beta}')\,\partial \mathrm{vec}(\boldsymbol{\Sigma}_u)'} \\
&= \sum_{t=1}^{n} \frac{1}{u_t} \left[(\boldsymbol{\varepsilon}_t + \boldsymbol{\gamma})' \boldsymbol{\Sigma}^{-1} \otimes \frac{\partial \boldsymbol{\varepsilon}_t'}{\partial \mathrm{vec}(\boldsymbol{\beta}')} \boldsymbol{\Sigma}^{-1} \right] - \sum_{t=1}^{n} \left[\boldsymbol{\gamma}' \boldsymbol{\Sigma}^{-1} \otimes \frac{\partial \boldsymbol{\varepsilon}_t'}{\partial \mathrm{vec}(\boldsymbol{\beta}')} \boldsymbol{\Sigma}^{-1} \right].
\end{aligned}$$

Thus

$$\frac{\partial^2 \ell_N}{\partial \mathrm{vec}(\boldsymbol{\beta}')\,\partial \mathrm{vech}(\boldsymbol{\Sigma})'} = \frac{\partial^2 \ell_N}{\partial \mathrm{vec}(\boldsymbol{\beta}')\,\partial \mathrm{vec}(\boldsymbol{\Sigma}_u)'} \boldsymbol{D}_d.$$

Cross derivative with respect to $(\boldsymbol{\gamma}, \mathbf{vech}\boldsymbol{\Sigma})$:

To obtain the cross derivative with respect to $(\boldsymbol{\gamma}, \mathrm{vech}\boldsymbol{\Sigma})$, differentiate (A.66) with respect to $\mathrm{vec}(\boldsymbol{\Sigma}_u)'$

$$\begin{aligned}
\frac{\partial^2 \ell_N}{\partial \boldsymbol{\gamma}\,\partial \mathrm{vec}(\boldsymbol{\Sigma}_u)'} = &\sum_{t=1}^{n} \frac{1}{u_t} \left[(\boldsymbol{\varepsilon}_t + \boldsymbol{\gamma})' \boldsymbol{\Sigma}^{-1} \otimes \left(\frac{\partial \boldsymbol{\varepsilon}_t'}{\partial \boldsymbol{\gamma}} + \boldsymbol{I}_d \right) \boldsymbol{\Sigma}^{-1} \right] + \sum_{t=1}^{n} u_t \left[\boldsymbol{\gamma}' \boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1} \right] \\
&- \sum_{t=1}^{n} \left[\boldsymbol{\gamma}' \boldsymbol{\Sigma}^{-1} \otimes \left(\frac{\partial \boldsymbol{\varepsilon}_t'}{\partial \boldsymbol{\gamma}} + \boldsymbol{I}_d \right) \boldsymbol{\Sigma}^{-1} \right] - \sum_{t=1}^{n} \left[(\boldsymbol{\varepsilon}_t + \boldsymbol{\gamma})' \boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1} \right].
\end{aligned}$$

Thus,

$$\frac{\partial^2 \ell_N}{\partial \boldsymbol{\gamma}\,\partial \mathrm{vech}(\boldsymbol{\Sigma})'} = \frac{\partial^2 \ell_N}{\partial \boldsymbol{\gamma}\,\partial \mathrm{vec}(\boldsymbol{\Sigma}_u)'} \boldsymbol{D}_d\,.$$

CHAPTER B

Fisher Information Matrix of VG Distribution

The Fisher information matrix measures the amount of information a random variable $\boldsymbol{Y}$ has about a parameter $\boldsymbol{\theta}$, and it can be used to obtain SEs of a parameter estimate $\hat{\boldsymbol{\theta}}$. However, it is typically more difficult to calculate as it requires taking expectations of the observed information matrix, and so numerical integration techniques needs to be employed to perform the calculation. For higher dimensional models, this computation is infeasible. In this appendix, we provide formulas that can accurately and efficiently compute the Fisher information matrix of the VG distribution by algebraically integrating out the first and second term of (2.20), and reduce the dimensions of the integral of the third term down to one which can be numerically integrated much more efficiently.

We first present some preliminary results necessary for the calculation of the Fisher information matrix in the next section.

B1 Preliminary results

To simplify the calculation of the first and second term of the Fisher information matrix, we introduce the expectation result.

Lemma B1.1. *Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be d_x-dimensional and d_y-dimensional random vectors respectively such that*

$$\mathbb{E}_{\boldsymbol{X},\boldsymbol{Y}}\big|g(\boldsymbol{X},\boldsymbol{Y})\big| < \infty$$

for some scalar function g, then

$$\mathbb{E}_{\boldsymbol{X},\boldsymbol{Y}}\big[g(\boldsymbol{X},\boldsymbol{Y})\big] = \mathbb{E}_{\boldsymbol{Y}}\mathbb{E}_{\boldsymbol{X}|\boldsymbol{Y}}\big[g(\boldsymbol{X},\boldsymbol{Y})\big] = \mathbb{E}_{\boldsymbol{X}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{X}}\big[g(\boldsymbol{X},\boldsymbol{Y})\big]$$

where we simplify the notation for conditional expectation $\mathbb{E}_{\boldsymbol{X}|\boldsymbol{Y}}\big[h(\boldsymbol{X})\big] = \mathbb{E}_{\boldsymbol{X}|\boldsymbol{Y}}\big[h(\boldsymbol{X})|\boldsymbol{Y}\big]$ *for some scalar function h.*

Proof. By definition of the expectation and using Bayes rule, we get that

$$\begin{aligned}\mathbb{E}_{\boldsymbol{X},\boldsymbol{Y}}\big[g(\boldsymbol{X},\boldsymbol{Y})\big] &= \int_{\mathbb{R}^{d_y}}\int_{\mathbb{R}^{d_x}} g(\boldsymbol{x},\boldsymbol{y}) f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})\, d\boldsymbol{x}\, d\boldsymbol{y} \\ &= \int_{\mathbb{R}^{d_y}}\int_{\mathbb{R}^{d_x}} g(\boldsymbol{x},\boldsymbol{y}) f_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}|\boldsymbol{y}) f_{\boldsymbol{Y}}(\boldsymbol{y})\, d\boldsymbol{x}\, d\boldsymbol{y} \\ &= \mathbb{E}_{\boldsymbol{Y}}\mathbb{E}_{\boldsymbol{X}|\boldsymbol{Y}}\big[g(\boldsymbol{X},\boldsymbol{Y})\big].\end{aligned}$$

Swapping the order of integration using Fubini's theorem, similarly we get that

$$\mathbb{E}_{\boldsymbol{X},\boldsymbol{Y}}\big[g(\boldsymbol{X},\boldsymbol{Y})\big] = \mathbb{E}_{\boldsymbol{X}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{X}}\big[g(\boldsymbol{X},\boldsymbol{Y})\big].$$

□

Theorem B1.2. *Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be matrices of same dimension, and $\boldsymbol{C}$ and $\boldsymbol{D}$ be matrices of same dimension. Then*

$$(\boldsymbol{A}\circ\boldsymbol{B})\otimes(\boldsymbol{C}\circ\boldsymbol{D}) = (\boldsymbol{A}\otimes\boldsymbol{C})\circ(\boldsymbol{B}\otimes\boldsymbol{D})\,. \tag{B.1}$$

Proof. Consider the $(i,j)^{\text{th}}$ entry of $(\boldsymbol{A}\circ\boldsymbol{C})$, and $(k,l)^{\text{th}}$ entry of $(\boldsymbol{C}\circ\boldsymbol{D})$. Then from the left hand side of (B.1),

$$\begin{aligned}&(\boldsymbol{A}\circ\boldsymbol{B})_{ij}\otimes(\boldsymbol{C}\circ\boldsymbol{D})_{kl} \\ &= (\boldsymbol{A}_{ij}\boldsymbol{B}_{ij})\otimes(\boldsymbol{C}_{kl}\boldsymbol{D}_{kl}) \\ &= \boldsymbol{A}_{ij}\boldsymbol{B}_{ij}\boldsymbol{C}_{kl}\boldsymbol{D}_{kl}\end{aligned}$$

where the last equality holds since it only involve scalars, rearranging the terms and adding the Hadamard and Kronecker product gives us

$$= (\boldsymbol{A}_{ij}\otimes\boldsymbol{C}_{kl})\circ(\boldsymbol{B}_{ij}\otimes\boldsymbol{D}_{kl})\,.$$

Since the equation holds for arbitrary i,j,k,l, we proved the result. Note that the dimensions of the result is consistent since the position of the terms with respect to the Kronecker product is kept consistent. □

Corollary B1.3. *Let* $\boldsymbol{a}$ *and* $\boldsymbol{b}$ *be vectors of same length, and* $\boldsymbol{c}$ *and* $\boldsymbol{d}$ *be vectors of same length. Then*

$$(\boldsymbol{a} \circ \boldsymbol{b})(\boldsymbol{c}' \circ \boldsymbol{d}') = (\boldsymbol{a}\boldsymbol{c}') \circ (\boldsymbol{b}\boldsymbol{d}') \,. \tag{B.2}$$

The following theorem facilitates the matrix representation of the first order derivatives in Section B2 especially for the derivative of vec$\boldsymbol{\Sigma}_u$ in equation (B.10).

Theorem B1.4. *Suppose that* $\boldsymbol{A}' = \begin{pmatrix} \boldsymbol{a}_1 & \cdots & \boldsymbol{a}_n \end{pmatrix}$ *and* $\boldsymbol{B}' = \begin{pmatrix} \boldsymbol{b}_1 & \cdots & \boldsymbol{b}_n \end{pmatrix}$ *are* $d \times n$ *matrices where* $\boldsymbol{a}_i$ *and* $\boldsymbol{b}_i$ *are both* d*-dimensional vectors for* $i = 1, \dots, n$*. Additionally, let* $\boldsymbol{v}_c = (c_1, \dots, c_n)$*. Then*

$$\operatorname{vec}\left(\sum_{i=1}^{n} c_i \boldsymbol{a}_i \boldsymbol{b}_i' \right) = [(\mathbb{1}_d \otimes \boldsymbol{A}') \circ (\boldsymbol{B}' \otimes \mathbb{1}_d)] \boldsymbol{v}_c \,. \tag{B.3}$$

Proof. By definition of matrix multiplication, we can represent the right hand side of equation (B.3) as

$$\begin{aligned} &= \sum_{i=1}^{n} c_i \big[(\mathbb{1}_d \otimes \boldsymbol{A}') \circ (\boldsymbol{B}' \otimes \mathbb{1}_d) \big]_{\cdot i} \\ &= \sum_{i=1}^{n} c_i \big[(\mathbb{1}_d \otimes \boldsymbol{A}')_{\cdot i} \circ (\boldsymbol{B}' \otimes \mathbb{1}_d)_{\cdot i} \big] \\ &= \sum_{i=1}^{n} c_i \big[(\mathbb{1}_d \otimes \boldsymbol{a}_i) \circ (\boldsymbol{b}_i \otimes \mathbb{1}_d) \big] \end{aligned}$$

where $\boldsymbol{X}_{\cdot i}$ represents the i^{th} column of a matrix $\boldsymbol{X}$. Using Theorem B1.2 gives us

$$\begin{aligned} &= \sum_{i=1}^{n} c_i \big[(\mathbb{1}_d \circ \boldsymbol{b}_i) \otimes (\boldsymbol{a}_i \circ \mathbb{1}_d) \big] \\ &= \sum_{i=1}^{n} c_i (\boldsymbol{b}_i \otimes \boldsymbol{a}_i) \,. \end{aligned}$$

Using the vectorisation property $\text{vec}(\boldsymbol{ABC}) = \big(\boldsymbol{C}^\top \otimes \boldsymbol{A}\big)\text{vec}(\boldsymbol{B})$,

$$\begin{aligned}
&= \sum_{i=1}^{n} c_i(\boldsymbol{b}_i \otimes \boldsymbol{a}_i)\text{vec}(1) \\
&= \sum_{i=1}^{n} c_i \text{vec}(\boldsymbol{a}_i \boldsymbol{b}_i') \\
&= \text{vec}\left(\sum_{i=1}^{n} c_i \boldsymbol{a}_i \boldsymbol{b}_i'\right).
\end{aligned}$$

□

Theorem B1.5 (Isserlis' theorem for odd moments)**.** *Suppose that* $(X_1, \dots, X_{2d})$ *is a zero mean multivariate normal random vector where* d *is some positive integer, then*

$$\mathbb{E}[X_1 X_2 \dots X_{2d-1}] = \boldsymbol{0}\,.$$

Proof. See Isserlis [55]. □

Theorem B1.6. *Suppose that* $\boldsymbol{X} \sim \mathcal{N}_d(\boldsymbol{0}, \boldsymbol{\Sigma})$*, then*

$$\mathbb{E}[\boldsymbol{X}\boldsymbol{X}' \otimes \boldsymbol{X}\boldsymbol{X}'] = \boldsymbol{K}^{(d,d)}(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma}) + \text{vec}(\boldsymbol{\Sigma})\text{vec}(\boldsymbol{\Sigma})' + (\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma}). \tag{B.4}$$

Proof. See Magnus and Neudecker [72, Theorem4.1(i)]. □

B2 Matrix representation of first order derivatives

Using the first order derivatives in Section A8.1 to calculate the observed information matrix in (2.20) directly is tedious as it requires taking expectation of the product of two summations which is difficult to implement. To simplify the computation, we first introduce the matrix representation of the first order derivatives.

For this appendix, let $\boldsymbol{Y}$ be a $n \times d$ data matrix such that

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{y}_1' \\ \vdots \\ \boldsymbol{y}_n' \end{pmatrix}. \tag{B.5}$$

Consider the derivative of the complete data log-likelihood with respect to $\boldsymbol{\mu}$.

$$\begin{aligned}\frac{\partial \ell_N}{\partial \boldsymbol{\mu}} &= \boldsymbol{\Sigma}^{-1}\sum_{i=1}^{n}\frac{1}{u_i}(\boldsymbol{y}_i - \boldsymbol{\mu} - u_i\boldsymbol{\gamma}) \\ &= \sum_{i=1}^{n}\frac{1}{u_i}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}) - \sum_{i=1}^{n}\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}. \end{aligned} \tag{B.6}$$

By defining the vector $\boldsymbol{v}_{1/u} = (\frac{1}{u_1}, \ldots, \frac{1}{u_n})$, and representing the sum using matrices, the derivative of the log-likelihood in (B.6) gives us the following representations:

$$\begin{aligned}\frac{\partial \ell_N}{\partial \boldsymbol{\mu}} &= \boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \mathbb{1}_n\boldsymbol{\mu}')'\boldsymbol{v}_{1/u} + (-n\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}) \\ &= \boldsymbol{C}^{\mu}_{1/u}\boldsymbol{v}_{1/u} + \boldsymbol{c}^{\mu} \end{aligned} \tag{B.7}$$

where $\boldsymbol{C}^{\mu}_{1/u} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \mathbb{1}_n\boldsymbol{\mu}')'$, and $\boldsymbol{c}^{\mu} = -n\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}$. Representing the sum using matrices to the other derivatives in Section (A8.1) gives us

$$\frac{\partial \ell_N}{\partial \boldsymbol{\gamma}} = \boldsymbol{C}^{\gamma}_{u}\boldsymbol{v}_u + \boldsymbol{c}^{\gamma} \tag{B.8}$$

where $\boldsymbol{C}^{\gamma}_{u} = -\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}\mathbb{1}_n'$ and $\boldsymbol{c}^{\gamma} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \mathbb{1}_n\boldsymbol{\mu}')'\mathbb{1}_n$, and

$$\frac{\partial \ell_N}{\partial \mathrm{vech}\boldsymbol{\Sigma}} = \boldsymbol{D}_d^{\top}\frac{\partial \ell_N}{\partial \mathrm{vec}\boldsymbol{\Sigma}_u} \tag{B.9}$$

such that

$$\frac{\partial \ell_N}{\partial \mathrm{vec}\boldsymbol{\Sigma}_u} = \boldsymbol{C}^{\mathrm{vec}\Sigma_u}_{1/u}\boldsymbol{v}_{1/u} + \boldsymbol{C}^{\mathrm{vec}\Sigma_u}_{u}\boldsymbol{v}_u + \boldsymbol{c}^{\mathrm{vec}\Sigma_u} \tag{B.10}$$

and

$$\begin{aligned}\boldsymbol{C}^{\mathrm{vec}\Sigma_u}_{1/u} &= \frac{1}{2}\big(\mathbb{1}_d \otimes \boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \mathbb{1}_n\boldsymbol{\mu}')'\big) \circ \big(\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \mathbb{1}_n\boldsymbol{\mu}')' \otimes \mathbb{1}_d\big)\,, \\ \boldsymbol{C}^{\mathrm{vec}\Sigma_u}_{u} &= \frac{1}{2}\mathrm{vec}\big(\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\big)\mathbb{1}_n'\,, \text{ and} \\ \boldsymbol{c}^{\mathrm{vec}\Sigma_u} &= -\frac{1}{2}\mathrm{vec}\big[n\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1}\big((\boldsymbol{Y} - \mathbb{1}_n\boldsymbol{\mu}')'\mathbb{1}_n\boldsymbol{\gamma}' + \boldsymbol{\gamma}\mathbb{1}_n'(\boldsymbol{Y} - \mathbb{1}_n\boldsymbol{\mu}')\big)\boldsymbol{\Sigma}^{-1}\big] \end{aligned}$$

where $C^{\mathrm{vec}\Sigma_u}_{1/u}$ follows from Theorem B1.2. Moreover,

$$\frac{\partial \ell_G}{\partial \nu} = \boldsymbol{C}^{\nu}_{u}\boldsymbol{v}_u + \boldsymbol{C}^{\nu}_{\log u}\boldsymbol{v}_{\log u} + \boldsymbol{c}^{\nu} \tag{B.11}$$

where $\boldsymbol{C}^{\nu}_{u} = -\mathbb{1}_n'$, $\boldsymbol{C}^{\nu}_{\log u} = \mathbb{1}_n'$, and $\boldsymbol{c}^{\nu} = n(1 + \log\nu - \psi(\nu))$.

By combining these matrices together, the derivatives of the complete data log-likelihood can be written in the form of

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{u}) = \boldsymbol{C}^{\theta}_{1/u} \boldsymbol{v}_{1/u} + \boldsymbol{C}^{\theta}_{u} \boldsymbol{v}_{u} + \boldsymbol{C}^{\theta}_{\log u} \boldsymbol{v}_{\log u} + \boldsymbol{c}^{\theta} \tag{B.12}$$

where $\boldsymbol{C}^{\theta}_{g(u)} = \begin{pmatrix} \boldsymbol{C}^{\mu}_{g(u)} \\ \boldsymbol{C}^{\gamma}_{g(u)} \\ \boldsymbol{C}^{\mathrm{vec}\Sigma_u}_{g(u)} \\ \boldsymbol{C}^{\nu}_{g(u)} \end{pmatrix}$, $\boldsymbol{v}_{g(u)} = \begin{pmatrix} g(u_1) \\ \vdots \\ g(u_n) \end{pmatrix}$, and $\boldsymbol{C}^{\mathrm{vech}\Sigma}_{g(u)} = \boldsymbol{D}_d^{\top} \boldsymbol{C}^{\mathrm{vec}\Sigma_u}_{g(u)}$ for some scalar function g.

B3 Simplification of missing information matrix calculation

The missing information matrix (1.12) in our context becomes

$$\mathcal{I}_{\mathrm{mis}}(\boldsymbol{\theta}; \boldsymbol{y}_{\mathrm{obs}}) = \mathrm{cov}_{\boldsymbol{u}|\boldsymbol{Y}} \left(\frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \boldsymbol{Y}, \boldsymbol{u}) \right)$$

where the covariance is taken with respect to $\boldsymbol{u}$ given data matrix $\boldsymbol{Y}$. Using the matrix representation of the first derivative in (B.12), this gives us

$$\begin{aligned}
&= \mathrm{cov}_{\boldsymbol{u}|\boldsymbol{Y}} \left(\boldsymbol{C}^{\theta}_{1/u} \boldsymbol{v}_{1/u} + \boldsymbol{C}^{\theta}_{u} \boldsymbol{v}_{u} + \boldsymbol{C}^{\theta}_{\log u} \boldsymbol{v}_{\log u} + \boldsymbol{c}^{\theta} \right) \\
&= \boldsymbol{C}_{1/u} \mathrm{cov}_{\boldsymbol{u}|\boldsymbol{Y}}(\boldsymbol{v}_{1/u}) \boldsymbol{C}_{1/u}^{\top} + \boldsymbol{C}_{u} \mathrm{cov}_{\boldsymbol{u}|\boldsymbol{Y}}(\boldsymbol{v}_{u}) \boldsymbol{C}_{u}^{\top} + \boldsymbol{C}_{\log u} \mathrm{cov}_{\boldsymbol{u}|\boldsymbol{Y}}(\boldsymbol{v}_{\log u}) \boldsymbol{C}_{\log u}^{\top} \\
&\quad + \boldsymbol{C}_{1/u} \mathrm{cov}_{\boldsymbol{u}|\boldsymbol{Y}}(\boldsymbol{v}_{1/u}, \boldsymbol{v}_{u}) \boldsymbol{C}_{u}^{\top} + \boldsymbol{C}_{u} \mathrm{cov}_{\boldsymbol{u}|\boldsymbol{Y}}(\boldsymbol{v}_{u}, \boldsymbol{v}_{1/u}) \boldsymbol{C}_{1/u}^{\top} \\
&\quad + \boldsymbol{C}_{1/u} \mathrm{cov}_{\boldsymbol{u}|\boldsymbol{Y}}(\boldsymbol{v}_{1/u}, \boldsymbol{v}_{\log u}) \boldsymbol{C}_{\log u}^{\top} + \boldsymbol{C}_{\log u} \mathrm{cov}_{\boldsymbol{u}|\boldsymbol{Y}}(\boldsymbol{v}_{\log u}, \boldsymbol{v}_{1/u}) \boldsymbol{C}_{1/u}^{\top} \\
&\quad + \boldsymbol{C}_{u} \mathrm{cov}_{\boldsymbol{u}|\boldsymbol{Y}}(\boldsymbol{v}_{u}, \boldsymbol{v}_{\log u}) \boldsymbol{C}_{\log u}^{\top} + \boldsymbol{C}_{\log u} \mathrm{cov}_{\boldsymbol{u}|\boldsymbol{Y}}(\boldsymbol{v}_{\log u}, \boldsymbol{v}_{u}) \boldsymbol{C}_{u}^{\top}
\end{aligned} \tag{B.13}$$

where we use the bilinearity property of covariance function. Note that the constant vector $c^{\boldsymbol{\theta}}$ can simply be ignored in the calculation of the observed information matrix as it does not depend on $\boldsymbol{u}$.

Using the mutual independence of the u's, we get that

$$\mathrm{cov}_{\boldsymbol{u}|\boldsymbol{Y}}(\boldsymbol{v}_u) = \begin{pmatrix} \mathrm{var}_{\boldsymbol{u}|\boldsymbol{Y}}(u_1) & & \boldsymbol{0} \\ & \ddots & \\ \boldsymbol{0} & & \mathrm{var}_{\boldsymbol{u}|\boldsymbol{Y}}(u_n) \end{pmatrix}.$$

This mutual independence property also applies to those other covariance matrices in equation (B.13). This simplification allows for a more efficient calculation of the observed information matrix.

Thus applying the simplification of the covariance matrices to the missing information matrix in (B.13) gives us

$$\begin{aligned}\mathcal{I}_{\text{mis}}(\boldsymbol{\theta};\boldsymbol{y}_{\text{obs}}) &= \boldsymbol{C}_{1/u}\text{cov}_{\boldsymbol{u}|\boldsymbol{Y}}\big(\boldsymbol{v}_{1/u}\big)\boldsymbol{C}_{1/u}^{\top} + ... \\ &= \sum_{i=1}^{n}\text{var}_{\boldsymbol{u}|\boldsymbol{Y}}\left(\frac{1}{u_i}\right)\boldsymbol{C}_{1/u,i}\boldsymbol{C}_{1/u,i}^{\top} + ...\end{aligned}$$

where $\boldsymbol{C}_{1/u,i}$ represents the i^{th} column of $\boldsymbol{C}_{1/u}$. Using (1.12) for the left hand side and the variance formula for the right hand side, then equating the terms gives us

$$\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\big[\ell'(\boldsymbol{\theta};\boldsymbol{Y},\boldsymbol{u})\ell'(\boldsymbol{\theta};\boldsymbol{Y},\boldsymbol{u})^{\top}\big] = \sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\left[\frac{1}{u_i^2}\right]\boldsymbol{C}_{1/u,i}\boldsymbol{C}_{1/u,i}^{\top} + ... \tag{B.14}$$

and

$$\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\big[\ell'(\boldsymbol{\theta};\boldsymbol{Y},\boldsymbol{u})\big]\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\big[\ell'(\boldsymbol{\theta};\boldsymbol{Y},\boldsymbol{u})\big]^{\top} = \sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\left[\frac{1}{u_i}\right]^2\boldsymbol{C}_{1/u,i}\boldsymbol{C}_{1/u,i}^{\top} + ...\tag{B.15}$$

This representation is relevant for the calculation of the second and third term of the Fisher Information matrix later in Section B6.

B4 Conditional normal moment results

Suppose that $\boldsymbol{y}|u \sim \mathcal{N}(\boldsymbol{\mu}+u\boldsymbol{\gamma}, u\boldsymbol{\Sigma})$ and $\tilde{\boldsymbol{y}} = \boldsymbol{y}-\boldsymbol{\mu}-u\boldsymbol{\gamma}$ for some given mixing variable u. Before the calculation of the first and second term of the Fisher information matrix, we need to first find the expressions for the following conditional moments:

(i) $\mathbb{E}_{\boldsymbol{y}|u}[\boldsymbol{y}]$,
(ii) $\mathbb{E}_{\boldsymbol{y}|u}\big[(\boldsymbol{y}-\boldsymbol{\mu})(\boldsymbol{y}-\boldsymbol{\mu})'\big]$,
(iii) $\mathbb{E}_{\boldsymbol{y}|u}\big[(\boldsymbol{y}-\boldsymbol{\mu})\otimes(\boldsymbol{y}-\boldsymbol{\mu})(\boldsymbol{y}-\boldsymbol{\mu})'\big]$
$= \mathbb{E}_{\boldsymbol{y}|u}\big[\text{vec}\big((\boldsymbol{y}-\boldsymbol{\mu})(\boldsymbol{y}-\boldsymbol{\mu})'\big)(\boldsymbol{y}-\boldsymbol{\mu})'\big]$,
(iv) $\mathbb{E}_{\boldsymbol{y}|u}\big[(\boldsymbol{y}-\boldsymbol{\mu})(\boldsymbol{y}-\boldsymbol{\mu})'\otimes(\boldsymbol{y}-\boldsymbol{\mu})(\boldsymbol{y}-\boldsymbol{\mu})'\big]$
$= \mathbb{E}_{\boldsymbol{y}|u}\Big[\text{vec}\big((\boldsymbol{y}-\boldsymbol{\mu})(\boldsymbol{y}-\boldsymbol{\mu})'\big)\text{vec}\big((\boldsymbol{y}-\boldsymbol{\mu})(\boldsymbol{y}-\boldsymbol{\mu})'\big)'\Big]$.

Without loss of generality, suppose $\boldsymbol{\mu} = \mathbf{0}$ so that $\boldsymbol{y}|u \sim \mathcal{N}(u\boldsymbol{\gamma}, u\boldsymbol{\Sigma})$ and $\tilde{\boldsymbol{y}} = \boldsymbol{y} - u\boldsymbol{\gamma}$, and thus $\tilde{\boldsymbol{y}}|u \sim \mathcal{N}(\mathbf{0}, u\boldsymbol{\Sigma})$.

First moment: We immediately get that

$$\mathbb{E}_{\boldsymbol{y}|u}[\boldsymbol{y}] = u\boldsymbol{\gamma}\,.$$

Second moment: Using $\boldsymbol{y} = \tilde{\boldsymbol{y}} - u\boldsymbol{\gamma}$ and expanding gives us

$$\begin{aligned}\mathbb{E}_{\boldsymbol{y}|u}[\boldsymbol{y}\boldsymbol{y}'] &= \mathbb{E}_{\boldsymbol{y}|u}\big[(\tilde{\boldsymbol{y}} + u\gamma)(\tilde{\boldsymbol{y}} + u\gamma)'\big] \\ &= \underbrace{\mathbb{E}_{\boldsymbol{y}|u}[\tilde{\boldsymbol{y}}\tilde{\boldsymbol{y}}']}_{=u\boldsymbol{\Sigma}} + u^2\boldsymbol{\gamma}\boldsymbol{\gamma}' + u\,\underbrace{\mathbb{E}_{\boldsymbol{y}|u}[\tilde{\boldsymbol{y}}]}_{=\mathbf{0}}\boldsymbol{\gamma}' + u\boldsymbol{\gamma}\,\underbrace{\mathbb{E}_{\boldsymbol{y}|u}[\tilde{\boldsymbol{y}}']}_{=\mathbf{0}'} \\ &= u\boldsymbol{\Sigma} + u^2\boldsymbol{\gamma}\boldsymbol{\gamma}'.\end{aligned}$$

Third moment:

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{y}|u}\big[\boldsymbol{y} \otimes \boldsymbol{y}\boldsymbol{y}'\big] \\ &= \mathbb{E}_{\boldsymbol{y}|u}\big[\big(\tilde{\boldsymbol{y}} + u\boldsymbol{\gamma}\big) \otimes (\tilde{\boldsymbol{y}} + u\boldsymbol{\gamma})(\tilde{\boldsymbol{y}} + u\boldsymbol{\gamma})'\big].\end{aligned}$$

Since the odd moments are zero by Isserlis' theorem (B1.5) we can disregard them in our calculation. Expanding gives us

$$= \mathbb{E}_{\boldsymbol{y}|u}\big[u\boldsymbol{\gamma} \otimes \tilde{\boldsymbol{y}}\tilde{\boldsymbol{y}}' + u\underbrace{\tilde{\boldsymbol{y}} \otimes \boldsymbol{\gamma}\tilde{\boldsymbol{y}}'}_{\tilde{\boldsymbol{y}}\tilde{\boldsymbol{y}}'\otimes\boldsymbol{\gamma}} + u\tilde{\boldsymbol{y}} \otimes \tilde{\boldsymbol{y}}\boldsymbol{\gamma}' + u^3\boldsymbol{\gamma} \otimes \boldsymbol{\gamma}\boldsymbol{\gamma}'\big]\,.$$

Evaluating the expectations gives us

$$= u^2\big(\boldsymbol{\gamma} \otimes \boldsymbol{\Sigma} + \boldsymbol{\Sigma} \otimes \boldsymbol{\gamma} + \mathrm{vec}(\boldsymbol{\Sigma})\boldsymbol{\gamma}'\big) + u^3\boldsymbol{\gamma} \otimes \boldsymbol{\gamma}\boldsymbol{\gamma}'.$$

Forth moment:

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{y}|u}[\boldsymbol{y}\boldsymbol{y}'\otimes\boldsymbol{y}\boldsymbol{y}'] \\
&= \mathbb{E}_{\boldsymbol{y}|u}\big[(\tilde{\boldsymbol{y}}+u\boldsymbol{\gamma})(\tilde{\boldsymbol{y}}+u\boldsymbol{\gamma})'\otimes(\tilde{\boldsymbol{y}}+u\boldsymbol{\gamma})(\tilde{\boldsymbol{y}}+u\boldsymbol{\gamma})'\big] \\
&= \mathbb{E}_{\boldsymbol{y}|u}\Big[\tilde{\boldsymbol{y}}\tilde{\boldsymbol{y}}'\otimes\tilde{\boldsymbol{y}}\tilde{\boldsymbol{y}}' + u^4\boldsymbol{\gamma}\boldsymbol{\gamma}'\otimes\boldsymbol{\gamma}\boldsymbol{\gamma}' + u^2\underbrace{\tilde{\boldsymbol{y}}\boldsymbol{\gamma}'\otimes\tilde{\boldsymbol{y}}\boldsymbol{\gamma}'}_{(\tilde{\boldsymbol{y}}\otimes\tilde{\boldsymbol{y}})(\boldsymbol{\gamma}'\otimes\boldsymbol{\gamma}')} + u^2\underbrace{\boldsymbol{\gamma}\tilde{\boldsymbol{y}}'\otimes\boldsymbol{\gamma}\tilde{\boldsymbol{y}}'}_{(\boldsymbol{\gamma}\otimes\boldsymbol{\gamma})(\tilde{\boldsymbol{y}}\otimes\tilde{\boldsymbol{y}})'} \\
&\quad + u\tilde{\boldsymbol{y}}\tilde{\boldsymbol{y}}'\otimes\boldsymbol{\gamma}\boldsymbol{\gamma}' + u^2\underbrace{\boldsymbol{\gamma}\boldsymbol{\gamma}'\otimes\tilde{\boldsymbol{y}}\tilde{\boldsymbol{y}}'}_{\boldsymbol{K}^{(d,d)}(\tilde{\boldsymbol{y}}\tilde{\boldsymbol{y}}')\otimes(\boldsymbol{\gamma}\boldsymbol{\gamma}')\boldsymbol{K}^{(d,d)}} + u^2\underbrace{\tilde{\boldsymbol{y}}\boldsymbol{\gamma}'\otimes\boldsymbol{\gamma}\tilde{\boldsymbol{y}}'}_{\tilde{\boldsymbol{y}}\tilde{\boldsymbol{y}}'\otimes(\boldsymbol{\gamma}\boldsymbol{\gamma}')\boldsymbol{K}^{(d,d)}} + u^2\underbrace{\boldsymbol{\gamma}\tilde{\boldsymbol{y}}'\otimes\tilde{\boldsymbol{y}}\boldsymbol{\gamma}'}_{\boldsymbol{K}^{(d,d)}(\tilde{\boldsymbol{y}}\tilde{\boldsymbol{y}}')\otimes\boldsymbol{\gamma}\boldsymbol{\gamma}'}\Big]
\end{aligned}
$$

where we used the property $(\boldsymbol{A}\otimes\boldsymbol{B}) = \boldsymbol{K}^{(\mathrm{row}\boldsymbol{A},\mathrm{row}\boldsymbol{B})}(\boldsymbol{B}\otimes\boldsymbol{A})\boldsymbol{K}^{(\mathrm{col}\boldsymbol{B},\mathrm{col}\boldsymbol{A})}$. Equating the expectations gives us

$$
\begin{aligned}
&= u^2\boldsymbol{K}^{(d,d)}(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma}) + u^2\mathrm{vec}(\boldsymbol{\Sigma})\mathrm{vec}(\boldsymbol{\Sigma})' + u^2(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma}) + u^4\boldsymbol{\gamma}\boldsymbol{\gamma}'\otimes\boldsymbol{\gamma}\boldsymbol{\gamma}' \\
&\quad + u^3\mathrm{vec}(\boldsymbol{\Sigma})(\boldsymbol{\gamma}'\otimes\boldsymbol{\gamma}') + u^3(\boldsymbol{\gamma}\otimes\boldsymbol{\gamma})\mathrm{vec}(\boldsymbol{\Sigma})' \\
&\quad + u^3\boldsymbol{\Sigma}\otimes\boldsymbol{\gamma}\boldsymbol{\gamma}' + u^3\boldsymbol{K}^{(d,d)}(\boldsymbol{\Sigma}\otimes\boldsymbol{\gamma}\boldsymbol{\gamma}')\boldsymbol{K}^{(d,d)} \\
&\quad + u^3\boldsymbol{\Sigma}\otimes\boldsymbol{\gamma}\boldsymbol{\gamma}'\boldsymbol{K}^{(d,d)} + u^3\boldsymbol{K}^{(d,d)}\boldsymbol{\Sigma}\otimes\boldsymbol{\gamma}\boldsymbol{\gamma}' \\
&= u^2\big(\boldsymbol{I}_d + \boldsymbol{K}^{(d,d)}\big)(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma}) + u^2\big[\mathrm{vec}\boldsymbol{\Sigma} + u\mathrm{vec}(\boldsymbol{\gamma}\boldsymbol{\gamma}')\big]\big[\mathrm{vec}\boldsymbol{\Sigma} + u\mathrm{vec}(\boldsymbol{\gamma}\boldsymbol{\gamma}')\big]' \\
&\quad + u^3\big(\boldsymbol{I}_d + \boldsymbol{K}^{(d,d)}\big)(\boldsymbol{\Sigma}\otimes\boldsymbol{\gamma}\boldsymbol{\gamma}')\big(\boldsymbol{I}_d + \boldsymbol{K}^{(d,d)}\big)\,.
\end{aligned}
$$

B5 Multidimensional integration

Theorem B5.1. *Suppose h is a continuous function, $\boldsymbol{y}$ is a vector in $\mathbb{R}^d$ and $r > 0$, then*

$$
\int_{\|\boldsymbol{y}\|=r} h(\boldsymbol{a}'\boldsymbol{y}, \|\boldsymbol{y}\|^2)\,dS = \frac{2r^{d-1}\pi^{(d-1)/2}}{\Gamma\big(\frac{d-1}{2}\big)}\int_0^\pi h\big(\|\boldsymbol{a}\|r\cos\phi, r^2\big)\sin^{d-2}\phi\;d\phi \tag{B.16}
$$

where dS represents the spherical differential, and $\boldsymbol{a}$ is a constant vector in $\mathbb{R}^d$.

Proof. See Blumenson [9]. □

This theorem transform the spherical integral on $\mathbb{R}^d$ to a one-dimensional integral which is much more feasible to compute. For our case, applying the transformation $\boldsymbol{z} = \boldsymbol{A}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})$ where $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}'$ to the density function of VG distribution in (1.36) gives

the following representation

$$f_Z(\boldsymbol{z}) = g(\|\boldsymbol{z}\|)\exp(\boldsymbol{\gamma}_z'\boldsymbol{z})$$

where

$$g(\|\boldsymbol{z}\|) = \frac{2\nu^\nu}{(2\pi)^{\frac{d}{2}}\Gamma(\nu)} \frac{K_{\nu-\frac{d}{2}}\left(\sqrt{(2\nu+\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma})\|\boldsymbol{z}\|^2}\right)}{\left(\sqrt{\|\boldsymbol{z}\|^2/(2\nu+\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma})}\right)^{\frac{d}{2}-\nu}}, \tag{B.17}$$

$\|\boldsymbol{z}\|^2 = (\boldsymbol{y}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})$, and $\boldsymbol{\gamma}_z = \boldsymbol{A}^{-1}\boldsymbol{\gamma}$. We apply Theorem B5.1 to calculate moments of $f_{VG}(\boldsymbol{y})$ which is used later to evaluate the third term of the Fisher information matrix.

B5.1 Higher order spherical moments

One problem is that calculating higher order moments of the VG distribution such as

$$\mathbb{E}_{\boldsymbol{y}}[\boldsymbol{y}] = \int_{\mathbb{R}^d} \boldsymbol{y} f_{VG}(\boldsymbol{y})\, d\boldsymbol{y} = |\boldsymbol{\Sigma}|^{-1/2} \int_0^\infty g(r) \int_{\|\boldsymbol{z}\|=r} \boldsymbol{z}\exp(\boldsymbol{\gamma}_z'\boldsymbol{z})\, dS\, dr$$

does not allow the inner integrand to have the representation as in Theorem B5.1. To get around this problem, we introduce the spherical MGF defined by

$$\mathscr{M}_r(\boldsymbol{s}) := \int_{\|\boldsymbol{z}\|=r} \exp(\boldsymbol{s}'\boldsymbol{z})\, dS$$

such that the first order derivative gives us

$$\mathscr{M}_r^{(1)}(\boldsymbol{s}) := \frac{\partial}{\partial \boldsymbol{s}}\mathscr{M}_r(\boldsymbol{s}) = \int_{\|\boldsymbol{z}\|=r} \boldsymbol{z}\exp(\boldsymbol{s}'\boldsymbol{z})\, dS.$$

Note that $\mathscr{M}_r(\boldsymbol{s})$ is well-defined since the integral is absolutely convergent for any $r > 0$.

Similarly, we can represent higher order spherical moments using higher order derivatives of the spherical MGF.

$$\mathscr{M}_r^{(2)}(\boldsymbol{s}) := \frac{\partial^2}{\partial \boldsymbol{s}\partial \boldsymbol{s}'}\mathscr{M}_r(\boldsymbol{s}) = \int_{\|\boldsymbol{z}\|=r} \boldsymbol{z}\boldsymbol{z}'\exp(\boldsymbol{s}'\boldsymbol{z})\, dS\,,$$

$$\mathscr{M}_r^{(3)}(\boldsymbol{s}) := \frac{\partial^3}{\partial \boldsymbol{s}\partial \boldsymbol{s}\partial \boldsymbol{s}'}\mathscr{M}_r(\boldsymbol{s}) = \int_{\|\boldsymbol{z}\|=r} \boldsymbol{z}\otimes\boldsymbol{z}\boldsymbol{z}'\exp(\boldsymbol{s}'\boldsymbol{z})\, dS\,,$$

$$\mathscr{M}_r^{(4)}(\boldsymbol{s}) := \frac{\partial^4}{\partial \boldsymbol{s}\partial \boldsymbol{s}'\partial \boldsymbol{s}\partial \boldsymbol{s}'}\mathscr{M}_r(\boldsymbol{s}) = \int_{\|\boldsymbol{z}\|=r} \boldsymbol{z}\boldsymbol{z}'\otimes\boldsymbol{z}\boldsymbol{z}'\exp(\boldsymbol{s}'\boldsymbol{z})\, dS\,.$$

The following theorem uses matrix derivative results in Appendix A to derive these spherical moments.

Theorem B5.2.

$$\begin{aligned}
\mathscr{M}_r(\boldsymbol{s}) &= (2\pi r)^{\frac{d}{2}} \|\boldsymbol{s}\|^{1-\frac{d}{2}} I_{\frac{d}{2}-1}(r\|\boldsymbol{s}\|), \\
\mathscr{M}_r^{(1)}(\boldsymbol{s}) &= (2\pi r)^{\frac{d}{2}} r\|\boldsymbol{s}\|^{-\frac{d}{2}} I_{\frac{d}{2}}(r\|\boldsymbol{s}\|)\boldsymbol{s}, \\
\mathscr{M}_r^{(2)}(\boldsymbol{s}) &= (2\pi r)^{\frac{d}{2}} r\|\boldsymbol{s}\|^{-\frac{d}{2}-1}\Big[I_{\frac{d}{2}}(r\|\boldsymbol{s}\|)\,\|\boldsymbol{s}\|\boldsymbol{I}_d + I_{\frac{d}{2}+1}(r\|\boldsymbol{s}\|)\,r\boldsymbol{s}\boldsymbol{s}'\Big], \\
\mathscr{M}_r^{(3)}(\boldsymbol{s}) &= (2\pi r)^{\frac{d}{2}} r^2\|\boldsymbol{s}\|^{-\frac{d}{2}-3}\Big[I_{\frac{d}{2}}(r\|\boldsymbol{s}\|)\boldsymbol{C}_{31} + I_{\frac{d}{2}+1}(r\|\boldsymbol{s}\|)\boldsymbol{C}_{32}\Big], \\
\mathscr{M}_r^{(4)}(\boldsymbol{s}) &= (2\pi r)^{\frac{d}{2}} r^2\|\boldsymbol{s}\|^{-\frac{d}{2}-5}\Big[I_{\frac{d}{2}}(r\|\boldsymbol{s}\|)\boldsymbol{C}_{41} + I_{\frac{d}{2}+1}(r\|\boldsymbol{s}\|)\boldsymbol{C}_{42}\Big],
\end{aligned}$$

where $I_\lambda(\cdot)$ is a modified Bessel function of the first kind,

$$\begin{aligned}
\boldsymbol{C}_{31} &= r\|\boldsymbol{s}\|\boldsymbol{A}_3, \\
\boldsymbol{C}_{32} &= \|\boldsymbol{s}\|^2\boldsymbol{A}_1 - (d+2)\boldsymbol{A}_3, \\
\boldsymbol{C}_{41} &= r\|\boldsymbol{s}\|\big(\|\boldsymbol{s}\|^2(\boldsymbol{A}_2+\boldsymbol{B}_2) - (d+4)\boldsymbol{A}_4\big), \\
\boldsymbol{C}_{42} &= \big((d+4)(d+2) + r^2\|\boldsymbol{s}\|^2\big)\boldsymbol{A}_4 - (d+2)\|\boldsymbol{s}\|^2(\boldsymbol{A}_2+\boldsymbol{B}_2) + \|\boldsymbol{s}\|^4\boldsymbol{A}_0,
\end{aligned}$$

and

$$\begin{aligned}
\boldsymbol{A}_4 &= \boldsymbol{s}\boldsymbol{s}'\otimes\boldsymbol{s}\boldsymbol{s}', \\
\boldsymbol{A}_3 &= \boldsymbol{s}\otimes\boldsymbol{s}\boldsymbol{s}', \\
\boldsymbol{A}_2 &= \boldsymbol{I}_d\otimes\boldsymbol{s}\boldsymbol{s}' + \boldsymbol{s}\otimes\boldsymbol{I}_d\otimes\boldsymbol{s}' + (\boldsymbol{s}\otimes\boldsymbol{s})\mathrm{vec}(\boldsymbol{I}_d)', \\
\boldsymbol{B}_2 &= \boldsymbol{s}\boldsymbol{s}'\otimes\boldsymbol{I}_d + \boldsymbol{s}'\otimes\boldsymbol{I}_d\otimes\boldsymbol{s} + \mathrm{vec}(\boldsymbol{I}_d)(\boldsymbol{s}'\otimes\boldsymbol{s}'), \\
\boldsymbol{A}_1 &= \mathrm{vec}(\boldsymbol{I}_d)\boldsymbol{s}' + \boldsymbol{I}_d\otimes\boldsymbol{s} + \boldsymbol{s}\otimes\boldsymbol{I}_d, \\
\boldsymbol{A}_0 &= \boldsymbol{I}_d\otimes\boldsymbol{I}_d + \boldsymbol{K}^{(d,d)} + \mathrm{vec}(\boldsymbol{I}_d)\mathrm{vec}(\boldsymbol{I}_d)'.
\end{aligned}$$

Proof. Using the formula in Theorem B5.1 with $h(\boldsymbol{s}'\boldsymbol{z}, \|z\|^2) = \exp(\boldsymbol{s}'\boldsymbol{z})$, and the integral formula in Gradshteyn and Ryzhik [41, 3.915.4], this gives us

$$\begin{aligned}
&\int_{\|\boldsymbol{z}\|=r} \exp(\boldsymbol{s}'\boldsymbol{z})\,dS \\
&= \frac{2r^{d-1}\pi^{(d-1)/2}}{\Gamma\big(\frac{d-1}{2}\big)} \int_0^\pi \exp(\|\boldsymbol{s}\|r\cos\phi)\sin^{d-2}\phi\,d\phi \\
&= (2\pi r)^{\frac{d}{2}}\|\boldsymbol{s}\|^{1-\frac{d}{2}} I_{\frac{d}{2}-1}(r\|\boldsymbol{s}\|).
\end{aligned}$$

Let $f(\boldsymbol{s}) = \|\boldsymbol{s}\|^{1-\frac{d}{2}} I_{\frac{d}{2}-1}(r\|\boldsymbol{s}\|)$, then it is sufficient to consider derivatives of $f(\boldsymbol{s})$. For the first order derivative, applying the chain rule gives us

$$\begin{aligned}\frac{\partial f}{\partial \boldsymbol{s}} &= \frac{\partial \|\boldsymbol{s}\|}{\partial \boldsymbol{s}} \frac{df}{d\|\boldsymbol{s}\|} \\ &= \frac{\boldsymbol{s}}{\|\boldsymbol{s}\|} \times r\|\boldsymbol{s}\|^{1-\frac{d}{2}} I_{\frac{d}{2}-1}(r\|\boldsymbol{s}\|) \\ &= r\|\boldsymbol{s}\|^{-\frac{d}{2}} I_{\frac{d}{2}}(r\|\boldsymbol{s}\|)\, \boldsymbol{s}\end{aligned}$$

where equations (A.39) and

$$\begin{aligned}\frac{d}{dz} z^{-\alpha} I_\lambda(\beta z) &= z^{-\alpha-1}(\beta z I_{\lambda+1}(\beta z) + (\alpha+\lambda) I_\lambda(\beta z)) \\ &= \beta z^{-\lambda} I_{\lambda+1}(\beta z) \quad \text{if } \alpha = \lambda\end{aligned}$$

for $\alpha, \lambda, \beta \in \mathbb{R}$ are used for the second equality.

Note that for the second order derivative, we have that $\frac{\partial^2 f}{\partial \boldsymbol{s} \partial \boldsymbol{s}'} = \frac{\partial^2 f}{\partial \boldsymbol{s}' \partial \boldsymbol{s}}$. Applying the product rule from Corollary A4.4

$$\begin{aligned}\frac{\partial^2 f}{\partial \boldsymbol{s}' \partial \boldsymbol{s}} &= \frac{\partial}{\partial \boldsymbol{s}'}\left[r\|\boldsymbol{s}\|^{-\frac{d}{2}} I_{\frac{d}{2}}(r\|\boldsymbol{s}\|)\, \boldsymbol{s}\right] \\ &= r\left[\frac{\partial}{\partial \boldsymbol{s}'}\left(\|\boldsymbol{s}\|^{-\frac{d}{2}} I_{\frac{d}{2}}(r\|\boldsymbol{s}\|)\right) \otimes \boldsymbol{s} + \|\boldsymbol{s}\|^{-\frac{d}{2}} I_{\frac{d}{2}}(r\|\boldsymbol{s}\|) \frac{\partial \boldsymbol{s}}{\partial \boldsymbol{s}'}\right].\end{aligned}$$

For the derivative in the first term,

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{s}'} \|\boldsymbol{s}\|^{-\frac{d}{2}} I_{\frac{d}{2}}(r\|\boldsymbol{s}\|) &= \frac{\partial \|\boldsymbol{s}\|}{\partial \boldsymbol{s}'} \frac{d\|\boldsymbol{s}\|^{-\frac{d}{2}} I_{\frac{d}{2}}(r\|\boldsymbol{s}\|)}{d\|\boldsymbol{s}\|} \\ &= \frac{\boldsymbol{s}'}{\|\boldsymbol{s}\|} \times r\|\boldsymbol{s}\|^{-\frac{d}{2}} I_{\frac{d}{2}+1}(r\|\boldsymbol{s}\|)\end{aligned}$$

and the derivative in the second term $\frac{\partial \boldsymbol{s}}{\partial \boldsymbol{s}'} = \boldsymbol{I}_d$. Applying these results back into the second order derivative and factoring out $\|\boldsymbol{s}\|^{-\frac{d}{2}-1}$ gives us the result

$$\frac{\partial^2 f}{\partial \boldsymbol{s}' \partial \boldsymbol{s}} = r\|\boldsymbol{s}\|^{-\frac{d}{2}-1}\left[\|\boldsymbol{s}\| I_{\frac{d}{2}}(r\|\boldsymbol{s}\|)\, \boldsymbol{I}_d + r I_{\frac{d}{2}+1}(r\|\boldsymbol{s}\|)\, \boldsymbol{s}\boldsymbol{s}'\right].$$

For the third order derivative, using the Kronecker product rule in Theorem A4.3 gives us

$$\frac{\partial^3 f}{\partial \boldsymbol{s} \partial \boldsymbol{s} \partial \boldsymbol{s}'} = r\frac{\partial}{\partial \boldsymbol{s}}\left[\|\boldsymbol{s}\|^{-\frac{d}{2}} I_{\frac{d}{2}}(r\|\boldsymbol{s}\|)\boldsymbol{I}_d\right] + r^2 \frac{\partial}{\partial \boldsymbol{s}}\left[\|\boldsymbol{s}\|^{-\frac{d}{2}-1} I_{\frac{d}{2}+1}(r\|\boldsymbol{s}\|)\boldsymbol{s}\boldsymbol{s}'\right]. \tag{B.18}$$

For the derivative of the first term

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{s}}\Big[\|\boldsymbol{s}\|^{-\frac{d}{2}} I_{\frac{d}{2}}(r\|\boldsymbol{s}\|)\boldsymbol{I}_d\Big] &= \left[\frac{\partial\|\boldsymbol{s}\|}{\partial \boldsymbol{s}}\frac{d\|\boldsymbol{s}\|^{-\frac{d}{2}} I_{\frac{d}{2}}(r\|\boldsymbol{s}\|)}{d\|\boldsymbol{s}\|}\right]\otimes \boldsymbol{I}_d \\ &= \frac{\boldsymbol{s}}{\|\boldsymbol{s}\|}\times r\|\boldsymbol{s}\|^{-\frac{d}{2}} I_{\frac{d}{2}+1}(r\|\boldsymbol{s}\|)\otimes \boldsymbol{I}_d \\ &= r\|\boldsymbol{s}\|^{-\frac{d}{2}-1} I_{\frac{d}{2}+1}(r\|\boldsymbol{s}\|)\,\boldsymbol{s}\otimes \boldsymbol{I}_d.\end{aligned}$$

For the derivative of the second term,

$$\begin{aligned}&\frac{\partial}{\partial \boldsymbol{s}}\Big[\|\boldsymbol{s}\|^{-\frac{d}{2}-1} I_{\frac{d}{2}+1}(r\|\boldsymbol{s}\|)\boldsymbol{s}\boldsymbol{s}'\Big] \\ &= \frac{\partial\|\boldsymbol{s}\|}{\partial \boldsymbol{s}}\frac{d\|\boldsymbol{s}\|^{-\frac{d}{2}-1} I_{\frac{d}{2}+1}(r\|\boldsymbol{s}\|)}{d\|\boldsymbol{s}\|}\otimes \boldsymbol{s}\boldsymbol{s}' + \|\boldsymbol{s}\|^{-\frac{d}{2}-1} I_{\frac{d}{2}+1}(r\|\boldsymbol{s}\|)\frac{\partial \boldsymbol{s}\boldsymbol{s}'}{\partial \boldsymbol{s}} \\ &= \frac{\boldsymbol{s}}{\|\boldsymbol{s}\|}\times \|\boldsymbol{s}\|^{-\frac{d}{2}-2}\Big(r\|\boldsymbol{s}\| I_{\frac{d}{2}}(r\|\boldsymbol{s}\|) - (d+2) I_{\frac{d}{2}+1}(r\|\boldsymbol{s}\|)\Big)\otimes \boldsymbol{s}\boldsymbol{s}' \\ &\qquad + \|\boldsymbol{s}\|^{-\frac{d}{2}-1} I_{\frac{d}{2}+1}(r\|\boldsymbol{s}\|)(\mathrm{vec}(\boldsymbol{I}_d)\boldsymbol{s}' + \boldsymbol{I}_d\otimes \boldsymbol{s})\end{aligned}$$

where we use (A.31). Combining these terms together in (B.18) gives us

$$\frac{\partial^3 f}{\partial \boldsymbol{s}\partial \boldsymbol{s}\partial \boldsymbol{s}'} = r^2\|\boldsymbol{s}\|^{\frac{d}{2}-3}\Big[\boldsymbol{I}_{\frac{d}{2}}(r\|\boldsymbol{s}\|)\boldsymbol{C}_{31} + \boldsymbol{I}_{\frac{d}{2}+1}(r\|\boldsymbol{s}\|)\boldsymbol{C}_{32}\Big]$$

where

$$\begin{aligned}\boldsymbol{C}_{31} &= r\|\boldsymbol{s}\|\,\boldsymbol{A}_3, \\ \boldsymbol{C}_{32} &= \|\boldsymbol{s}\|^2\boldsymbol{A}_1 - (d+2)\boldsymbol{A}_3\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{A}_3 &= \boldsymbol{s}\otimes \boldsymbol{s}\boldsymbol{s}', \\ \boldsymbol{A}_1 &= \mathrm{vec}(\boldsymbol{I}_d)\boldsymbol{s}' + \boldsymbol{I}_d\otimes \boldsymbol{s} + \boldsymbol{s}\otimes \boldsymbol{I}_d\,.\end{aligned}$$

Note that for the forth order derivative, we have that $\frac{\partial^4 f}{\partial \boldsymbol{s}\partial \boldsymbol{s}'\partial \boldsymbol{s}\partial \boldsymbol{s}'} = \frac{\partial^4 f}{\partial \boldsymbol{s}'\partial \boldsymbol{s}\partial \boldsymbol{s}\partial \boldsymbol{s}'}$. Applying the product rule from Corollary A4.4 gives us

$$\begin{aligned}\frac{\partial^4 f}{\partial \boldsymbol{s}'\partial \boldsymbol{s}\partial \boldsymbol{s}\partial \boldsymbol{s}'} = r^3\frac{\partial}{\partial \boldsymbol{s}'}\Big[\|\boldsymbol{s}\|^{-\frac{d}{2}-2} I_{\frac{d}{2}}(r\|\boldsymbol{s}\|)\boldsymbol{A}_3\Big] + r^2\frac{\partial}{\partial \boldsymbol{s}'}\Big[\|\boldsymbol{s}\|^{-\frac{d}{2}-1} I_{\frac{d}{2}+1}(r\|\boldsymbol{s}\|)\boldsymbol{A}_1\Big] \\ - (d+2)r^2\frac{\partial}{\partial \boldsymbol{s}'}\Big[\|\boldsymbol{s}\|^{-\frac{d}{2}-3} I_{\frac{d}{2}+1}(r\|\boldsymbol{s}\|)\boldsymbol{A}_3\Big]. \quad \text{(B.19)}\end{aligned}$$

For the derivative of the first term,

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{s}'}\Big[\|\boldsymbol{s}\|^{-\frac{d}{2}-2} I_{\frac{d}{2}}(r\|\boldsymbol{s}\|)\boldsymbol{A}_3\Big] \\
&= \frac{\partial\|\boldsymbol{s}\|}{\partial \boldsymbol{s}'}\frac{d\|\boldsymbol{s}\|^{-\frac{d}{2}-2} I_{\frac{d}{2}}(r\|\boldsymbol{s}\|)}{d\|\boldsymbol{s}\|} \otimes \boldsymbol{A}_3 + \|\boldsymbol{s}\|^{-\frac{d}{2}-2} I_{\frac{d}{2}}(r\|\boldsymbol{s}\|)\frac{\partial \boldsymbol{A}_3}{\partial \boldsymbol{s}'} \\
&= \|\boldsymbol{s}\|^{-\frac{d}{2}-4}\Big(r\|\boldsymbol{s}\| I_{\frac{d}{2}+1}(r\|\boldsymbol{s}\|) - 2I_{\frac{d}{2}}(r\|\boldsymbol{s}\|)\Big)\boldsymbol{A}_4 + \|\boldsymbol{s}\|^{-\frac{d}{2}-2} I_{\frac{d}{2}}(r\|\boldsymbol{s}\|)\boldsymbol{A}_2 \\
&= \|\boldsymbol{s}\|^{-\frac{d}{2}-4}\Big[I_{\frac{d}{2}}(r\|\boldsymbol{s}\|)\left(\|\boldsymbol{s}\|^2\boldsymbol{A}_2 - 2\boldsymbol{A}_4\right) + I_{\frac{d}{2}+1}(r\|\boldsymbol{s}\|)\, r\|\boldsymbol{s}\|\boldsymbol{A}_4\Big]
\end{aligned}$$

where we let

$$\begin{aligned}
\boldsymbol{A}_2 &= \frac{\partial \boldsymbol{A}_3}{\partial \boldsymbol{s}'} = \boldsymbol{I}_d \otimes \boldsymbol{s}\boldsymbol{s}' + \boldsymbol{s} \otimes \boldsymbol{I}_d \otimes \boldsymbol{s}' + (\boldsymbol{s} \otimes \boldsymbol{s})\mathrm{vec}(\boldsymbol{I}_d)' \\
\boldsymbol{A}_4 &= \boldsymbol{s}\boldsymbol{s}' \otimes \boldsymbol{s}\boldsymbol{s}'
\end{aligned}$$

from equation (A.32). For the derivative of the second term,

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{s}'}\Big[\|\boldsymbol{s}\|^{-\frac{d}{2}-1} I_{\frac{d}{2}+1}(r\|\boldsymbol{s}\|)\boldsymbol{A}_1\Big] \\
&= \frac{\partial\|\boldsymbol{s}\|}{\partial \boldsymbol{s}'}\frac{d\|\boldsymbol{s}\|^{-\frac{d}{2}-1} I_{\frac{d}{2}+1}(r\|\boldsymbol{s}\|)}{d\|\boldsymbol{s}\|} \otimes \boldsymbol{A}_1 + \|\boldsymbol{s}\|^{-\frac{d}{2}-1} I_{\frac{d}{2}+1}(r\|\boldsymbol{s}\|)\frac{\partial \boldsymbol{A}_1}{\partial \boldsymbol{s}'} \\
&= \|\boldsymbol{s}\|^{-\frac{d}{2}-3}\Big(r\|\boldsymbol{s}\| I_{\frac{d}{2}}(r\|\boldsymbol{s}\|) - (d+2)I_{\frac{d}{2}+1}(r\|\boldsymbol{s}\|)\Big)\boldsymbol{B}_2 + \|\boldsymbol{s}\|^{-\frac{d}{2}-1} I_{\frac{d}{2}+1}(r\|\boldsymbol{s}\|)\boldsymbol{A}_0 \\
&= \|\boldsymbol{s}\|^{-\frac{d}{2}-3}\Big[I_{\frac{d}{2}}(r\|\boldsymbol{s}\|)\, r\|\boldsymbol{s}\|\boldsymbol{B}_2 + I_{\frac{d}{2}+1}(r\|\boldsymbol{s}\|)\big(\|\boldsymbol{s}\|^2\boldsymbol{A}_0 - (d+2)\boldsymbol{B}_2\big)\Big]
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{A}_0 &= \frac{\partial \boldsymbol{A}_1}{\partial \boldsymbol{s}'} = \boldsymbol{I}_d \otimes \boldsymbol{I}_d + \boldsymbol{K}^{(d,d)} + \mathrm{vec}(\boldsymbol{I}_d)\mathrm{vec}(\boldsymbol{I}_d)', \\
\boldsymbol{B}_2 &= \boldsymbol{s}' \otimes \boldsymbol{A}_1 = \boldsymbol{s}\boldsymbol{s}' \otimes \boldsymbol{I}_d + \boldsymbol{s}' \otimes \boldsymbol{I}_d \otimes \boldsymbol{s} + \mathrm{vec}(\boldsymbol{I}_d)(\boldsymbol{s}' \otimes \boldsymbol{s}').
\end{aligned}$$

For the derivative of the final term,

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{s}'}\Big[\|\boldsymbol{s}\|^{-\frac{d}{2}-3} I_{\frac{d}{2}+1}(r\|\boldsymbol{s}\|)\boldsymbol{A}_3\Big] \\
&= \frac{\partial\|\boldsymbol{s}\|}{\partial \boldsymbol{s}'}\frac{d\|\boldsymbol{s}\|^{-\frac{d}{2}} I_{\frac{d}{2}+1}(r\|\boldsymbol{s}\|)}{d\|\boldsymbol{s}\|} + \|\boldsymbol{s}\|^{-\frac{d}{2}-3} I_{\frac{d}{2}+1}(r\|\boldsymbol{s}\|)\frac{\partial \boldsymbol{A}_3}{\partial \boldsymbol{s}'} \\
&= \|\boldsymbol{s}\|^{-\frac{d}{2}-5}\Big(r\|\boldsymbol{s}\| I_{\frac{d}{2}}(r\|\boldsymbol{s}\|) - (d+4)I_{\frac{d}{2}+1}(r\|\boldsymbol{s}\|)\Big)\boldsymbol{A}_4 + \|\boldsymbol{s}\|^{-\frac{d}{2}-3} I_{\frac{d}{2}+1}(r\|\boldsymbol{s}\|)\boldsymbol{A}_2 \\
&= \|\boldsymbol{s}\|^{-\frac{d}{2}-5}\Big[I_{\frac{d}{2}}(r\|\boldsymbol{s}\|)\, r\|\boldsymbol{s}\|\boldsymbol{A}_4 + I_{\frac{d}{2}+1}(r\|\boldsymbol{s}\|)\left(\|\boldsymbol{s}\|^2\boldsymbol{A}_2 - (d+4)\boldsymbol{A}_4\right)\Big].
\end{aligned}$$

Combining these terms together in (B.19) gives us

$$\frac{\partial^4 f}{\partial \boldsymbol{s}' \partial \boldsymbol{s} \partial \boldsymbol{s} \partial \boldsymbol{s}'} = r^2 \|\boldsymbol{s}\|^{-\frac{d}{2}-5} \Big(\boldsymbol{I}_{\frac{d}{2}}(r\|\boldsymbol{s}\|) \boldsymbol{C}_{41} + \boldsymbol{I}_{\frac{d}{2}+1}(r\|\boldsymbol{s}\|) \boldsymbol{C}_{42} \Big)$$

where

$$\begin{aligned} \boldsymbol{C}_{41} &= r\|\boldsymbol{s}\| \big(\|\boldsymbol{s}\|^2 (\boldsymbol{A}_2 + \boldsymbol{B}_2) - (d+4) \boldsymbol{A}_4 \big), \\ \boldsymbol{C}_{42} &= \big((d+4)(d+2) + r^2 \|\boldsymbol{s}\|^2 \big) \boldsymbol{A}_4 - (d+2) \|\boldsymbol{s}\|^2 (\boldsymbol{A}_2 + \boldsymbol{B}_2) + \|\boldsymbol{s}\|^4 \boldsymbol{A}_0. \end{aligned}$$

□

B6 Fisher information matrix

Let $\boldsymbol{y}_i \sim \mathcal{N}_d(\boldsymbol{\mu} + u_i \boldsymbol{\gamma}, u_i \boldsymbol{\Sigma})$ where $\boldsymbol{y}_i$'s are independent, $u_i \sim \mathcal{G}(\nu, \nu)$ where u_i's are independent and identically distributed, and $\ell_c(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}; \boldsymbol{Y}, \boldsymbol{u})$ be the complete data likelihood where $\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{y}_1 & \cdots & \boldsymbol{y}_n \end{pmatrix}'$ and $\boldsymbol{u} = (u_1, \ldots, u_n)$. Under certain regularity conditions, the Fisher information matrix is the expected value of the observed information matrix with respect to the data matrix $\boldsymbol{Y}$. That is

$$\mathcal{I}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{Y}} \big[\boldsymbol{I}_{\text{obs}}(\boldsymbol{\theta}; \boldsymbol{Y}) \big] \tag{B.20}$$

where the observed information matrix from (2.20) is given by

$$\boldsymbol{I}_{\text{obs}}(\boldsymbol{\theta}; \boldsymbol{Y}) = -\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}[\ell_{\text{c}}''(\boldsymbol{\theta})] - \mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}} \big[\ell_{\text{c}}'(\boldsymbol{\theta}) \ell_{\text{c}}'(\boldsymbol{\theta})^\top \big] - \mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}[\ell_{\text{c}}'(\boldsymbol{\theta})] \mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}[\ell_{\text{c}}'(\boldsymbol{\theta})]^\top \tag{B.21}$$

where we let $\ell_c(\boldsymbol{\theta}) = \ell_{\text{com}}(\boldsymbol{\theta}; \boldsymbol{Y}, \boldsymbol{u})$. So under certain regularity conditions (the same ones as in Louis' formula), the Fisher information matrix can be calculated by taking expectation of the observed information matrix.

For the one-dimensional case, the Fisher information matrix can be calculated by numerical integration directly. However, in the multidimensional case, this direct approach is infeasible as it requires multidimensional integration which is computationally demanding. To simplify the multidimensional integration, we use the matrix representations in Section B2 to calculate the expectation of the three terms in (B.21). The expectation of the first and second term can be evaluated using Lemma B1.1 and the conditional normal moment results in Section B4. The expectation of the third term can be simplified using d-dimensional spherical coordinates to integrate over a sphere of radius r, then numerically integrating the radius over $\mathbb{R}^+$.

The following sections provide the derivation to evaluate the expectation of the three terms in (B.21) with respect to $\boldsymbol{Y}$.

B6.1 First term of Fisher information matrix

Applying Lemma B1.1 to the first term of the Fisher information matrix in (B.20) gives us

$$\mathbb{E}_{\boldsymbol{Y}}\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}[\ell_c''(\boldsymbol{\theta})] = \mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}[\ell_c''(\boldsymbol{\theta})]. \tag{B.22}$$

From the NMVM representation of the VG distribution in (1.37), we have that $\boldsymbol{y}_i|u_i \sim \mathcal{N}(\boldsymbol{\mu} + u_i\boldsymbol{\gamma}, u_i\boldsymbol{\Sigma})$ and $u_i \sim \mathcal{G}(\nu,\nu)$. Using the general and log-moment formulas from (1.26) to (1.28), we have that

$$\begin{aligned}
\mathbb{E}_u[u^m] &= \frac{\Gamma(\nu+m)}{\nu^m\,\Gamma(\nu)} \qquad \text{for } \nu+m>0,\\
\mathbb{E}_u[u^m\log u] &= \frac{\Gamma(\nu+m)}{\nu^m\Gamma(\nu)}\big(\psi(\nu+m)-\log\nu\big) \qquad \text{for } \nu+m>0,\\
\mathbb{E}_u\big[(\log u)^2\big] &= \big(\psi(\nu)-\log\nu\big)^2+\psi'(\nu).
\end{aligned}$$

where $u \sim \mathcal{G}(\nu,\nu)$.

B6.1.1 Diagonal entries

($\boldsymbol{\mu}$,$\boldsymbol{\mu}$) entry:

$$\mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\left[\frac{\partial^2\ell_c}{\partial\boldsymbol{\mu}\partial\boldsymbol{\mu}^\top}\right] = -\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{u}}\left[\frac{1}{u_i}\right] = -n\boldsymbol{\Sigma}^{-1}\mathbb{E}_u\left[\frac{1}{u}\right].$$

($\boldsymbol{\gamma}$,$\boldsymbol{\gamma}$) entry:

$$\mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\left[\frac{\partial^2\ell_c}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^\top}\right] = -\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{u}}[u_i] = -n\boldsymbol{\Sigma}^{-1}\mathbb{E}_u[u].$$

(ν,ν) entry:

$$\mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\left[\frac{\partial^2\ell_c}{\partial\nu^2}\right] = \frac{n}{\nu} - n\psi'(\nu).$$

(vech$\boldsymbol{\Sigma}$,vech$\boldsymbol{\Sigma}$) entry:

$$\mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\left[\frac{\partial^2\ell_c}{\partial\text{vech}\boldsymbol{\Sigma}\partial\text{vech}\boldsymbol{\Sigma}^\top}\right] = \boldsymbol{D}_d^\top\mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\left[\frac{\partial^2\ell_c}{\partial\text{vec}\boldsymbol{\Sigma}_u\partial\text{vec}\boldsymbol{\Sigma}_u^\top}\right]\boldsymbol{D}_d$$

where

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\left[\frac{\partial^2\ell_c}{\partial\text{vec}\boldsymbol{\Sigma}_u\partial\text{vec}\boldsymbol{\Sigma}_u^\top}\right]\\
&= \mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\left[\frac{n}{2}\big(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\big) - \frac{1}{2}\big(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}S_{\tilde{\boldsymbol{y}}\tilde{\boldsymbol{y}}/u}\boldsymbol{\Sigma}^{-1}\big) - \frac{1}{2}\big(\boldsymbol{\Sigma}^{-1}S_{\tilde{\boldsymbol{y}}\tilde{\boldsymbol{y}}/u}\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\big)\right]\\
&= \frac{n}{2}\big(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\big) - \frac{1}{2}\big(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\big[S_{\tilde{\boldsymbol{y}}\tilde{\boldsymbol{y}}/u}\big]\boldsymbol{\Sigma}^{-1}\big) - \frac{1}{2}\big(\boldsymbol{\Sigma}^{-1}\mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\big[S_{\tilde{\boldsymbol{y}}\tilde{\boldsymbol{y}}/u}\big]\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\big)
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\big[S_{\tilde{\boldsymbol{y}}\tilde{\boldsymbol{y}}/u}\big] &= \mathbb{E}_{\boldsymbol{u}}\left(\sum_{i=1}^{n}\frac{1}{u_i}\underbrace{\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\Big[(\boldsymbol{y}_i-\boldsymbol{\mu}-u_i\boldsymbol{\gamma})(\boldsymbol{y}_i-\boldsymbol{\mu}-u_i\boldsymbol{\gamma})^\top\Big]}_{u_i\boldsymbol{\Sigma}}\right)\\
&= n\boldsymbol{\Sigma}\,.
\end{aligned}$$

This leads to $\mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\left[\frac{\partial^2\ell_c}{\partial\text{vec}\boldsymbol{\Sigma}_u\partial\text{vec}\boldsymbol{\Sigma}_u^\top}\right] = -\frac{n}{2}(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1})$ thus giving us the result

$$\mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\left[\frac{\partial^2\ell_c}{\partial\text{vech}\boldsymbol{\Sigma}\partial\text{vech}\boldsymbol{\Sigma}^\top}\right] = -\frac{n}{2}\boldsymbol{D}_d^\top\big(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\big)\boldsymbol{D}_d.$$

B6.1.2 Off-diagonal entries

($\boldsymbol{\mu}$,$\boldsymbol{\gamma}$) entry:

$$\mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\left[\frac{\partial^2\ell_c}{\partial\boldsymbol{\mu}\partial\boldsymbol{\gamma}^\top}\right] = -n\boldsymbol{\Sigma}^{-1}.$$

(vech$\boldsymbol{\Sigma}$,$\boldsymbol{\mu}$) entry:

$$\mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\left[\frac{\partial^2\ell_c}{\partial\text{vech}\boldsymbol{\Sigma}\partial\boldsymbol{\mu}^\top}\right] = \boldsymbol{D}_d^\top\mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\left[\frac{\partial^2\ell_c}{\partial\text{vec}\boldsymbol{\Sigma}_u\partial\boldsymbol{\mu}^\top}\right]\boldsymbol{D}_d$$

where

$$\mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\left[\frac{\partial^2\ell_c}{\partial\text{vec}\boldsymbol{\Sigma}_u\partial\boldsymbol{\mu}^\top}\right] = -\mathbb{E}_{\boldsymbol{u}}\left[\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{n}\frac{1}{u_i}\underbrace{\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}(\boldsymbol{y}_i-\boldsymbol{\mu}-u_i\boldsymbol{\gamma})}_{=\mathbf{0}}\otimes\boldsymbol{\Sigma}^{-1}\right]$$
$$= \mathbf{0}.$$

Thus we have that

$$\mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\left[\frac{\partial^2\ell_c}{\partial\text{vech}\boldsymbol{\Sigma}\partial\boldsymbol{\mu}^\top}\right] = \mathbf{0}.$$

(vech$\boldsymbol{\Sigma}$,$\boldsymbol{\gamma}$) entry:

$$\mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\left[\frac{\partial^2\ell_c}{\partial\text{vech}\boldsymbol{\Sigma}\partial\boldsymbol{\gamma}^\top}\right] = \boldsymbol{D}_d^\top\mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\left[\frac{\partial^2\ell_c}{\partial\text{vec}\boldsymbol{\Sigma}_u\partial\boldsymbol{\mu}^\top}\right]\boldsymbol{D}_d$$

where

$$\mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\left[\frac{\partial^2\ell_c}{\partial\text{vec}\boldsymbol{\Sigma}_u\partial\boldsymbol{\mu}^\top}\right] = -\mathbb{E}_{\boldsymbol{u}}\left[\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{n}\underbrace{\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}(\boldsymbol{y}_i-\boldsymbol{\mu}-u_i\boldsymbol{\gamma})}_{=\mathbf{0}}\otimes\boldsymbol{\Sigma}^{-1}\right]$$
$$= \mathbf{0}.$$

Thus we have that

$$\mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\left[\frac{\partial^2\ell_c}{\partial\text{vech}\boldsymbol{\Sigma}\partial\boldsymbol{\gamma}^\top}\right] = \mathbf{0}.$$

Note that the cross derivatives involving ν are all zero.

B6.1.3 Final result

Combining these derivatives together gives us the expectation of the first term of the observed information matrix in (B.21).

$$-\mathbb{E}_{\boldsymbol{Y}}\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}(\ell_c'') = \begin{pmatrix} n\boldsymbol{\Sigma}^{-1}\mathbb{E}_u\left(\frac{1}{u}\right) & n\boldsymbol{\Sigma}^{-1} & \mathbf{0} & \mathbf{0} \\ n\boldsymbol{\Sigma}^{-1} & n\boldsymbol{\Sigma}^{-1}\mathbb{E}_u(u) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \frac{n}{2}\boldsymbol{D}_d^\top(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1})\boldsymbol{D}_d & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & n\psi'(\nu)-\frac{n}{\nu} \end{pmatrix}.$$

B6.2 Second term of Fisher information matrix

Applying Lemma B1.1 to the expectation of the simplified representation of the second term (B.14), this gives us the second term of the Fisher information matrix as

$$\begin{aligned}\mathbb{E}_{\boldsymbol{Y}}\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\left[\ell_c'(\boldsymbol{\theta})\ell_c'(\boldsymbol{\theta})^\top\right] &= \sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{Y}}\left[\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\left[\frac{1}{u_i^2}\right]\boldsymbol{C}_{1/u,i}\boldsymbol{C}_{1/u,i}^\top\right] + ... \\ &= \sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{u}}\left[\frac{1}{u_i^2}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\left[\boldsymbol{C}_{1/u,i}\boldsymbol{C}_{1/u,i}^\top\right]\right] \\ &= n\,\mathbb{E}_u\left[\frac{1}{u^2}\mathbb{E}_{\boldsymbol{y}|u}\left[\boldsymbol{c}_{1/u}\boldsymbol{c}_{1/u}^\top\right]\right] + ... \end{aligned} \tag{B.23}$$

where we let $\boldsymbol{y} \sim \mathcal{VG}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}, \nu)$, $u \sim \mathcal{G}(\nu, \nu)$ and $\boldsymbol{c}_{1/u}$ represent the column of $\boldsymbol{C}_{1/u}$ without the index i. We use this representation to simplify the calculation of the second term of the Fisher information matrix. Recall that the additive constants with respect to $\boldsymbol{u}$ can simply be ignored in the following calculation. That is,

$$\frac{\partial \ell_c}{\partial \boldsymbol{\theta}} = \boldsymbol{C}_{1/u}^{\theta}\boldsymbol{v}_{1/u} + \boldsymbol{C}_{u}^{\theta}\boldsymbol{v}_{u} + \boldsymbol{C}_{\log u}^{\theta}\boldsymbol{v}_{\log u}\,. \tag{B.24}$$

$(\boldsymbol{\mu}, \boldsymbol{\mu})$ entry:

$$\mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\left(\frac{\partial \ell_c}{\partial \boldsymbol{\mu}}\frac{\partial \ell_c}{\partial \boldsymbol{\mu}^\top}\right) = n\boldsymbol{\Sigma}^{-1}\mathbb{E}_u\left[\frac{1}{u^2}\mathbb{E}_{\boldsymbol{y}|u}[(\boldsymbol{y}-\boldsymbol{\mu})(\boldsymbol{y}-\boldsymbol{\mu})']\right]\boldsymbol{\Sigma}^{-1}$$

Using the second order moment result in Section B4

$$= n\boldsymbol{\Sigma}^{-1}\mathbb{E}_u\left[\frac{1}{u^2}\left(u\boldsymbol{\Sigma} + u^2\boldsymbol{\gamma}\boldsymbol{\gamma}'\right)\right]\boldsymbol{\Sigma}^{-1} \tag{B.25}$$

$$= n\,\mathbb{E}_u\left[\frac{1}{u}\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\right]. \tag{B.26}$$

$(\boldsymbol{\gamma}, \boldsymbol{\gamma})$ entry:

$$\mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\left(\frac{\partial \ell_c}{\partial \boldsymbol{\gamma}}\frac{\partial \ell_c}{\partial \boldsymbol{\gamma}'}\right) = n\,\mathbb{E}_u\left[u^2\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\right].$$

$(\mathrm{vech}\boldsymbol{\Sigma}, \mathrm{vech}\boldsymbol{\Sigma})$ entry:

$$\mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\left(\frac{\partial \ell_c}{\partial \mathrm{vech}\boldsymbol{\Sigma}}\frac{\partial \ell_c}{\partial \mathrm{vech}\boldsymbol{\Sigma}^\top}\right) = \boldsymbol{D}_d^\top\, \mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\left[\frac{\partial \ell_c}{\partial \mathrm{vec}\boldsymbol{\Sigma}_u}\frac{\partial \ell_c}{\partial \mathrm{vec}\boldsymbol{\Sigma}_u^\top}\right]\boldsymbol{D}_d.$$

Focusing on $\mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\left[\frac{\partial \ell_c}{\partial \mathrm{vec}\boldsymbol{\Sigma}_u}\frac{\partial \ell_c}{\partial \mathrm{vec}\boldsymbol{\Sigma}_u^\top}\right]$, and using the matrix representation in (B.24),

$$\mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\left[\frac{\partial \ell_c}{\partial \mathrm{vec}\boldsymbol{\Sigma}_u}\frac{\partial \ell_c}{\partial \mathrm{vec}\boldsymbol{\Sigma}_u^\top}\right] = n\,\mathbb{E}_u\Bigg[\underbrace{\mathbb{E}_{\boldsymbol{y}|u}\left[\frac{1}{u^2}\boldsymbol{c}_{1/u}\boldsymbol{c}_{1/u}^\top\right]}_{(i)} + \underbrace{\mathbb{E}_{\boldsymbol{y}|u}\left[u^2\boldsymbol{c}_u\boldsymbol{c}_u^\top\right]}_{(ii)} + \underbrace{\mathbb{E}_{\boldsymbol{y}|u}\left[\boldsymbol{c}_{1/u}\boldsymbol{c}_u^\top\right]}_{(iii)} + \mathbb{E}_{\boldsymbol{y}|u}\left[\boldsymbol{c}_u\boldsymbol{c}_{1/u}^\top\right]\Bigg]$$

where

(i) $$\begin{aligned}&\frac{1}{4u^2}\mathbb{E}_{\boldsymbol{y}|u}\Big[\mathrm{vec}\big(\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}\big)\mathrm{vec}\big(\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}\big)'\Big]\\ &= \frac{1}{4u^2}\big(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\big)\mathbb{E}_{\boldsymbol{y}|u}\Big[\mathrm{vec}\big((\boldsymbol{y}-\boldsymbol{\mu})(\boldsymbol{y}-\boldsymbol{\mu})'\big)\mathrm{vec}\big((\boldsymbol{y}-\boldsymbol{\mu})(\boldsymbol{y}-\boldsymbol{\mu})'\big)'\Big]\big(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\big)\\ &= \frac{1}{4u^2}\big(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\big)\Big[\big(\boldsymbol{I}_d+\boldsymbol{K}^{(d,d)}\big)(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma}) + [\mathrm{vec}\boldsymbol{\Sigma}+u\,\mathrm{vec}(\boldsymbol{\gamma}\boldsymbol{\gamma}')][\mathrm{vec}\boldsymbol{\Sigma}+u\,\mathrm{vec}(\boldsymbol{\gamma}\boldsymbol{\gamma}')]'\\ &\qquad + u\big(\boldsymbol{I}_d+\boldsymbol{K}^{(d,d)}\big)(\boldsymbol{\Sigma}\otimes\boldsymbol{\gamma}\boldsymbol{\gamma}')\big(\boldsymbol{I}_d+\boldsymbol{K}^{(d,d)}\big)\Big]\big(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\big)\end{aligned}$$

using the forth order moment result in Section (B4) for the last equality,

(ii) $$\begin{aligned}&\frac{u^2}{4}\,\mathbb{E}_{\boldsymbol{y}|u}\Big[\mathrm{vec}\big(\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\big)\mathrm{vec}\big(\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\big)'\Big]\\ &= \frac{u^2}{4}\big(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\big)\mathrm{vec}(\boldsymbol{\gamma}\boldsymbol{\gamma}')\mathrm{vec}(\boldsymbol{\gamma}\boldsymbol{\gamma}')'\big(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\big),\end{aligned}$$

(iii) $$\begin{aligned}&\frac{1}{4}\mathbb{E}_{\boldsymbol{y}|u}\Big[\mathrm{vec}\big(\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}\big)\mathrm{vec}\big(\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\big)'\Big]\\ &= \frac{1}{4}\big(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\big)\mathbb{E}_{\boldsymbol{y}|u}\big[\mathrm{vec}\big((\boldsymbol{y}-\boldsymbol{\mu})(\boldsymbol{y}-\boldsymbol{\mu})'\big)\big]\mathrm{vec}(\boldsymbol{\gamma}\boldsymbol{\gamma}')'\big(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\big)\\ &= \frac{1}{4}\big(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\big)\big[u\,\mathrm{vec}\boldsymbol{\Sigma}+u^2\mathrm{vec}(\boldsymbol{\gamma}\boldsymbol{\gamma}')\big]\mathrm{vec}(\boldsymbol{\gamma}\boldsymbol{\gamma}')'\big(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\big).\end{aligned}$$

Combining it together gives us

$$\mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\left[\frac{\partial \ell_c}{\partial \mathrm{vec}\boldsymbol{\Sigma}_u}\frac{\partial \ell_c}{\partial \mathrm{vec}\boldsymbol{\Sigma}_u^\top}\right]$$
$$= \frac{n}{4}\big(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\big)\mathbb{E}_u\Big[\big(\boldsymbol{I}_d+\boldsymbol{K}^{(d,d)}\big)(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma}) + [\mathrm{vec}\boldsymbol{\Sigma}+2u\,\mathrm{vec}(\boldsymbol{\gamma}\boldsymbol{\gamma}')][\mathrm{vec}\boldsymbol{\Sigma}+2u\,\mathrm{vec}(\boldsymbol{\gamma}\boldsymbol{\gamma}')]' + u\big(\boldsymbol{I}_d+\boldsymbol{K}^{(d,d)}\big)(\boldsymbol{\Sigma}\otimes\boldsymbol{\gamma}\boldsymbol{\gamma}')\big(\boldsymbol{I}_d+\boldsymbol{K}^{(d,d)}\big)\Big]\big(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\big).$$

$(\boldsymbol{\nu},\boldsymbol{\nu})$ entry:

$$\mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\left(\frac{\partial \ell_c}{\partial \nu}\frac{\partial \ell_c}{\partial \nu}\right) = n\,\mathbb{E}_u\big[u^2+(\log u)^2-2u\log u\big].$$

$(\boldsymbol{\mu},\boldsymbol{\gamma})$ entry:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\left(\frac{\partial \ell_c}{\partial \boldsymbol{\mu}}\frac{\partial \ell_c}{\partial \boldsymbol{\gamma}'}\right) &= n\boldsymbol{\Sigma}^{-1}\mathbb{E}_{\boldsymbol{u}}\big[\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}[(\boldsymbol{y}-\boldsymbol{\mu})\boldsymbol{\gamma}']\big]\boldsymbol{\Sigma}^{-1}\\ &= n\boldsymbol{\Sigma}^{-1}\mathbb{E}_u\big[u\boldsymbol{\gamma}\boldsymbol{\gamma}'\big]\boldsymbol{\Sigma}^{-1}\\ &= n\,\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\mathbb{E}_u[u].\end{aligned}$$

$(\mathrm{vech}\boldsymbol{\Sigma},\boldsymbol{\mu})$ entry:

$$\mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\left(\frac{\partial \ell_c}{\partial \mathrm{vech}\boldsymbol{\Sigma}}\frac{\partial \ell_c}{\partial \boldsymbol{\mu}^\top}\right) = \boldsymbol{D}_d^\top\,\mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\left[\frac{\partial \ell_c}{\partial \mathrm{vec}\boldsymbol{\Sigma}_u}\frac{\partial \ell_c}{\partial \boldsymbol{\mu}^\top}\right].$$

Focusing on $\mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\left[\frac{\partial \ell_c}{\partial \mathrm{vec}\boldsymbol{\Sigma}_u}\frac{\partial \ell_c}{\partial \boldsymbol{\mu}^\top}\right]$ gives us

$$\mathbb{E}_u\mathbb{E}_{\boldsymbol{y}|u}\left[\frac{\partial \ell_c}{\partial \mathrm{vec}\boldsymbol{\Sigma}_u}\frac{\partial \ell_c}{\partial \boldsymbol{\mu}^\top}\right] = n\,\mathbb{E}_u\left[\mathbb{E}_{\boldsymbol{y}|u}\left[\frac{1}{u^2}\boldsymbol{c}_{1/u}^{\mathrm{vec}\Sigma}\boldsymbol{c}_{1/u}^{\mu}\right]+\mathbb{E}_{\boldsymbol{y}|u}\left[\boldsymbol{c}_u^{\mathrm{vec}\Sigma}\boldsymbol{c}_{1/u}^{\mu}\right]\right]$$

where

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{y}|u}\left[\frac{1}{u^2}\boldsymbol{c}_{1/u}^{\mathrm{vec}\Sigma}\boldsymbol{c}_{1/u}^{\mu}\right] \\
&= \frac{1}{2}\mathbb{E}_{\boldsymbol{y}|u}\left[\frac{1}{u^2}\mathrm{vec}\big(\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}\big)(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}\right] \\
&= \frac{1}{2u^2}\big(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\big)\mathbb{E}_{\boldsymbol{y}|u}\big[\mathrm{vec}\big((\boldsymbol{y}-\boldsymbol{\mu})(\boldsymbol{y}-\boldsymbol{\mu})'\big)(\boldsymbol{y}-\boldsymbol{\mu})'\big]\boldsymbol{\Sigma}^{-1} \\
&= \frac{1}{2u^2}\big(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\big)\big[u^2\boldsymbol{\gamma}\otimes\boldsymbol{\Sigma}+u^2\boldsymbol{\Sigma}\otimes\boldsymbol{\gamma}+u^2\mathrm{vec}(\boldsymbol{\Sigma})\boldsymbol{\gamma}'+u^3\boldsymbol{\gamma}\otimes\boldsymbol{\gamma}\boldsymbol{\gamma}'\big]\boldsymbol{\Sigma}^{-1} \\
&= \frac{1}{2}\big(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\big)\big[\boldsymbol{\gamma}\otimes\boldsymbol{\Sigma}+\boldsymbol{\Sigma}\otimes\boldsymbol{\gamma}+\mathrm{vec}(\boldsymbol{\Sigma})\boldsymbol{\gamma}'+u\boldsymbol{\gamma}\otimes\boldsymbol{\gamma}\boldsymbol{\gamma}'\big]\boldsymbol{\Sigma}^{-1},
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{y}|u}\Big[\boldsymbol{c}_{u}^{\mathrm{vec}\Sigma}\boldsymbol{c}_{1/u}^{\mu}\Big] &= \frac{1}{2}\mathbb{E}_{\boldsymbol{y}|u}\big[\mathrm{vec}\big(\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\big)(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}\big] \\
&= \frac{1}{2}\big(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\big)\mathrm{vec}(\boldsymbol{\gamma}\boldsymbol{\gamma}')\mathbb{E}_{\boldsymbol{y}|u}[(\boldsymbol{y}-\boldsymbol{\mu})']\boldsymbol{\Sigma}^{-1} \\
&= \frac{1}{2}\big(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\big)\mathrm{vec}(\boldsymbol{\gamma}\boldsymbol{\gamma}')u\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1} \\
&= \frac{u}{2}\big(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\big)(\boldsymbol{\gamma}\otimes\boldsymbol{\gamma}\boldsymbol{\gamma}')\boldsymbol{\Sigma}^{-1}.
\end{aligned}
$$

Combining the terms gives us

$$
\mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\left[\frac{\partial\ell_c}{\partial\mathrm{vec}\boldsymbol{\Sigma}_u}\frac{\partial\ell_c}{\partial\boldsymbol{\mu}^\top}\right] = \frac{n}{2}\left(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\right)\big(\boldsymbol{\gamma}\otimes\boldsymbol{\Sigma}+\boldsymbol{\Sigma}\otimes\boldsymbol{\gamma}+\mathrm{vec}(\boldsymbol{\Sigma})\boldsymbol{\gamma}'+2\,\mathbb{E}_u[u]\boldsymbol{\gamma}\otimes\boldsymbol{\gamma}\boldsymbol{\gamma}'\big)\boldsymbol{\Sigma}^{-1}.
$$

$(\mathrm{vech}\boldsymbol{\Sigma},\boldsymbol{\gamma})$ entry:

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\left(\frac{\partial\ell_c}{\partial\mathrm{vech}\boldsymbol{\Sigma}}\frac{\partial\ell_c}{\partial\boldsymbol{\gamma}'}\right) \\
&= \boldsymbol{D}_d^\top\,\mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\left[\frac{\partial\ell_c}{\partial\mathrm{vec}\boldsymbol{\Sigma}_u}\frac{\partial\ell_c}{\partial\boldsymbol{\gamma}'}\right].
\end{aligned}
$$

Focusing on $\mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\left[\frac{\partial\ell_c}{\partial\mathrm{vec}\boldsymbol{\Sigma}_u}\frac{\partial\ell_c}{\partial\boldsymbol{\gamma}'}\right]$ gives us

$$\mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\left[\frac{\partial \ell_c}{\partial \mathrm{vec}\boldsymbol{\Sigma}_u}\frac{\partial \ell_c}{\partial \boldsymbol{\gamma}'}\right] = n\,\mathbb{E}_u\left[\mathbb{E}_{\boldsymbol{y}|u}\left[\frac{1}{u^2}\boldsymbol{c}_{1/u}^{\mathrm{vec}\Sigma}\boldsymbol{c}_u^{\gamma}\right] + \mathbb{E}_{\boldsymbol{y}|u}\left[\boldsymbol{c}_u^{\mathrm{vec}\Sigma}\boldsymbol{c}_u^{\gamma}\right]\right]$$

where

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{y}|u}\left[\boldsymbol{c}_{1/u}^{\mathrm{vec}\Sigma}\boldsymbol{c}_{1/u}^{\gamma}\right] \\
&= \frac{1}{2}\mathbb{E}_{\boldsymbol{y}|u}\left[\mathrm{vec}\left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}\right)\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\right] \\
&= \frac{1}{2}\left(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\right)\mathbb{E}_{\boldsymbol{y}|u}\left[\mathrm{vec}\left((\boldsymbol{y}-\boldsymbol{\mu})(\boldsymbol{y}-\boldsymbol{\mu})'\right)\right]\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1} \\
&= \frac{1}{2}\left(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\right)\left[u\,\mathrm{vec}\boldsymbol{\Sigma} + u^2\mathrm{vec}(\boldsymbol{\gamma}\boldsymbol{\gamma}')\right]\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{y}|u}\left[u^2\boldsymbol{c}_u^{\mathrm{vec}\Sigma}\boldsymbol{c}_{1/u}^{\gamma}\right] &= \frac{u^2}{2}\mathbb{E}_{\boldsymbol{y}|u}\left[\mathrm{vec}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\right)\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\right] \\
&= \frac{u^2}{2}\left(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\right)\mathrm{vec}(\boldsymbol{\gamma}\boldsymbol{\gamma}')\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}.
\end{aligned}$$

Combining the terms gives us

$$\mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\left[\frac{\partial \ell_c}{\partial \mathrm{vec}\boldsymbol{\Sigma}_u}\frac{\partial \ell_c}{\partial \boldsymbol{\mu}^\top}\right] = \frac{n}{2}\left(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\right)\mathbb{E}_u\left[u\,\mathrm{vec}\boldsymbol{\Sigma} + 2\,u^2\mathrm{vec}(\boldsymbol{\gamma}\boldsymbol{\gamma}')\right]\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}.$$

$(\boldsymbol{\mu}, \nu)$ entry:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\left[\frac{\partial \ell_c}{\partial \boldsymbol{\mu}}\frac{\partial \ell_c}{\partial \nu}\right] &= n\,\mathbb{E}_u\left[\mathbb{E}_{\boldsymbol{y}|u}\left[\boldsymbol{c}_{1/u}^{\mu}\boldsymbol{c}_u^{\nu}\right] + \frac{1}{u}\log u\mathbb{E}_{\boldsymbol{y}|u}\left[\boldsymbol{c}_{1/u}^{\mu}\boldsymbol{c}_{\log u}^{\nu}\right]\right] \\
&= n\,\mathbb{E}_u\left[\mathbb{E}_{\boldsymbol{y}|u}\left[-\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right] + \frac{1}{u}\log u\mathbb{E}_{\boldsymbol{y}|u}\left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right]\right] \\
&= n\,\mathbb{E}_u\left[-\boldsymbol{\Sigma}^{-1}\underbrace{\mathbb{E}_{\boldsymbol{y}|u}\left[(\boldsymbol{y}-\boldsymbol{\mu})\right]}_{u\boldsymbol{\gamma}} + \frac{1}{u}\log u\boldsymbol{\Sigma}^{-1}\underbrace{\mathbb{E}_{\boldsymbol{y}|u}\left[(\boldsymbol{y}-\boldsymbol{\mu})\right]}_{u\boldsymbol{\gamma}}\right] \\
&= n\,\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}\,\mathbb{E}_u[\log u - u].
\end{aligned}$$

$(\boldsymbol{\gamma}, \nu)$ entry:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\left[\frac{\partial \ell_c}{\partial \boldsymbol{\gamma}}\frac{\partial \ell_c}{\partial \nu}\right] &= n\,\mathbb{E}_u\big[u^2\,\mathbb{E}_{\boldsymbol{y}|u}[\boldsymbol{c}_u^{\gamma}\boldsymbol{c}_u^{\nu}] + u\log u\,\mathbb{E}_{\boldsymbol{y}|u}[\boldsymbol{c}_u^{\gamma}\boldsymbol{c}_{\log u}^{\nu}]\big]\\ &= n\,\mathbb{E}_u\big[u^2\,\mathbb{E}_{\boldsymbol{y}|u}[-\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}] + u\log u\,\mathbb{E}_{\boldsymbol{y}|u}[\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}]\big]\\ &= n\,\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}\,\mathbb{E}_u\big[u(\log u - u)\big].\end{aligned}$$

$(\mathbf{vech}\boldsymbol{\Sigma}, \nu)$ entry:

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\left(\frac{\partial \ell_c}{\partial \mathrm{vech}\boldsymbol{\Sigma}}\frac{\partial \ell_c}{\partial \nu}\right)\\ &= \boldsymbol{D}_d^{\top}\,\mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\left[\frac{\partial \ell_c}{\partial \mathrm{vec}\boldsymbol{\Sigma}_u}\frac{\partial \ell_c}{\partial \nu}\right].\end{aligned}$$

Focusing at $\mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\left[\frac{\partial \ell_c}{\partial \mathrm{vec}\boldsymbol{\Sigma}_u}\frac{\partial \ell_c}{\partial \nu}\right]$ gives us

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{u}}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{u}}\left[\frac{\partial \ell_c}{\partial \mathrm{vec}\boldsymbol{\Sigma}_u}\frac{\partial \ell_c}{\partial \nu}\right]\\ &= n\,\mathbb{E}_u\left[\frac{1}{u^2}\mathbb{E}_{\boldsymbol{y}|u}\big[\boldsymbol{c}_{1/u}^{\mathrm{vec}\Sigma}\boldsymbol{c}_u^{\nu}\big] + \frac{1}{u}\log u\mathbb{E}_{\boldsymbol{y}|u}\big[\boldsymbol{c}_{1/u}^{\mathrm{vec}\Sigma}\boldsymbol{c}_{\log u}^{\nu}\big]\right.\\ &\qquad\left. + u^2\mathbb{E}_{\boldsymbol{y}|u}\big[\boldsymbol{c}_u^{\mathrm{vec}\Sigma}\boldsymbol{c}_u^{\nu}\big] + u\log u\mathbb{E}_{\boldsymbol{y}|u}\big[\boldsymbol{c}_u^{\mathrm{vec}\Sigma}\boldsymbol{c}_{\log u}^{\nu}\big]\right]\\ &= n\,\mathbb{E}_u\left[\frac{1}{u^2}\mathbb{E}_{\boldsymbol{y}|u}\left[-\frac{1}{2}\big(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\big)\mathrm{vec}\big((\boldsymbol{y}-\boldsymbol{\mu})(\boldsymbol{y}-\boldsymbol{\mu})'\big)\right]\right.\\ &\qquad + \frac{1}{u}\log u\mathbb{E}_{\boldsymbol{y}|u}\left[\frac{1}{2}\big(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\big)\mathrm{vec}\big((\boldsymbol{y}-\boldsymbol{\mu})(\boldsymbol{y}-\boldsymbol{\mu})'\big)\right]\\ &\qquad\left. + u^2\mathbb{E}_{\boldsymbol{y}|u}\left[-\frac{1}{2}\big(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\big)\mathrm{vec}(\boldsymbol{\gamma}\boldsymbol{\gamma}')\right] + u\log u\mathbb{E}_{\boldsymbol{y}|u}\left[\frac{1}{2}\big(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\big)\mathrm{vec}(\boldsymbol{\gamma}\boldsymbol{\gamma}')\right]\right]\\ &= \frac{n}{2}\big(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\big)\mathbb{E}_u\left[-\frac{1}{u^2}\underbrace{\mathbb{E}_{\boldsymbol{y}|u}\big[\mathrm{vec}\big((\boldsymbol{y}-\boldsymbol{\mu})(\boldsymbol{y}-\boldsymbol{\mu})'\big)\big]}_{\mathrm{vec}(u\boldsymbol{\Sigma}+u^2\boldsymbol{\gamma}\boldsymbol{\gamma}')} + \frac{1}{u}\log u\underbrace{\mathbb{E}_{\boldsymbol{y}|u}\big[\mathrm{vec}\big((\boldsymbol{y}-\boldsymbol{\mu})(\boldsymbol{y}-\boldsymbol{\mu})'\big)\big]}_{\mathrm{vec}(u\boldsymbol{\Sigma}+u^2\boldsymbol{\gamma}\boldsymbol{\gamma}')}\right.\\ &\qquad\left. - u^2\mathrm{vec}(\boldsymbol{\gamma}\boldsymbol{\gamma}') + u\log u\,\mathrm{vec}(\boldsymbol{\gamma}\boldsymbol{\gamma}')\right]\\ &= \frac{n}{2}\big(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\big)\mathbb{E}_u\left[-u\,\mathrm{vec}(\boldsymbol{\Sigma}+2u\boldsymbol{\gamma}\boldsymbol{\gamma}') + \log u\,\mathrm{vec}(\boldsymbol{\Sigma}+2u\boldsymbol{\gamma}\boldsymbol{\gamma})\right]\\ &= \frac{n}{2}\big(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\big)\mathbb{E}_u\left[(\log u - u)\mathrm{vec}(\boldsymbol{\Sigma}+2u\boldsymbol{\gamma}\boldsymbol{\gamma})\right].\end{aligned}$$

B6.3 Third term of Fisher information matrix

Let $\boldsymbol{z}_i = \boldsymbol{A}^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu})$ where $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}^\top$, $\boldsymbol{Z} = \begin{pmatrix} \boldsymbol{z}_1 & \cdots & \boldsymbol{z}_n \end{pmatrix}^\top$ and $\boldsymbol{z} = \boldsymbol{A}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})$ where $\boldsymbol{y}$ has the same distribution as $\boldsymbol{y}_i$. The first order derivatives of the complete data log-likelihood in terms of $\boldsymbol{z}_i$'s while ignoring additive constants with respect to $\boldsymbol{u}$ can be written as

$$\begin{aligned}
\frac{\partial \ell_c}{\partial \boldsymbol{\mu}} &= \boldsymbol{\Sigma}^{-1} \sum_{i=1}^n \frac{1}{u_i}(\boldsymbol{y}_i - \boldsymbol{\mu}) = \boldsymbol{A}^{-\top} \sum_{i=1}^n \frac{1}{u_i}\boldsymbol{z}_i, \\
\frac{\partial \ell_c}{\partial \boldsymbol{\gamma}} &= \boldsymbol{\Sigma}^{-1} \sum_{i=1}^n u_i \boldsymbol{\gamma} = \boldsymbol{A}^{-\top} \sum_{i=1}^n u_i \boldsymbol{\gamma}_z, \\
\frac{\partial \ell_c}{\partial \nu} &= \sum_{i=1}^n \log u_i - \sum_{i=1}^n u_i
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial \ell_c}{\partial \text{vec}\boldsymbol{\Sigma}_u} &= \frac{1}{2}\text{vec}\left(\boldsymbol{\Sigma}^{-1} \left[\sum_{i=1}^n \frac{1}{u_i}(\boldsymbol{y}_i - \boldsymbol{\mu})(\boldsymbol{y}_i - \boldsymbol{\mu})' + \sum_{i=1}^n u_i \boldsymbol{\gamma}\boldsymbol{\gamma}' \right] \boldsymbol{\Sigma}^{-1} \right) \\
&= \frac{1}{2}\text{vec}\left(\boldsymbol{A}^{-\top} \left[\sum_{i=1}^n \frac{1}{u_i}\boldsymbol{z}_i\boldsymbol{z}_i' + \sum_{i=1}^n u_i \boldsymbol{\gamma}_z\boldsymbol{\gamma}_z' \right] \boldsymbol{A}^{-1} \right) \\
&= \frac{1}{2}(\boldsymbol{A}^{-\top} \otimes \boldsymbol{A}^{-\top}) \left(\sum_{i=1}^n \frac{1}{u_i}\boldsymbol{z}_i \otimes \boldsymbol{z}_i + \sum_{i=1}^n u_i \boldsymbol{\gamma}_z \otimes \boldsymbol{\gamma}_z \right),
\end{aligned}$$

where it is sufficient to obtain derivatives with respect to $\text{vec}\boldsymbol{\Sigma}_u$ since the formulas to obtain derivatives with respect to $\text{vech}\boldsymbol{\Sigma}$ are given in Section A7.1 and A7.2.

$(\boldsymbol{\mu}, \boldsymbol{\mu})$ entry:

$$\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\left(\frac{\partial \ell_c}{\partial \boldsymbol{\mu}}\right)\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\left(\frac{\partial \ell_c}{\partial \boldsymbol{\mu}^\top}\right) = \boldsymbol{A}^{-\top}\left[\sum_{i=1}^n \mathbb{E}_{\boldsymbol{u}|\boldsymbol{Z}}\left[\tfrac{1}{u_i}\right]^2 \boldsymbol{z}_i\boldsymbol{z}_i^\top\right]\boldsymbol{A}^{-1}.$$

Taking expectation with respect to $\boldsymbol{Y}$,

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{Y}}\left[\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\left(\frac{\partial \ell_c}{\partial \boldsymbol{\mu}}\right)\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\left(\frac{\partial \ell_c}{\partial \boldsymbol{\mu}^\top}\right)\right] \\
&= n\boldsymbol{A}^{-\top}\left[\int_{\mathbb{R}^d}\mathbb{E}_{u|\boldsymbol{z}}\left[\tfrac{1}{u}\right]^2\boldsymbol{z}\boldsymbol{z}^\top f_{\boldsymbol{Z}}(\boldsymbol{z})\,d\boldsymbol{z}\right]\boldsymbol{A}^{-1} \\
&= n\boldsymbol{A}^{-\top}\left[\int_0^\infty \mathbb{E}_{u|r}\left[\tfrac{1}{u}\right]^2 g(r)\underbrace{\int_{\|\boldsymbol{z}\|=r}\boldsymbol{z}\boldsymbol{z}^\top\exp(\boldsymbol{\gamma}_z\boldsymbol{z})\,dS}_{\mathscr{M}_r^{(2)}(\boldsymbol{\gamma}_z)}\,dr\right]\boldsymbol{A}^{-1}.
\end{aligned}$$

where $\mathscr{M}_r^{(k)}(\boldsymbol{s})$ for $k = 0, \ldots, 4$ are defined in Section B5.1.

$(\boldsymbol{\gamma}, \boldsymbol{\gamma})$ entry:

$$\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\left(\frac{\partial \ell_c}{\partial \boldsymbol{\gamma}}\right)\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\left(\frac{\partial \ell_c}{\partial \boldsymbol{\gamma}^\top}\right) = n\mathbb{E}_{u|\boldsymbol{z}}[u]^2\boldsymbol{A}^{-\top}\boldsymbol{\gamma}_z\boldsymbol{\gamma}_z^\top\boldsymbol{A}^{-1}.$$

Taking expectation with respect to $\boldsymbol{Y}$

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{Y}}\left[\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\left(\frac{\partial \ell_c}{\partial \boldsymbol{\gamma}}\right)\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\left(\frac{\partial \ell_c}{\partial \boldsymbol{\gamma}^\top}\right)\right] \\
&= n\left[\int_0^\infty \mathbb{E}_{u|r}[u]^2 g(r)\mathscr{M}_r^{(0)}(\boldsymbol{\gamma}_z)\,dr\right]\boldsymbol{A}^{-\top}\boldsymbol{\gamma}_z\boldsymbol{\gamma}_z^\top\boldsymbol{A}^{-1}.
\end{aligned}$$

$(\mathbf{vech}\boldsymbol{\Sigma}, \mathbf{vech}\boldsymbol{\Sigma})$ entry:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\left(\frac{\partial \ell_c}{\partial \mathrm{vec}\boldsymbol{\Sigma}_u}\right)\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\left(\frac{\partial \ell_c}{\partial \mathrm{vec}\boldsymbol{\Sigma}_u'}\right) \\
&= \frac{1}{4}\left(\boldsymbol{A}^{-\top}\otimes\boldsymbol{A}^{-\top}\right)\left(\sum_{i=1}^n \mathbb{E}_{\boldsymbol{u}|\boldsymbol{Z}}\left[\tfrac{1}{u_i}\right]\boldsymbol{z}_i\otimes\boldsymbol{z}_i + \sum_{i=1}^n \mathbb{E}_{\boldsymbol{u}|\boldsymbol{Z}}[u_i]\boldsymbol{\gamma}_z\otimes\boldsymbol{\gamma}_z\right) \\
&\qquad\left(\sum_{i=1}^n \mathbb{E}_{\boldsymbol{u}|\boldsymbol{Z}}\left[\tfrac{1}{u_i}\right]\boldsymbol{z}_i\otimes\boldsymbol{z}_i + \sum_{i=1}^n \mathbb{E}_{\boldsymbol{u}|\boldsymbol{Z}}[u_i]\boldsymbol{\gamma}_z\otimes\boldsymbol{\gamma}_z\right)^\top\left(\boldsymbol{A}^{-1}\otimes\boldsymbol{A}^{-1}\right).
\end{aligned}$$

Taking expectation with respect to $\boldsymbol{Y}$,

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{Y}}\left[\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\left(\frac{\partial \ell_c}{\partial \mathrm{vec}\boldsymbol{\Sigma}_u}\right)\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\left(\frac{\partial \ell_c}{\partial \mathrm{vec}\boldsymbol{\Sigma}_u^\top}\right)\right] \\
&= \frac{n}{4}\big(\boldsymbol{A}^{-\top}\otimes\boldsymbol{A}^{-\top}\big)\mathbb{E}_{\boldsymbol{z}}\Big[\mathbb{E}_{u|\boldsymbol{z}}\big[\tfrac{1}{u}\big]^2\boldsymbol{z}\boldsymbol{z}'\otimes\boldsymbol{z}\boldsymbol{z}' + \mathbb{E}_{u|\boldsymbol{z}}[u]^2\boldsymbol{\gamma}_z\boldsymbol{\gamma}_z'\otimes\boldsymbol{\gamma}_z\boldsymbol{\gamma}_z' \\
&\qquad + \mathbb{E}_{u|\boldsymbol{z}}\big[\tfrac{1}{u}\big]\mathbb{E}_{u|\boldsymbol{z}}[u]\big\{(\boldsymbol{z}\otimes\boldsymbol{z})(\boldsymbol{\gamma}_z'\otimes\boldsymbol{\gamma}_z') + (\boldsymbol{\gamma}_z\otimes\boldsymbol{\gamma}_z)(\boldsymbol{z}'\otimes\boldsymbol{z}')\big\}\Big]\big(\boldsymbol{A}^{-1}\otimes\boldsymbol{A}^{-1}\big) \\
&= \frac{n}{4}\big(\boldsymbol{A}^{-\top}\otimes\boldsymbol{A}^{-\top}\big)\int_0^\infty g(r)\Big[\mathbb{E}_{u|r}\big[\tfrac{1}{u}\big]^2\mathscr{M}_r^{(4)}(\boldsymbol{\gamma}_z) + \mathbb{E}_{u|r}[u]^2\mathscr{M}_r^{(0)}(\boldsymbol{\gamma}_z)\boldsymbol{\gamma}_z\boldsymbol{\gamma}_z'\otimes\boldsymbol{\gamma}_z\boldsymbol{\gamma}_z' \\
&\quad + \mathbb{E}_{u|r}\big[\tfrac{1}{u}\big]\mathbb{E}_{u|r}[u]\big\{\mathrm{vec}\big(\mathscr{M}_r^{(2)}(\boldsymbol{\gamma}_z)\big)(\boldsymbol{\gamma}_z'\otimes\boldsymbol{\gamma}_z') + (\boldsymbol{\gamma}_z\otimes\boldsymbol{\gamma}_z)\mathrm{vec}\big(\mathscr{M}_r^{(2)}(\boldsymbol{\gamma}_z)\big)'\big\}\Big]dr\big(\boldsymbol{A}^{-1}\otimes\boldsymbol{A}^{-1}\big).
\end{aligned}$$

(ν,ν) **entry:**

$$\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\left(\frac{\partial \ell_c}{\partial \nu}\right)^2 = \sum_{i=1}^n \mathbb{E}_{\boldsymbol{u}|\boldsymbol{Z}}[\log u_i]^2 + \sum_{i=1}^n \mathbb{E}_{\boldsymbol{u}|\boldsymbol{Z}}[u_i]^2 - 2\sum_{i=1}^n \mathbb{E}_{\boldsymbol{u}|\boldsymbol{Z}}[\log u_i]\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Z}}[u_i].$$

Taking expectation with respect to $\boldsymbol{Y}$,

$$\mathbb{E}_{\boldsymbol{Y}}\left[\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\left(\frac{\partial \ell_c}{\partial \nu}\right)^2\right] = n\int_0^\infty g(r)\big(\mathbb{E}_{u|r}[\log u] - \mathbb{E}_{u|r}[u]\big)^2\mathscr{M}_r^{(0)}(\boldsymbol{\gamma}_z)dr.$$

$(\boldsymbol{\mu},\boldsymbol{\gamma})$ **entry:**

$$\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\left(\frac{\partial \ell_c}{\partial \boldsymbol{\mu}}\right)\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\left(\frac{\partial \ell_c}{\partial \boldsymbol{\gamma}^\top}\right) = -\boldsymbol{A}^{-\top}\left[\sum_{i=1}^n \mathbb{E}_{\boldsymbol{u}|\boldsymbol{Z}}\left[\tfrac{1}{u_i}\right]\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Z}}[u_i]\boldsymbol{z}_i\right]\boldsymbol{\gamma}_z^\top\boldsymbol{A}^{-1}.$$

Taking expectation with respect to $\boldsymbol{Y}$,

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{Y}}\left[\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\left(\frac{\partial \ell_c}{\partial \boldsymbol{\mu}}\right)\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\left(\frac{\partial \ell_c}{\partial \boldsymbol{\gamma}^\top}\right)\right] \\
&= -n\boldsymbol{A}^{-\top}\left[\int_0^\infty \mathbb{E}_{u|r}\big[\tfrac{1}{u}\big]\mathbb{E}_{u|r}[u]g(r)\mathscr{M}_r^{(1)}(\boldsymbol{\gamma}_z)\,dr\right]\boldsymbol{\gamma}_z^\top\boldsymbol{A}^{-1}.
\end{aligned}$$

$(\mathbf{vech}\boldsymbol{\Sigma}, \boldsymbol{\mu})$ entry:

$$\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\left(\frac{\partial \ell_c}{\partial \mathrm{vec}\boldsymbol{\Sigma}_u}\right)\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\left(\frac{\partial \ell_c}{\partial \boldsymbol{\mu}^\top}\right)$$
$$= \frac{1}{2}\left(\boldsymbol{A}^{-\top}\otimes\boldsymbol{A}^{-\top}\right)\left(\sum_{i=1}^n \mathbb{E}_{\boldsymbol{u}|\boldsymbol{Z}}\left[\tfrac{1}{u_i}\right]\boldsymbol{z}_i\otimes\boldsymbol{z}_i + \sum_{i=1}^n \mathbb{E}_{\boldsymbol{u}|\boldsymbol{Z}}[u_i]\boldsymbol{\gamma}_z\otimes\boldsymbol{\gamma}_z\right)\left(\sum_{i=1}^n \mathbb{E}_{\boldsymbol{u}|\boldsymbol{Z}}\left[\tfrac{1}{u_i}\right]\boldsymbol{z}_i\right)^\top \boldsymbol{A}^{-1}.$$

Taking expectation with respect to $\boldsymbol{Y}$,

$$\mathbb{E}_{\boldsymbol{Y}}\left[\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\left(\frac{\partial \ell_c}{\partial \mathrm{vec}\boldsymbol{\Sigma}_u}\right)\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\left(\frac{\partial \ell_c}{\partial \boldsymbol{\mu}^\top}\right)\right]$$
$$= \frac{n}{2}\left(\boldsymbol{A}^{-\top}\otimes\boldsymbol{A}^{-\top}\right)\mathbb{E}_{\boldsymbol{z}}\left[\mathbb{E}_{u|\boldsymbol{z}}\left[\tfrac{1}{u}\right]^2\boldsymbol{z}\otimes\boldsymbol{z}\boldsymbol{z}' + \mathbb{E}_{u|\boldsymbol{z}}\left[\tfrac{1}{u}\right]\mathbb{E}_{u|\boldsymbol{z}}[u](\boldsymbol{\gamma}_z\otimes\boldsymbol{\gamma}_z)\boldsymbol{z}'\right]\boldsymbol{A}^{-1}$$
$$= \frac{n}{2}\left(\boldsymbol{A}^{-\top}\otimes\boldsymbol{A}^{-\top}\right)\int_0^\infty g(r)\left[\mathbb{E}_{u|r}\left[\tfrac{1}{u}\right]^2\mathscr{M}_r^{(3)}(\boldsymbol{\gamma}_z) + \mathbb{E}_{u|r}\left[\tfrac{1}{u}\right]\mathbb{E}_{u|r}[u](\boldsymbol{\gamma}_z\otimes\boldsymbol{\gamma}_z)\mathscr{M}_r^{(1)}(\boldsymbol{\gamma}_z)^\top\right]dr\,\boldsymbol{A}^{-1}.$$

$(\boldsymbol{\mu}, \nu)$ entry:

$$\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\left(\frac{\partial \ell_c}{\partial \boldsymbol{\mu}}\right)\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\left(\frac{\partial \ell_c}{\partial \nu}\right) = \boldsymbol{A}^{-\top}\sum_{i=1}^n \mathbb{E}_{\boldsymbol{u}|\boldsymbol{Z}}\left[\tfrac{1}{u_i}\right]\left(\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Z}}[\log u_i] - \mathbb{E}_{\boldsymbol{u}|\boldsymbol{Z}}[u_i]\right)\boldsymbol{z}_i\,.$$

Taking expectation with respect to $\boldsymbol{Y}$,

$$\mathbb{E}_{\boldsymbol{Y}}\left[\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\left(\frac{\partial \ell_c}{\partial \boldsymbol{\mu}}\right)\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\left(\frac{\partial \ell_c}{\partial \nu}\right)\right]$$
$$= n\boldsymbol{A}^{-\top}\int_0^\infty \mathbb{E}_{u|\boldsymbol{z}}\left[\tfrac{1}{u}\right]\left(\mathbb{E}_{u|\boldsymbol{z}}[\log u] - \mathbb{E}_{u|\boldsymbol{z}}[u]\right)g(r)\mathscr{M}_r^{(1)}(\boldsymbol{\gamma}_z)\,dr\,.$$

$(\mathbf{vech}\boldsymbol{\Sigma}, \boldsymbol{\gamma})$ entry:

$$\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\left(\frac{\partial \ell_c}{\partial \mathrm{vec}\boldsymbol{\Sigma}_u}\right)\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\left(\frac{\partial \ell_c}{\partial \boldsymbol{\gamma}^\top}\right)$$
$$= -\frac{1}{2}\left(\boldsymbol{A}^{-\top}\otimes\boldsymbol{A}^{-\top}\right)\left(\sum_{i=1}^n \mathbb{E}_{\boldsymbol{u}|\boldsymbol{Z}}\left[\tfrac{1}{u_i}\right]\boldsymbol{z}_i\otimes\boldsymbol{z}_i + \sum_{i=1}^n \mathbb{E}_{\boldsymbol{u}|\boldsymbol{Z}}[u_i]\boldsymbol{\gamma}_z\otimes\boldsymbol{\gamma}_z\right)\left(\sum_{i=1}^n \mathbb{E}_{\boldsymbol{u}|\boldsymbol{Z}}[u_i]\right)\boldsymbol{\gamma}_z'\boldsymbol{A}^{-1}.$$

Taking expectation with respect to $\boldsymbol{Y}$,

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{Y}}\left[\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\left(\frac{\partial \ell_c}{\partial \mathrm{vec}\boldsymbol{\Sigma}_u}\right)\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\left(\frac{\partial \ell_c}{\partial \boldsymbol{\gamma}^\top}\right)\right] \\
&= -\frac{n}{2}\big(\boldsymbol{A}^{-\top}\otimes\boldsymbol{A}^{-\top}\big)\mathbb{E}_{\boldsymbol{z}}\Big[\mathbb{E}_{u|\boldsymbol{z}}\big[\tfrac{1}{u}\big]\mathbb{E}_{u|\boldsymbol{z}}[u]\boldsymbol{z}\otimes\boldsymbol{z} + \mathbb{E}_{u|\boldsymbol{z}}[u]^2\boldsymbol{\gamma}_z\otimes\boldsymbol{\gamma}_z\Big]\boldsymbol{\gamma}_z'\boldsymbol{A}^{-1} \\
&= -\frac{n}{2}\big(\boldsymbol{A}^{-\top}\otimes\boldsymbol{A}^{-\top}\big)\int_0^\infty g(r)\Big[\mathbb{E}_{u|r}\big[\tfrac{1}{u}\big]\mathbb{E}_{u|r}[u]\mathrm{vec}\mathscr{M}_r^{(2)}(\boldsymbol{\gamma}_z) + \mathbb{E}_{u|r}[u]^2\mathscr{M}_r^{(0)}(\boldsymbol{\gamma}_z)\boldsymbol{\gamma}_z\otimes\boldsymbol{\gamma}_z\Big]dr\,\boldsymbol{\gamma}_z'\boldsymbol{A}^{-1}.
\end{aligned}$$

$(\mathbf{vech}\boldsymbol{\Sigma}, \nu)$ entry:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\left(\frac{\partial \ell_c}{\partial \mathrm{vec}\boldsymbol{\Sigma}_u}\right)\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\left(\frac{\partial \ell_c}{\partial \nu}\right) \\
&= \frac{1}{2}\Big(\boldsymbol{A}^{-\top}\otimes\boldsymbol{A}^{-\top}\Big)\mathrm{vec}\left(\sum_{i=1}^n \mathbb{E}_{\boldsymbol{u}|\boldsymbol{Z}}\Big[\tfrac{1}{u_i}\Big]\boldsymbol{z}_i\boldsymbol{z}_i' + \sum_{i=1}^n \mathbb{E}_{\boldsymbol{u}|\boldsymbol{Z}}[u_i]\boldsymbol{\gamma}_z\boldsymbol{\gamma}_z'\right)\left(\sum_{i=1}^n \mathbb{E}_{\boldsymbol{u}|\boldsymbol{Z}}[\log u_i] - \sum_{i=1}^n \mathbb{E}_{\boldsymbol{u}|\boldsymbol{Z}}[u_i]\right)^\top.
\end{aligned}$$

Taking expectation with respect to $\boldsymbol{Y}$,

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{Y}}\left[\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\left(\frac{\partial \ell_c}{\partial \mathrm{vec}\boldsymbol{\Sigma}_u}\right)\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\left(\frac{\partial \ell_c}{\partial \nu}\right)\right] \\
&= \frac{n}{2}\big(\boldsymbol{A}^{-\top}\otimes\boldsymbol{A}^{-\top}\big)\mathbb{E}_{\boldsymbol{z}}\Big[\mathbb{E}_{u|\boldsymbol{z}}\big[\tfrac{1}{u}\big]\big(\mathbb{E}_{u|\boldsymbol{z}}[\log u] - \mathbb{E}_{u|\boldsymbol{z}}[u]\big)(\boldsymbol{z}\otimes\boldsymbol{z}) \\
&\qquad\qquad + \mathbb{E}_{u|\boldsymbol{z}}[u]\big(\mathbb{E}_{u|\boldsymbol{z}}[\log u] - \mathbb{E}_{u|\boldsymbol{z}}[u]\big)\boldsymbol{\gamma}_z\otimes\boldsymbol{\gamma}_z\Big] \\
&= \frac{n}{2}\big(\boldsymbol{A}^{-\top}\otimes\boldsymbol{A}^{-\top}\big)\int_0^\infty g(r)\Big[\mathbb{E}_{u|r}\big[\tfrac{1}{u}\big]\big(\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Z}}[\log u] - \mathbb{E}_{\boldsymbol{u}|\boldsymbol{Z}}[u]\big)\mathrm{vec}\mathscr{M}_r^{(2)}(\boldsymbol{\gamma}_z) \\
&\qquad\qquad + \mathbb{E}_{u|r}[u]\big(\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Z}}[\log u] - \mathbb{E}_{\boldsymbol{u}|\boldsymbol{Z}}[u]\big)\mathscr{M}_r^{(0)}(\boldsymbol{\gamma}_z)\boldsymbol{\gamma}_z\otimes\boldsymbol{\gamma}_z\Big]dr\,.
\end{aligned}$$

$(\boldsymbol{\gamma}, \nu)$ entry:

$$\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\left(\frac{\partial \ell_c}{\partial \boldsymbol{\gamma}}\right)\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\left(\frac{\partial \ell_c}{\partial \nu}\right) = -\boldsymbol{A}^{-\top}\sum_{i=1}^n \mathbb{E}_{\boldsymbol{u}|\boldsymbol{Z}}[u_i]\big(\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Z}}[\log u_i] - \mathbb{E}_{\boldsymbol{u}|\boldsymbol{Z}}[u_i]\big)\boldsymbol{\gamma}_z\,.$$

Taking expectation with respect to $\boldsymbol{Y}$,

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{Y}}\left[\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\left(\frac{\partial \ell_c}{\partial \boldsymbol{\gamma}}\right)\mathbb{E}_{\boldsymbol{u}|\boldsymbol{Y}}\left(\frac{\partial \ell_c}{\partial \nu}\right)\right] \\
&= -n\boldsymbol{A}^{-\top}\int_0^\infty \mathbb{E}_{u|\boldsymbol{z}}[u]\big(\mathbb{E}_{u|\boldsymbol{z}}[\log u] - \mathbb{E}_{u|\boldsymbol{z}}[u]\big)g(r)\mathscr{M}_r^{(0)}(\boldsymbol{\gamma}_z)\,dr\,\boldsymbol{\gamma}_z\,.
\end{aligned}$$

CHAPTER C

Other Related Functions and Distributions

C1 Modified Bessel function of the second kind

The modified Bessel function of the second kind appears in the density function (1.36) and E-step of VG distribution in Section 2.1.1. Some useful asymptotic properties are presented to improve numeric stability when applying EM algorithms. See Abramowitz and Stegun [1] for more information.

The modified Bessel function of the second kind has the following integral representations:

$$K_\lambda(z) = \frac{1}{2}\int_0^\infty w^{\lambda-1}\exp\left(-\frac{z}{2}\left(\frac{1}{w}+w\right)\right)dw,$$
$$K_\lambda(z) = \int_0^\infty \exp(-z\cosh t)\cosh(\lambda t)dt$$

for $z > 0$. The following integral formula is useful to obtain the density of the GH distribution,

$$\int_0^\infty w^{\lambda-1}\exp\left(-\frac{1}{2}\left(\frac{\chi}{w}+\psi w\right)\right)dw = 2\left(\frac{\chi}{\psi}\right)^{\frac{\lambda}{2}}K_\lambda\left(\sqrt{\chi\psi}\right)$$

where λ, χ, and ψ satisfy the parameter conditions for the GIG distribution in (1.20).

Some symmetry properties includes

$$K_{-\lambda}(z) = K_\lambda(z),$$
$$K^{(1,0)}_{-\lambda}(z) = -K^{(1,0)}_\lambda(z)$$

where $K_\lambda^{(1,0)}(\omega) = \frac{\partial}{\partial\alpha} K_\alpha(\omega)\big|_{\alpha=\lambda}$.

Some asymptotic properties when $\lambda > 0$ is fixed and $z \to 0$ includes

$$\begin{aligned}
K_0(z) &\sim -\big(\log\big(\tfrac{z}{2}\big) + \Upsilon\big) \\
&\sim -\log(z), \\
K_\lambda(z) &\sim \frac{1}{2}\left(\Gamma(\lambda)\Big(\frac{z}{2}\Big)^{-\lambda} + \Gamma(-\lambda)\Big(\frac{z}{2}\Big)^{\lambda}\right) \text{ for non-integer } \lambda \\
&\sim 2^{\lambda-1}\Gamma(\lambda) z^{-\lambda}, \\
K_\lambda^{(1,0)}(z) &\sim 2^{\lambda-1}\Gamma(\lambda) z^{-\lambda}\big(\psi(\lambda) - \log(\tfrac{z}{2})\big), \\
K_\lambda^{(2,0)}(z) &\sim 2^{\lambda-1}\Gamma(\lambda) z^{-\lambda}\Big[\psi'(\lambda) + \big(\psi(\lambda) - \log\big(\tfrac{z}{2}\big)\big)^2\Big]
\end{aligned}$$

where Υ represents a Euler-Mascheroni constant, and $\psi(x) = \frac{\partial}{\partial x}\log\Gamma(x)$ represents a digamma function.

Some asymptotic properties when $z > 0$ is fixed and $\lambda \to 0$ includes

$$\begin{aligned}
K_\lambda(z) &\sim \int_0^\infty \exp(-z\cosh(t))\, dt, \\
K_\lambda^{(1,0)}(z) &\sim \lambda \int_0^\infty t^2 \exp(-z\cosh(t))\, dt, \\
K_\lambda^{(2,0)} &\sim \int_0^\infty t^2 \exp(-z\cosh(t))\, dt.
\end{aligned}$$

The asymptotic property when $z \to \infty$ is given by

$$K_\lambda(z) \sim \sqrt{\frac{\pi}{2z}} e^{-z}.$$

Note that $K_{\frac{1}{2}}(z) = \sqrt{\frac{\pi}{2z}} e^{-z}$.

C2 Student's t distribution

Setting $\psi = 0$, $\lambda = -\upsilon/2$, and $\chi = \upsilon$ from the GH distribution in (1.20) gives us the multivariate skewed Student's t distribution with density function

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{\upsilon^{\frac{\upsilon}{2}}(\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma})^{\frac{\upsilon+d}{2}}}{(2\pi)^{\frac{d}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}\Gamma\left(\frac{\upsilon}{2}\right)2^{\frac{\upsilon}{2}-1}} \times \frac{K_{\frac{\upsilon+d}{2}}\left(\sqrt{(\upsilon+z^2)\boldsymbol{\gamma}\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}}\right)e^{(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}}}{\left(\sqrt{(\upsilon+z^2)\boldsymbol{\gamma}\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}}\right)^{\frac{\upsilon+d}{2}}}$$

where z^2 represents the Mahalanobis distance in (1.35), and for the symmetric case as $\boldsymbol{\gamma} \to \mathbf{0}$

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{\upsilon^{\frac{\upsilon}{2}}\Gamma\left(\frac{\upsilon+d}{2}\right)}{\pi^{\frac{d}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}\Gamma\left(\frac{\upsilon}{2}\right)(\upsilon+z^2)^{\frac{\upsilon+d}{2}}}.$$

It is interesting to note that this parametrisation corresponds to the case where the mixing variable follows $\mathcal{IG}(\alpha = \frac{\upsilon}{2}, \beta = \frac{\upsilon}{2})$ which has expected value of $\frac{\upsilon}{\upsilon-2}$ when $\upsilon > 2$, but the expected value does not exist when $\upsilon \leq 2$. Additionally, from the variance of the mixture representation in (1.32), the variance of the symmetric Student's t distribution does not exist when $\upsilon \leq 2$, while for the skewed case the variance does not exist when $\upsilon \leq 4$ since it involves the variance of the mixing distribution.

Alternatively, choosing the mixing variable to instead follow $\mathcal{IG}(\alpha = \frac{\upsilon}{2}, \beta = \frac{\upsilon}{2} - 1)$ for $\upsilon > 2$ and gives the the expected value of 1. This is equivalent to setting $\psi = 0$, $\lambda = -\upsilon/2$, and $\chi = \upsilon - 2$ from the GH distribution

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{(\upsilon-2)^{\frac{\upsilon}{2}}(\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma})^{\frac{\upsilon+d}{2}}}{(2\pi)^{\frac{d}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}\Gamma\left(\frac{\upsilon}{2}\right)2^{\frac{\upsilon}{2}-1}} \times \frac{K_{\frac{\upsilon+d}{2}}\left(\sqrt{(\upsilon-2+z^2)\boldsymbol{\gamma}\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}}\right)e^{(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}}}{\left(\sqrt{(\upsilon-2+z^2)\boldsymbol{\gamma}\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}}\right)^{\frac{\upsilon+d}{2}}} \tag{C.1}$$

and for the symmetric case

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{(\upsilon-2)^{\frac{\upsilon}{2}}\Gamma\left(\frac{\upsilon+d}{2}\right)}{\pi^{\frac{d}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}\Gamma\left(\frac{\upsilon}{2}\right)(\upsilon-2+z^2)^{\frac{\upsilon+d}{2}}}.$$

The Student's t random variable using this parametrisation is denoted by $\boldsymbol{Y} \sim t_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}, \upsilon)$ and is used for the implementation of the VARMA-t model in Section 5.6.

C3 Generalised Gumbel distribution

The pdf of a generalised Gumbel (GG) distribution is given by

$$f_{\mathrm{GG}}(x) = \frac{m^m}{\sigma\Gamma(m)} \exp\left(m\frac{x-\mu}{\sigma} - m\exp\left(\frac{x-\mu}{\sigma}\right)\right),\ x \in \mathbb{R} \tag{C.2}$$

where $\mu \in \mathbb{R}$ is the location parameter, $\sigma > 0$ is the scale parameter, and $m > 0$ is the shape parameter. Note that we consider the reflected version of the GG distribution given in [2].

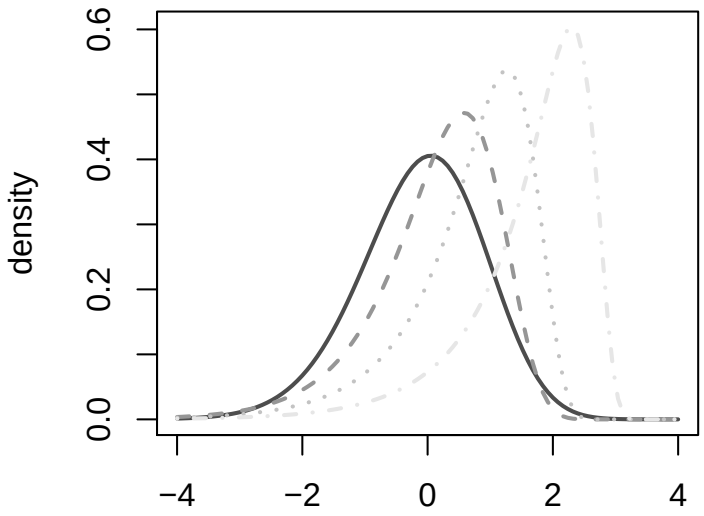


Figure C.1. Density plot of the GG distribution with $m = 10$ (solid black), $m = 1$ (dashed dark grey), $m = 0.5$ (dotted grey), and $m = 0.3$ (dot-dashed light grey) such that the location and scale is standardised using the mean and variance formula in equation (C.4).

We can easily generate GG random variables based on gamma random variables using the following theorem.

Theorem C3.1. *If* $X \sim \mathcal{G}(m, m)$, *then* $Y = \frac{\log X - \mu}{\sigma}$ *follows GG distribution with density function in equation (C.2).*

Proof. The idea of the proof is similar to the proof in Adeyemi [2]. □

Using this transformation, we can compute the CDF as

$$F_{\text{GG}}(x) = F_{\text{gamma}}(\exp(\mu + \sigma x); m, m), \ x \in \mathbb{R} \tag{C.3}$$

and quantile function

$$Q_{\text{GG}}(p) = \mu + \sigma \log Q_{\text{gamma}}(p; m, m), \ p \in [0, 1]$$

where $F_{\text{gamma}}(x; a, b)$ and $Q_{\text{gamma}}(p; a, b)$ are CDF and quantile function of $\mathcal{G}(a, b)$. In other words, we just need to calculate the CDF and quantiles of the gamma distribution.

The mean and variance of a GG random variable X is given by

$$\mathbb{E}(X) = \mu + \sigma(\psi(m) - \log m) \quad \text{and} \quad \text{Var}(X) = \sigma^2 \psi'(m). \tag{C.4}$$

C4 Double generalised gamma distribution

After fitting $\log|\hat{\mu}_n|$ using a GG distribution, we can deduce the distribution of $\hat{\mu}_n$ follows a double generalised gamma [62] using the following theorem.

Theorem C4.1. *Suppose that X follows a symmetric distribution such that $\log|X|$ follows GG distribution with pdf in equation (C.2), then X follows a double generalised gamma distribution with pdf*

$$\frac{\gamma\beta^\alpha}{2\Gamma(\alpha)}|x|^{\gamma\alpha-1}\exp(-\beta|x|^\alpha), \ x \in \mathbb{R} \tag{C.5}$$

where $\alpha > 0$, $\beta > 0$, and $\gamma > 0$.

Proof. Suppose that $Y = \log|X|$ follows a GG distribution, then Y has pdf

$$f_Y(y) \propto \exp\Big(\frac{my}{\sigma} - m\exp\Big(-\frac{\mu}{\sigma}\Big)\exp\Big(\frac{y}{\sigma}\Big)\Big)$$

where it is sufficient to consider the functional form. Now applying the transformation of the random variable $Y = \log W$ where $W = |X|$, we get the pdf for W,

$$\begin{aligned} f_W(w) &\propto \exp\left(\frac{m}{\sigma}\log w - m\exp\Big(-\frac{\mu}{\sigma}\Big)\exp\left(\frac{\log w}{\sigma}\right)\right)\frac{1}{w} \\ &\propto w^{m/\sigma-1}\exp\Big(-m\exp\Big(-\frac{\mu}{\sigma}\Big)w^{1/\sigma}\Big) \end{aligned}$$

which has the functional form of the generalised gamma distribution. Reverse the transformation of $W = |X|$ by reflecting the pdf at 0 gives us the result. □

By setting $\alpha = 2$, $\beta = 1/2$, and $\gamma = 1/2$ in equation (C.5) gives the standard normal density as a special case.

As the simulation results show that the GG distribution fits $\log|\hat{\mu}_n|$ reasonably well, we can model $\hat{\mu}_n$ using a double generalised gamma distribution. Suppose that $\hat{\boldsymbol{\Sigma}}$ is the scale parameter estimate of VG distribution with diagonals $\hat{\Sigma}_{ii}$ for $i = 1, \dots, d$. Then the SE of $\hat{\boldsymbol{\mu}}$ can be approximated using the formula

$$SE(\hat{\mu}_i) \approx \sqrt{\frac{\hat{\Sigma}_{ii}\beta^{-2/\gamma}\Gamma(\alpha + 2/\gamma)}{\Gamma(\alpha)}}$$

where $\alpha = m_{\text{GG}}$, $\beta = m_{\text{GG}} \exp(-\mu_{\text{GG}}/\sigma_{\text{GG}})$, $\gamma = 1/\sigma_{\text{GG}}$, and $(\mu_{\text{GG}}, \sigma_{\text{GG}}, m_{\text{GG}})$ are estimates of GG distribution extrapolated from Figure 3.7. Since the SE is sensitive to outliers, the MAD can instead be used as a robust measure of spread for $\hat{\boldsymbol{\mu}}$,

$$\text{MAD}(\hat{\mu}_i) \approx Q_{\text{gengamma}}(0.5) = \sqrt{\hat{\Sigma}_{ii}} Q_{\text{gamma}}(0.5; m_{\text{GG}}, m_{\text{GG}} \exp(-\mu_{\text{GG}}/\sigma_{\text{GG}}))^{\sigma_{\text{GG}}} \quad \text{(C.6)}$$

for $i = 1, \dots, d$ where $Q_{\text{gengamma}}(\cdot)$ represents the quantile function for the generalised gamma distribution where the pdf has functional form in (C.5) with support on $x > 0$.

Bibliography

[1] Abramowitz, M. and Stegun, I. A. (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York, ninth dover printing, tenth gpo printing edition.

[2] Adeyemi, S. (2002). On a generalization of the gumbel distribution. *Inter-Stat (London)*, 11(4):1–7.

[3] Ahmad, K. E., Moustafa, H. M., and Abd-Elrahman, A. M. (1997). Approximate Bayes estimation for mixtures of two Weibull distributions under type-2 censoring. *J. Statist. Comput. Simulation*, 58(3):269–285.

[4] Archambeau, C., Lee, J. A., and Verleysen, M. (2003). On convergence problems of the em algorithm for finite gaussian mixtures. In *In Proc. 11th European Symposium on Artificial Neural Networks*, pages 99–106.

[5] Baker, S. G. (1992). A simple method for computing the observed information matrix when using the EM algorithm with categorical data. *J. Comput. Graph. Statist.*, 1(1):63–76.

[6] Barndorff-Nielsen, O. (1978). Hyperbolic distributions and distributions on hyperbolae. *Scand. J. Statist.*, 5(3):151–157.

[7] Barndorff-Nielsen, O., Kent, J., and Sørensen, M. (1982). Normal variance-mean mixtures and z distributions. *Internat. Statist. Rev.*, 50(2):145–159.

[8] Barndorff-Nielsen, O. E. and Stelzer, R. (2005). Absolute moments of generalized hyperbolic distributions and approximate scaling of normal inverse Gaussian Lévy processes. *Scand. J. Statist.*, 32(4):617–637.

[9] Blumenson, L. (1960). A derivation of n-dimensional sperical coordinates. *The American Mathematical Monthly*, 67(1):63–66.

[10] Böhning, D. and Lindsay, B. G. (1988). Monotonicity of quadratic-approximation algorithms. *Ann. Inst. Statist. Math.*, 40(4):641–663.

[11] Boik, R. (2006). Lecture notes: Statistics 550 spring 2006.

[12] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *J. Econometrics*, 31(3):307–327.

[13] Bonato, M. (2011). Robust estimation of skewness and kurtosis in distributions with infinite higher moments. *Finance Research Letters*, 8(2):77 – 87.

[14] Boris Choy, S. and Chan, J. S. (2008). Scale mixtures distributions in statistical modelling. *Australian & New Zealand Journal of Statistics*, 50(2):135–146.

[15] Box, G. E. P. and Pierce, D. A. (1970). Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American Statistical Association*, 65(332):1509–1526.

[16] Breymann, W. and Lüthi, D. (2013). ghyp: A package on generalized hyperbolic distributions. *Available in http://cran. r-project. org/web/packages/ghyp/vignettes/Generalized_Hyperbolic_Distribution. pdf*.

[17] Celeux, G., Hurn, M., and Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *J. Amer. Statist. Assoc.*, 95(451):957–970.

[18] Chan, J. S. and Kuk, A. Y. (1997). Maximum likelihood estimation for probit-linear mixed models with correlated random effects. *Biometrics*, pages 86–97.

[19] Chan, J. S., Kuk, A. Y., and Yam, C. H. (2005). Monte carlo approximation through gibbs output in generalized linear mixed models. *Journal of Multivariate Analysis*, 94(2):300–312.

[20] Cheng, R. C. H. and Amin, N. A. K. (1983). Estimating parameters in continuous univariate distributions with a shifted origin. *J. Roy. Statist. Soc. Ser. B*, 45(3):394–403.

[21] Cheng, R. C. H. and Iles, T. C. (1987). Corrected maximum likelihood in nonregular problems. *J. Roy. Statist. Soc. Ser. B*, 49(1):95–101.

[22] Cheng, R. C. H. and Traylor, L. (1995). Non-regular maximum likelihood problems. *J. Roy. Statist. Soc. Ser. B*, 57(1):3–44. With discussion and a reply by the authors.

[23] Choy, S. T. B., Chen, C. W. S., and Lin, E. M. H. (2014). Bivariate asymmetric garch models with heavy tails and dynamic conditional correlations. *Quantitative Finance*, 14(7):1297–1313.

[24] Cohen, Jr., A. C. (1951). Estimating parameters of logarithmic-normal distributions by maximum likelihood. *J. Amer. Statist. Assoc.*, 46:206–212.

[25] Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1(2):223–236.

[26] de Frutos, R. F. and Serrano, G. R. (1997). A generalized least squares estimation method for invertible vector moving average models. *Economics Letters*, 57(2):149–156.

[27] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39(1):1–38. With discussion.

[28] Deuflhard, P. (2004). *Newton methods for nonlinear problems*, volume 35 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin. Affine invariance and adaptive algorithms.

[29] Efron, B. (2003). Second thoughts on the bootstrap. *Statist. Sci.*, 18(2):135–140. Silver anniversary of the bootstrap.

[30] Efron, B. and Tibshirani, R. J. (1993). *An introduction to the bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, New York.

[31] Embrechts, P. (1983). A property of the generalized inverse Gaussian distribution with some applications. *J. Appl. Probab.*, 20(3):537–544.

[32] Erdem, E. and Shi, J. (2011). Arma based approaches for forecasting the tuple of wind speed and direction. *Applied Energy*, 88(4):1405–1414.

[33] Fama, E. F. (1965). The Behavior of Stock-Market Prices. *The Journal of Business*, 38(1):34–105.

[34] Finlay, R. and Seneta, E. (2008). Stationary-increment variance-gamma and "t" models: Simulation and parameter estimation. *Int. Statist. Rev.*, 76(2):167–186.

[35] Fung, T. and Seneta, E. (2010). Modelling and estimation for bivariate financial returns. *Int. Statist. Rev.*, 78(1):117–133.

[36] Giesbrecht, F. and Kempthorne, O. (1976). Maximum likelihood estimation in the three-parameter lognormal distribution. *J. Roy. Statist. Soc. Ser. B*, 38(3):257–264.

[37] Gohberg, I. and Lerer, L. (1976). Resultants of matrix polynomials. *Bulletin of the American Mathematical Society*, 82(4):565–567.

[38] Gouriéroux, C. and Monfort, A. (2013). Revisiting identification in structural var models. *Journal of Banking & Finance*, 37(10):3843–3854.

[39] Gourieroux, C., Monfort, A., and Renault, E. (1989). Testing for common roots. *Econometrica: Journal of the Econometric Society*, pages 171–185.

[40] Gouriéroux, C. and Zakoïan, J.-M. (2015). On uniqueness of moving average representations of heavy-tailed stationary processes. *Journal of Time Series Analysis*, 36(6):876–887.

[41] Gradshteyn, I. and Ryzhik, I. M. (2007). Table of integrals, series, and products (academic, new york, 1965). *Google Scholar*.

[42] Graff, P. and Feroz, F. (2013). BAMBI: Blind Accelerated Multimodal Bayesian Inference. Astrophysics Source Code Library.

[43] Granger, C. W. and Joyeux, R. (1980). An introduction to long-memory time series models and fractional differencing. *Journal of time series analysis*, 1(1):15–29.

[44] Hamilton, J. D. (1994). *Time series analysis*, volume 2. Princeton university press Princeton.

[45] Hannan, E. and Deistler, M. (1988). The statistical theory of linear systemswiley. *New York*.

[46] He, Y. (2009). *Improving the EM algorithm for maximum likelihood inference*. PhD thesis, Purdue University.

[47] Heracleous, M. (2003). Volatility modeling using the student's t distribution.

[48] Hillmer, S. C. and Tiao, G. C. (1979). Likelihood function of stationary multiple autoregressive moving average models. *Journal of the American Statistical Association*, 74(367):652–660.

[49] Hosking, J. R. (1981). Fractional differencing. *Biometrika*, 68(1):165–176.

[50] Hossain, M., Kozubowski, T., and Podgorski, K. (2015). A novel weighted likelihood estimation with empirical bayes flavor. Working Paper.

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

[52] Hu, W. and Kercheval, A. N. (2008). The skewed t distribution for portfolio credit risk. In *Econometrics and risk management*, volume 22 of *Adv. Econom.*, pages 55–83. Emerald/JAI, Bingley.

[53] Ibragimov, I. A. and Khasminskii, R. Z. (1981a). Asymptotic behavior of statistical estimates of the location parameter for samples with unbounded density. *J. Sov. Math.*, 16(2):1035–1041.

[54] Ibragimov, I. A. and Khasminskii, R. Z. (1981b). *Statistical estimation*, volume 16 of *Applications of Mathematics*. Springer-Verlag, New York-Berlin. Asymptotic theory, Translated from the Russian by Samuel Kotz.

[55] Isserlis, L. (1918). On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, 12(1/2):134–139.

[56] Jørgensen, B. (1982). *Statistical properties of the generalized inverse Gaussian distribution*, volume 9 of *Lecture Notes in Statistics*. Springer-Verlag, New York-Berlin.

[57] Kawai, R. (2015). On the likelihood function of small time variance gamma Lévy processes. *Statistics*, 49(1):63–83.

[58] Klein, A., Mélard, G., and Spreij, P. (2005). On the resultant property of the fisher information matrix of a vector arma process. *Linear algebra and its applications*, 403:291–313.

[59] Koreisha, S. and Pukkila, T. (1990). A generalized least-squares approach for estimation of autoregressive moving-average models. *J. Time Ser. Anal.*, 11(2):139–151.

[60] Kotz, S., Kozubowski, T. J., and Podgrski, K. (2001). *The Laplace distribution and generalizations : a revisit with applications to communications, economics, engineering, and finance*. Birkhuser, Boston.

[61] Labelle, G. (2010). On extensions of the Newton-Raphson iterative scheme to arbitrary orders. In *22nd International Conference on Formal Power Series and Algebraic Combinatorics (FPSAC 2010)*, Discrete Math. Theor. Comput. Sci. Proc., AN, pages 845–856. Assoc. Discrete Math. Theor. Comput. Sci., Nancy.

[62] Lin, G. D. and Huang, J. S. (1997). The cube of a logistic distribution is indeterminate. *Austral. J. Statist.*, 39(3):247–252.

[63] Liu, C. (1997). ML estimation of the multivariate t distribution and the EM algorithm. *J. Multivariate Anal.*, 63(2):296–312.

[64] Liu, C. (1998). Information matrix computation from conditional information via normal approximation. *Biometrika*, 85(4):pp. 973–979.

[65] Liu, C. and Rubin, D. B. (1994). The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence. *Biometrika*, 81(4):633–648.

[66] Liu, C. and Rubin, D. B. (1995). ML estimation of the t distribution using EM and its extensions, ECM and ECME. *Statist. Sinica*, 5(1):19–39.

[67] Liu, S., Wu, H., and Meeker, W. Q. (2015). Understanding and addressing the unbounded "likelihood" problem. *Amer. Statist.*, 69(3):191–200.

[68] Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 44(2):226–233.

[69] Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer-Verlag, Berlin.

[70] Madan, D. B. and Seneta, E. (1987). Chebyshev polynomial approximations and characteristic function estimation. *J. Roy. Statist. Soc. Ser. B*, 49(2):163–169.

[71] Madan, D. B. and Seneta, E. (1990). The variance gamma (V.G.) model for share market returns. *J. Bus.*, 63(4):511–524.

[72] Magnus, J. R. and Neudecker, H. (1979). The commutation matrix: some properties and applications. *Ann. Statist.*, 7(2):381–394.

[73] Mandelbrot, B. B. (1963). The variation of certain speculative prices. *Journal of Business*, 36(4):394–419.

[74] McLachlan, G. J. and Krishnan, T. (2008). *The EM algorithm and extensions*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition.

[75] McNeil, A. J., Frey, R., and Embrechts, P. (2005). *Quantitative risk management*. Princeton Series in Finance. Princeton University Press, Princeton, NJ. Concepts, techniques and tools.

[76] Meng, X.-L. (1994). On the rate of convergence of the ECM algorithm. *Ann. Statist.*, 22(1):326–339.

[77] Meng, X.-L. and Rubin, D. B. (1991). Using em to obtain asymptotic variance-covariance matrices: The sem algorithm. *Journal of the American Statistical Association*, 86(416):899–909.

[78] Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika*, 80(2):267–278.

[79] Meng, X.-L. and van Dyk, D. (1997). The EM algorithm—an old folk-song sung to a fast new tune. *J. Roy. Statist. Soc. Ser. B*, 59(3):511–567. With discussion and a reply by the authors.

[80] Metaxoglou, K. and Smith, A. (2007). Maximum likelihood estimation of VARMA models using a state-space EM algorithm. *J. Time Ser. Anal.*, 28(5):666–685.

[81] Minka, T. P. (2000). Old and new matrix algebra useful for statistics. Technical report.

[82] Nakamoto, S. (2009). Bitcoin: A peer-to-peer electronic cash system.

[83] Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. In *Handbook of econometrics, Vol. IV*, volume 2 of *Handbooks in Econom.*, pages 2111–2245. North-Holland, Amsterdam.

[84] Nocedal, J. and Wright, S. J. (2006). *Numerical optimization*. Springer Series in Operations Research and Financial Engineering. Springer, New York, second edition.

[85] Oakes, D. (1999). Direct calculation of the information matrix via the EM algorithm. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 61(2):479–482.

[86] Orchard, T. and Woodbury, M. A. (1972). A missing information principle: theory and applications. pages 697–715.

[87] Pai, P.-F. and Lin, C.-S. (2005). A hybrid arima and support vector machines model in stock price forecasting. *Omega*, 33(6):497–505.

[88] Petersen, K. B. and Pedersen, M. S. (2012). The matrix cookbook.

[89] Podgórski, K. and Wallin, J. (2015). Maximizing leave-one-out likelihood for the location parameter of unbounded densities. *Ann. Inst. Statist. Math.*, 67(1):19–38.

[90] Protassov, R. S. (2004). EM-based maximum likelihood parameter estimation for multivariate generalized hyperbolic distributions with fixed λ. *Stat. Comput.*, 14(1):67–77.

[91] Rao, B. (1968). Estimation of the location of the cusp of a continuous density. *Ann. Math. Statist.*, 39:76–87.

[92] Reinsel, G. C., Basu, S., and Yap, S. F. (1992). Maximum likelihood estimators in the multivariate autoregressive moving-average model from a generalized least squares viewpoint. *J. Time Ser. Anal.*, 13(2):133–145.

[93] Richardson, L. F. (1911). The approximate arithmetical solution by finite differences of physical problems involving differential equations, with an application to the stresses in a masonry dam. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 210:307–357.

[94] Robert V. Hogg, Joeseph McKean, A. T. C. (2012). *Introduction to Mathematical Statistics*. 7th Edition. Pearson, 7 edition.

[95] Roy, A., McElroy, T. S., and Linton, P. (2014). Estimation of causal invertible varma models. *arXiv preprint arXiv:1406.4584*.

[96] Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592. With comments by R. J. A. Little and a reply by the author.

[97] Salakhutdinov, R. and Roweis, S. (2003). Adaptive overrelaxed bound optimization methods. In *Proceedings of the International Conference on Machine Learning*, volume 20, pages 664–671.

[98] Schonemann, P. H. (1985). On the formal differentiation of traces and determinants. *Multivariate Behavioral Research*, 20(2):113–139.

[99] Schruth, D. (2013). *caroline: A Collection of Database, Data Structure, Visualization, and Utility Functions for R.* R package version 0.7.6.

[100] Seo, B. and Kim, D. (2012). Root selection in normal mixture models. *Comput. Statist. Data Anal.*, 56(8):2454–2470.

[101] Shanno, D. F. (1970). Conditioning of quasi-Newton methods for function minimization. *Math. Comp.*, 24:647–656.

[102] Tjetjep, A. and Seneta, E. (2006). Skewed normal variance-mean models for asset pricing and the method of moments. *Int. Statist. Rev.*, 74(1):109–126.

[103] Tsay, R. S. (2005). *Analysis of financial time series*, volume 543. John Wiley & Sons.

[104] Tsay, R. S. (2014). *Multivariate time series analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ. With R and financial applications.

[105] Wang, Y. and Tsay, R. S. (2013). On diagnostic checking of vector arma-garch models with gaussian and student-t innovations. *Econometrics*, 1(1):1–31.

[106] White, M., Wen, J., Bowling, M., and Schuurmans, D. (2015). Optimal estimation of multivariate arma models. In Bonet, B. and Koenig, S., editors, *AAAI*, pages 3080–3086. AAAI Press.

[107] Wilson, G. T. (1973). The estimation of parameters in multivariate time series models. *J. Roy. Statist. Soc. Ser. B*, 35:76–85.

[108] Wraith, D. and Forbes, F. (2015). Location and scale mixtures of gaussians with flexible tail behaviour: Properties, inference and application to multivariate clustering. *Computational Statistics & Data Analysis*, 90:61–73.

[109] Wu, C.-F. J. (1983). On the convergence properties of the EM algorithm. *Ann. Statist.*, 11(1):95–103.

[110] Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American statistical Association*, 57(298):348–368.

[111] Zou, T., Lan, W., Wang, H., and Tsai, C.-L. (2017). Covariance regression analysis. *Journal of the American Statistical Association*, 112(517):266–281.