Spatio-Temporal Generalized Autoregressive Conditional Heteroskedasticity models

A circular approach to spatio-temporal estimation

Master's thesis in Statistics

Sondre Hølleland

Supervisors

Hans Arnfinn Karlsen

Bård Støve

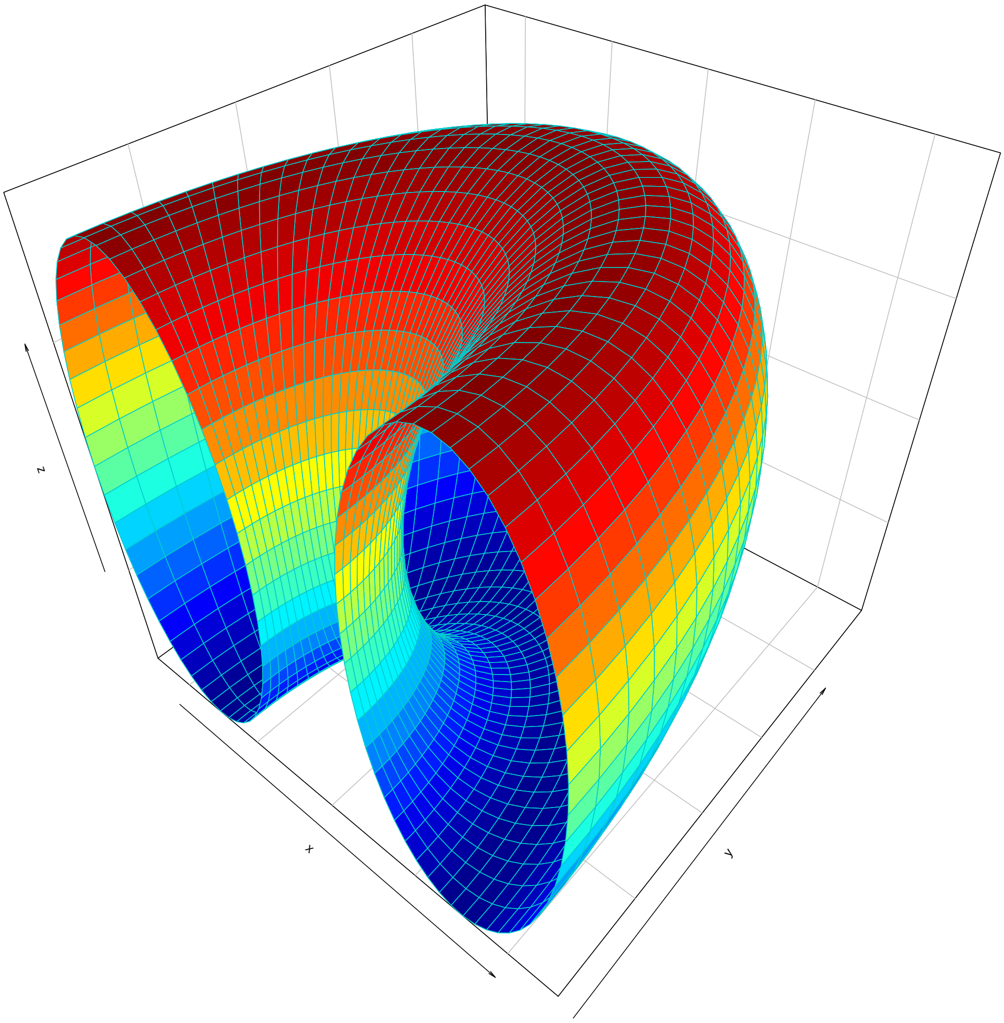
Department of Mathematics

University of Bergen

June 2016

Spatio-Temporal Generalized Autoregressive Conditional Heteroskedasticity models

A circular approach to spatio-temporal estimation



Sondre Hølleland

with supervisors

Hans Arnfinn Karlsen and Bård Støve

There should be no boundary to human endevour. However bad life may seem, while there is life, there is hope.

Stephen Hawking

Abstract

This thesis presents a spatio-temporal extension of the GARCH process with a specific spatial dependence structure. Different simulation and estimation methods are developed. Assuming a circular spatial structure at each time point, gives a closed and finite set of variables at each point in time, making the spatio-temporal process adapted in the temporal dimension. This assumption makes likelihood estimation trivial and we obtain an analytical expression for estimators – both using maximum likelihood and least squares estimation. On non-circular data, this procedure leads to biased estimates, but we suggest doing a parametric bootstrap bias correction, which turns out to be very effective and improve estimates substantially. We also suggest another approach to apply the circular model to non-circular data, by using a Gibbs sampler and an EM-algorithm.

Acknowledgements

First of all, thanks to my supervisors, Hans Arnfinn Karlsen and Bård Støve, for excellent guidance and for always pushing me in the right direction. Many thanks to my partner, Christi Elin M. Nedreås, for being patient with me and to my family for support. I would also like to like to thank my fellow students, Håvard Frøysa, Berent Lunde, Kristine Sivertsen and Sigve Dyrhovden, for moral support and interesting lunch break discussions.

Contents

1 Introduction **1**

2 Preliminaries **3**

2.1 Spatial Statistics . . . 3

2.1.1 Notation . . . 4

2.1.2 Examples: Gaussian CAR and SAR models . . . 6

2.2 Time Series . . . 7

2.2.1 ARMA . . . 8

2.3 GARCH . . . 8

2.3.1 Moments . . . 9

2.3.2 Stationarity . . . 10

2.3.3 ARMA representation . . . 11

2.3.4 Quasi Maximum Likelihood Estimation . . . 11

2.3.5 Asymptotic Results of the QMLE . . . 14

2.3.6 Applications . . . 14

2.4 Multivariate GARCH . . . 16

2.4.1 The BEKK model for MGARCH . . . 16

2.4.2 Asymptotic Results . . . 17

2.5 Circular AR(1) . . . 18

3 Spatio-Temporal GARCH process **23**

3.1 The process . . . 23

3.2 Properties of STGARCH . . . 24

3.2.1 Moments . . . 24

3.2.2 Stationarity . . . 25

3.2.3 ARMA Representation . . . 25

3.3 Boundary Issues . . . 26

3.3.1 Conditioning on the boundary . . . 26

3.3.2 The Circular model 27
3.4 Applications 28

4 Simulation 31
4.1 Vector notation 32
4.2 Initiating and Burn-In 34
4.3 Simulation Algorithm 34
4.3.1 Circular Simulation 34
4.3.2 Regular Simulation 35
4.4 Example: ACF and Marginal Density 36

5 Estimation 39
5.1 Maximum Likelihood Estimation 39
5.1.1 Likelihood derivation 40
5.1.2 Optimization of the Likelihood 42
5.1.3 Asymptotic Normality 44
5.2 Least Squares Estimation 49
5.2.1 The Method 50
5.3 MLE vs LSE 53

6 Empirical Experiments 55
6.1 Convergence efficiency of MLE 57
6.2 Circular vs Non-circular 57
6.3 Comparing settings 61
6.4 Student's t-distributed innovations 62
6.5 STGARCH(1,1) 64

7 Parametric bootstrap bias correction 67
7.1 Bootstrap Bias Correction 68
7.2 Example: STARCH(1) 68

8 Gibbs sampler and EM algorithm 73
8.1 Gibbs sampler 73
8.2 EM algorithm 75

9 Concluding Remarks 77

Bibliography 79

Appendices 83

A Spatial consistency 85

B Covariance matrix experiments 89

B.1 Testing of (5.14) 89

B.2 Convergence of (5.16) 91

C R-Code 93

C.1 Simulation algorithm 93

C.2 Estimation 95

C.3 Parametric Bootstrap Bias Correction 98

List of Tables

2.1 Results from estimating a GARCH$(1, 1)$ for the GIG stock returns. 16

6.1 Comparison of different methods and approaches in a 1D STARCH(1) 59

6.2 Comparison of different methods and approaches in a 2D STARCH(1) 60

6.3 Results from settings experiment 62

6.4 Results from estimation of STGARCH$(1, 1)$ 65

6.5 Correlation matrices from estimation of STGARCH$(1, 1)$ 66

7.1 Parametric Bootstrap bias correction experiment results 70

A.1 Results from expanding space experiment of 2D STARCH(1), with $\alpha_1 = 0.01$ . . 86

A.2 Standard deviations from expanding space experiment with $\alpha_1 = 0.01$ 86

A.3 Results from expanding space experiment of 2D STARCH(1), with $\alpha_1 = 0.1$ . . . 87

A.4 Standard deviations from expanding space experiment with $\alpha_1 = 0.1$ 87

B.1 Monte Carlo covariance matrix comparison: $5 \times 5 \times 1000$ 90

B.2 Monte Carlo covariance matrix comparison: $10 \times 10 \times 1000$ 90

B.3 Monte Carlo covariance matrix comparison: $20 \times 20 \times 1000$ 90

B.4 Resulting information matrices of STGARCH$(1, 1)$ for increasing T 92

List of Figures

2.1 Population growth in Norway by counties 5

2.2 Fitted GARCH$(1, 1)$ to Gaming Innovation Group stock data 15

2.3 Circular AR model illustration 19

2.4 ACF of Circular AR(1) 21

3.1 Conditioning on the boundary illustration 27

3.2 Snake biting its own tail analogy to a circular model 28

3.3 Circular spatio-temporal process illustration in one and two dimensions 29

3.4 Spatial slice of simulated STARCH 30

4.1 Simulated circular 1D STARCH(1) 3D plots 31

4.2 Simulated 1D circular STARCH(1) matrix plots 32

4.3 Regular simulation in 2D 35

4.4 Circular and regular ACF of 1D STARCH(1) 37

4.5 Marginal density of circular 1D STARCH(1) 38

5.1 Asymptotic normality of ML estimators in a 2D circular ARCH(1) 47

6.1 Neighbourhood structure in one and two dimensions used in experiments 56

6.2 Convergence rate of ML routines 58

6.3 Simulation – Estimation experiment scheme 59

6.4 Student’s t distributed innovations experiment results 63

7.1 Bias in circular estimation 69

7.2 Comparison of parametric BBC estimates and regular estimates 71

8.1 Gibbs procedure illustration 73

Chapter 1

Introduction

Autoregressive conditional hetereoskedasticity (ARCH) models were first introduced by Engle (1982). In his paper, he is using the models to estimate means and variances of inflation in the U.K. Bollerslev (1986) saw from empirical applications of the ARCH models that it seemed of «*immediate practical interest to extend the ARCH class of models to allow for both a longer memory and a more flexible lag structure*» (Bollerslev, 1986, pp. 308) and suggested the generalized ARCH(GARCH) models. Today, the main field of application for these models is finance and the models have become extremely popular, both among practitioners and academics, since the mid 1980s. In 2003, Robert F. Engle shared a Nobel Prize in Economics with Clive W.J. Granger, «*for methods of analyzing economic time series with time-varying volatility (ARCH)*»(Nobelprize.org, 2003).

Many contributions have been made to these models. Nelson (1990) found necessary and sufficient conditions for a (strictly) stationary, ergodic GARCH(1,1) process and Bougerol and Picard (1992) generalized the result for GARCH(p,q) models. The asymptotic properties of quasi maximum likelihood estimators in ARCH were first established by Weiss (1986) under finite fourth moment assumptions. Later Francq et al. (2004) proved consistency and asymptotic normality under weaker assumptions for the GARCH(p,q) models.

Engle and Kroner (1995) presented a multivariate GARCH, found sufficient constraints to guarantee the positive definiteness of the conditional covariance matrices and necessary and sufficient conditions for weak stationarity of the process. Comte and Lieberman (2003) developed asymptotic results under regularity conditions for the BEKK process presented by Engle and Kroner (1995).

Over the last four decades there have been made numerous extensions to the original model. There are actually so many extensions, all with their own acronym, that Bollerslev (2008) made an extensive list of what all the acronyms mean along with a short explanation. The list is 45 pages long. This confirms that the models became popular with academics.

In this thesis we present an extension of the well-known ARCH- and GARCH-models, which we call spatio-temporal GARCH (STGARCH). This extension was first suggested by Karlsen (2015b). We will be using the acronym ST(G)ARCH throughout this thesis. This should however not be confused with structural ARCH and smooth transition GARCH with acronyms STARCH and STGARCH, respectively (Bollerslev, 2008).

For readers not familiar with the standard (G)ARCH models, we present them in Chapter 2. We also introduce spatial statistical notation and give two examples of pure spatial models and some definitions from time series theory. We also introduce circular models by an example. In Chapter 3 we introduce the STGARCH model and derive some of its properties. After introducing the model, we make some comments on how to simulate from the model in Chapter 4 and in Chapter 5 we talk about estimation theory, developing estimators using both least squares– and quasi maximum likelihood estimation under a circular space assumption. We also derive asymptotic normality of the quasi maximum likelihood estimators. By Chapter 6 we have methods for simulating data and estimating parameters, so we conduct some Monte Carlo experiments with these methods. The main approach for estimation we suggest, assumes a circular spatial process, which in most cases will be an erroneous assumption leading to biased estimates. We therefore develop procedures to do a parametric bootstrap bias correction in Chapter 7. In Chapter 8 we discuss another possible approach to use the circular model on non-circular data, by a combination of Gibbs sampling and the EM-algorithm. Up till now, we have not managed to make this approach work satisfactory. We present it to suggest future work.

The thesis focuses on parameter estimation of these models under some simplifying assumptions. These simplifying assumptions are employed to reduce the number of parameters to a plausible situation without having to turn to Bayesian methods. Compared to the matrices of parameters in multivariate GARCH, the STGARCH manages with less parameters. In fact, we show in Chapter 5 that STGARCH is a parametrization of a multivariate GARCH defined by the BEKK represenation of Engle and Kroner (1995).

Chapter 2

Preliminaries

The goal of this chapter is to present the theory required for understanding the main part of the thesis. This will include headlines such as spatial statistics and time series.

We introduce spatial statistical notation, how to define neighbourhood structure and introduce the CAR- and SAR models as examples of spatial stochastic processes. Since spatio-temporal GARCH is a process that develops in the temporal dimension with spatial dependencies, pure spatial statistics is not that important. The essential part is to learn how to define neighbourhood structures in spatial statistics.

For time series, we define ARMA-, ARCH- and GARCH models. Since GARCH models are very important for this thesis, we will learn how to perform quasi maximum likelihood estimation of these models, give asymptotic results, applications and illustrate usage of these models by an example. We introduce the multivariate GARCH BEKK processes and present some asymptotic results related to these as well.

At the end of the chapter we consider a circular AR(1) process, as an example of circular models. This is to get the reader familiar with the concept of circular modelling. These kinds of models will be more thoroughly explained in the main part of the thesis.

2.1. Spatial Statistics

Spatial statistics is a vast discipline of statistical science and contributions and new developments are being made each year. In this section, the aim is not to give a book-length introduction to spatial statistics, but provide some knowledge about the subject. We introduce the CAR and SAR models as examples of spatial processes. Hopefully, this section will give us insight in the spatial aspect of the spatio-temporal GARCH (STGARCH) models, to be introduced in Chapter 3.

The main reference for this section is Cressie and Wikle (2011). The spatial world is fundamentally different from the temporal. First of all, the temporal dimension has a clear ordering and direction. Time goes forward and the order $0 = t_0 \leq t_1 \leq t_2 \leq \ldots \leq t_n$ is very natural. In space however, there is no preferred direction nor ordering.

According to Cressie and Wikle (2011) there are three different spatial processes. They call them

- Temporal snapshot
- Temporal aggregation
- Temporally frozen state

The temporal snapshot is exactly what you would expect. You measure something at different locations at the same time, giving you an instantaneous image of the process for the entire area at a specific time point. Aggregation can mean that you measure the process over some time period, and then aggregate the process over time. For instance, by taking the mean of several measurements at each location as your spatial process. The temporally frozen state Cressie and Wikle mentions is a process that does not really evolve through time. Their example is an ore deposit deep underground.

Spatial processes can be divided into continuous and discrete processes. We only consider spatial processes with continuous state space, defined on a grid or a lattice. This is because these are the only kind of spatial processes that can be related to the spatial part of STGARCH models.

2.1.1. Notation

Imagine we have a set of regions or points on a grid $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$ where $\boldsymbol{s}_i \in \mathbb{Z}^d$, $i = 1, \ldots, n$. For each of these locations, we observe the process $\boldsymbol{X} = \{X(\boldsymbol{s}_i) : i = 1, \ldots, n\}$. Figure 2.1 shows the country of Norway with its 19 counties (Norwegians call it «Fylker»). The 19 counties form a lattice of Norway and data for each county is aggregated to analyse the process of interest. The colour pallet represents the relative population growth of each county from January 1st 2014 to January 1st 2015, in percentage. This is an example of a temporal aggregation process, because we have aggregated the change in population over a time period of one year. In this setting it would be natural to form a vector of the population growth process, $\boldsymbol{X} = [X(s_1), \ldots, X(s_{19})]^T$ where s_i represents the different counties for $i = 1, \ldots, 19$.

Consider $\boldsymbol{s}_i \in \mathbb{Z}^2, i = 1, \ldots, n$ forming an equidistant grid. In this situation it might seem useful to let $\boldsymbol{X} = \{X(\boldsymbol{s}_i) : i = 1, \ldots, n\}$ form a matrix. However, we insist on using the vector notation. By enumerating the n location vectors with the subscripts $1, \ldots, n$ we keep consistency

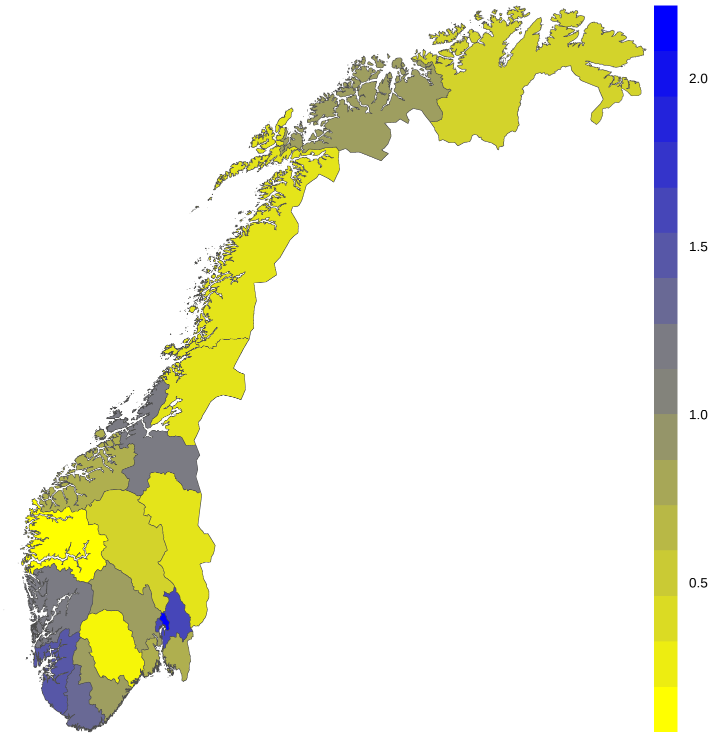


Figure 2.1: *Relative population growth (in percent) from January* 1st 2014 *to January* 1st 2015 *in Norway by counties (Fylker). Data is collected by SSB (2015) and the map shape files are provided by Kartverket (2015).*

with the lattice region case, where the matrix notation is less helpful. This makes the notation more general and what follows will yield both these cases.

According to Wall (2004) there are two fundamentally different ways to model spatial structure underlying lattice data. She says that one way is to treat the lattice data as if the summary statistic for the region was measured in the center of the region and then distances between centroids can be used to develop the spatial covariance structure. The other way (which is used for SAR and CAR) is to define a neighbourhood structure. Instead of measuring distances, this approach simply states that since region A and B are neighbours, they should be correlated in some way. The most common way of doing this is to define A and B as neighbours if the two regions share a border.

Definition 2.1.1. ***(Neighbourhood)***
Let $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ be two regions or grid points and define $\mathcal{N}(\boldsymbol{s}_j) = \{\boldsymbol{s}_k : \boldsymbol{s}_k$ and $\boldsymbol{s}_j$ are neighbours$\}$. Then $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ are neighbours if $\boldsymbol{s}_i \in \mathcal{N}(\boldsymbol{s}_j)$.

We use the common convention that a point or region is not its own neighbour, hence $\boldsymbol{s}_i \notin \mathcal{N}(\boldsymbol{s}_i)$. Also neighbour relations are mutual, meaning that $\boldsymbol{s}_i \in \mathcal{N}(\boldsymbol{s}_j) \Leftrightarrow \boldsymbol{s}_j \in \mathcal{N}(\boldsymbol{s}_i)$. A convenient way of representing the neighbourhood structure, is by defining a $n \times n$ neighbourhood matrix, $\boldsymbol{\mathcal{W}}$. If we let $\boldsymbol{\mathcal{W}} = \{\omega_{ij} : 1 \leq i, j \leq n\}$, we can define $\boldsymbol{\mathcal{W}}$ by

$$\omega_{ij} = \mathbf{1}(\boldsymbol{s}_j \in \mathcal{N}(\boldsymbol{s}_i)), \tag{2.1}$$

where $\mathbf{1}(\cdot)$ is the indicator function and $\mathcal{N}(\boldsymbol{s}_i)$ is the neighbourhood of $\boldsymbol{s}_i$. This means that ω_{ij} will be one if $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ are neighbours and zero if not. Since neighbour relations are mutual and regions are not their own neighbour, $\boldsymbol{\mathcal{W}}$ is a symmetric matrix with zeroes along its diagonal.

2.1.2. Examples: Gaussian CAR and SAR models

Whittle (1954) first introduced simultaneous autoregressive (SAR) models, and twenty years later conditional autoregressive (CAR) models were introduced by Besag (1974). The models are highly related. We primarily use the presentation of Wall (2004) and Cressie and Wikle (2011) in our development of these models.

Let $\mathrm{MVN}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ denote a multivariate normal distribution with expectation vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Let $\{X(\boldsymbol{s}_i) : \boldsymbol{s}_i \in (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n)\}$ be a Gaussian random process. We employ the common notation that $X(\boldsymbol{s}_{-i}) = \{X(\boldsymbol{s}_j) : j \neq i\}$. One can think of $\{\boldsymbol{s}_i : i = 1, \ldots, n\}$ as points on a grid or as regions forming a lattice of the area of interest, D_s, such that $\boldsymbol{s}_1 \cup \cdots \cup \boldsymbol{s}_n = D_s$ and $\boldsymbol{s}_i \cap \boldsymbol{s}_j = \emptyset$ for all $i \neq j$.

The SAR model can be defined as

$$X(\boldsymbol{s}_i) = \mu_i + \sum_{j=1}^{n} b_{ij}(X(\boldsymbol{s}_j) - \mu_j) + Z_i, \tag{2.2}$$

where $\boldsymbol{Z} = (Z_1, \ldots, Z_n)^T \sim \mathrm{MVN}\left(\mathbf{0}, \boldsymbol{\Lambda}\right)$ with $\boldsymbol{\Lambda}$ diagonal, $\mathrm{E}\, X(\boldsymbol{s}_i) = \mu_i$ and b_{ij} are known or unknown constants and $b_{ii} = 0$, $i = 1, \ldots, n$. The innovation terms, Z_i, will in general be correlated with $\{X(\boldsymbol{s}_j) : j \neq i\}$, and this is why the model is called simultaneous. If n is finite, let $\mathbb{B} = \{b_{ij} : 1 \leq i, j \leq n\}$ be a $n \times n$ matrix containing the constants b_{ij}. Then the joint distribution of $\boldsymbol{X} = [X(\boldsymbol{s}_1), \ldots, X(\boldsymbol{s}_n)]^T$ is given by

$$\boldsymbol{X} \sim \mathrm{MVN}\left(\boldsymbol{\mu}, (\mathbb{I}_n - \mathbb{B})^{-1}\boldsymbol{\Lambda}(\mathbb{I}_n - \mathbb{B})^{-T}\right), \tag{2.3}$$

where $\mathbb{I}_n$ is an $n \times n$ identity matrix and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$ and $(\mathbb{I}_n - \mathbb{B})^{-1}\boldsymbol{\Lambda}(\mathbb{I}_n - \mathbb{B})^{-T}$ must be symmetric and positive definite.

The CAR model is an alternative to the SAR model. Both models describe the same kind of process, so we use the same notation.

$$X(\boldsymbol{s}_i)|X(\boldsymbol{s}_{-i}) \sim N\Big(\mu_i + \sum_{\boldsymbol{s}_j \in N(\boldsymbol{s}_i)} c_{ij}(X(\boldsymbol{s}_j) - \mu_j), \tau_i^2\Big). \tag{2.4}$$

Here τ_i^2 is the conditional variance and c_{ij} are known or unknown constants, and just like for b_{ij}, we set $c_{ii} = 0$ for $i = 1, \ldots, n$. We can also here form a $n \times n$ matrix $\mathbb{C} = \{c_{ij} : 1 \leq i, j \leq n\}$

and $\mathbb{T} = \operatorname{diag}\left\{\tau_1^2, \ldots, \tau_n^2\right\}$ if n is finite. (Besag, 1974) showed that if $(\mathbb{I}_n - \mathbb{C})^{-1}\mathbb{T}$ is symmetric and positive definite, we have that

$$\boldsymbol{X} \sim \operatorname{MVN}\left(\boldsymbol{\mu}, (\mathbb{I}_n - \mathbb{C})^{-1}\mathbb{T}\right). \tag{2.5}$$

The next question now is how do we specify $\mathbb{B}$ and $\mathbb{C}$ for the two models? A common way of doing this is to let $\mathbb{B} = \rho_s\boldsymbol{\mathcal{W}}$ for the SAR or $\mathbb{C} = \rho_c\boldsymbol{\mathcal{W}}$ for the CAR model, where $\boldsymbol{\mathcal{W}}$ is a user defined $n \times n$ neighbourhood matrix, defined by (2.1).

For the CAR model, $\boldsymbol{\mathcal{W}}$ and $\mathbb{T}$ need to satisfy the symmetry condition

$$\omega_{ij}\tau_j^2 = \omega_{ji}\tau_i^2.$$

Clayton and Bernardinelli (1992) suggested using a weighting scheme for the neighbourhood matrix to obtain consistency. This is done by

$$\boldsymbol{\mathcal{W}}^* = \{\omega_{ij}/\omega_{i\cdot} : 1 \le i, j \le n\}, \tag{2.6}$$

where $\omega_{i\cdot} = \sum_{j=1}^n \omega_{ij}$ is the i^{th} row sum of $\boldsymbol{\mathcal{W}}$.

2.2. Time Series

Compared to spatial statistics, time series is an even larger scientific discipline. We start out with some essential definitions before moving on to introducing the central time series models of this thesis. We have kept the general time series section short, since the concepts defined here are well known.

The two following definitions are from Brockwell and Davis (2006, p. 15).

Definition 2.2.1. ***Mean and covariance functions***
Let $\{X_t : t \in \mathbb{Z}\}$, be a time series with $\operatorname{E}\left(X_t^2\right) < \infty$. The mean function of $\{X_t\}$ is $\mu_t = \mu_X(t) = \operatorname{E}(X_t)$. The covariance function of $\{X_t\}$ is

$$\gamma_X(r, s) = \operatorname{Cov}(X_r, X_s) = \operatorname{E}(X_r - \mu_r)(X_s - \mu_s), \tag{2.7}$$

for all integers r and s.

Definition 2.2.2. ***Weak stationarity***
*$\{X_t\}$ is weakly stationary if **i)** $\operatorname{E} X_t^2 < \infty$, **ii)** μ_t is independent of t, **iii)** $\gamma_X(t + h, t)$ is independent of t for each h.*

Stationarity, as opposed to weak, requires that $(X_1, X_2, \ldots, X_n)$ and $(X_{1+h}, X_{2+h}, \ldots, X_{n+h})$ have the same distribution for all integers $h > 0$. If a time series is stationary and $\operatorname{E}\left(X_t^2\right) < \infty$, the time series is also weakly stationary. What we here call stationary is often referred to as strictly stationary. If a process is weakly stationary, we write $\gamma_X(s, t) = \gamma_X(h)$, where $h = |t - s|$.

Example 2.2.3. ***White Noise Process***

The time series $\{Z_t\}$ is a white noise process if the variables are ***uncorrelated*** *with mean zero and variance σ_Z^2. We write $Z_t \sim \mathrm{WN}(0, \sigma_Z^2)$. One may also require independence, but in general a white noise process is only uncorrelated. If the distribution of $\{Z_t\}$ is Gaussian, the variables $\{Z_t\}$ are also independent. For $\{Z_t\} \sim \mathrm{WN}(0, \sigma_Z^2)$ we have $\mu_Z = 0$ and $\gamma_Z(h) = \sigma_Z^2 \delta_{h,0}$, where δ is the Kronecker delta and $h = |r - s|$ is the lag. Since the mean function is constant (not dependent on t) and the covariance function only depends on the lag h, we say that $\{Z_t\}$ is weakly stationary.*

2.2.1. ARMA

ARMA models, or autoregressive moving average models, are the cornerstone models of time series analysis. You will find the following definition in any text book about time series, but we use Brockwell and Davis (2006, p. 83).

Definition 2.2.4. ***ARMA models***

The time series $\{X_t : t \in \mathbb{Z}\}$ is an $ARMA(p, q)$ process if it is stationary and satisfies

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_1 + \theta_1 Z_{t-1} + \cdots \theta_q Z_{t-q}, \tag{2.8}$$

where $\{Z_t\} \sim WN(0, \sigma^2)$ is given and the polynomials $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$ and $\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q$ have no common factors.

The time series $\{X_t\}$ is said to be an ARMA(p,q) process with mean μ if $(X_t - \mu)$ is an ARMA(p,q) process. Using the backshift operator we can represent an ARMA process in a more concice form by

$$\phi(B) X_t = \theta(B) Z_t, \tag{2.9}$$

where B is defined as $BX_t = X_{t-1}$ and $B^j X_t = X_{t-j}$. The process is an AR(p) if $q = 0$ and a MA(q) if $p = 0$. The $\{X_t\}$ process is causal if all the roots of the polynomial $\phi(z)$ are outside the unit circle and invertible if all roots of the polynomial $\theta(z)$ are outside the unit circle.

2.3. GARCH

Engle (1982) introduced the autoregressive conditional heteroscedasticity (ARCH) process. The models were expanded to GARCH or generalized ARCH by Bollerslev (1986). Both ARCH and GARCH are used to model processes with varying volatility. We define the processes and derive some of their properties.

Definition 2.3.1. ARCH(p)
Let $\{Z_t\}$ be iid $\mathrm{WN}(0,1)$ *and σ_t^2 be a positive function of $\{X_s, s<t\}$. Then $\{X_t\}$ is an* ARCH*(p) process if for each $t \in \mathbb{Z}$,*

$$\begin{aligned} X_t &= \sigma_t Z_t, \\ \sigma_t^2 &= \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2, \end{aligned} \tag{2.10}$$

where p determine the order of the process.

Definition 2.3.2. GARCH(p, q)
Let $\{Z_t\}$ be iid $\mathrm{WN}(0,1)$*. Then $\{X_t\}$ is a* GARCH*(p,q) process if for each $t \in \mathbb{Z}$,*

$$\begin{aligned} X_t &= \sigma_t Z_t, \\ \sigma_t^2 &= \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2, \end{aligned} \tag{2.11}$$

where p and q determine the order of the process.

Notice that a GARCH(p, 0) is the same as an ARCH(p) process. The most common models are the ARCH(1) and GARCH(1, 1).

Remark to (2.11)**:** Notice the role of p and q in our notation. There seem to be no clear convention if p should regard the ARCH or GARCH term of σ_t^2 in (2.11). All seem to agree that p should be used for the order of an ARCH process, but for GARCH they make q determine the order of the ARCH term (see Bollerslev (1986), Bougerol and Picard (1992), Comte and Lieberman (2003)). We find this practice illogical and have support from e.g. Davis and Mikosch (2009) and McNeil et al. (2005).

In what follows, let $\{X_t\}$ be a GARCH(p, q) process unless otherwise specified.

2.3.1. Moments

The first properties we look at are the first two central moments. Since σ_t only depend on past values of $\{X_s\}$ and $\{\sigma_s\}$, σ_t and Z_t are independent. Therefore,

$$\mathrm{E}\left(X_t\right) = \mathrm{E}\left(\sigma_t Z_t\right) = \mathrm{E}\left(\sigma_t\right)\mathrm{E}\left(Z_t\right) = 0. \tag{2.12}$$

The next central moment is the variance. This simplifies to calculating the second moment of X_t by the zero expectation we just showed. Hence,

$$\mathrm{Var}\left(X_t\right) = \mathrm{E}X_t^2 = \mathrm{E}\left(\sigma_t^2 Z_t^2\right) = \mathrm{E}\left(\sigma_t^2\right)\mathrm{E}\left(Z_t^2\right) = \mathrm{E}\,\sigma_t^2, \tag{2.13}$$

since σ_t^2 and Z_t^2 are independent and $\mathrm{E}\left(Z_t^2\right) = \mathrm{Var}\left(Z_t\right) = 1$. The variance of X_t is finite if and only if $\mathrm{E}\,\sigma_t^2$ is finite.

2.3.2. Stationarity

If we assume that $\{X_t\}$ is a weakly stationary process, we can derive a necessary condition for weak stationarity of GARCH processes. The second moment must be finite and $\mu_X = \mathrm{E}\left(X_t\right) = 0$. Since $\{X_t\}$ is weakly stationary by assumption, point (iii) of definition 2.2.2 also holds. The covariance function only depend on the lag h and we have that $\gamma_X(0) = \mathrm{Var}\left(X_t\right) = \mathrm{E}\left(\sigma_t^2\right)$ for all $t \in \mathbb{Z}$. Therefore

$$\begin{aligned}\gamma_X(0) = \mathrm{Var}\left(X_t\right) = \mathrm{E}\left(\sigma_t^2\right) &= \alpha_0 + \sum_{i=1}^{p} \alpha_i \, \mathrm{E}\left(X_{t-i}^2\right) + \sum_{j=1}^{q} \beta_j \, \mathrm{E}\left(\sigma_{t-j}^2\right) \\ &= \alpha_0 + \sum_{i=1}^{p} \alpha_i \gamma_X(0) + \sum_{j=1}^{q} \beta_j \gamma_X(0),\end{aligned} \tag{2.14}$$

which leads to

$$\sigma_X^2 = \mathrm{Var}\left(X_t\right) = \frac{\alpha_0}{1 - \sum_{i=1}^{p} \alpha_i - \sum_{j=1}^{q} \beta_j}. \tag{2.15}$$

We see that in order for $\mathrm{Var}\left(X_t\right) < \infty$, we need that

$$\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j < 1. \tag{2.16}$$

This is a necessary and sufficient condition for weak stationarity of a GARCH process, and if X_t is weakly stationary (2.15) gives the asymptotic variance of the process.

Applications to financial data have shown that the processes we want to model in many cases do not have the weakly stationary property, but are stationary. Nelson (1990) found that a GARCH$(1,1)$ process is stationary if

$$\mathrm{E} \log(\alpha_1 Z_1^2 + \beta_1) < 0. \tag{2.17}$$

Using Jensen's inequality on (2.17), we get

$$\mathrm{E} \log(\alpha_1 Z_1^2 + \beta_1) \leq \log \mathrm{E}(\alpha_1 Z_1^2 + \beta_1) = \log(\alpha_1 + \beta_1) < 0, \tag{2.18}$$

which is equivalent to $\alpha_1 + \beta_1 < 1$. This is the same criterion as for weakly stationarity above. Bougerol and Picard (1992) found conditions for stationarity of GARCH(p, q) models by requiring that the top Lyapunov exponent of a sequence of random matrices is strictly negative (Bougerol and Picard, 1992, Th.1.3). Calculating the top Lyapunov exponent is quite complicated, so we will not go into the details here, but Bougerol and Picard (1992, pp. 122) proved that if (2.16) holds, the Lyapunov exponent is strictly negative and there exist an ergodic, stationary solution to (2.11). We will for the rest of this chapter assume that $\{X_t\}$ is stationary.

2.3.3. ARMA representation

We now show a way of representing a squared GARCH process as an ARMA process. Notice that we can write

$$X_t^2 = \sigma_t^2 Z_t^2 = \sigma_t^2 + \sigma_t^2(Z_t^2 - 1) = \sigma_t^2 + V_t, \tag{2.19}$$

where $V_t = \sigma_t^2(Z_t^2 - 1)$. First of all, we need to show that V_t is a martingale difference, and therefore a white noise process. We do this by showing that $\mathrm{E}\left(|V_t|\right) < \infty$ and $\mathrm{E}\left(V_t | \mathcal{F}_{t-1}^V\right) = 0$, where $\mathcal{F}_t^V = \bigvee\{V_s : s \leq t\}$ is the sigma algebra consisting of the infinite history of $\{V_t\}$. The first condition is satisfied if $\mathrm{E}|\sigma_t^2| < \infty$. Since $\sigma_t^2 > 0$ always, we can remove the absolute symbols. If (2.16) holds, $\mathrm{E}\,\sigma_t^2 < \infty$. The second condition is shown by using the independence between σ_t and Z_t like in (2.12).

$$\mathrm{E}\left(V_t | \mathcal{F}_{t-1}^V\right) = \mathrm{E}\left(\sigma_t^2 | \mathcal{F}_{t-1}^V\right) \mathrm{E}\left((Z_t^2 - 1) | \mathcal{F}_{t-1}^V\right) = 0. \tag{2.20}$$

Hence, V_t is a martingale difference. These are per definition uncorrelated and therefore a white noise process.

In order to simplify notation, let α_i and β_i be defined for all $i \in \mathbb{Z}$. This is achieved if $\alpha_i = 0$ for $i \notin \{1, \ldots, p\}$ and $\beta_j = 0$ for $j \notin \{1, \ldots, q\}$. We can write (2.19) as

$$X_t^2 = \sigma_t^2 + V_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2 + V_t. \tag{2.21}$$

By (2.19) we also have that $\sigma_t^2 = X_t^2 - V_t$. We insert this into (2.21).

$$\begin{aligned} X_t^2 &= \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2 + \sum_{j=1}^{q} \beta_j \left(X_{t-j}^2 - V_{t-j}\right) + V_t \\ &= \alpha_0 + \sum_{i=1}^{p \vee q} (\alpha_i + \beta_i) X_{t-i}^2 - \sum_{j=1}^{q} \beta_j V_{t-j} + V_t. \end{aligned} \tag{2.22}$$

Here we have used the notation $p \vee q$ meaning $\max\{p, q\}$. The squared process represented by (2.22) will be an $\mathrm{ARMA(p \vee q, q)}$ process with expectation σ_X^2 given by (2.15). This enables us to use regular ARMA-methods for estimation of GARCH.

2.3.4. Quasi Maximum Likelihood Estimation

In a GARCH(p,q) model there is $1+p+q$ parameters to estimate. The vector of parameters, $\boldsymbol{\theta}$, is defined as

$$\boldsymbol{\theta} = [\alpha_0, \alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q]^T. \tag{2.23}$$

Maximum likelihood estimation is a common approach also in estimation of GARCH models. Let $\{x_1, \ldots, x_n\}$ denote a realization of the process we want to model as a GARCH process. We will develop the Gaussian conditional quasi-likelihood, conditioned on $x_0, \dots, x_{1-p}$ and $\widetilde{\sigma}_0^2, \dots, \widetilde{\sigma}_{1-q}^2$. Quasi maximum likelihood (QML) means that the likelihood is somehow misspecified and should be treated more as an objective function to be maximized rather than a proper likelihood (McNeil et al., 2005). A Gaussian QML means that we perhaps erroneously assume Gaussian innovations. If the innovations are truly Gaussian, the QML is a proper likelihood.

Let $\mathcal{X}_0$ be the set of initial values, defined by

$$\mathcal{X}_0 = \{x_0, \dots, x_{1-p}, \widetilde{\sigma}_0, \dots, \widetilde{\sigma}_{1-q}\}. \tag{2.24}$$

Francq et al. (2004) suggests that the initial values can be chosen as

$$x_0^2 = \dots = x_{1-p}^2 = \widetilde{\sigma}_0^2 = \dots = \widetilde{\sigma}_{1-q}^2 = \alpha_0 \tag{2.25}$$

or

$$x_0^2 = \dots = x_{1-p}^2 = \widetilde{\sigma}_0^2 = \dots = \widetilde{\sigma}_{1-q}^2 = x_1^2. \tag{2.26}$$

Since $\{\sigma_t\}$ is an unobserved process, we estimate the process using $\{\widetilde{\sigma}_t\}$, which is defined recursively, for $t \geq 1$, by

$$\widetilde{\sigma}_t^2 = \widetilde{\sigma}_t^2(\boldsymbol{\theta}) = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2 + \sum_{j=1}^{q} \beta_j \widetilde{\sigma}_{t-j}^2. \tag{2.27}$$

The conditional likelihood function, or the simultaneous density of $\{X_1, \dots, X_n\}$ given $\mathcal{X}_0$ and $\boldsymbol{\theta}$, can be factorized as

$$f_{X_1,\dots,X_n|\mathcal{X}_0,\boldsymbol{\theta}} = f_{X_1|\mathcal{X}_0,\boldsymbol{\theta}} f_{X_2|X_1,\mathcal{X}_0,\boldsymbol{\theta}} f_{X_3|X_2,X_1,\mathcal{X}_0,\boldsymbol{\theta}} \cdots f_{X_n|X_{n-1},\dots,X_1,\mathcal{X}_0,\boldsymbol{\theta}}. \tag{2.28}$$

By conditioning on the initial values, $\mathcal{X}_0$, $\widetilde{\sigma}_1^2(\boldsymbol{\theta})$ is a deterministic function only of the parameter vector and we have that

$$X_1|\mathcal{X}_0 = \widetilde{\sigma}_1 Z_1|\mathcal{X}_0 \sim N(0, \widetilde{\sigma}_1^2), \tag{2.29}$$

by the quasi Gaussian assumption. We have established the first density function of the factorization in (2.28). Since $\widetilde{\sigma}_1$ is deterministic given $\mathcal{X}_0$, we have, by (2.27)

$$\widetilde{\sigma}_2^2|X_1, \mathcal{X}_0 = \widetilde{\sigma}_2^2|X_1, X_0, \dots, X_{-p+1}, \widetilde{\sigma}_1, \widetilde{\sigma}_0, \dots, \widetilde{\sigma}_{-q+1}.$$

This is also a deterministic function of $\boldsymbol{\theta}$. Therefore

$$X_2|X_1, \mathcal{X}_0 = \widetilde{\sigma}_2 Z_2|X_1, \mathcal{X}_0 \sim N(0, \widetilde{\sigma}_2^2). \tag{2.30}$$

By repeating this argument we have that

$$\begin{aligned} X_1|\mathcal{X}_0 &\sim N(0,\widetilde{\sigma}_1^2(\boldsymbol{\theta})) \\ X_2|X_1,\mathcal{X}_0 &\sim N(0,\widetilde{\sigma}_2^2(\boldsymbol{\theta})) \\ X_3|X_2,X_1,\mathcal{X}_0 &\sim N(0,\widetilde{\sigma}_3^2(\boldsymbol{\theta})) \\ &\vdots \\ X_n|X_{n-1},\dots,X_1,\mathcal{X}_0 &\sim N(0,\widetilde{\sigma}_n^2(\boldsymbol{\theta})). \end{aligned} \tag{2.31}$$

The likelihood can therefore be expressed as

$$L_n(\boldsymbol{\theta}) = f_{X_1,\dots,X_n|\mathcal{X}_0,\boldsymbol{\theta}} = \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi\widetilde{\sigma}_t^2}} \exp\Big(-\frac{X_t^2}{2\widetilde{\sigma}_t^2}\Big). \tag{2.32}$$

The QML estimator of $\boldsymbol{\theta}$ can then be defined as any measureable solution $\widehat{\boldsymbol{\theta}}_n$ of

$$\widehat{\boldsymbol{\theta}}_n = \operatorname*{argmax}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L_n(\boldsymbol{\theta}) = \operatorname*{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \widetilde{I}_n(\boldsymbol{\theta}), \tag{2.33}$$

where $\widetilde{I}_n = n^{-1}\sum_{t=1}^{n}\widetilde{l}_t$ is the negative conditional quasi log-likelihood function and

$$\widetilde{l}_t = \widetilde{l}_t(\boldsymbol{\theta}) = -\log\left[\frac{1}{\sqrt{2\pi\widetilde{\sigma}_t^2}} \exp\Big(-\frac{X_t^2}{2\widetilde{\sigma}_t^2}\Big)\right] = \frac{1}{2}\log 2\pi + \frac{1}{2}\log\widetilde{\sigma}_t^2 + \frac{X_t^2}{2\widetilde{\sigma}_t^2}, \tag{2.34}$$

where $\widetilde{\sigma}_t = \widetilde{\sigma}_t(\boldsymbol{\theta})$. Francq et al. (2004) points out that the choice of initial conditions does not matter for the asymptotic properties of the QMLE, but it may yet be important from a practicle point of view. In fact, we can show that if the true parameters are known, $\widetilde{\sigma}_t^2(\boldsymbol{\theta}_0) \to \sigma_t^2$ with exponential almost sure convergence for increasing t, where $\boldsymbol{\theta}_0$ is the true parameter for a GARCH$(1,1)$ process with initialization $\widetilde{\sigma}_0^2 \geq 0$.

$$|\widetilde{\sigma}_t^2 - \sigma_t^2| = |\alpha_0 + \alpha_1 X_{t-1}^2 + \beta_1\widetilde{\sigma}_{t-1}^2 - \alpha_0 - \alpha_1 X_{t-1}^2 - \beta_1\sigma_{t-1}^2| = \beta_1|\widetilde{\sigma}_{t-1}^2 - \sigma_{t-1}^2| \tag{2.35}$$

By repeatedly using the recursive relation, we find that $|\widetilde{\sigma}_t^2 - \sigma_t^2| = (\beta_1)^t|\widetilde{\sigma}_0^2 - \sigma_0^2|$. Since $0 \leq \beta_1 < 1$, we have that

$$\lim_{t\to\infty}|\widetilde{\sigma}_t^2 - \sigma_t^2| = 0 \tag{2.36}$$

Straumann et al. (2006) gives this simple proof for the GARCH$(1,1)$ case, but also discuss the general GARCH(p,q) model. This characteristic that the $\widetilde{\sigma}_t^2$ process converge to the true process, regardless of the initial values is called the invertibility of the process. This means that the choice of initial values are asymptotically insignificant.

We find $\widehat{\boldsymbol{\theta}}_n$ that satisfy Equation (2.33), by differentiating $\widetilde{I}_n$ with respect to $\boldsymbol{\theta}$, equating the derivative to zero and solve for $\boldsymbol{\theta}$.

$$\frac{\partial\widetilde{I}_n}{\partial\boldsymbol{\theta}} = n^{-1}\sum_{t=1}^{n}\frac{\partial\widetilde{l}_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = n^{-1}\sum_{t=1}^{n}\Big(\frac{1}{2\widetilde{\sigma}_t^2}\frac{\partial\widetilde{\sigma}_t^2}{\partial\boldsymbol{\theta}} - \frac{X_t^2}{2\widetilde{\sigma}_t^4}\frac{\partial\widetilde{\sigma}_t^2}{\partial\boldsymbol{\theta}}\Big) = n^{-1}\sum_{t=1}^{n}\frac{1}{2\widetilde{\sigma}_t^4}\frac{\partial\widetilde{\sigma}_t^2}{\partial\boldsymbol{\theta}}(\widetilde{\sigma}_t^2 - X_t^2) = \mathbf{0}. \tag{2.37}$$

Solving (2.37) for $\boldsymbol{\theta}$ gives us an estimate $\widehat{\boldsymbol{\theta}}$. This will typically be done by an iterative algorithm, since $\widetilde{\sigma}_t = \widetilde{\sigma}_t(\widehat{\boldsymbol{\theta}})$ one will have to update the estimated $\{\widetilde{\sigma}_t\}$ process for each iteration. When a solution has been found, one should make sure that the solution is in fact a maximum of the log likelihood function.

2.3.5. Asymptotic Results of the QMLE

Francq et al. (2004) proves asymptotic results for the QML estimator. They assume that the true parameter vector $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ is a compact parameter space. They also require Bougerol and Picard's (1992) criterion for the existence of a unique non-anticipative strictly stationary solution to be fulfilled. Z_t^2 must have a non-degenerate distribution with $\mathrm{E}\left(Z_t^2\right) = 1$. The polynomials $A_\theta(z) = \sum_{i=1}^{p} \alpha_i z^i$ and $B_\theta(z) = 1 - \sum_{j=1}^{q} \beta_j z^j$ must have no common root, $A_{\theta_0}(1) \neq 0$ and $\alpha_p + \beta_q \neq 0$. If all these assumptions are fulfilled, the following result holds.

Theorem 2.3.3. *(Francq et al., 2004, Th. 2.1, pp. 609) Let $\{\widehat{\boldsymbol{\theta}}_n\}$ be a sequence of QML estimators satisfying (2.33), with initial conditions (2.25) or (2.26). Then under the assumptions stated above, almost surely $\widehat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}_0$ as $n \to \infty$.*

We have asymptotic normality of the estimators if we in addition to the above assumptions, assume that the true parameter vector, $\boldsymbol{\theta}_0$, lies in the interior of $\boldsymbol{\Theta}$ and that $\kappa_z := \mathrm{E}\left(Z_t^4\right) < \infty$.

Theorem 2.3.4. *(Francq et al., 2004, Th. 2.2, pp. 610) Under the assumptions above,*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}_n} - \boldsymbol{\theta}_0) \xrightarrow[n]{\mathcal{D}} \mathrm{MVN}\left(\mathbf{0}, (\kappa_z - 1)J^{-1}\right), \tag{2.38}$$

where

$$J := \mathrm{E}_{\boldsymbol{\theta}_0}\left(\frac{\partial^2 l_t(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right). \tag{2.39}$$

For proofs of theorems 2.3.3 and 2.3.4, see Francq et al. (2004). Notice how weak assumptions are being made for these asymptotic properties to hold with the exemption of the compactness of $\boldsymbol{\Theta}$. This is currently not a luxury that generalizes to the multivariate case.

2.3.6. Applications

McNeil et al. (2005) talks about *stylized facts about financial time series*. This is a collection of empirical observations that seem to apply to the majority of daily series of risk-factor changes, such as log-returns on equities, indexes and exchange rates. These observations have been so entrenched in the econometric society that they have been elevated to the status of facts. McNeil et al. (2005, pp. 117) lists the following:

1) Return series are not iid although they show little serial correlation

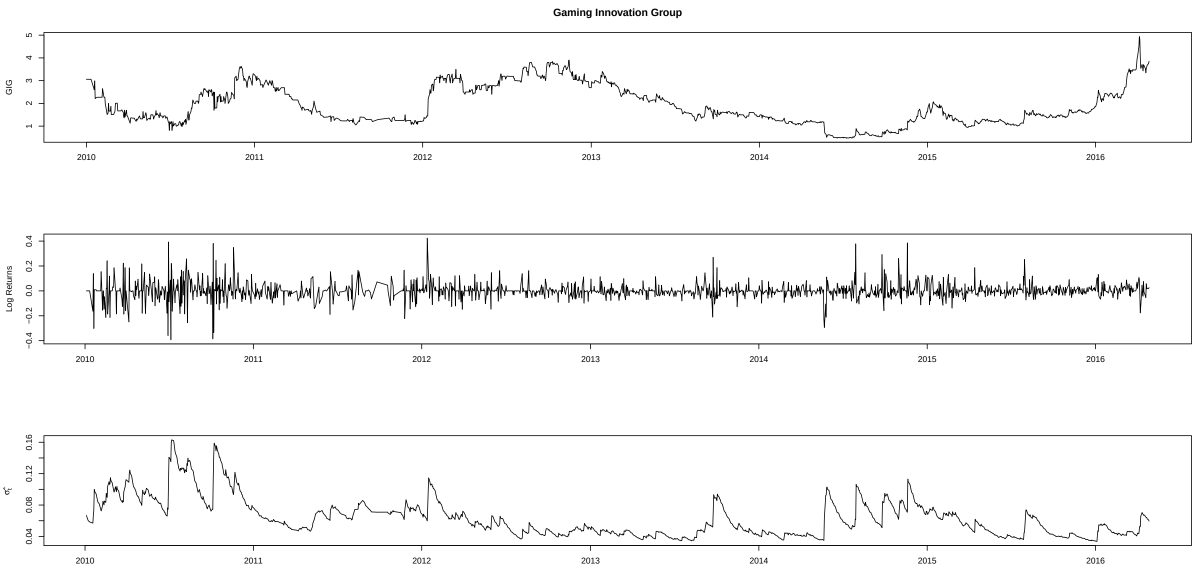


Figure 2.2: *The first panel shows the closing price of the Gaming Innovation Group (GIG.OL) stock from* 1st *of January 2010 to* 26th *of April 2016. The second panel shows the logarithmic returns of GIG over the same time period. In the final panel we have plotted a estimated* GARCH(1, 1) *volatility process* (σ_t^2).

2) Series of absolute or squared returns show profound serial correlation

3) Conditional expected returns are close to zero

4) Volatility appears to vary over time

5) Return series are leptokurtic or heavy-tailed

6) Extreme returns appear in clusters

Empirical studies have shown that GARCH models fit all of these stylized facts and this is what makes them convenient to model financial return data. Figure 2.2 shows the resulting $\{\sigma_t^2\}$ process from fitting a GARCH(1, 1) to the daily logarithmic return time series of the Gaming Innovation Group stock, traded at Oslo Stock Exchange, dating from January 1st 2010 to April 26th 2016. The estimated parameters, with their corresponding standard deviations and p-values are given in Table 2.1. These results come from running the *garchFit* function from the R package **fGarch** (Wuertz et al., 2013). We also tested higher order models, but this lead to insignificant estimates.

According to McNeil et al. (2005), volatility clustering is the «tendency for extreme returns to be followed by other extreme returns». We can see the volatility clustering effect in the GIG logarithmic return data in Figure 2.2 by noticing that a peak is often closely followed by other peaks. It is easier to see the volatility clustering in the estimated volatility process. Here we

Parameter	Estimate	Standard deviation	p-value
α_0	$4.784 \cdot 10^{-5}$	$1.454 \cdot 10^{-6}$	0.001
α_1	$5.421 \cdot 10^{-2}$	$1.061 \cdot 10^{-2}$	$3.2 \cdot 10^{-7}$
β_1	$9.366 \cdot 10^{-1}$	$1.166 \cdot 10^{-2}$	$< 2 \cdot 10^{-16}$

Table 2.1: *Results from estimating a* GARCH(1, 1) *for the GIG stock returns.*

see that towards a peak there is a rapid increase, while after the peak, it decays slowly. The volatility stays high for period of time following a peak.

2.4. Multivariate GARCH

Comte and Lieberman (2003) proved asymptotic normality of quasi maximum likelihood estimates in a multivariate GARCH (MGARCH) model under some regularity conditions. They also proved strong consistency of the estimators. We define MGARCH models and give the results from this article. The proofs are very technical, and the interested reader can find them in the original article.

We use a notation similar to Comte and Lieberman (2003), but alter it somewhat to uphold consistency with the thesis notation. The sole purpose of introducing MGARCH here, is to use the results for the spatio-temporal GARCH in Chapter 5, and therefore we have limited this section to a minimum.

2.4.1. The BEKK model for MGARCH

Let $\{\boldsymbol{X}_t\}_{t\in\mathbb{Z}}$ be a sequence of random variables of $\mathbb{R}^d$ and $\mathcal{F}_t$ be the sigma field generated by the history of $\boldsymbol{X}_t$. Assuming $\boldsymbol{X}_t$ is square integrable and such that

$$\boldsymbol{X}_t = \mathbb{H}_t^{1/2}\boldsymbol{Z}_t, \tag{2.40}$$

with $\mathbb{H}_t$ being a symmetric, positive definite $d \times d$ matrix defined using BEKK's (Baba, Engle, Kraft and Kroner) representation (Engle and Kroner, 1995)

$$\mathbb{H}_t = \mathbb{C} + \sum_{i=1}^{p}\Big(\sum_{j=1}^{k}\mathbb{A}_{ij}\boldsymbol{X}_{t-i}\boldsymbol{X}_{t-i}^T\mathbb{A}_{ij}^T\Big) + \sum_{i=1}^{q}\Big(\sum_{j=1}^{k}\mathbb{B}_{ij}\mathbb{H}_{t-i}\mathbb{B}_{ij}^T\Big), \tag{2.41}$$

where the matrix $\mathbb{C}$ is positive definite and the matrices $\mathbb{A}_{ij}$, for $i = 1,\dots,p$, $j = 1,\dots,k$ and $\mathbb{B}_{ij}$ for $i = 1,\dots,q$ and $j = 1,\dots,k$ are real $d \times d$ matrices and k is an integer satisfying $k \le d(d+1)/2$. All the matrices are functions of the parameter vector, $\boldsymbol{\theta}$.

We let the innovation vector be defined as

$$\boldsymbol{Z}_t \sim \text{iid}(\boldsymbol{0}, \mathbb{I}_\text{d}), \tag{2.42}$$

$\mathbb{I}_d$ being the $d \times d$ identity matrix. The process $\{\boldsymbol{X}_t\}$ is a martingale difference

$$\mathrm{E}\left(\boldsymbol{X}_t | \mathcal{F}_{t-1}\right) = \mathbf{0}, \tag{2.43}$$

with conditional covariance matrix

$$\mathrm{E}\left(\boldsymbol{X}_t \boldsymbol{X}_t^T | \mathcal{F}_{t-1}\right) = \mathbb{H}_t. \tag{2.44}$$

Equation (2.41) can be written using the vec operator. This is an operator that stacks the columns of a matrix to form a vector.

$$\mathrm{vec}(\mathbb{H}_t) = \mathrm{vec}(\mathbb{C}) + \sum_{i=1}^{p} \mathbb{A}_i^\star \mathrm{vec}(\boldsymbol{X}_{t-i}\boldsymbol{X}_{t-i}^T) + \sum_{i=1}^{q} \mathbb{B}_i^\star \mathrm{vec}(\mathbb{H}_{t-i}^T), \tag{2.45}$$

where $\mathbb{A}_i^\star = \sum_{j=1}^{k} \mathbb{A}_{ij} \otimes \mathbb{A}_{ij}$, $i = 1, \ldots, p$, and $\mathbb{B}_i^\star = \sum_{j=1}^{k} \mathbb{B}_{ij} \otimes \mathbb{B}_{ij}$, $i = 1, \ldots, q$, and $\otimes$ is the Kronecker product. Since the matrices involved are symmetric, we can also write (2.45) using the vech operator that stacks the lower-triangular portion of a symmetric matrix to form a vector.

$$\mathrm{vech}(\mathbb{H}_t) = \mathrm{vech}(\mathbb{C}) + \sum_{i=1}^{p} \widetilde{\mathbb{A}}_i \mathrm{vech}(\boldsymbol{X}_{t-i}\boldsymbol{X}_{t-i}^T) + \sum_{i=1}^{q} \widetilde{\mathbb{B}}_i \mathrm{vech}(\mathbb{H}_{t-i}^T), \tag{2.46}$$

where $\mathbb{L}_d$ and $\mathbb{K}_d$ are matrices of dimension $d(d+1) \times d^2$, given by $\{\mathbb{A}_i^\star\}$ and $\{\mathbb{B}_i^\star\}$, satisfying $\widetilde{\mathbb{A}}_i = \mathbb{L}_d \mathbb{A}_i^\star \mathbb{K}_d^T$ for $i = 1, \ldots, p$ and $\widetilde{\mathbb{B}}_i = \mathbb{L}_d \mathbb{B}_i^\star \mathbb{K}_d^T$ for $i = 1, \ldots, q$.

Equation (2.46) is used to define the constraints of the theorems in the consecutive section, but it is our understanding that (2.45) is on a par with (2.46), but due to symmetry of the matrices, (2.45) will have more identical elements. Stelzer (2008) notes that all BEKK processes are VEC, but the converse is not always true. He uses linear algebra to prove relations between the VEC and BEKK models.

2.4.2. Asymptotic Results

In their article, Comte and Lieberman (2003) proved asymptotic results for the multivariate GARCH(p,q) model under some assumptions. Let $\widehat{\boldsymbol{\theta}}_n$ be QML estimator and $\boldsymbol{\theta}_0$ the true parameter vector, both belonging to the parameter space $\boldsymbol{\Theta}$.

Theorem 2.4.1. *(Comte and Lieberman, 2003, Th. 2) For the* MGARCH(p, q) *process defined by* (2.40), (2.41) *and* (2.42)*, assume that:*

1. *$\boldsymbol{\Theta}$ is compact, $\mathbb{C}$, $\widetilde{\mathbb{A}}_i$, $\widetilde{\mathbb{B}}_i$ are continuous functions of $\boldsymbol{\theta}$, and there exists $c > 0$ such that* $\inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \det \mathbb{C}(\boldsymbol{\theta}) \geq c > 0$.
2. *The model is identifiable.*
3. *The rescaled errors $\boldsymbol{Z}_t$ admit a density absolutely continuous w.r.t. the Lebesgue measure and positive in a neighbourhood of the origin.*
4. $\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$, $\rho\big(\sum_{i=1}^{p} \widetilde{\mathbb{A}}_i(\boldsymbol{\theta}) + \sum_{i=1}^{q} \widetilde{\mathbb{B}}_i(\boldsymbol{\theta})\big) < 1$.

Then $\widehat{\boldsymbol{\theta}}_n$ is strongly consistent, that is,

$$\widehat{\boldsymbol{\theta}}_n \xrightarrow[n]{a.s.} \boldsymbol{\theta}_0, \textit{ under } \mathbb{P}_{\boldsymbol{\theta}_0}.$$

Theorem 2.4.2. *(Comte and Lieberman, 2003, Th. 3) Under the assumptions:*

- *(i) (1)-(4) from Theorem 2.4.1 and $\mathbb{C}(\boldsymbol{\theta})$, $\widetilde{\mathbb{A}}_i(\boldsymbol{\theta})$, $\widetilde{\mathbb{B}}_i(\boldsymbol{\theta})$ admit continuous derivates up to order 3 on $\boldsymbol{\Theta}$,*
- *(ii) The components of $\boldsymbol{Z}_t$ are independent, $\boldsymbol{X}_t$ admits bounded moments of order 8,*
- *(iii) The initial value (in $\mathbb{H}$) is drawn for the stationary ergodic law,*

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \xrightarrow[n]{\mathcal{D}} \text{MVN}\left(\mathbf{0}, \mathbb{C}_1^{-1}\mathbb{C}_0\mathbb{C}_1^{1}\right), \textit{ under } \mathbb{P}_{\boldsymbol{\theta}_0}. \tag{2.47}$$

where $\mathbb{C}_0 = \text{E}\left(\frac{\partial l_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \frac{\partial l_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^T}\right)$ is finite and $\mathbb{C}_1 = \text{E}\left(\left[\frac{\partial^2 l_t(\boldsymbol{\theta}_0)}{\partial \theta_i \partial \theta_j}\right]_{1 \le i,j \le r}\right)$ is finite and positive definite and $l_t(\boldsymbol{\theta}_0)$ is the log likelihood function.

Note that if moreover $\boldsymbol{Z}_t \sim \text{MVN}(\mathbf{0}, \mathbb{I})$, then $\mathbb{C}_0 = 2\mathbb{C}_1$ and the asymptotic law reduces to $\text{MVN}\left(\mathbf{0}, 2\mathbb{C}_1^{-1}\right)$.

Theorems 2.4.1 and 2.4.2 are only valid under a random initial condition in the stationary law, $\mathbb{P}_{\boldsymbol{\theta}_0}$. However, Comte and Lieberman (2003) also establish an extension to the fixed initial value case (Comte and Lieberman, 2003, Th. 4, pp. 68).

2.5. Circular AR(1)

A main point of this thesis is the use of a circular model. Therefore, we will introduce this type of model, by considering a circular autoregressive model (not to be confused with CAR models). In this circular model, we assume that we have a finite number of variables, and these go in loop. If you run through the variables and get to the end, you are back at the beginning. Figure 2.3 illustrates how this works.

Let us first define the modulus operator. We say that $d = \text{mod}(a, b)$ if a, b, c, d are integers and c is the integer part of a/b, noted by $c = \lfloor a/b \rfloor$, and $a/b = c + d/b$. The modulus is the leftovers from an integer division. For example, $\text{mod}(10, 3) = 1$, since $10/3 = 3 + 1/3$.

A circular AR(1) can be defined as

$$\begin{aligned} X_t &= \phi X_{t-1} + Z_t, \qquad t = 1, \dots, n, \\ X_0 &= X_n, \\ Z_0 &= Z_n, \end{aligned} \tag{2.48}$$

where $\{Z_t\}$ is iid $\text{WN}(0, \sigma^2)$. Due to the circularity, for $t > n$, we have that $X_t = X_{\text{mod}(t,n)}$ and

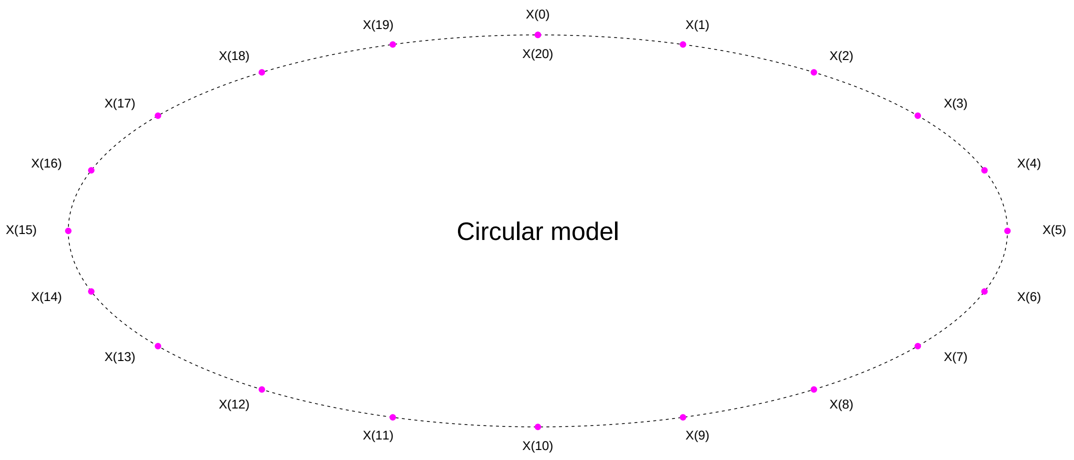


Figure 2.3: *Circular* AR *model illustration.*

$Z_t = Z_{\text{mod}(t,n)}$. If we iterate the first equation of this definition backwards in time, we get

$$\begin{aligned} X_t &= \phi X_{t-1} + Z_t \\ &= \phi^2 X_{t-2} + \phi Z_{t-1} + Z_t \\ &\vdots \\ &= \phi^j X_{t-j} + \sum_{k=0}^{j-1} \phi^k Z_{t-k}, \qquad j \geq 0. \end{aligned} \tag{2.49}$$

We want to exploit that $X_t = X_{t-n}$. If we let $j = n$ in (2.49), we get

$$\begin{aligned} X_t &= \phi^n X_{t-n} + \sum_{k=0}^{n-1} \phi^k Z_{t-k} \\ &= \phi^n X_t + \sum_{k=0}^{n-1} \phi^k Z_{t-k}. \end{aligned} \tag{2.50}$$

Solving for X_t we get

$$X_t = \frac{1}{1-\phi^n} \sum_{k=0}^{n-1} \phi^k Z_{t-k}. \tag{2.51}$$

By (2.51) and since $\operatorname{E} Z_t = 0$, we have that $\operatorname{E} X_t = 0$. We derive the auto correlation function (ACF) of a circular AR(1) model. Due to the representation in (2.51), we can write

$$\begin{aligned} \gamma(h) &= \operatorname{Cov}\left(X_t, X_{t+h}\right) = \operatorname{E}\left(X_t X_{t+h}\right) \\ &= \operatorname{E}\left[\left(\frac{1}{1-\phi^n} \sum_{k=0}^{n-1} \phi^k Z_{t-k}\right)\left(\frac{1}{1-\phi^n} \sum_{l=0}^{n-1} \phi^l Z_{t+h-l}\right)\right] \\ &= \frac{1}{(1-\phi^n)^2} \sum_{k=0}^{n-1} \sum_{l=0}^{n-1} \phi^k \phi^l \operatorname{E}\left(Z_{t-k} Z_{t+h-l}\right). \end{aligned} \tag{2.52}$$

Since $\{Z_t\}$ is an iid white noise process, it is an independent sequence of variables. This means that $\mathrm{E}\left(Z_{t-k}Z_{t+h-l}\right) = 0$ if $l \neq \mathrm{mod}(k+h,n)$.

$$\begin{aligned}\gamma(h) &= \frac{1}{(1-\phi^n)^2}\sum_{k=0}^{n-1}\sum_{l=0}^{n-1}\phi^k\phi^l\sigma^2\delta_{l,\mathrm{mod}(k+h,n)} \\ &= \frac{\sigma^2}{(1-\phi^n)^2}\sum_{k=0}^{n-h-1}\sum_{l=0}^{n-1}\phi^k\phi^l\delta_{l,\mathrm{mod}(k+h,n)} + \frac{\sigma^2}{(1-\phi^n)^2}\sum_{k=n-h}^{n-1}\sum_{l=0}^{n-1}\phi^k\phi^l\delta_{l,\mathrm{mod}(k+h,n)}.\end{aligned} \tag{2.53}$$

The reason for splitting the sum into these two parts, is that

$$\mathrm{mod}(k+h,n) = \begin{cases} h+k, & \text{if } 0 \leq k \leq n-h-1 \\ h+k-n, & \text{if } n-h \leq k \leq n-1 \end{cases}.$$

We therefore get

$$\begin{aligned}\gamma(h) &= \frac{\sigma^2}{(1-\phi^n)^2}\sum_{k=0}^{n-h-1}\phi^{k+h+k} + \frac{\sigma^2}{(1-\phi^n)^2}\sum_{k=n-h}^{n-1}\phi^{k+h+k-n} \\ &= \frac{\sigma^2\phi^h}{(1-\phi^n)^2}\sum_{k=0}^{n-h-1}\phi^{2k} + \frac{\sigma^2\phi^{h-n}}{(1-\phi^n)^2}\sum_{u=0}^{h-1}\phi^{2(u+n-h)} \\ &= \frac{\sigma^2\phi^h}{(1-\phi^n)^2}\sum_{k=0}^{n-h-1}\phi^{2k} + \frac{\sigma^2\phi^{n-h}}{(1-\phi^n)^2}\sum_{u=0}^{h-1}\phi^{2u}.\end{aligned} \tag{2.54}$$

We have for $a < 1$, that $\sum_{k=0}^{n-1} a^k = \frac{a^n-1}{a-1}$. Letting $a = \phi^2$, we get

$$\gamma(h) = \frac{\sigma^2\left(\phi^h(\phi^{2(n-h)}-1) + \phi^{n-h}(\phi^{2h}-1)\right)}{(1-\phi^n)^2(\phi^2-1)} = \frac{\sigma^2\left(\phi^{2n-h}-\phi^h+\phi^{n+h}-\phi^{n-h}\right)}{(1-\phi^n)^2(\phi^2-1)}, \tag{2.55}$$

where $h \in \{0, 1, \ldots, n-1\}$. We find the variance by inserting $h = 0$.

$$\mathrm{Var}\left(X_t\right) = \gamma(0) = \frac{\sigma^2\left(\phi^{2n}-1\right)}{(1-\phi^n)^2(\phi^2-1)}. \tag{2.56}$$

The ACF is found by dividing (2.55) by (2.56).

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \frac{\phi^{2n-h}-\phi^h+\phi^{n+h}-\phi^{n-h}}{\phi^{2n}-1}, \qquad h = 0, 1 \ldots, n-1. \tag{2.57}$$

In order to define $\rho(h)$ for all lag $h > n$, we let $\rho(h) = \rho(\mathrm{mod}(h,n))$. Below (Figure 2.4), we have plotted the ACF for a circular AR(1) process for lags up to 60 with $\phi = 0.7$ and $n = 20$. The figure illustrates how the circularity influence the dependence structure between the variables quite well.

We return to circular models in the next chapter.

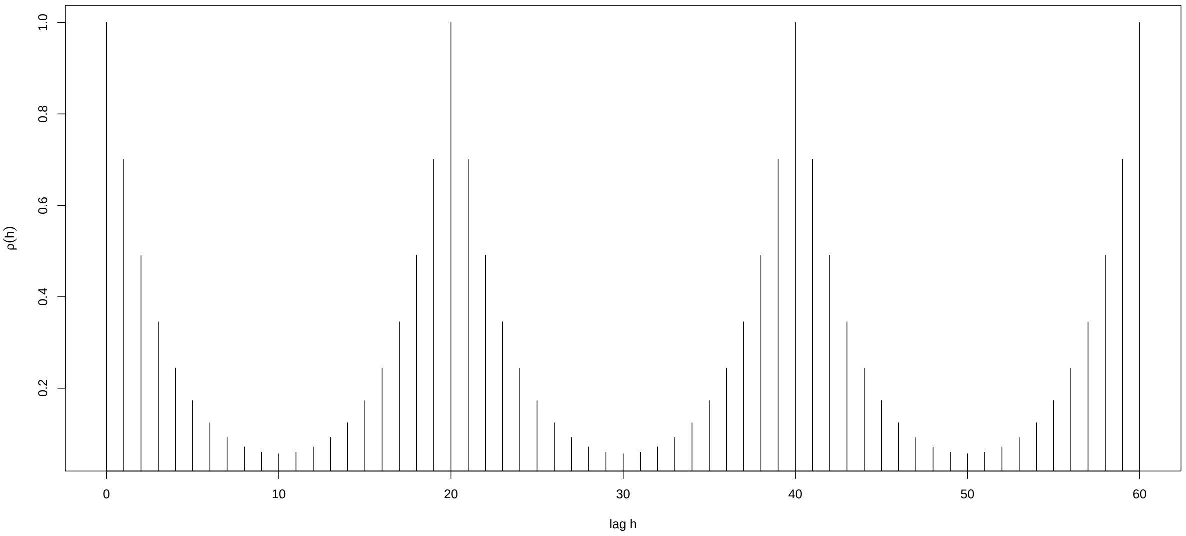


Figure 2.4: *ACF of Circular* AR(1) *with* $n = 20$ *and* $\phi = 0.7$.

Chapter 3

Spatio-Temporal GARCH process

In this chapter we introduce the family of spatio-temporal GARCH (STGARCH) models and view some of their properties, such as moments, stationary conditions and derive an ARMA representation. We have a small discussion about the boundary issues concerning STGARCH and finally we mention possible real life applications.

3.1. The process

After introducing GARCH models at the end of Chapter 2, we can take what we learned into the spatio-temporal expansion. As one will see from the following definition, the structure of the process is the same, but we let every component depend on where the process is in space.

Definition 3.1.1. STGARCH(p, q)
Let $\{Z_t(\boldsymbol{u}) : t = 1, \ldots, T, \boldsymbol{u} \in \mathbb{Z}^d\}$ *be a sequence of iid* $WN(0, 1)$ *random variables. Then* $\{X_t(\boldsymbol{u})\}$ *is a* STGARCH(p, q) *process, if for every* $t \in \{1, \ldots, T\}$ *and* $\boldsymbol{u} \in \mathbb{Z}^d$

$$\begin{aligned} X_t(\boldsymbol{u}) &= \sigma_t(\boldsymbol{u}) Z_t(\boldsymbol{u}) \\ \sigma_t^2(\boldsymbol{u}) &= \alpha_0 + \sum_{s=1}^{p} \sum_{\boldsymbol{v} \in \mathbb{V}_a} \alpha_s(\boldsymbol{v}) X_{t-s}^2(\boldsymbol{u} - \boldsymbol{v}) + \sum_{s=1}^{q} \sum_{\boldsymbol{v} \in \mathbb{V}_b} \beta_s(\boldsymbol{v}) \sigma_{t-s}^2(\boldsymbol{u} - \boldsymbol{v}). \end{aligned} \tag{3.1}$$

If nothing else is stated, we will assume that $\{Z_t(\boldsymbol{u})\}$ is standard normally distributed, but in general we only require it to be a unit variance independent white noise process. We assume $\mathbb{V}_a \subseteq \mathbb{V}$ and $\mathbb{V}_b \subseteq \mathbb{V}$, but $\mathbb{V}_a \neq \mathbb{V}_b$ is allowed. Let $\alpha_s(\boldsymbol{v})$ and $\beta_s(\boldsymbol{v})$ be defined for all $s \in \mathbb{Z}$ and $\boldsymbol{v} \in \mathbb{V}$ by defining them to be zero if $s \notin \{1, \ldots, p\}$ (or q) and $\boldsymbol{v} \notin \mathbb{V}_a$ (or $\mathbb{V}_b$).

A STGARCH model is, just like the regular time series GARCH, a parametric model. One problem that may arise is a high number of parameters. In the general form of (3.1), the parameter vector is

$$\boldsymbol{\theta} = [\alpha_0, \alpha_1(a_1), \ldots, \alpha_1(a_n), \alpha_2(a_1), \ldots, \alpha_p(a_n), \beta_1(b_1), \ldots, \beta_1(b_m), \beta_2(b_1), \ldots, \beta_q(b_m)]^T ,$$

where $\mathbb{V}_a = \{a_1, \dots a_n\}$ and $\mathbb{V}_b = \{b_1, \dots b_m\}$ and p and q determine the temporal order of the process. If we want a high p and q, and in addition let $\mathbb{V}_a$ and $\mathbb{V}_b$ include many points, the number of parameters can potentially become enormous. This is why we will always include some symmetry assumptions. We have summarized these assumptions below.

Assumption 3.1.1.
In this thesis we will usually assume the following symmetry conditions

- $\star$ $\mathbb{V} = \mathbb{V}_a = \mathbb{V}_b$
- $\star$ $\alpha_i = \alpha_i(a_1) = \alpha_i(a_2) = \dots = \alpha_i(a_n)$
- $\star$ $\beta_j = \beta_j(b_1) = \beta_j(b_2) = \dots = \beta_j(b_m)$

The first assumption makes the neighbourhood structure equal for the ARCH and GARCH terms of the process. The second and third assumptions reduces the number of parameters to one per ARCH or GARCH term, in contrast to one per neighbour in every ARCH and GARCH term.

3.2. Properties of STGARCH

In this section we view some of the key properties of a STGARCH process. One will see that in most cases, the results from the time series GARCH generalizes to the spatio-temporal version. We start out by calculating the two first central moments of the process, the expectation and variance. Then we continue to find stationarity conditions and finally illustrate that also the STGARCH has an ARMA representation.

3.2.1. Moments

Equivalent to the time series GARCH, showing that $\mathrm{E}X_t(\boldsymbol{u}) = 0$ follows by the independence between $Z_t(\boldsymbol{u})$ and $\sigma_t(\boldsymbol{u})$.

$$\mathrm{E}X_t(\boldsymbol{u}) = \mathrm{E}\sigma_t(\boldsymbol{u})Z_t(\boldsymbol{u}) = \mathrm{E}\sigma_t(\boldsymbol{u})\mathrm{E}Z_t(\boldsymbol{u}) = 0. \tag{3.2}$$

Since $\mathrm{E}\,X_t(\boldsymbol{u}) = 0$, finding the variance of $X_t(\boldsymbol{u})$ reduces to finding the second moment. The second moment of a STGARCH is given by

$$\mathrm{E}X_t^2(\boldsymbol{u}) = \mathrm{E}\sigma_t^2(\boldsymbol{u})Z_t^2(\boldsymbol{u}) = \mathrm{E}\sigma_t^2(\boldsymbol{u})\mathrm{E}Z_t^2(\boldsymbol{u}) = \mathrm{E}\sigma_t^2(\boldsymbol{u}), \tag{3.3}$$

using the same argument. Both these results follow exactly from the standard time series case.

3.2.2. Stationarity

A STGARCH will not be stationary for all values of the parameters. We suspect that the stationarity conditions of Bougerol and Picard (1992) generalizes to STGARCH, but we have not proven it. Instead, we search for some weakly stationary condition with respect to the parameter space. We have already shown that $\mathrm{E}\,(X_t(\boldsymbol{u})) = 0$, so condition (ii) in Definition 2.2.2 is fulfilled.

Let us, like for the time series GARCH, assume that $\{X_t(\boldsymbol{u})\}$ is a weakly stationary process and use the definition to find a necessary condition for weak stationarity. By assumption, $\mathrm{E}X_t^2(\boldsymbol{u}) = \mathrm{E}\sigma_t^2(\boldsymbol{u}) < \infty$ by definition 2.2.2(i), and the condition (iii) that $\gamma_X(s,t) = \gamma_X(h)$ ensures that the variance is constant. Hence,

$$\begin{aligned}\sigma^2 &= \operatorname{Var} X_t(\boldsymbol{u}) = \mathrm{E}\,\sigma_t^2(\boldsymbol{u}) \\ &= \alpha_0 + \sum_{i=1}^{p}\sum_{\boldsymbol{V}\in\mathbb{V}_a} \alpha_s(\boldsymbol{v})\,\mathrm{E}\,X_{t-s}^2(\boldsymbol{u}-\boldsymbol{v}) + \sum_{i=1}^{q}\sum_{\boldsymbol{V}\in\mathbb{V}_b} \beta_s(\boldsymbol{v})\,\mathrm{E}\,\sigma_{t-s}^2(\boldsymbol{u}-\boldsymbol{v}).\end{aligned} \tag{3.4}$$

Since $\mathrm{E}\,X_t^2(\boldsymbol{u}) = \mathrm{E}\,\sigma_t^2(\boldsymbol{u})$ and $\sigma^2 = \mathrm{E}\,X_t^2(\boldsymbol{u})$ for all $t \in \mathbb{Z}$, we have that

$$\sigma^2 = \alpha_0 + \sum_{i=1}^{p}\sum_{\boldsymbol{V}\in\mathbb{V}_a} \alpha_s(\boldsymbol{v})\sigma^2 + \sum_{i=1}^{q}\sum_{\boldsymbol{V}\in\mathbb{V}_b} \beta_s(\boldsymbol{v})\sigma^2. \tag{3.5}$$

Solving for σ^2 we get

$$\sigma^2 = \alpha_0\Big(1 - \sum_{i=1}^{p}\sum_{\boldsymbol{V}\in\mathbb{V}_a} \alpha_s(\boldsymbol{v}) - \sum_{i=1}^{q}\sum_{\boldsymbol{V}\in\mathbb{V}_b} \beta_s(\boldsymbol{v})\Big)^{-1}. \tag{3.6}$$

In order for $\{X_t(\boldsymbol{u})\}$ to be weakly stationary, we must have that $\mathrm{E}\,X_t^2(\boldsymbol{u}) = \sigma^2 < \infty$ and therefore

$$\sum_{i=1}^{p}\sum_{\boldsymbol{V}\in\mathbb{V}_a} \alpha_s(\boldsymbol{v}) + \sum_{i=1}^{q}\sum_{\boldsymbol{V}\in\mathbb{V}_b} \beta_s(\boldsymbol{v}) < 1. \tag{3.7}$$

We see that the result from Equation (2.16) generalizes and if (3.7) is fulfilled, $\{X_t(\boldsymbol{u})\}$ is a weakly stationary process. For future reference, we always assume (3.7) to hold.

3.2.3. ARMA Representation

Just like the regular GARCH, the STGARCH has an ARMA representation. The arguments and procedure for developing this representation is (almost) identical to the regular case. We start out by squaring the process and using that the $\sigma_t^2(\boldsymbol{u})$ process is linear in the parameters.

$$X_t^2(\boldsymbol{u}) = \sigma_t^2(\boldsymbol{u})Z_t^2(\boldsymbol{u}) = \sigma_t^2(\boldsymbol{u}) + \sigma_t^2(\boldsymbol{u})(Z_t^2(\boldsymbol{u}) - 1) = \sigma_t^2(\boldsymbol{u}) + V_t(\boldsymbol{u}), \tag{3.8}$$

where V_t is a martingale difference and hence white noise process. Remember that $\alpha_i(\boldsymbol{u})$ is zero if $i \notin \{1,\dots,p\}$ or $\boldsymbol{u} \notin \mathbb{V}$ and correspondingly $\beta_j(\boldsymbol{u})$ is zero if $j \notin \{1,\dots,q\}$ or $\boldsymbol{u} \notin \mathbb{V}$. By insertion of the definition of $\sigma_t^2(\boldsymbol{u})$, (3.8) can be written as

$$X_t^2(\boldsymbol{u}) = \alpha_0 + \sum_{s=1}^{p} \sum_{\boldsymbol{v} \in \mathbb{V}} \alpha_s(\boldsymbol{v}) X_{t-s}^2(\boldsymbol{u}-\boldsymbol{v}) + \sum_{s=1}^{q} \sum_{\boldsymbol{v} \in \mathbb{V}} \beta_s(\boldsymbol{v}) \sigma_{t-s}^2(\boldsymbol{u}-\boldsymbol{v}) + V_t(\boldsymbol{u}). \tag{3.9}$$

By (3.8) we also have that $\sigma_t^2(\boldsymbol{u}) = X_t^2(\boldsymbol{u}) - V_t(\boldsymbol{u})$. Inserting this in (3.9) we get

$$\begin{aligned} X_t^2(\boldsymbol{u}) &= \alpha_0 + \sum_{s=1}^{p} \sum_{\boldsymbol{v} \in \mathbb{V}} \alpha_s(\boldsymbol{v}) X_{t-s}^2(\boldsymbol{u}-\boldsymbol{v}) + \sum_{s=1}^{q} \sum_{\boldsymbol{v} \in \mathbb{V}} \beta_s(\boldsymbol{v}) (X_{t-s}^2(\boldsymbol{u}-\boldsymbol{v}) - V_{t-s}(\boldsymbol{u}-\boldsymbol{v})) + V_t(\boldsymbol{u}) \\ &= \alpha_0 + \sum_{s=1}^{p \vee q} \sum_{\boldsymbol{v} \in \mathbb{V}} (\alpha_s(\boldsymbol{v}) + \beta_s(\boldsymbol{u})) X_{t-s}^2(\boldsymbol{u}-\boldsymbol{v}) - \sum_{s=1}^{q} \sum_{\boldsymbol{v} \in \mathbb{V}} \beta_s(\boldsymbol{v}) V_{t-s}(\boldsymbol{u}-\boldsymbol{v}) + V_t(\boldsymbol{u}). \end{aligned}$$

As we can see, the squared process can be written as a linear combination of the parameters. On this form, $\{X_t^2(\boldsymbol{u})\}$ is an ARMA process with expectation σ^2, given by (3.6).

This ARMA representation can be quite useful in estimation – especially for the STARCH models. As for the regular GARCH, the ARMA representation allows us to use ARMA procedures in estimation of parameters, however not as straight forward as for regular GARCH. We have not done this explicitly, but in the least squares approach to estimation we present in Chapter 5, we make use of (3.8) in a vectorized form.

3.3. Boundary Issues

Up till now, the space where the process is defined has not been limited in any way. From a theoretical perspective we can have an unlimited space. However, in applications it will not be possible. We usually have to limit the model to some area of interest. This will then lead to some issues along the boundary of the area. There are many ways of dealing with these issues. We suggest two solutions.

3.3.1. Conditioning on the boundary

One way to solve the boundary problem is to condition on the boundary observations. The issue arises, because we can not calculate $\sigma_t(\boldsymbol{u})$ on the boundary, due to lack of observations outside the boundary. The idea here is that we only use the boundary observations for calculating $\sigma_t(\boldsymbol{u})$, but they are not included in the parameter estimation. Usually, one will have to condition on the initial state of the process, but here we also condition on the boundary. To illustrate this approach, consider the situation where the space is a line ($d = 1$), and $X_t(u) \sim$ STARCH(1) where $\sigma_t(u)$ only depends on $X_{t-1}(u-1)$, $X_{t-1}(u)$ and $X_{t-1}(u+1)$. Let v be the point on the edge of the area of interest. We will not be able to apply the model to this point, because we do not have observations outside the area of interest. We can then reduce the area of interest by using the values along the boundary, only as the dependent values for the observations one

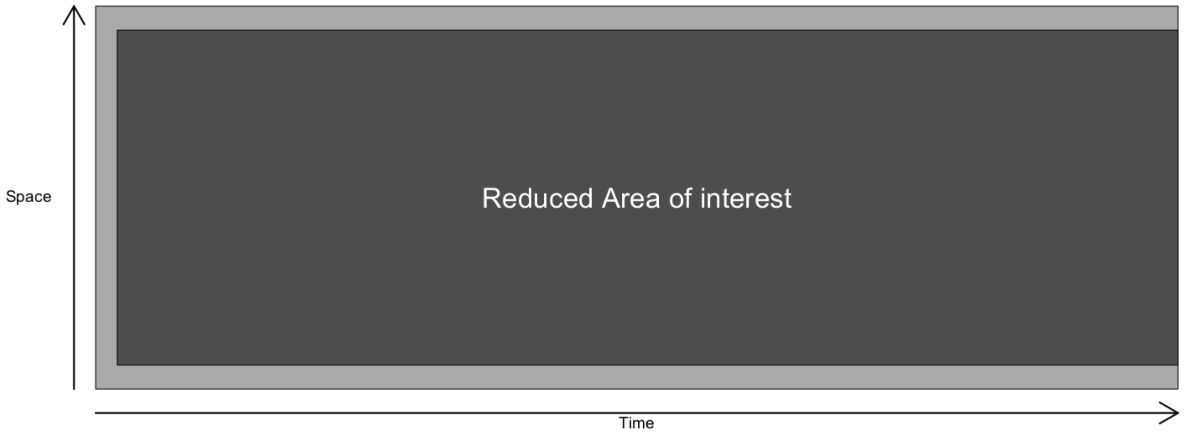


Figure 3.1: *Illustration of how the reduction of area of interest can be done in one spatial dimension by conditioning on the boundary. The lighter grey area is what we condition on during the estimation.*

point from the boundary. Figure 3.1 illustrates this example. Imagine we have observations for the entire rectangle, but we only base our parameter estimation on the dark grey area, while the lighter grey is only used to approximate $\{\sigma_t(u)\}$.

In a model with closest neighbour structure and one spatial dimension, this approach will cut away $2 \cdot (T-1)$ observations, T being the temporal sample size. The number of points lost does not depend on the number of spatial observations you have. In one dimensional space, this is not a great loss. However, in two dimensions, the number of «lost» variables is $2(d_1+d_2-2)(T-1)$, where d_1 is the number of points on the first axes and d_2 on the second. Hence, in 2D the size of the area of interest has a big influence on the performance of the procedure. This can of course be a big drawback in many cases.

In 1D this can be a useful way of dealing with the boundary issue, but in higher dimension the loss of data can potentially be too great. Conditioning on the boundary would therefore require sufficient amount of data, in order to afford surrendering some it. It is also not obvious how this procedure can be used for STGARCH processes, since estimating the unobserved process on the boundary can potentially be a problem. We return to this approach in Chapter 6.

3.3.2. The Circular model

Another method, maybe more sophisticated, is to assume that the process is circular in the spatial dimension(s). By circular we mean that we assume that the boundaries are connected. We got this idea from the old mobile phone game called «snake», popular in the 1990s. The snake moves around in a plane and the goal is to catch food lying around. The snake increases in size for each meal and if you crash into yourself or a wall, the game is over. In some game modes, the snake is allowed to go through the walls and come back out of the wall on the opposite side of the board. This notion of going through a wall and coming back out on the opposite side, is what we call the circularity of our models.

In one dimension, the process is distributed on a line for each time point. Imagine we bend the line, connecting the two end points and creating a circle, making the two endpoints neighbours. You can think of it as a snake bending its body to bite its own tail (not a reference to the game above).

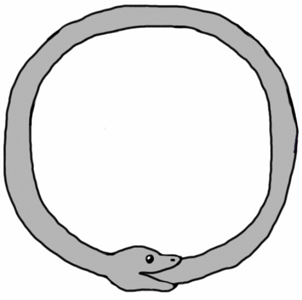


Figure 3.2: *Analogy to a circular model in one dimension – a snake biting its own tail.*
Illustration: Christi Elin M. Nedreås

The obvious downside to this approach is that one may create a dependency between two points that are in fact very far from each other. The argument against this critic is that (hopefully) this effect will be minor. Advantages with this approach is that we do not get any truncation effect. We get a closed model and under the circular assumption, we have observed the entire process. There are no missing data outside our boundaries, because there are no boundaries. The process becomes an adapted process in time. As we will see in Chapter 5, the fact that the process is adapted is essential for expressing the likelihood function at all. You can see how the circular process looks like in one and two spatial dimensions in Figure 3.3. In one dimension the circular model is defined on circles and in two dimensions on the surface of toruses or donuts.

In Chapter 5, we develop estimation theory under the circular assumption and we create simulation procedures to simulate STGARCH processes using the same assumption in Chapter 4. It turns out that assuming a circular space makes both simulation and estimation of the spatio-temporal process easier.

3.4. Applications

We have presented a theoretical process – but what do we use it for? GARCH models have traditionally been used for modelling the behaviour of the stock market or other types of financial returns. We have already mentioned volatility clustering as a key feature for GARCH models.

Continuing the traditional use of GARCH models, one can try to model the world's different

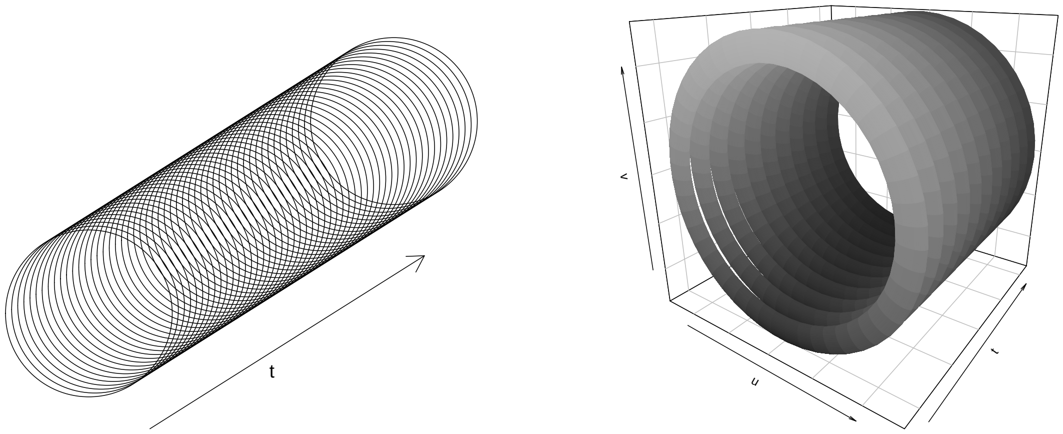


Figure 3.3: *Circular process: In one spatial dimension the lines at each time step are connected at the ends forming circles (left hand side), while in two dimensions connecting the boundaries will map the space onto surfaces of toruses (right hand side).*

stock exchanges' pricing indexes as a STGARCH. We could say that the neighbouring stock exchanges influence each other the most. This would also give a natural interpretation of the circular model. However, when the stock exchanges in USA open, they already know what happened at the stock exchange in China the same day, hence this will influence their trading day much more than what happened the day before. Some alteration of the model will be required. Perhaps one could only consider neighbouring stock exchanges in both space and time. By this, we mean that USA stock exchanges' volatility would depend on yesterdays results for the closest neighbours, but for the neighbours across the Pacific Ocean we could use today's price. This complicates the model somewhat, but should not influence the results presented in this thesis substantially.

Modelling the European stock exchanges might be a better idea. One could think that the volatility in neighbouring countries' (countries that share a border) markets might effect each other, due to spatial proximity. Fitting a STGARCH model could be a way to test if there is a spatial effect.

Satellite data has been suggested as a possible application. Cressie and Wikle (2011) says *«satellite observations of an atmospheric quantity might be affected by the presence of clouds, so that measurements in cloudy regions have a different measurement-error variance than those in clear regions.»* They suggest using a spatially varying GARCH or stochastic volatility model to capture these spatially and temporally explicit variances. Figure 3.4 shows a temporal snapshot of a simulated squared STARCH process. There is a striking resemblance to clouds in the sky – some more dense than others.

An obvious application of STGARCH is to combine this modelling with spatial temporal regression or autoregression type models. An example is a STARMA(p,q) model. For such a model the innovations could be a STGARCH process.

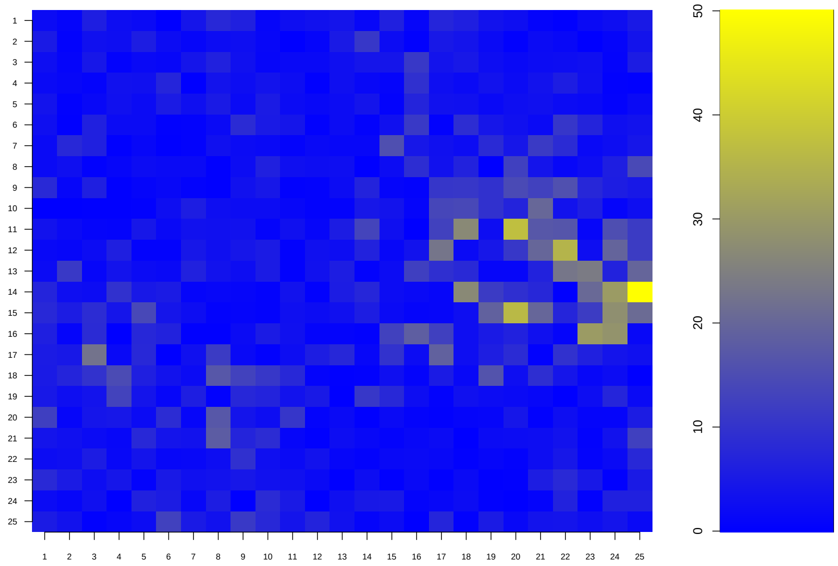


Figure 3.4: *Slice of simulated* STARCH *absolute process at a specific point in time. Dimension is* 25×25. *Function created by Seidel (2016).*

Chapter 4

Simulation

Different simulation techniques for the STGARCH model are developed in this chapter. One may think that simulation in this model is straightforward, but as mentioned in the final part of the previous chapter we meet trouble on the boundary. We discuss the techniques of «circular» and «regular» simulation. The plan is to introduce simulation techniques in this chapter and estimation in the next, so that we can do Monte Carlo studies of the properties of the estimators in Chapter 6.

Figure 4.1 shows a simulated one dimensional STARCH process with parameters $\alpha_0 = 0.3$ and $\alpha_1 = 0.33$. For details about the process specification, see the introduction to Chapter 6. We have plotted both the process and the absolute process. It is easier to see that the extremes are clustered together in both time and space when considering the absolute process, due to the nature of three-dimensional plots. The R-package used for creating these plots are **plot3D** published by Soetaert (2014).

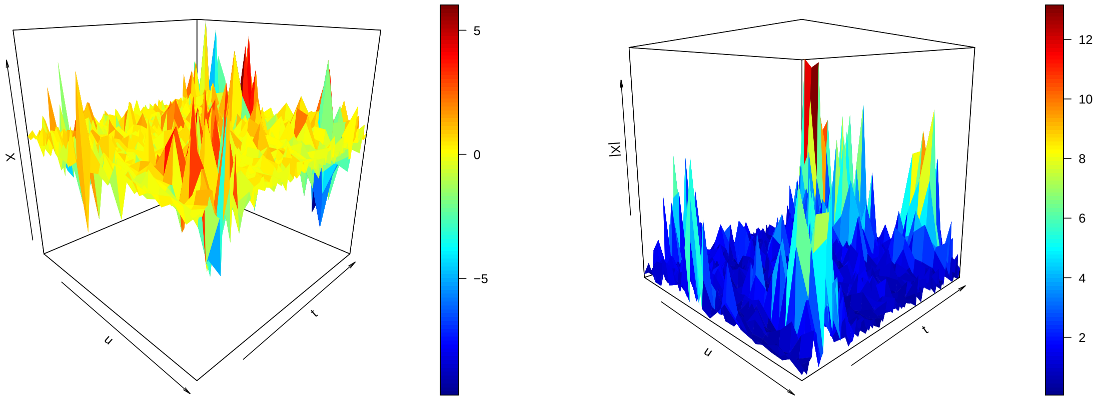


Figure 4.1: *Simulated 1D circular* STARCH(1) *process with parameters* $\alpha_0 = 0.3$ *and* $\alpha_1 = 0.33$. *Data size is* 20×100. *The figure to your left is a 3D plot of the process while the plot to your right is the absolute process.*

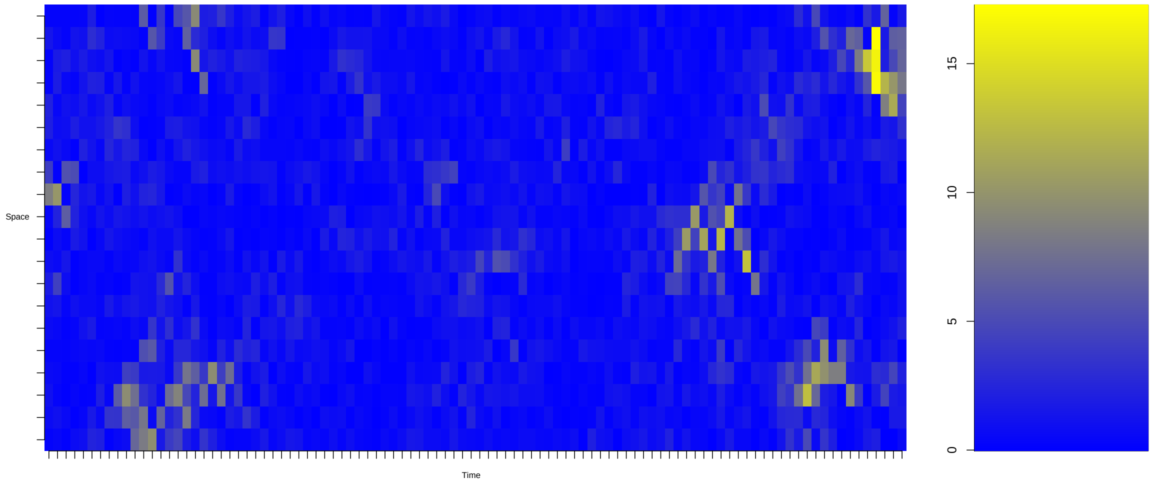


Figure 4.2: *Circular simulated absolute 1D* STARCH(1) *process with parameters* $\alpha_0 = 0.3$ *and* $\alpha_1 = 0.33$. *Data size is* 20×100 *and it is the same absolute process as in Figure 4.1. The plot illustrates the clustering of extremes. R function created by Seidel (2016).*

Figure 4.2 is a matrix plot of the same absolute process as in Figure 4.1. This plot makes it easier to see the clustering effect in both space and time. You can see how the yellow coloured rectangles are grouped together. You can also see the circularity in this plot. There is a large cluster of extremes at the beginning of this time interval and you can see that it is both on the top and bottom of the plot. This is because these two clusters are connected through the circularity.

This chapter is dedicated to showing you how we made this simulation.

4.1. Vector notation

For this chapter and the next, a vector notation is quite convenient. Up till now, it has been useful to consider the process at each point in both time and space, but now it is time to introduce a notation that vectorize the process at each time point. The reason why we have not introduced this notation earlier, is that per definition 3.1.1, the process $\{X_t(\boldsymbol{u}) : t = 1, \dots, T, \boldsymbol{u} \in \mathbb{Z}^d\}$ is in principle spatially unlimited. From this chapter onwards we go from the theoretical process to trying model something as a STGARCH process. Instead of having an unlimited spatial process, we have to consider the process on a finite area. This means that we can form a finite vector process at each point in time.

Instead of using the vector $\boldsymbol{u}$ to determine the location in space, we numerate each location by $i = 1, \dots, M$, where M is the total number of spatial points. In this way we can create the vectorized process, defined for $t = 1, \dots, T$ as

$$\boldsymbol{X}_t = \begin{bmatrix} X_{t,1}, & X_{t,2}, & \dots, & X_{t,M} \end{bmatrix}^T . \tag{4.1}$$

Just like for the vectorized $\{\boldsymbol{X}_t\}$ process, we need a convenient vector formulation of $\sigma_t^2(\boldsymbol{u})$. To make notation effective, we need a way of squaring a vector. For that purpose we define the Hadamard product of vectors. Let $\boldsymbol{a} = [a_1, \dots, a_n]^T$ and $\boldsymbol{b} = [b_1, \dots, b_n]^T$, then the Hadamard product of $\boldsymbol{a}$ and $\boldsymbol{b}$ is defined as

$$\boldsymbol{a} \circ \boldsymbol{b} = \begin{bmatrix} a_1 b_1, & a_2 b_2, & \dots, & a_n b_n \end{bmatrix}^T . \tag{4.2}$$

We call it the Hadamard squared of a vector when you Hadamard multiply the vector with itself and note this as a squared vector; $\boldsymbol{a}^2$. In practise, this means squaring each element of the vector. We can then define the vector processes $\boldsymbol{\sigma}_t$ and $\boldsymbol{\sigma}_t^2$ for each $t = 1, \dots, T$ by

$$\begin{aligned} \boldsymbol{\sigma}_t &= \begin{bmatrix} \sigma_{t,1}, & \sigma_{t,2}, & \dots, & \sigma_{t,M} \end{bmatrix}^T , \\ \boldsymbol{\sigma}_t^2 &= \begin{bmatrix} \sigma_{t,1}^2, & \sigma_{t,2}^2, & \dots, & \sigma_{t,M}^2 \end{bmatrix}^T . \end{aligned} \tag{4.3}$$

We use the $M \times M$ neighbourhood matrix $\boldsymbol{\mathcal{W}}$ (defined by Equation (2.1)) to determine which points are considered neighbours. In this situation, we would also like the neighbourhood matrix to include the diagonal, and let $\boldsymbol{\mathcal{W}}^\star = \boldsymbol{\mathcal{W}} + \mathbb{I}$. This matrix will replace the sets $\mathbb{V}_a$ and $\mathbb{V}_b$, from definition 3.1.1. Like mentioned in Chapter 3, we assume the same neighbourhood structure for both the ARCH and GARCH terms and one parameter per lag. This is a practical, but unnecessary limitation. This means that

$$\boldsymbol{\theta} = \begin{bmatrix} \alpha_0, & \alpha_1, & \dots, & \alpha_p, & \beta_1, & \dots, & \beta_q \end{bmatrix}^T . \tag{4.4}$$

By the notation above, we can now form a design matrix, $\mathbb{X}_t$ with $1 + p + q$ columns, defined as

$$\mathbb{X}_t = \begin{bmatrix} \boldsymbol{1}, & \boldsymbol{\mathcal{W}}^\star \boldsymbol{X}_{t-1}^2, & \dots, & \boldsymbol{\mathcal{W}}^\star \boldsymbol{X}_{t-p}^2, & \boldsymbol{\mathcal{W}}^\star \boldsymbol{\sigma}_{t-1}^2, & \dots, & \boldsymbol{\mathcal{W}}^\star \boldsymbol{\sigma}_{t-q}^2 \end{bmatrix} . \tag{4.5}$$

Now we are ready to state Equation (3.1) in the vector notation.

$$\begin{aligned} \boldsymbol{X}_t &= \operatorname{diag}\{\boldsymbol{\sigma}_t\} \boldsymbol{Z}_t, \\ \boldsymbol{\sigma}_t^2 &= \mathbb{X}_t \boldsymbol{\theta}, \end{aligned} \tag{4.6}$$

where diag is the operator that creates a diagonal matrix with the argument vector along its diagonal. In this notation, we can treat all dimensions equally. The only thing that gets more complicated form 1D to 2D is the neighbourhood matrix $\boldsymbol{\mathcal{W}}^\star$.

4.2. Initiating and Burn-In

The STGARCH model develops through time in an iterative matter. If we know everything that happened in the past, we can calculate the next $\boldsymbol{\sigma}_t$. However, we have to start at some point. We need to set some initial values for $\boldsymbol{\sigma}_0$ and $\boldsymbol{X}_0$. In fact, if we are to simulate a STGARCH(p,q) process, we need to set values for $\{\boldsymbol{\sigma}_0, \boldsymbol{\sigma}_{-1}, \dots, \boldsymbol{\sigma}_{-q+1}\}$ and $\{\boldsymbol{X}_0, \boldsymbol{X}_{-1}, \dots, \boldsymbol{X}_{-p+1}\}$, where the vectors consist of the process values at all spatial points. There are different ways of dealing with this. One way is to draw random initial values for the $\boldsymbol{X}$'s and absolute random initial values for the $\boldsymbol{\sigma}$'s, both from the innovation distribution. Another approach is to set everything to zero, so that the first simulated observation will yield $\boldsymbol{X}_1 = \text{diag}\left\{\mathbf{1}\sqrt{\alpha_0}\right\} \boldsymbol{Z}_1$, because $\boldsymbol{\sigma}_1 = \mathbf{1}\sqrt{\alpha_0}$. After simulating $\max\{p, q\} + 1$ time steps, we have non-zero values on all the elements going into the calculation of $\boldsymbol{\sigma}_t$.

Which method we choose, does not really matter, as long as we find a way to get the process started. The reason for this is that we do a temporal burn-in of the process. A burn-in means that you simulate many time steps of the process before you start on the part you will actually use as your simulated dataset. You give the process time to stabilize, making your initial value set-up insignificant. We hold the space fixed and simulate forward in time. Unless dimensions are extremely high, simulating data is not a very demanding task for the computer.

4.3. Simulation Algorithm

Like we have mentioned in Chapter 3 the circular model can be a good approximation to a non-circular setting. However, when we simulate, we can create a truly circular dataset. From a theoretical perspective, this gives us the opportunity to test the performance of circular estimation (Chapter 5) – both on circular and non-circular data. This is our motivation for simulating truly circular models.

4.3.1. Circular Simulation

When we are in a circular setting, points on each side of the boundaries are neighbours, just like neighbouring points in the middle of the grid. This means we have a finite number of spatial points and observations for all neighbours. The circular assumption creates a closed system.

Using the notation we set up in section 4.1, forming a simulation algorithm is almost straightforward. Define the neighbourhood matrix, $\boldsymbol{\mathcal{W}}^{\star}$, in such a manner that the points on opposite boundaries are defined as neighbours. We then need to set the parameter vector given in (4.4) and the initial values discussed in the previous section. Then you start iterating, by first calculating the next $\boldsymbol{\sigma}_t$ and draw $\boldsymbol{Z}_t$ from the iid innovation distribution and calculate the next $\boldsymbol{X}_t$ by (4.6). We summarize the routine in Algorithm 1.

Algorithm 1 Circular Simulation

Let B be the size of the burn-in and T the temporal size of the simulated dataset, both in the temporal dimension.

```
Set the neighbourhood matrix $\boldsymbol{\mathcal{W}}^\star$
Set parameter vector $\boldsymbol{\theta} = [\alpha_0, \boldsymbol{\alpha}^T, \boldsymbol{\beta}^T]^T$
Do (for instance) one of the following:
    • Initialize $\boldsymbol{\sigma}_0 = \cdots = \boldsymbol{\sigma}_{-q+1} = \boldsymbol{X}_0 = \cdots = \boldsymbol{X}_{-p+1} = \mathbf{0}$
    • Draw $\boldsymbol{Z}_j \sim \text{MVN}(\mathbf{0}, \mathbb{I})$ and set $\boldsymbol{\sigma}_j = |\boldsymbol{Z}_j|$, $j = 0, \dots, -q+1$. Draw $\boldsymbol{X}_i \sim \text{MVN}(\mathbf{0}, \mathbb{I})$, $i = 0, \dots, -p+1$.
for each node $t = 1, \dots, B+T$ do
    Set
        $\mathbb{X}_t = \left[\mathbf{1}, \ \boldsymbol{\mathcal{W}}^\star \boldsymbol{X}_{t-1}^2, \ \dots, \ \boldsymbol{\mathcal{W}}^\star \boldsymbol{X}_{t-p}^2, \ \boldsymbol{\mathcal{W}}^\star \boldsymbol{\sigma}_{t-1}^2, \ \dots, \ \boldsymbol{\mathcal{W}}^\star \boldsymbol{\sigma}_{t-q}^2\right].$
    Calculate $\boldsymbol{\sigma}_t^2 = \mathbb{X}_t \boldsymbol{\theta}$
    Draw $\boldsymbol{Z}_t \sim \text{MVN}(\mathbf{0}, \mathbb{I})$
    Calculate $\boldsymbol{X}_t = \boldsymbol{\sigma}_t \boldsymbol{Z}_t$
end for
Return $\{\boldsymbol{X}_{B+1}, \dots, \boldsymbol{X}_{B+T}\}$
```

4.3.2. Regular Simulation

Now, we leave the circular assumption. The alternative to a circular model, is that we have an unlimited area where the process takes place, but we only observe a part of it. This means the

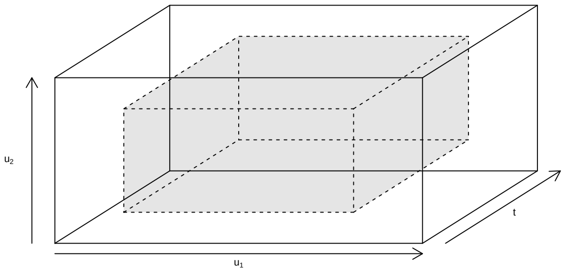


Figure 4.3: Regular simulation in 2D: Simulate a large circular dataset and truncate it to create a non-circular dataset (grey area).

process is truncated. This truncation has some side effects, but in this chapter we only consider how we generate this kind of process.

There are probably many ways of doing a regular simulation. We suggest using the circular algorithm (Algorithm 1). We illustrate how this can be done in a one spatial dimension model, but the procedure is equivalent for higher dimension. Assume that our area of interest is of size $M \times T$ (spatial × temporal). We can simulate a $(M + 2L) \times T$ circular model using Algorithm 1, which will give us $(M + 2L) \times T$ simulated observations of the process $\{X_t(u)\}$. If we discard the L furthest points in each spatial direction, we get a non-circular dataset of size $M \times T$. In higher dimensions there are simply more directions to truncate the circular dataset in. Figure 4.3 illustrates how we do this in two spatial dimensions. The larger white box illustrates the circularly simulated data, while the grey smaller box is the truncated dataset. We need to make L sufficiently large, to make sure we do not get a circular effect, especially if T is large or M is small.

4.4. Example: ACF and Marginal Density

An example of useful things we can do with simulated datasets is to look at the auto correlation function (ACF) of the squared process. With the opportunity to simulate data, we can look at the sample ACF instead of deriving a theoretical version. This will also illustrate some of the differences between circular and regular simulation.

We have plotted the ACF of a circular and a regular 1D STARCH process with the «closest neighbour» structure we use in Chapter 6 (see introduction to Chapter 6 for more details). The parameter vector is $\boldsymbol{\theta} = [\alpha_0, \alpha_1]^T = [1.4, 0.3]^T$. The dimension is set to 10 × 100. The sample auto covariance functions for both circular and regular models are given by

$$\begin{aligned}
\widehat{\gamma}_C(h,s) &= \frac{1}{M(T-s)} \sum_{t=1}^{T-s} \sum_{u=1}^{M} \left(x_t^2(u) - \widehat{\mu}\right)\left(x_{t+s}^2(\mathrm{mod}[u+h,n]) - \widehat{\mu}\right) \\
\widehat{\gamma}_R(h,s) &= \frac{1}{(T-s)(M-h)} \sum_{t=1}^{T-s} \sum_{u=1}^{M-h} \left(x_t^2(u) - \widehat{\mu}\right)\left(x_{t+s}^2(u+h) - \widehat{\mu}\right)
\end{aligned} \tag{4.7}$$

where $\widehat{\mu} = \frac{1}{MT} \sum_{t=1}^{T} \sum_{u=1}^{M} x_t(u)^2$ and the auto correlation functions are

$$\widehat{\rho}_C(h,s) = \frac{\widehat{\gamma}_C(h,s)}{\widehat{\gamma}_C(0,0)} \qquad \widehat{\rho}_R(h,s) = \frac{\widehat{\gamma}_R(h,s)}{\widehat{\gamma}_R(0,0)} \tag{4.8}$$

We simulated two datasets, one circular and one regular, using the same set-up, and calculated the ACF for both cases. The results are presented in Figure 4.4. This visualizes the circular effect very nicely. We also see that the correlation fades as lags increase, both spatially and temporally.

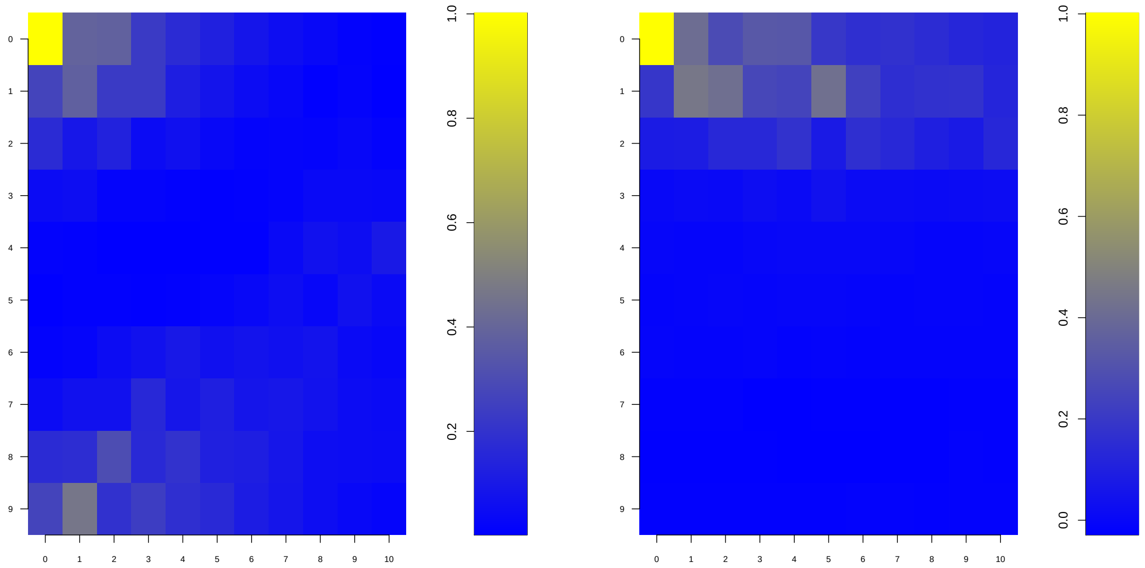


Figure 4.4: *Circular and Regular ACF of STARCH(1): Temporal lag on the x-axis and spatial lag on the y-axis. Function is created by Seidel (2016).*

For the same datasets, we would also like to view the marginal densities. Figure 4.5 shows the non-parametric marginal density estimate of the simulated datasets. We used the *density* function from the R-package **stats** (R Core Team, 2015). We can see from the figure that the marginal densities are practically the same. We have also included a normal density using the sample mean and variance as parameters, for comparison purposes. One of the stylized facts about time series GARCH, is that it is heavier tailed than the normal distribution. This is not easily seen in Figure 4.5. We have therefore included minimum and maximum of the observations, to show the reader that extreme events have occurred. Estimating the kurtosis of the observations, we get approximately 5.5 in both cases, which exceeds the normal kurtosis of 3. This means that the leptokurtic property from time series seems to be preserved for STGARCH, at least for this realization.

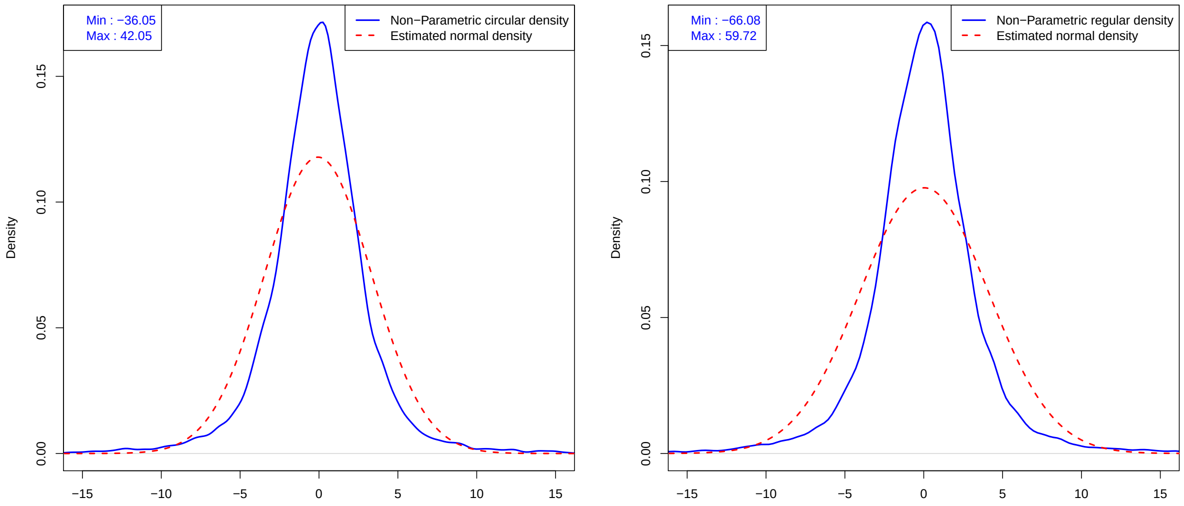


Figure 4.5: *Marginal density of circular 1D* STARCH(1) *process with* $\alpha_0 = 1.4$ *and* $\alpha_1 = 0.3$ *and dimensions are* 10×100.

Chapter 5

Estimation

In this chapter we look at estimation of the STGARCH model. We consider two different methods:

- Maximum Likelihood,
- Least Squares.

The least squares method comes directly from Equation (3.8) used in the ARMA-representation of the squared STGARCH process. Maximum likelihood is the go-to method for estimation of GARCH and we generalize the results to our model. We argue that the maximum likelihood estimators are consistent and asymptotic normally distributed, using results from multivariate GARCH. A sketched proof of asymptotic normality of the maximum likelihood estimators is presented, but this proof is not completely rigorous at this point.

For this entire chapter we assume our process to be circular in the spatial dimension. You may be wondering how we can assume a circular model for a problem that obviously is not circular. First of all, we can use a circular model to approximate the non-circular situation. This will be tested on simulated data in Chapter 6. In Chapter 8 we assume the observed data originates from a circular process, but we only observe a subset of it. A Gibbs sampler is used to get a pseudo sample of the missing area and in collaboration with an EM-algorithm (Estimation-Maximization), we can estimate circularly on non-circular data. In order to do so, we must first establish estimation methods.

5.1. Maximum Likelihood Estimation

Maximum likelihood is one of the main methods in estimation of parametric models - especially when there is a specified innovation distribution involved. We assume standard normally distributed innovations and the formulas and theory we present here relies heavily on the Gaussian assumption. This assumption leads to analytical equations and formulas, which might not be the case for other distributions. If the innovations are not truly Gaussian, the likelihood can be viewed as a Gaussian quasi likelihood.

5.1.1. Likelihood derivation

The first step to any likelihood inference, is to derive the likelihood function and this will be our goal for this section. Deriving the likelihood is sometimes uneasy and this is the case here. By assuming a circular STGARCH(p,q) model, the boundary issues are solved by the non-existence of boundaries and derivation of a likelihood function is possible.

We separate between the two sigma algebras, $\mathcal{F}_t$ and $\mathcal{G}_t$ defined by

$$\begin{aligned} \mathcal{F}_t &= \bigvee \left\{X_s(\boldsymbol{u}); \boldsymbol{u} \in \mathbb{Z}^d, s \leq t\right\}, \\ \mathcal{G}_t &= \bigvee \left\{X_s(\boldsymbol{u}); \boldsymbol{u} \in \mathbb{Z}^d, 1 \leq s \leq t\right\}, \end{aligned} \tag{5.1}$$

where $\mathcal{F}_t$ covers the infinite past of $\{\boldsymbol{X}_t\}$, while $\mathcal{G}_t$ only cover the history dating back to $t = 1$. The reason for this distinction is that when we do likelihood theory, we will condition on some initial values which in theory already are included in $\mathcal{F}_t$.

In a circular model, we have $\boldsymbol{X}_t$ finite. Let M be the number of spatial points and let $t = 1, \ldots, T$. We use the notation introduced in section 4.1 for the STGARCH process. Let

$$\begin{aligned} \mathbb{X}_t &= \begin{bmatrix} \mathbf{1} & \boldsymbol{\mathcal{W}}^\star \boldsymbol{X}_{t-1}^2 & \cdots & \boldsymbol{\mathcal{W}}^\star \boldsymbol{X}_{t-p}^2 & \boldsymbol{\mathcal{W}}^\star \boldsymbol{\sigma}_{t-1}^2 & \cdots & \boldsymbol{\mathcal{W}}^\star \boldsymbol{\sigma}_{t-q}^2 \end{bmatrix}, \\ \boldsymbol{\theta} &= \begin{bmatrix} \alpha_0 & \boldsymbol{\alpha}^T & \boldsymbol{\beta}^T \end{bmatrix}^T, \\ \boldsymbol{\sigma}_t^2 &= \mathbb{X}_t \boldsymbol{\theta}, \\ \boldsymbol{X}_t &= \operatorname{diag}\left\{\boldsymbol{\sigma}_t\right\} \boldsymbol{Z}_t. \end{aligned} \tag{5.2}$$

Here we have assumed symmetry for all neighbours. We use the same α or β parameter for all neighbours, but different parameters for each lag. Also, we have assumed the same neighbourhood relationships for each lag.

Let

$$\mathcal{A}_0 = \bigvee\{\boldsymbol{X}_0, \ldots, \boldsymbol{X}_{-p+1}, \boldsymbol{\sigma}_0, \ldots, \boldsymbol{\sigma}_{-q+1}\} \tag{5.3}$$

be the sigma algebra of initial values. In a circular model, we have a finite space and the process develops through time as an adapted process. The process $\{\boldsymbol{\sigma}_t\}$ is $\mathcal{F}_{t-1}$ measurable. This might not seem obvious, but by iterating the definition of $\{\boldsymbol{\sigma}_t\}$ (see definition 3.1.1) you will end up with an infinite sum of the infinite history of $\boldsymbol{X}_t$. For the purpose of estimation, an infinite history is not possible to come by. Another approach that makes $\{\boldsymbol{\sigma}_t\}$ measurable, is by conditioning on the initial values, $\mathcal{A}_0$, and $\mathcal{G}_{t-1}$. That means, $\boldsymbol{\sigma}_t | \mathcal{A}_0 \vee \mathcal{G}_{t-1}$, for $t \geq 1$, is measurable. Due to the independence of $\boldsymbol{Z}_t \sim \text{MVN}\left(\mathbf{0}, \mathbb{I}\right)$, we have that $\boldsymbol{X}_t | \mathcal{A}_0 \vee \mathcal{G}_{t-1} \sim \text{MVN}\left(\mathbf{0}, \operatorname{diag}\left\{\boldsymbol{\sigma}_t^2\right\}\right)$. We have conditionally independent multivariate random variables, which means that

$$f(\boldsymbol{X}_t|\mathcal{G}_{t-1} \vee \mathcal{A}_0, \boldsymbol{\theta}) = (2\pi)^{-M/2} |\operatorname{diag}\left\{\boldsymbol{\sigma}_t^2\right\}|^{-1/2} \exp\{-\frac{1}{2}\boldsymbol{X}_t^T \operatorname{diag}\left\{\boldsymbol{\sigma}_t^2\right\}^{-1} \boldsymbol{X}_t\}. \tag{5.4}$$

In order to optimize the likelihood function we find the log-likelihood by taking the logarithm of (5.4). We use that the logarithm of the determinant of a diagonal matrix is simply the sum of the logarithm of the diagonal elements.

$$\begin{aligned} \log f(\boldsymbol{X}_t|\mathcal{G}_{t-1} \vee \mathcal{A}_0, \boldsymbol{\theta}) &= -\frac{M}{2}\log(2\pi) - \frac{1}{2}\log|\operatorname{diag}\left\{\boldsymbol{\sigma}_t^2\right\}| - \frac{1}{2}\boldsymbol{X}_t^T \operatorname{diag}\left\{\boldsymbol{\sigma}_t^2\right\}^{-1} \boldsymbol{X}_t \\ &= -\frac{M}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{M}\log\left(\mathbb{X}_{t,i}\boldsymbol{\theta}\right) - \frac{1}{2}\sum_{i=1}^{M} X_{t,i}(\mathbb{X}_{t,i}\boldsymbol{\theta})^{-1}X_{t,i} \end{aligned} \tag{5.5}$$

We have here used $\mathbb{X}_{t,i}$ to denote the ith row of $\mathbb{X}_t$. Now, we have an expression for the distribution of $\boldsymbol{X}_t|\mathcal{A}_0 \vee \mathcal{G}_{t-1}$, but we want the simultaneous distribution of $\boldsymbol{X} = \{\boldsymbol{X}_t : t = 1, \ldots, T\}$. Let $\pi_0 = f(\mathcal{A}_0|\boldsymbol{\theta})$. Then

$$L(\boldsymbol{\theta}) = f(\boldsymbol{X}|\boldsymbol{\theta}) = \pi_0 f\big(\boldsymbol{X}|\mathcal{A}_0, \boldsymbol{\theta}\big) = \pi_0 f\big(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T|\mathcal{A}_0, \boldsymbol{\theta}\big). \tag{5.6}$$

Due to conditional independence between each time step and by using the definition of conditional probability, we have that

$$\begin{aligned} L(\boldsymbol{\theta}) &= \pi_0 f(\boldsymbol{X}_1|\mathcal{A}_0, \boldsymbol{\theta}) f(\boldsymbol{X}_2, \ldots, \boldsymbol{X}_T|\boldsymbol{X}_1, \mathcal{A}_0, \boldsymbol{\theta}) \\ &= \pi_0 f(\boldsymbol{X}_1|\mathcal{A}_0, \boldsymbol{\theta}) f(\boldsymbol{X}_2, \ldots, \boldsymbol{X}_T|\mathcal{G}_1 \vee \mathcal{A}_0, \boldsymbol{\theta}) \\ &= \pi_0 f(\boldsymbol{X}_1|\mathcal{A}_0, \boldsymbol{\theta}) f(\boldsymbol{X}_2|\mathcal{G}_1 \vee \mathcal{A}_0, \boldsymbol{\theta}) f(\boldsymbol{X}_3, \ldots, \boldsymbol{X}_T|\boldsymbol{X}_2, \mathcal{G}_1 \vee \mathcal{A}_0, \boldsymbol{\theta}) \\ &= \ldots \\ &= \pi_0 \prod_{t=1}^{T} f(\boldsymbol{X}_t|\mathcal{G}_{t-1} \vee \mathcal{A}_0, \boldsymbol{\theta}), \end{aligned} \tag{5.7}$$

where $f(\boldsymbol{X}_t|\mathcal{G}_{t-1} \vee \mathcal{A}_0, \boldsymbol{\theta})$ is given by (5.4). We find the log-likelihood by taking the logarithm of (5.7) and insert (5.5).

$$\begin{aligned} l(\boldsymbol{\theta}) &= \log \pi_0 + \sum_{t=1}^{T} \log f(\boldsymbol{X}_t|\mathcal{G}_{t-1} \vee \mathcal{A}_0, \boldsymbol{\theta}) \\ &= \log \pi_0 - \frac{MT}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\sum_{i=1}^{M}\log\left(\mathbb{X}_{t,i}\boldsymbol{\theta}\right) - \frac{1}{2}\sum_{t=1}^{T}\sum_{i=1}^{M} X_{t,i}(\mathbb{X}_{t,i}\boldsymbol{\theta})^{-1}X_{t,i}. \end{aligned} \tag{5.8}$$

Since π_0 is unknown and difficult to estimate we employ the conditional likelihood instead, $l(\boldsymbol{\theta}|\mathcal{A}_0)$, where conditioning on $\mathcal{A}_0$ makes $\log \pi_0$ non-stochastic.

5.1.2. Optimization of the Likelihood

We can express the corresponding score function by finding the derivative of $l(\boldsymbol{\theta}|\mathcal{A}_0)$ with respect to $\boldsymbol{\theta}$.

$$\begin{aligned}\boldsymbol{U}(\boldsymbol{\theta}) = \frac{\partial}{\partial\boldsymbol{\theta}} l(\boldsymbol{\theta}|\mathcal{A}_0) &= -\frac{1}{2}\sum_{t=1}^{T}\sum_{i=1}^{M}\frac{\mathbb{X}_{t,i}^T}{\mathbb{X}_{t,i}\boldsymbol{\theta}} + \frac{1}{2}\sum_{t=1}^{T}\sum_{i=1}^{M}\frac{\mathbb{X}_{t,i}^T X_{t,i}^2}{(\mathbb{X}_{t,i}\boldsymbol{\theta})^2} \\ &= \frac{1}{2}\sum_{t=1}^{T}\sum_{i=1}^{M}\frac{\mathbb{X}_{t,i}^T(X_{t,i}^2 - \mathbb{X}_{t,i}\boldsymbol{\theta})}{(\mathbb{X}_{t,i}\boldsymbol{\theta})^2}.\end{aligned} \tag{5.9}$$

To find the parameter vector $\boldsymbol{\theta}$ that maximizes the conditional log-likelihood function we need to solve the equation $\boldsymbol{U}(\boldsymbol{\theta}) = \mathbf{0}$. Let $\mathbb{W}_t(\boldsymbol{\theta}) = \operatorname{diag}\left\{(\sqrt{2}\mathbb{X}_t\boldsymbol{\theta})^{-2}\right\}$ be a diagonal weight matrix for $t = 1, \ldots, T$ and let $\Delta_{t,i}(\boldsymbol{\theta}) = \mathbb{X}_{t,i}^T(X_{t,i}^2 - \mathbb{X}_{t,i}\boldsymbol{\theta})$. Then our score function is of the form

$$\boldsymbol{U}(\boldsymbol{\theta}) = \sum_{t=1}^{T}\sum_{i=1}^{M}\mathbb{W}_{t,i}(\boldsymbol{\theta})\Delta_{t,i}(\boldsymbol{\theta}),$$

where $\mathbb{W}_{t,i}$ is the i^{th} diagonal element of $\mathbb{W}_t$. In this notation we have a weighted least squares problem, and this can be solved iteratively.

Given $\boldsymbol{\theta}^{(m)}$, calculate $\mathbb{W}_t(\boldsymbol{\theta}^{(m)})$ and solve

$$\sum_{t=1}^{T}\sum_{i=1}^{N}\mathbb{W}_{t,i}(\boldsymbol{\theta}^{(m)})\Delta_{t,i}(\boldsymbol{\theta}^{(m+1)}) = \mathbf{0}. \tag{5.10}$$

The $\boldsymbol{\theta}^{(m+1)}$ that solves this equation, becomes our new parameter vector, and we repeat the routine until convergence. We go on to find the solution of Equation (5.10). To simplify notation, let $\mathbb{W}_t(\boldsymbol{\theta}^{(m)}) = \mathbb{W}_t^{(m)}$.

$$\begin{aligned}\sum_{t=1}^{T}\sum_{i=1}^{M}\mathbb{W}_{t,i}^{(m)}\Delta_{t,i}(\boldsymbol{\theta}^{(m+1)}) &= \sum_{t=1}^{T}\sum_{i=1}^{M}\mathbb{W}_{t,i}^{(m)}\mathbb{X}_{t,i}^T(\mathbb{X}_{t,i}\boldsymbol{\theta}^{(m+1)} - X_{t,i}^2) = \mathbf{0} \\ \sum_{t=1}^{T}\mathbb{X}_t^T\mathbb{W}_t^{(m)}\mathbb{X}_t\boldsymbol{\theta}^{(m+1)} &= \sum_{t=1}^{T}\mathbb{X}_t^T\mathbb{W}_t^{(m)}\boldsymbol{X}_t^2 \\ \Big(\sum_{t=1}^{T}\mathbb{X}_t^T\mathbb{W}_t^{(m)}\mathbb{X}_t\Big)\boldsymbol{\theta}^{(m+1)} &= \sum_{t=1}^{T}\mathbb{X}_t^T\mathbb{W}_t^{(m)}\boldsymbol{X}_t^2.\end{aligned} \tag{5.11}$$

Now, if $\left(\sum_{t=1}^{T}\mathbb{X}_t^T\mathbb{W}_t^{(m)}\mathbb{X}_t\right)$ is an invertible matrix, we can find an analytical expression for the m^{th} iteration maximum likelihood estimator, $\widehat{\boldsymbol{\theta}}_{\text{ML}}^{(m+1)}$.

$$\widehat{\boldsymbol{\theta}}_{\text{ML}}^{(m+1)} = \Big(\sum_{t=1}^{T}\mathbb{X}_t^T\mathbb{W}_t^{(m)}\mathbb{X}_t\Big)^{-1}\sum_{t=1}^{T}\mathbb{X}_t^T\mathbb{W}_t^{(m)}\boldsymbol{X}_t^2. \tag{5.12}$$

We have not proven it, but we must assume that $\widehat{\boldsymbol{\theta}}_{\text{ML}}^{(m+1)}$ converge to $\widehat{\boldsymbol{\theta}}_{\text{ML}}$ and that $\widehat{\boldsymbol{\theta}}_{\text{ML}}$ is a solution to (5.11). We should also have proven that $\widehat{\boldsymbol{\theta}}_{\text{ML}}$ in fact is a maximum for the likelihood function, but we suspect that the results from Francq et al. (2004) and Davis and Mikosch (2009) will generalize to our situation. When the iteration has converged, we can write

$$\widehat{\boldsymbol{\theta}}_{\mathrm{ML}} = \Big(\sum_{t=1}^{T} \mathbb{X}_t^T \mathbb{W}_t \mathbb{X}_t \Big)^{-1} \sum_{t=1}^{T} \mathbb{X}_t^T \mathbb{W}_t \boldsymbol{X}_t^2. \tag{5.13}$$

The variance of the converged estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}$, can be approximated by

$$\widehat{\mathrm{Var}}\big(\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}\big) = N^{-1} \Big(\sum_{t=1}^{T} \mathbb{X}_t^T \mathbb{W}_t \mathbb{X}_t \Big)^{-1}, \tag{5.14}$$

where $N = MT$ is the total number of data. In what follows, we will derive (5.14).

Notice that Equation (5.13) can be written as

$$\widehat{\boldsymbol{\theta}}_{\mathrm{ML}} = \Big(N^{-1} \sum_{t=1}^{T} \mathbb{X}_t^T \mathbb{W}_t \mathbb{X}_t \Big)^{-1} N^{-1} \sum_{t=1}^{T} \mathbb{X}_t^T \mathbb{W}_t \boldsymbol{X}_t^2. \tag{5.15}$$

The reason for normalizing with N and not T is that when you multiply the matrices, you will get a sum over the spatial vector as well. Our hypothesis is that asymptotically the matrix

$$N^{-1} \sum_{t=1}^{T} \mathbb{X}_t^T \mathbb{W}_t \mathbb{X}_t \to \mathcal{J} \tag{5.16}$$

will converge to a constant matrix. We have not proven this theoretically, but empirical experiments substantiates the claim (see Appendix B.2). We therefore concentrate on the term $N^{-1} \sum_{t=1}^{T} \mathbb{X}_t^T \mathbb{W}_t \boldsymbol{X}_t^2$. Let $\boldsymbol{S}_t = \mathbb{X}_t^T \mathbb{W}_t \boldsymbol{X}_t^2$. We want an asymptotic expression for

$$\mathcal{I} = \mathrm{Var}\left(\boldsymbol{S}_1\right). \tag{5.17}$$

The ergodic theorem (Billingsley, 1995, Th.24.1, pp. 314) tells us that

$$N^{-1} \sum_{t=1}^{T} \mathrm{E}\Big(\big(\boldsymbol{S}_t - \mathrm{E}\left(\boldsymbol{S}_t\right)\big) \big(\boldsymbol{S}_t - \mathrm{E}\left(\boldsymbol{S}_t\right)\big)^T | \mathcal{F}_{t-1} \Big) \to \mathrm{Var}\left(\boldsymbol{S}_1\right), \tag{5.18}$$

if the process is stationary, ergodic and has finite moments. For $\boldsymbol{S}_t$ to be stationary, $\boldsymbol{X}_t^2$ must be. We must therefore assume a finite 4th moment of $\boldsymbol{X}_t$. By assuming the conditions for using the Ergodic theorem fulfilled, we can use the sum of conditional variances to obtain an asymptotic expression.

$$\begin{aligned} \sum_{t=1}^{T} \mathrm{Var}\left(\boldsymbol{S}_t | \mathcal{F}_{t-1}\right) &= \sum_{t=1}^{T} \mathrm{Var}\left(\mathbb{X}_t^T \mathbb{W}_t \boldsymbol{X}_t^2 | \mathcal{F}_{t-1}\right) \\ &= \sum_{t=1}^{T} \mathrm{Var}\left(\mathbb{X}_t^T \mathbb{W}_t \,\mathrm{diag}\left\{\mathbb{X}_t \boldsymbol{\theta}\right\} \boldsymbol{Z}_t^2 | \mathcal{F}_{t-1}\right) \\ &= \sum_{t=1}^{T} \mathbb{X}_t^T \mathbb{W}_t \,\mathrm{diag}\left\{\mathbb{X}_t \boldsymbol{\theta}\right\} \mathrm{Var}\left(\boldsymbol{Z}_t^2 | \mathcal{F}_{t-1}\right) \left(\mathbb{X}_t^T \mathbb{W}_t \,\mathrm{diag}\left\{\mathbb{X}_t \boldsymbol{\theta}\right\}\right)^T. \end{aligned} \tag{5.19}$$

Since $\boldsymbol{Z}_t$ is independent of $\mathcal{F}_{t-1}$ and standard normally distributed, $\text{Var}\left(\boldsymbol{Z}_t^2|\mathcal{F}_{t-1}\right) = \text{Var}\left(\boldsymbol{Z}_t^2\right) = 2\mathbb{I}$.

$$\begin{aligned}\sum_{t=1}^{T}\text{Var}\left(\boldsymbol{S}_t|\mathcal{F}_{t-1}\right) &= \sum_{t=1}^{T}\mathbb{X}_t^T\mathbb{W}_t\,\text{diag}\left\{\mathbb{X}_t\boldsymbol{\theta}\right\}2\mathbb{I}\,\text{diag}\left\{\mathbb{X}_t\boldsymbol{\theta}\right\}^T\mathbb{W}_t^T\mathbb{X}_t \\ &= 2\sum_{t=1}^{T}\mathbb{X}_t^T\mathbb{W}_t\,\text{diag}\left\{(\mathbb{X}_t\boldsymbol{\theta})^2\right\}\mathbb{W}_t^T\mathbb{X}_t.\end{aligned} \tag{5.20}$$

We remember that $\mathbb{W}_t = \text{diag}\left\{\frac{1}{2}(\mathbb{X}_t\boldsymbol{\theta})^{-2}\right\}$ and by including the normalization, we get

$$\begin{aligned}N^{-1}\sum_{t=1}^{T}\text{Var}\left(\boldsymbol{S}_t|\mathcal{F}_{t-1}\right) &= 2N^{-1}\sum_{t=1}^{T}\mathbb{X}_t^T\,\text{diag}\left\{\frac{1}{2}(\mathbb{X}_t\boldsymbol{\theta})^{-2}\right\}\text{diag}\left\{(\mathbb{X}_t\boldsymbol{\theta})^2\right\}\mathbb{W}_t^T\mathbb{X}_t \\ &= N^{-1}\sum_{t=1}^{T}\mathbb{X}_t^T\mathbb{W}_t^T\mathbb{X}_t.\end{aligned} \tag{5.21}$$

We recognize (5.21) as the matrix we claimed asymptotically will converge to a constant matrix $\mathcal{J}$ as sample size increase and this constant matrix is by (5.18) $\mathcal{I}$. This means that $\mathcal{J} = \mathcal{I}$ and

$$\begin{aligned}\left(N^{-1}\sum_{t=1}^{T}\mathbb{X}_t^T\mathbb{W}_t^T\mathbb{X}_t\right)^{-1}\left(N^{-1}\sum_{t=1}^{T}\mathbb{X}_t^T\mathbb{W}_t^T\mathbb{X}_t\right)\left(N^{-1}\sum_{t=1}^{T}\mathbb{X}_t^T\mathbb{W}_t^T\mathbb{X}_t\right)^{-1} &= \left(N^{-1}\sum_{t=1}^{T}\mathbb{X}_t^T\mathbb{W}_t^T\mathbb{X}_t\right)^{-1} \\ &\to \mathcal{I}^{-1} = \text{Var}\left(\widehat{\boldsymbol{\theta}}_{\text{ML}}\right)\end{aligned}$$

We can therefore estimate the inverse information matrix $\mathcal{I}^{-1}$ by (5.14).

Empirical experiments are indicating that the approximation is quite good. The way these experiments have been carried out is by simulating 1000 datasets and estimate the parameters. In addition, for each dataset we estimate the covariance matrix using (5.14). When this is done for all the datasets, we calculate the mean formula covariance matrix. We also use an empirical covariance matrix estimator on the set of estimated parameters. This gives us a Monte Carlo covariance matrix estimate. If the formula estimate is good, it should be close to the Monte Carlo estimate. We did this for larger sample sizes as well to see if the variance decreased and this was the case. For more details on the experiment see Appendix B.1.

5.1.3. Asymptotic Normality

We will here sketch a derivation of asymptotic normality of the maximum likelihood estimators of a circular STGARCH process. An expression for the Hessian matrix (Equation (5.34)) is obtained through some technical calculations and we argue for the asymptotic normality by use of the Ergodic theorem and the central limit theorem. Since this «proof» is more of a sketch of a proof, we also have sought support from the asymptotic results developed by Comte and Lieberman (2003), included in section 2.4, for the multivariate GARCH (MGARCH). We show that STGARCH is a special case of MGARCH by showing that it can be written as a BEKK process (2.41) and then argue that Comte and Lieberman's results also holds for STGARCH.

Let $N = MT$ be the total sample size. By performing a Taylor approximation of the score function around the true parameter $\boldsymbol{\theta}$ and using that $\boldsymbol{U}(\widehat{\boldsymbol{\theta}}) = \mathbf{0}$, we have

$$\mathbf{0} = \boldsymbol{U}(\widehat{\boldsymbol{\theta}}) = \boldsymbol{U}(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}} \boldsymbol{U}(\boldsymbol{\theta})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + \mathcal{O}_P(N^{-1}). \tag{5.22}$$

A more rigorous proof is needed for showing that the remainder is $\mathcal{O}_P(N^{-1})$, but here we assume that it is. Asymptotically, the remainder term will tend to zero as N increase. To ease notation, we neglect this term in what follows. We reorder (5.22) and get

$$\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = -\sqrt{N}\big(\nabla_{\boldsymbol{\theta}} \boldsymbol{U}(\boldsymbol{\theta})\big)^{-1} \boldsymbol{U}(\boldsymbol{\theta}). \tag{5.23}$$

We have from Equation (5.9) that

$$\boldsymbol{U}(\boldsymbol{\theta}) = \sum_{t=1}^{T} \mathbb{X}_t^T \mathbb{W}_t \Big(\boldsymbol{X}_t^2 - \mathbb{X}_t \boldsymbol{\theta}\Big) = \sum_{t=1}^{T} \sum_{i=1}^{M} \frac{\mathbb{X}_{t,i}^T (X_{t,i}^2 - \mathbb{X}_{t,i}\boldsymbol{\theta})}{2(\mathbb{X}_{t,i}\boldsymbol{\theta})^2} = \sum_{t=1}^{T} \sum_{i=1}^{M} \boldsymbol{U}_{i,t}(\boldsymbol{\theta}). \tag{5.24}$$

As we will need it later, we show that $\mathrm{E}\,\boldsymbol{U}(\boldsymbol{\theta}) = \mathbf{0}$. By using the law of iterated expectations, we have

$$\mathrm{E}\left(\boldsymbol{U}(\boldsymbol{\theta})\right) = \mathrm{E}\Big[\sum_{t=1}^{T} \mathbb{X}_t^T \mathbb{W}_t \Big(\boldsymbol{X}_t^2 - \mathbb{X}_t \boldsymbol{\theta}\Big)\Big] = \mathrm{E}\,\mathrm{E}\Big[\sum_{t=1}^{T} \mathbb{X}_t^T \mathbb{W}_t \Big(\boldsymbol{X}_t^2 - \mathbb{X}_t \boldsymbol{\theta}\Big)|\mathcal{F}_{t-1}\Big]. \tag{5.25}$$

The only term in this sum that is not measurable given $\mathcal{F}_{t-1}$ is $\boldsymbol{X}_t^2$. We have that

$$\mathrm{E}\left(\boldsymbol{X}_{t-1}^2 | \mathcal{F}_{t-1}\right) = \mathrm{diag}\left\{\mathbb{X}_t \boldsymbol{\theta}\right\} \mathrm{E}\left(\boldsymbol{Z}_t^2\right) = \mathrm{diag}\left\{\mathbb{X}_t \boldsymbol{\theta}\right\} \mathbf{1} = \mathbb{X}_t \boldsymbol{\theta}. \tag{5.26}$$

Therefore,

$$\mathrm{E}\,\boldsymbol{U}(\boldsymbol{\theta}) = \mathrm{E}\Big[\sum_{t=1}^{T} \mathbb{X}_t^T \mathbb{W}_t \Big(\mathrm{E}\left(\boldsymbol{X}_t^2 | \mathcal{F}_{t-1}\right) - \mathbb{X}_t \boldsymbol{\theta}\Big)\Big] = \mathbf{0}. \tag{5.27}$$

We go on to find an explicit expression for the Hessian matrix, or the second derivative of the log likelihood function.

$$\begin{aligned} \nabla_{\boldsymbol{\theta}} \boldsymbol{U}(\boldsymbol{\theta}) &= \nabla_{\boldsymbol{\theta}} \Big(\sum_{t=1}^{T} \mathbb{X}_t^T \mathbb{W}_t \boldsymbol{X}_t^2 - \mathbb{X}_t^T \mathbb{W}_t \mathbb{X}_t \boldsymbol{\theta}\Big) \\ &= \sum_{t=1}^{T} \nabla_{\boldsymbol{\theta}} \big(\mathbb{X}_t^T \mathbb{W}_t \boldsymbol{X}_t^2\big) - \sum_{t=1}^{T} \mathbb{X}_t^T \nabla_{\boldsymbol{\theta}} \big(\mathbb{W}_t \mathbb{X}_t \boldsymbol{\theta}\big) \end{aligned} \tag{5.28}$$

We do the differentiation term by term. First notice that $\mathbb{W}_t \boldsymbol{X}_t^2 = \mathrm{diag}\left\{2^{-1}(\mathbb{X}_t\boldsymbol{\theta})^{-2}\right\} \boldsymbol{X}_t^2$ can be written as the vector $\left[2^{-1}\boldsymbol{X}_{t,1}^2(\mathbb{X}_{t,1}\boldsymbol{\theta})^{-2}, \quad \ldots, \quad 2^{-1}\boldsymbol{X}_{t,M}^2(\mathbb{X}_{t,M}\boldsymbol{\theta})^{-2}\right]$. Then

$$\nabla_{\boldsymbol{\theta}} \left(\mathbb{X}_t \mathbb{W}_t \boldsymbol{X}_t^2\right) = 2^{-1} \mathbb{X}_t^T \nabla_{\boldsymbol{\theta}} \begin{bmatrix} \boldsymbol{X}_{t,1}^2 (\mathbb{X}_{t,1}\boldsymbol{\theta})^{-2} \\ \vdots \\ \boldsymbol{X}_{t,M}^2 (\mathbb{X}_{t,M}\boldsymbol{\theta})^{-2} \end{bmatrix} = 2^{-1} \mathbb{X}_t^T \begin{bmatrix} -2\boldsymbol{X}_{t,1}^2 \mathbb{X}_{t,1} (\mathbb{X}_{t,1}\boldsymbol{\theta})^{-3} \\ \vdots \\ -2\boldsymbol{X}_{t,M}^2 \mathbb{X}_{t,M} (\mathbb{X}_{t,M}\boldsymbol{\theta})^{-3} \end{bmatrix} \tag{5.29}$$

which we can rewrite as $(-\mathbb{X}_t^T \operatorname{diag}\left\{\boldsymbol{X}_t^2\right\} \operatorname{diag}\left\{(\mathbb{X}_t\boldsymbol{\theta})^{-3}\right\} \mathbb{X}_t)$. The other term becomes

$$\mathbb{X}_t^T \nabla_{\boldsymbol{\theta}}\big(\mathbb{W}_t\mathbb{X}_t\boldsymbol{\theta}\big) = \mathbb{X}_t^T \nabla_{\boldsymbol{\theta}}\big(\operatorname{diag}\left\{2^{-1}(\mathbb{X}_t\boldsymbol{\theta})^{-2}\right\}\mathbb{X}_t\boldsymbol{\theta}\big) = 2^{-1}\mathbb{X}_t^T\nabla_{\boldsymbol{\theta}} \begin{bmatrix} (\mathbb{X}_{1,t}\boldsymbol{\theta})^{-1} \\ \vdots \\ (\mathbb{X}_{M,t}\boldsymbol{\theta})^{-1} \end{bmatrix}. \tag{5.30}$$

We have that $\frac{\partial}{\partial\boldsymbol{\theta}}(\mathbb{X}_{i,t}\boldsymbol{\theta})^{-1} = -\mathbb{X}_{i,t}(\mathbb{X}_{i,t}\boldsymbol{\theta})^{-2}$ and therefore

$$\mathbb{X}_t^T \nabla_{\boldsymbol{\theta}}\big(\mathbb{W}_t\mathbb{X}_t\boldsymbol{\theta}\big) = -2^{-1}\mathbb{X}_t^T \begin{bmatrix} \mathbb{X}_{1,t}/(\mathbb{X}_{1,t}\boldsymbol{\theta})^{2} \\ \vdots \\ \mathbb{X}_{M,t}/(\mathbb{X}_{M,t}\boldsymbol{\theta})^{2} \end{bmatrix} = -2^{-1}\mathbb{X}_t^T \operatorname{diag}\left\{(\mathbb{X}_t\boldsymbol{\theta})^{-2}\right\}\mathbb{X}_t, \tag{5.31}$$

which we recognize as $-\mathbb{X}_t^T\mathbb{W}_t\mathbb{X}_t$. Putting (5.29) and (5.31) into (5.28), we get that

$$\nabla_{\boldsymbol{\theta}}\boldsymbol{U}(\boldsymbol{\theta}) = -\sum_{t=1}^{T}\mathbb{X}_t^T \operatorname{diag}\left\{\boldsymbol{X}_t^2\right\}\operatorname{diag}\left\{(\mathbb{X}_t\boldsymbol{\theta})^{-3}\right\}\mathbb{X}_t + \sum_{t=1}^{T}\mathbb{X}_t^T\mathbb{W}_t\mathbb{X}_t \tag{5.32}$$

Now taking the expectation of $\nabla_{\boldsymbol{\theta}}\boldsymbol{U}(\boldsymbol{\theta})$, we get

$$\begin{aligned} \mathrm{E}\left(\nabla_{\boldsymbol{\theta}}\boldsymbol{U}(\boldsymbol{\theta})\right) &= \mathrm{E}\left(-\sum_{t=1}^{T}\mathbb{X}_t^T\mathbb{W}_t\mathbb{X}_t + \sum_{t=1}^{T}\mathbb{X}_t^T\,\mathrm{E}\left(\operatorname{diag}\left\{\boldsymbol{X}_t^2\right\}|\mathcal{F}_{t-1}\right)\operatorname{diag}\left\{(\mathbb{X}_t\boldsymbol{\theta})^{-3}\right\}\mathbb{X}_t\right) \\ &= \mathrm{E}\left(-\sum_{t=1}^{T}\mathbb{X}_t^T\mathbb{W}_t\mathbb{X}_t + \sum_{t=1}^{T}\mathbb{X}_t^T\operatorname{diag}\left\{\mathbb{X}_t\boldsymbol{\theta}\right\}\operatorname{diag}\left\{(\mathbb{X}_t\boldsymbol{\theta})^{-3}\right\}\mathbb{X}_t\right) \\ &= \mathrm{E}\left(-\sum_{t=1}^{T}\mathbb{X}_t^T\mathbb{W}_t\mathbb{X}_t + \sum_{t=1}^{T}\mathbb{X}_t^T\operatorname{diag}\left\{(\mathbb{X}_t\boldsymbol{\theta})^{-2}\right\}\mathbb{X}_t\right) \end{aligned} \tag{5.33}$$

We recognize $\operatorname{diag}\left\{(\mathbb{X}_t\boldsymbol{\theta})^{-2}\right\}$ as $2\mathbb{W}_t$ and therefore

$$\mathrm{E}\left(\nabla_{\boldsymbol{\theta}}\boldsymbol{U}(\boldsymbol{\theta})\right) = -\mathrm{E}\left(\sum_{t=1}^{T}\mathbb{X}_t^T\mathbb{W}_t\mathbb{X}_t\right) + 2\,\mathrm{E}\left(\sum_{t=1}^{T}\mathbb{X}_t^T\mathbb{W}_t\mathbb{X}_t\right) = \mathrm{E}\left(\sum_{t=1}^{T}\mathbb{X}_t^T\mathbb{W}_t\mathbb{X}_t\right). \tag{5.34}$$

Notice that this is the not normalized inverse of the expression we found for $\widehat{\operatorname{Var}}(\widehat{\boldsymbol{\theta}})$.

Let $\boldsymbol{U}_N(\boldsymbol{\theta}) = N^{-1}\boldsymbol{U}(\boldsymbol{\theta})$ be the normalized score function. Normalizing it will obviously make no difference for the maximum likelihood estimation. The law of large numbers ensures that

$$\lim_{N\to\infty} -\nabla_{\boldsymbol{\theta}}\boldsymbol{U}_N(\widehat{\boldsymbol{\theta}}) \to \mathrm{E}\left(-\nabla_{\boldsymbol{\theta}}\boldsymbol{U}_{i,t}(\boldsymbol{\theta})\right) = \mathcal{I}(\boldsymbol{\theta}). \tag{5.35}$$

We have that $\operatorname{Var}\left(\boldsymbol{U}_N(\boldsymbol{\theta})\right) = \mathrm{E}\left(\boldsymbol{U}_N(\boldsymbol{\theta})\boldsymbol{U}_N(\boldsymbol{\theta})^T\right) = \mathrm{E}\left(-\nabla_{\boldsymbol{\theta}}\boldsymbol{U}_N(\boldsymbol{\theta})\right) = \mathcal{I}(\boldsymbol{\theta})$. If we use the normalized Taylor representation from the beginning of this section

$$\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = -\big(\nabla_{\boldsymbol{\theta}}\boldsymbol{U}_N(\boldsymbol{\theta})\big)^{-1}\sqrt{N}\boldsymbol{U}_N(\boldsymbol{\theta}). \tag{5.36}$$

We have, by the Ergodic theorem (Billingsley, 1995, Th.24.1, pp.314), that

$$\nabla_{\boldsymbol{\theta}}\boldsymbol{U}_N(\boldsymbol{\theta}) = -N^{-1}\sum_{t,i}\nabla^2_{\boldsymbol{\theta}}\log f(X_{i,t}|\boldsymbol{\theta}) \to -\mathrm{E}_{\boldsymbol{\theta}}\left(\nabla^2_{\boldsymbol{\theta}}\log f(X|\boldsymbol{\theta})\right) = \mathcal{I}(\boldsymbol{\theta}) \tag{5.37}$$

and

$$\sqrt{N}\boldsymbol{U}_N(\boldsymbol{\theta})^T = \sqrt{N}\Big(N^{-1}\sum_{t,i}\boldsymbol{U}_{t,i} - \mathbf{0}\Big) = \sqrt{N}\Big(N^{-1}\sum_{t,i}\boldsymbol{U}_{t,i} - \mathrm{E}\left(\boldsymbol{U}_{1,1}\right)\Big) \tag{5.38}$$
$$\to \mathrm{MVN}\left(\mathbf{0}, \mathrm{Var}_{\boldsymbol{\theta}}\left(\boldsymbol{U}_{1,1}\right)\right),$$

by the central limit theorem (Billingsley, 1995, Th.27.1, pp.357). Hence,

$$\mathrm{Var}\left(\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})\right) \to \mathcal{I}(\boldsymbol{\theta})^{-1} N\,\mathrm{Var}_{\boldsymbol{\theta}}\left(\boldsymbol{U}_N\right)\mathcal{I}(\boldsymbol{\theta})^{-1} = \mathcal{I}(\boldsymbol{\theta})^{-1}, \tag{5.39}$$

because $N\,\mathrm{Var}_{\boldsymbol{\theta}}\left(\boldsymbol{U}_N\right) = \mathrm{Var}_{\boldsymbol{\theta}}\left(\boldsymbol{U}_{1,1}\right) = \mathcal{I}(\boldsymbol{\theta})$. To summarize, we have argued that the normalized score function is multivariate normal distributed with expectation zero and covariance matrix $\mathcal{I}(\boldsymbol{\theta})$. Because of the Taylor representation (5.36), we may conclude that

$$\sqrt{N}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \to \mathrm{MVN}\left(\mathbf{0}, \mathcal{I}^{-1}\right). \tag{5.40}$$

We did a simulation experiment to check normality of the estimators. Here we present the marginal normalized densities of α_0 and α_1 in the STARCH(1) situation described in the introduction to Chapter 6. The result is presented in Figure 5.1 and we see that the densities and scatter plot support (5.40).

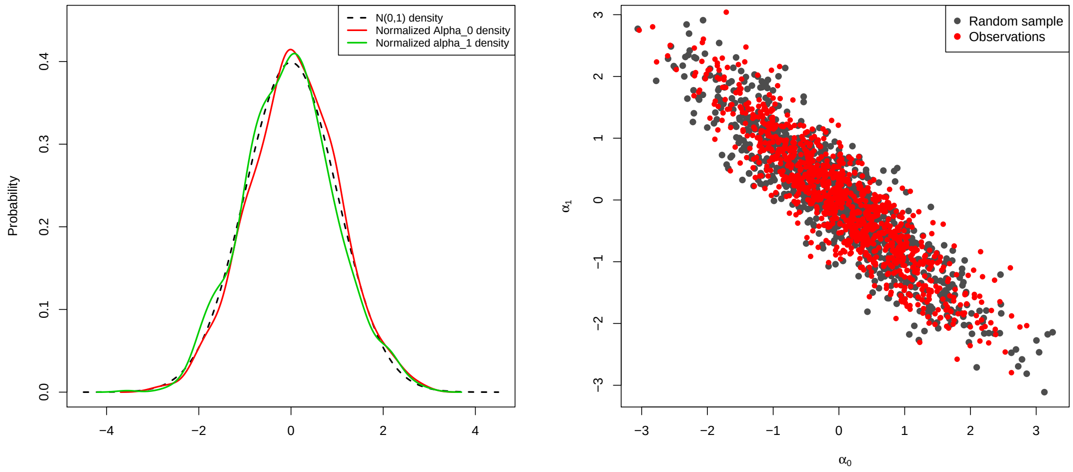


Figure 5.1: *Resulting marginal densities from estimating parameters from 1000 simulations to the left and two dimensional scatter plot to the right comparing the normalized observations to a random bivariate normal sample with the same correlations coefficient.*

MGARCH approach

In section 2.4, we introduced the multivariate GARCH (MGARCH) process. In particular, we considered the BEKK processes given by Equation (2.41). Comte and Lieberman (2003) proved strong consistency and asymptotic normality of these processes under some regularity conditions. We will in this section show that the STGARCH process can be formulated as a special case of the BEKK process, and in that way, we can use Comte and Lieberman's results also for the STGARCH. For the reader's convenience, we restate the BEKK representation given in Equation (2.41).

$$\mathbb{H}_t = \mathbb{C} + \sum_{i=1}^{p} \Big(\sum_{j=1}^{k} \mathbb{A}_{ij} \boldsymbol{X}_{t-i} \boldsymbol{X}_{t-i}^T \mathbb{A}_{ij}^T \Big) + \sum_{i=1}^{q} \Big(\sum_{j=1}^{k} \mathbb{B}_{ij} \mathbb{H}_{t-i} \mathbb{B}_{ij}^T \Big), \tag{5.41}$$

where the matrix $\mathbb{C}$ is positive definite and the matrices $\mathbb{A}_{ij}$, for $i = 1, \dots, p$, $j = 1, \dots, k$ and $\mathbb{B}_{ij}$ for $i = 1, \dots, q$ and $j = 1, \dots, k$ are real $d \times d$ matrices and k is an integer satisfying $k \leq d(d+1)/2$. All the matrices are functions of the parameter vector, $\boldsymbol{\theta}$.

We want to illustrate that STGARCH can be represented as (5.41). First of all, we let the matrix $\mathbb{C}$ be defined as

$$\mathbb{C} = \begin{bmatrix} \alpha_0 & 0 & \dots & 0 \\ 0 & \alpha_0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \alpha_0 \end{bmatrix} = \alpha_0 \mathbb{I}_M, \tag{5.42}$$

which obviously is positive definite by the requirement that $\alpha_0 > 0$.

The matrices $\mathbb{A}_{ij}$ and $\mathbb{B}_{ij}$ need to be defined in such a way that they create a diagonal matrix. Let $\boldsymbol{e}_j$ be the j^{th} unit vector of $\mathbb{R}^M$. Then $\mathbb{E}_{ij} = \boldsymbol{e}_i \boldsymbol{e}_j^T$ will be a matrix of zeroes, except for the (i, j) element, which will be 1. We then have that $\sum_{j=1}^{M} \mathbb{E}_{jj} \boldsymbol{X}_t \boldsymbol{X}_t^T \mathbb{E}_{jj}^T$ forms a diagonal matrix consisting only of the diagonal elements of $\boldsymbol{X}_t \boldsymbol{X}_t^T$. For STGARCH to be presented by (5.41) we need $\sum_{j=1}^{k} \mathbb{A}_{ij} \boldsymbol{X}_t \boldsymbol{X}_t^T \mathbb{A}_{ij}^T$ to be a diagonal matrix where each non-zero element is the sum of neighbouring diagonal elements from $\boldsymbol{X}_t \boldsymbol{X}_t^T$. This can be achieved by letting j be the sum over all elements of the user defined neighbourhood matrix, $\boldsymbol{\mathcal{W}}^\star$, that are non-zero. For one spatial dimension with closest neighbour structure, $\mathbb{A}_{ij}$ can be defined as

$$\mathbb{A}_{ij} = \begin{cases} \alpha_i^{1/2} \mathbb{E}_{jj}, & \text{if } j = 1, \dots, M, \\ \alpha_i^{1/2} \mathbb{E}_{(j-M),(j-M+1)} & \text{if } j = M+1 \dots 2M-1, \\ \alpha_i^{1/2} \mathbb{E}_{M,1} & \text{if } j = 2M, \\ \alpha_i^{1/2} \mathbb{E}_{(j+1-2M),(j-2M)} & \text{if } j = 2M+1 \dots 3M-1, \\ \alpha_i^{1/2} \mathbb{E}_{1,M} & \text{if } j = 3M, \end{cases} \tag{5.43}$$

where $k = 3M \leq M(M+1)/2$ if $M \geq 5$. The third and fifth line of (5.43) are the circular connections. In the two dimensional models we consider in Chapter 6 we can define $\mathbb{A}_{ij}$ in a similar manner with $k = 9M \leq M(M+1)/2$ if $M \geq 17$.

For the GARCH part of (5.41), $\{\mathbb{H}_{t-i} : i = 1, \ldots, q\}$ will already be diagonal matrices for the STGARCH model and hence $\mathbb{B}_{i1} = \beta_i^{1/2}\mathbb{I}$. We can then define $\mathbb{B}_{ij}$ similarly to (5.43) in one spatial dimension with closest neighbourhood structure as

$$\mathbb{B}_{ij} = \begin{cases} \beta_i^{1/2}\mathbb{I} & \text{if } j = 1, \\ \beta_i^{1/2}\mathbb{E}_{j,j-1} & \text{if } j = 2, \ldots, M, \\ \beta_i^{1/2}\mathbb{E}_{j-M,j+1-M} & \text{if } j = M+1, \ldots, 2M, \\ \beta_i^{1/2}\mathbb{E}_{1,M} & \text{if } j = 2M+1, \\ \beta_i^{1/2}\mathbb{E}_{M,1} & \text{if } j = 2M+2, \\ \mathbf{0} & \text{otherwise.} \end{cases} \tag{5.44}$$

We have defined $\mathbb{B}_{ij}$ to be zero if $j > 2M + 2$, because $2M + 2 \leq 3M$ if $M \geq 2$ and in (5.41) k is the same number for both the ARCH and GARCH term, so by letting $k = 3M$ for $M \geq 5$ we fulfil (5.41).

We have shown that a one dimensional closest neighbour structured STGARCH model can be represented by the BEKK representation and thus the results from Comte and Lieberman (2003) presented in section 2.4 holds. We also see how this can be done for higher dimensional models, but showing this explicitly is beyond the point. Perhaps a more general formulation of (5.43) and (5.44) can be found. Therefore, under Comte and Lieberman's conditions, the STGARCH quasi maximum likelihood estimators are consistent and asymptotic normally distributed. The conditions for consistency is quite weak, making this a strong result. However, for the asymptotic normality, Comte and Lieberman require $\mathrm{E}\left(\boldsymbol{X}_t^8\right) < \infty$, which is quite restrictive. They also require the parameter space to be compact, which remains unjustified in this situation. It is our understanding that these assumptions are too restrictive for the STGARCH, and that asymptotic normality holds under less demanding conditions, but this remains to be examined.

5.2. Least Squares Estimation

While maximum likelihood is the most common way of estimating GARCH models, simple least squares can be a good alternative. Least squares estimation is especially useful for the low ordered models (small p or q). This method is a direct consequence of the ARMA representation of STGARCH. By squaring the process, we achieve a linear relation between the process and the parameters. In fact, this becomes a linear regression problem. But it is not a standard linear regression, because we do not have independent and identically distributed observations, which complicates things.

5.2.1. The Method

As was suggested above, by squaring the STGARCH, we obtain an expression that is linear in the parameters. Although the data itself is normal, when we square it, the distribution becomes somewhat complicated. A simple regression technique that is commonly used is least squares (LS) estimation. LS estimation does not take the distribution into account, and only seeks to minimize the sum of the squared residuals. One can read more about least squares regression in Dobson and Barnett (2008). In a STGARCH, the residuals are $Z_t(\boldsymbol{u}) = \sigma_t(\boldsymbol{u})^{-1} X_t(\boldsymbol{u})$, but exploiting the ARMA representation of the STGARCH (3.8), we can define new residuals and minimize the sum of squares of these.

We continue to use the notation introduced in section 4.1, and define $\boldsymbol{Y}_t$ to be the Hadamard squared process of $\boldsymbol{X}_t$. We may therefore write

$$\boldsymbol{Y}_t = \boldsymbol{X}_t^2 = \operatorname{diag}\left\{\boldsymbol{\sigma}_t^2\right\} \boldsymbol{Z}_t^2. \tag{5.45}$$

If we then add zero by adding $\boldsymbol{\sigma}_t^2$ and subtracting $\operatorname{diag}\left\{\boldsymbol{\sigma}_t^2\right\} \mathbf{1}$, we get

$$\boldsymbol{Y}_t = \operatorname{diag}\left\{\boldsymbol{\sigma}_t^2\right\} \boldsymbol{Z}_t^2 + \boldsymbol{\sigma}_t^2 - \operatorname{diag}\left\{\boldsymbol{\sigma}_t^2\right\} \mathbf{1} = \boldsymbol{\sigma}_t^2 + \operatorname{diag}\left\{\boldsymbol{\sigma}_t^2\right\} (\boldsymbol{Z}_t^2 - \mathbf{1}) = \boldsymbol{\sigma}_t^2 + \boldsymbol{V}_t \tag{5.46}$$

where $\boldsymbol{V}_t = \operatorname{diag}\left\{\boldsymbol{\sigma}_t^2\right\} (\boldsymbol{Z}_t^2 - \mathbf{1})$ is considered a noise term or a residual vector. It was shown in Chapter 3 that this $\boldsymbol{V}_t$ is a white noise process.

We have from the model design that $\boldsymbol{\sigma}_t^2$ is linear in the parameters. By constructing a design matrix in the same manner as for the maximum likelihood estimator (see Equation (5.2)) we can write $\boldsymbol{\sigma}_t^2 = \mathbb{X}_t \boldsymbol{\theta}$. This means that (5.46) can be written as

$$\boldsymbol{Y}_t = \mathbb{X}_t \boldsymbol{\theta} + \boldsymbol{V}_t. \tag{5.47}$$

We recognize (5.47) as a multiple linear regression problem, with $\mathbb{X}_t$ consisting of observations from both observed and unobserved processes $\{\boldsymbol{Y}_t\}$ and $\{\boldsymbol{\sigma}_t^2\}$, respectively. Since $\{\boldsymbol{\sigma}_t^2\}$ is unobserved, it will need to be estimated, but we will return to this issue shortly.

The following derivation of least squares estimators follow the standard linear regression procedure. We want to find the parameter vector that minimizes the sum of squared residuals, $Q(\boldsymbol{\theta})$, often referred to as the objective function.

$$\begin{aligned} Q(\boldsymbol{\theta}) &= \sum_{t=1}^{T} \boldsymbol{V}_t^T \boldsymbol{V}_t = \sum_{t=1}^{T} (\boldsymbol{Y}_t - \boldsymbol{\sigma}_t^2)^T (\boldsymbol{Y}_t - \boldsymbol{\sigma}_t^2) \\ &= \sum_{t=1}^{T} (\boldsymbol{Y}_t - \mathbb{X}_t \boldsymbol{\theta})^T (\boldsymbol{Y}_t - \mathbb{X}_t \boldsymbol{\theta}) \\ &= \sum_{t=1}^{T} \left[\boldsymbol{Y}_t^T \boldsymbol{Y}_t - 2\boldsymbol{\theta}^T \mathbb{X}_t^T \boldsymbol{Y}_t + \boldsymbol{\theta}^T \mathbb{X}_t^T \mathbb{X}_t \boldsymbol{\theta} \right]. \end{aligned} \tag{5.48}$$

To find the minimum of Q, we set the derivative $\partial Q/\partial\boldsymbol{\theta}$ equal to zero and solve for $\widehat{\boldsymbol{\theta}}_{\mathrm{LS}}$.

$$\begin{aligned}\frac{\partial Q(\widehat{\boldsymbol{\theta}}_{\mathrm{LS}})}{\partial\boldsymbol{\theta}} &= -2\sum_{t=1}^{T}\mathbb{X}_t^T\boldsymbol{Y}_t + 2\sum_{t=1}^{T}\mathbb{X}_t^T\mathbb{X}_t\widehat{\boldsymbol{\theta}}_{\mathrm{LS}} = \mathbf{0}.\\ \Big(\sum_{t=1}^{T}\mathbb{X}_t^T\mathbb{X}_t\Big)\widehat{\boldsymbol{\theta}}_{\mathrm{LS}} &= \sum_{t=1}^{T}\mathbb{X}_t^T\boldsymbol{Y}_t.\end{aligned} \tag{5.49}$$

If $\sum_{t=1}^{T}\mathbb{X}_t^T\mathbb{X}_t$ is invertible, we can find the analytical solution

$$\widehat{\boldsymbol{\theta}}_{\mathrm{LS}} = \Big(\sum_{t=1}^{T}\mathbb{X}_t^T\mathbb{X}_t\Big)^{-1}\sum_{t=1}^{T}\mathbb{X}_t^T\boldsymbol{Y}_t. \tag{5.50}$$

As we can see, the least squares estimator (LSE) is of the same form as the maximum likelihood estimator (MLE). This is a common feature of the normal distribution. The difference is that the MLE is a weighted LSE. In fact, if we set $\mathbb{W}_t = \mathbb{I}$ for all t, the estimators are the same.

There are two levels of complexity in this situation. If we are considering a pure ARCH process, this approach will be effective estimating the parameters using (5.50). However, with a GARCH process, the unobserved process $\{\boldsymbol{\sigma}_t^2\}$ depend on both observed and unobserved processes. We therefore suggest iterating the estimation routine.

$$\widehat{\boldsymbol{\theta}}_{\mathrm{LS}}^{(m+1)} = \Big[\sum_{t=1}^{T}\mathbb{X}_t^T(\widehat{\boldsymbol{\theta}}_{\mathrm{LS}}^{(m)})\mathbb{X}_t(\widehat{\boldsymbol{\theta}}_{\mathrm{LS}}^{(m)})\Big]^{-1}\sum_{t=1}^{T}\mathbb{X}_t^T(\widehat{\boldsymbol{\theta}}_{\mathrm{LS}}^{(m)})\boldsymbol{Y}_t \tag{5.51}$$

where the notation $\mathbb{X}_t(\widehat{\boldsymbol{\theta}}_{\mathrm{LS}}^{(m)})$ means that the estimated $\{\widetilde{\boldsymbol{\sigma}}_s^2 : s = t-1,\ldots,t-q\}$ of $\mathbb{X}_t$ was calculated using the previous estimate $\widehat{\boldsymbol{\theta}}_{\mathrm{LS}}^{(m)}$. This requires an initial value $\widehat{\boldsymbol{\theta}}_{\mathrm{LS}}^{(0)}$ to start off the routine. In the next chapter, a one lag STARCH process is much used. We therefore stress that for this situation (5.51) is not necessary and (5.50) can be used directly without the need for any initial values.

If we assume that either our process is STARCH and we have calculated the estimate using (5.50) or we have a STGARCH process and the estimate of (5.51) has converged, we can estimate the covariance matrix of the estimators using

$$\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\theta}}_{\mathrm{LS}}) = 2\Big(\sum_{t=1}^{T}\widetilde{\mathbb{X}}_t^T\widetilde{\mathbb{X}}_t\Big)^{-1}\Big(\sum_{t=1}^{T}\widetilde{\mathbb{X}}_t^T\,\mathrm{diag}\Big\{(\widetilde{\mathbb{X}}_t\widehat{\boldsymbol{\theta}}_{\mathrm{LS}})^2\Big\}^T\widetilde{\mathbb{X}}_t\Big)\Big(\sum_{t=1}^{T}\widetilde{\mathbb{X}}_t^T\widetilde{\mathbb{X}}_t\Big)^{-1}, \tag{5.52}$$

where the estimated $\widetilde{\mathbb{X}}_t$ are based on the estimated $\widetilde{\boldsymbol{\sigma}}_t^2$, just like for the maximum likelihood. If we are considering a pure STARCH process, $\mathbb{X}_t$ will only be based on the observations $\{\boldsymbol{x}_t\}$ and do not need to be estimated in the same manner. The covariance matrix expression of (5.52) is closely related to the maximum likelihood formula (5.14) , except for the least square estimator we do not have the nice cancelling property.

Note that (5.50) (and (5.51)) can be written as

$$\widehat{\boldsymbol{\theta}}_{\mathrm{LS}} = \left(N^{-1}\sum_{t=1}^{T}\mathbb{X}_t^T\mathbb{X}_t\right)^{-1} N^{-1}\sum_{t=1}^{T}\mathbb{X}_t^T\boldsymbol{Y}_t. \tag{5.53}$$

Just like for the equivalent expression from the maximum likelihood procedure, we assume that asymptotically

$$\lim_{N\to\infty} N^{-1}\sum_{t=1}^{T}\mathbb{X}_t^T\mathbb{X}_t = \mathcal{J}^{-1}, \tag{5.54}$$

for some constant invertible matrix $\mathcal{J}$. This enables us to concentrate on the numerator of (5.53). Using the ergodic theorem (Billingsley, 1995, Th.24.1, pp.314) in the same manner as for the maximum likelihood expression, we get that

$$\mathcal{I}^{-1} \leftarrow N^{-1}\sum_{t=1}^{T}\operatorname{Var}\left(\mathbb{X}_t^T\boldsymbol{Y}_t|\mathcal{F}_{t-1}\right) = N^{-1}\sum_{t=1}^{T}\mathbb{X}_t^T\operatorname{Var}\left(\operatorname{diag}\left\{\mathbb{X}_t\boldsymbol{\theta}\right\}\boldsymbol{Z}_t^2|\mathcal{F}_{t-1}\right)\mathbb{X}_t \tag{5.55}$$

We have that $\operatorname{Var}\left(\operatorname{diag}\left\{\mathbb{X}_t\boldsymbol{\theta}\right\}\boldsymbol{Z}_t^2|\mathcal{F}_{t-1}\right) = \operatorname{diag}\left\{\mathbb{X}_t\boldsymbol{\theta}\right\}\operatorname{Var}\left(\boldsymbol{Z}_t^2\right)\operatorname{diag}\left\{\mathbb{X}_t\boldsymbol{\theta}\right\}^T$ because $\mathbb{X}_t$ is measurable when we condition on $\mathcal{F}_{t-1}$. One of the perks of the least squares estimation routine is that it is distribution independent. It assumes no distribution on the innovations. In our context, however, the innovations are the process $\{\boldsymbol{V}_t\}$, and we may still assume $\boldsymbol{Z}_t \sim \mathrm{MVN}\left(\boldsymbol{0}, \mathbb{I}\right)$. The variance of a Hadamard squared independent multivariate normal distributed vector is easily obtained. Since all elements of the vector $\boldsymbol{Z}_t$ are iid we can consider them marginally. If $Z \sim N(0,1)$, we have that $\mathrm{E}Z^2 = 1$ and $\mathrm{E}Z^4 = 3$, so that $\mathrm{Var}Z^2 = \mathrm{E}Z^4 - (\mathrm{E}Z^2)^2 = 3 - 1^2 = 2$. Since this yields for all elements of $\boldsymbol{Z}_t$ and due to the independence, $\mathrm{Var}\boldsymbol{Z}_t^2 = 2\mathbb{I}$. By these arguments and assumptions, we get that

$$\mathcal{I}^{-1} \leftarrow N^{-1}\sum_{t=1}^{T}\operatorname{Var}\left(\mathbb{X}_t^T\boldsymbol{Y}_t|\mathcal{F}_{t-1}\right) = 2\sum_{t=1}^{T}\mathbb{X}_t^T\operatorname{diag}\left\{(\mathbb{X}_t\boldsymbol{\theta})^2\right\}^T\mathbb{X}_t. \tag{5.56}$$

Asymptotically, we have that

$$\operatorname{Var}\left(\widehat{\boldsymbol{\theta}}_{\mathrm{LS}}\right) = \mathcal{J}\mathcal{I}^{-1}\mathcal{J}, \tag{5.57}$$

but in approximation of $\operatorname{Var}\left(\widehat{\boldsymbol{\theta}}_{\mathrm{LS}}\right)$, we estimate $\mathcal{I}^{-1}$ and $\mathcal{J}$ separately by

$$\begin{aligned}\widehat{\mathcal{I}}^{-1} &= 2\sum_{t=1}^{T}\widetilde{\mathbb{X}}_t^T\operatorname{diag}\left\{(\widetilde{\mathbb{X}}_t\widehat{\boldsymbol{\theta}}_{\mathrm{LS}})^2\right\}^T\widetilde{\mathbb{X}}_t,\\ \widehat{\mathcal{J}}^{-1} &= N^{-1}\sum_{t=1}^{T}\widetilde{\mathbb{X}}_t^T\widetilde{\mathbb{X}}_t.\end{aligned} \tag{5.58}$$

By inserting the estimates of (5.58) for the corresponding theoretical measures of (5.57), we obtain (5.52).

We summarize this method with an algorithm. The algorithm is presented under the usual assumptions, but can be generalized to other settings as well. For instance, if one do not want to assume symmetry between spatial neighbours, this is of course possible. This will only make the design matrix somewhat different.

Algorithm 2 Least squares estimation of STGARCH

Let $\{x_t(\boldsymbol{u})\}$ represent the data. We assume a symmetric model in space (only one α_i and one β_j for each time lag). We also assume the same neighbourhood matrix for each time lag and both the observed and unobserved processes.

$$X_t(\boldsymbol{u})^2 = \sigma_t(\boldsymbol{u})^2 + V_t(\boldsymbol{u}) = \alpha_0 + \sum_{s=1}^{p} \alpha_s \sum_{\boldsymbol{v} \in \boldsymbol{V}} X_{t-s}^2(\boldsymbol{u}-\boldsymbol{v}) + \sum_{s=1}^{q} \beta_s \sum_{\boldsymbol{v} \in \boldsymbol{V}} \widetilde{\sigma}_{t-s}^2(\boldsymbol{u}-\boldsymbol{v}) + V_t(\boldsymbol{u})$$

or on vector form

$$\boldsymbol{X}_t^2 = \mathbb{X}_t \boldsymbol{\theta} + \boldsymbol{V}_t$$

```
Square the data, $y_t(\boldsymbol{u}) = x_t^2(\boldsymbol{u})$
Initiate the neighbourhood matrix $\boldsymbol{\mathcal{W}}^\star$.
Set $\widetilde{\boldsymbol{\sigma}}_p = \ldots = \widetilde{\boldsymbol{\sigma}}_{-q+1} = \mathbf{0}$ and initial value for $\widehat{\boldsymbol{\theta}}_{\mathrm{LS}}^{(0)}$
for each node $m = 1, \ldots, M$ do
    for each node $t = p+1, \ldots, T$ do
        Set up the design matrix: $\widetilde{\mathbb{X}}_t = \begin{bmatrix} \mathbf{1} & \boldsymbol{\mathcal{W}}^\star \boldsymbol{y}_{t-1} & \ldots & \boldsymbol{\mathcal{W}}^\star \boldsymbol{y}_{t-p} & \boldsymbol{\mathcal{W}}^\star \widetilde{\boldsymbol{\sigma}}_{t-1}^2 & \ldots & \boldsymbol{\mathcal{W}}^\star \widetilde{\boldsymbol{\sigma}}_{t-q}^2 \end{bmatrix}$
        Calculate the next $\widetilde{\boldsymbol{\sigma}}_t^2 = \widetilde{\mathbb{X}}_t \widehat{\boldsymbol{\theta}}_{\mathrm{LS}}^{(m-1)}$
    end for
    Calculate the estimate
        $$\widehat{\boldsymbol{\theta}}_{\mathrm{LS}}^{(m)} = \left( \sum_{t=p+1}^{N} \widetilde{\mathbb{X}}_t^T \widetilde{\mathbb{X}}_t \right)^{-1} \sum_{t=p+1}^{N} \widetilde{\mathbb{X}}_t^T \boldsymbol{y}_t$$
end for
Return the estimated parameter vector: $\widehat{\boldsymbol{\theta}}_{\mathrm{LS}}^{(M)} = \left[\widehat{\alpha}_0, \widehat{\alpha}_1, \ldots, \widehat{\alpha}_p, \widehat{\beta}_1, \ldots \widehat{\beta}_q\right]^T$
```

5.3. MLE vs LSE

We have now produced closed form estimators for $\boldsymbol{\theta}$, both using (quasi) maximum likelihood and least squares estimation methods. We have also seen that they are of the same form. In fact, both are closely related to multiple linear regression.

For a STARCH process, the LSE routine does not require initial values to estimate the parameters. This makes using the LSE to calculate the initial values for the MLE a brilliant choice. The MLE refines the LS estimate.

What separates the MLE from the LSE is the weighting matrix $\mathbb{W}_t$. Our hypothesis is that the MLE will be less influenced by extremes or outliers, because an extreme observation will have a high conditional variance according to the model. Since the weighting matrix is proportional to the conditional variance of power minus two, it will give less weight to an observation with large conditional variance, making it less influential.

In the next chapter we compare the two approaches on simulated datasets using the most common examples; STARCH(1) and STGARCH(1,1).

Chapter 6

Empirical Experiments

In this chapter we will investigate some of the properties of the estimation routines we developed in the previous chapter. The goal is to test theory in practice. We use simulation for generating data using the methods we developed in Chapter 4. This enables us to test the circular estimation on truly circular data, which can be hard to come by in real situations. We also test the performance of the circular estimation on non-circular data.

There are two competing approaches to estimation in this chapter – the circular model and by conditioning on the boundary. We explained in Chapter 3 what the idea behind conditioning on the boundary is, but have not presented estimation theory for conditioning on the boundary in this thesis. One reason is that we have focused on the circular model. When we say that we condition on the boundary observations, we mean that these observations only contribute to estimating $\boldsymbol{\sigma}_t$. Since we have all observations necessary to estimate $\boldsymbol{\sigma}_t$, the likelihood can be based on

$$\boldsymbol{Z}_t = \operatorname{diag}\left\{\boldsymbol{\sigma}_t\right\}^{-1} \boldsymbol{X}_t \sim \operatorname{MVN}\left(\mathbf{0}, \mathbb{I}\right), \tag{6.1}$$

in a similar way as for the circular likelihood derivation. There will be some consequences due to the conditioning on the entire boundary, that we have not addressed and we are unsure how this will influence the likelihood. Therefore, we do not claim the likelihood we use to be the true likelihood, but a composite likelihood (Varin et al., 2011). Conditioning on the boundary can only be used for STARCH models, due to the difficulty of estimating $\widetilde{\boldsymbol{\sigma}}_t$ on the spatial boundaries. One can approximate a STGARCH with a high order STARCH(p), but this is not considered here. Therefore we only use this procedure in competition with the circular model in STARCH experiments. We reefer to the conditioning on the boundary approach as regular estimation, as opposed to circular estimation. The term regular simulation is used for non-circular simulation.

For the most part of this chapter we use the same model. The model is a STARCH(1) with independent standard normally distributed innovations, only two parameters and closest neighbourhood structure. Each neighbour influence the process equally and we assume only

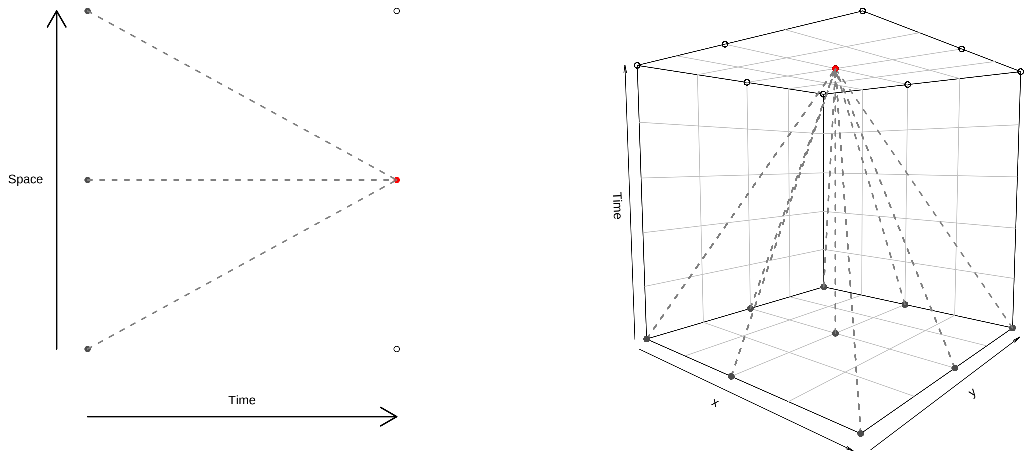


Figure 6.1: *Neighbourhood structure in 1D and 2D models used in this chapter. In 1D we let $\sigma_t(u)$ depend on the process in the same point u at one time lag, and the two neighbouring points. In 2D, we use the 8 closest neighbours in addition to the process in the same point at one time lag.*

the closest neighbours at one time lag have influence in addition to the point itself. In one spatial dimension there will be two neighbours and in two spatial dimensions there will be eight. Figure 6.1 illustrates the neighbourhood relationships in both one and two dimensions. Let $\boldsymbol{\mathcal{W}}^\star$ be the neighbourhood matrix structuring the neighbourhood relationships, defined in section 4.1. For every $t = 1, \ldots, T$ the model can be formulated as

$$\begin{aligned} \boldsymbol{X}_t &= \operatorname{diag}\left\{\boldsymbol{\sigma}_t\right\} \boldsymbol{Z}_t, & \mathbb{X}_t &= \begin{bmatrix} \mathbf{1}, & \boldsymbol{\mathcal{W}}^\star \boldsymbol{X}_{t-1}^2 \end{bmatrix}, \\ \boldsymbol{\sigma}_t^2 &= \mathbb{X}_t \boldsymbol{\theta}, & \boldsymbol{\theta} &= \begin{bmatrix} \alpha_0, & \alpha_1 \end{bmatrix}^T. \end{aligned}$$

The model is simple, but if the methods developed is to work for the complicated models they first need to pass the simplest ones.

We want to see how many iterations are required for the maximum likelihood routine to converge. The next experiment is to test the performance of the different estimation methods on both circular and regular data. This is done in both one and two dimensions with different parameter settings and sample sizes. A main question is consitency. We also consider a modification of the innovation distribution from standard normal to student's t-distribution.

In Appendix A we keep the time dimension fixed while gradually increasing the spatial sample size. In this way we try to examine spatial consistency. In Appendix B.1-B.2 simulation experiments are performed for evaluating the variance of the circular ML estimator given by formula (5.14) and likewise the asymptotic limit given by (5.16). Some R-code are included in Appendix C.

6.1. Convergence efficiency of MLE

We do this experiment first, because it is nice to have some idea of how many iterations is necessary for convergence of the estimates. If an estimate has converged after k iterations all the iterations after this point are totally unnecessary and a waste of computational power and time. This experiment is easily explained. We let the program pick parameters and dimension sizes at random, in the stationary domain and within some intervals, respectively. The spatial regions are squares. In each simulation experiment both the circular and conditioning on the boundary method are implemented. In both routines for each experiment we save the complete iteration sequence of estimated values. Then we check after how many iterations does the absolute difference between subsequent estimates first become less than some threshold. We are quite conservative and pick a difference of 10^{-10} as the convergence threshold.

After running this routine 10,000 times, we found the results presented in Figure 6.2. The figure shows the cumulative rate of convergence of each estimator, both by circular and regular approach. On the x-axis we have number of iterations, while on the y-axis we have the relative frequency of the 10,000 estimates that have converged after x iterations. To ensure that all estimates converged we set the maximum to 50 iterations, but the experiment show that 22 iterations suffice for convergence. For the circular estimation 95% of repetitions converge after 8 iterations. The corresponding number was 9 for regular estimation. If the true parameters are close to singularity or data dimensions are low, the convergence tend to be slower. We therefore suggest using a logical test that stops the iteration if the difference between subsequent estimates are below some threshold.

6.2. Circular vs Non-circular

In the next experiment we test the performance of the two estimation approaches, circular and regular, on both circular and regular models for one and two spatial dimensions. Our hypothesis is that in one dimension the difference will be small between the two approaches. At each point in time, there are two points in space that erroneously become neighbours (circular) or two points that are excluded from the estimation (regular). The circular assumption will affect a bit more than the two out-most points, but as spatial sample sizes increase, the relative amount of data influenced by either choice of approach will decrease and the estimates unite. This may also hold in two dimensions, but now the number of boundary points is increasing with an increasing spatial region. A possible convergence is therefore suspected to have a lower rate.

Figure 6.3 illustrates how this experiment is conducted. First we simulate a dataset, ei-

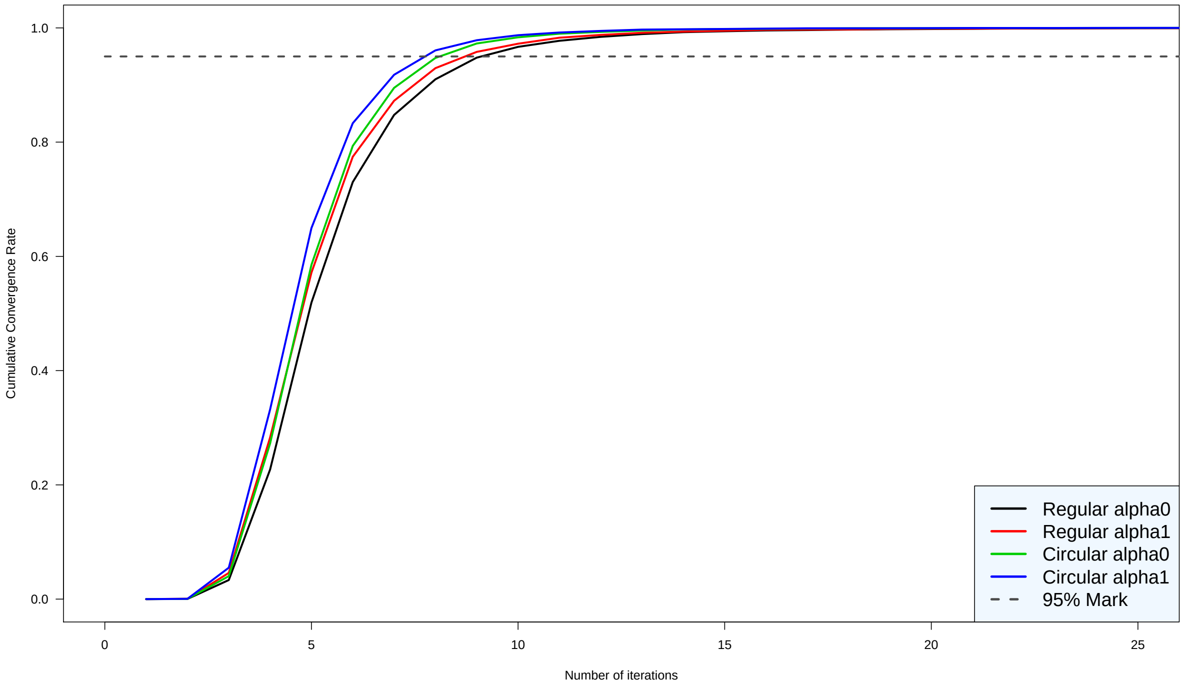


Figure 6.2: *Convergence of ML routine: The grey dashed line indicates the area where at least 95% of the estimators have converged. This is after 9 for the regular estimation (black and red) and 8 for the circular (green and blue).*

ther circular or regular. Then we estimate the parameters for this dataset, both circularly and regularly. Each of these approaches are done both by least squares– and maximum likelihood estimation. In the maximum likelihood estimation the least squares estimates are used as initial values.

We first look at the situation with one spatial dimension. The set-up is that the process has 10 spatial points which we observe 100 times, making the total number of observations $10 \times 100 = 1{,}000$. We use a burn-in of 100 time steps and for the regular simulation we cut off 200 points in each direction ($L = 200$). The true parameters here are $\alpha_0 = 0.3$ and $\alpha_1 = 0.05$.

The results from a simulation experiment with 1,000 repetitions are presented in Table 6.1. The abbreviations in the table need explaining: C is circular, R is regular, ML is maximum likelihood and LS is least squares. The position is also important. The first letter is the way the data was generated, either by circular or regular simulation. The second letter tells us which estimation approach was used. Also here the alternatives are circular or regular. Finally, the two last letters tells us which method was used, either maximum likelihood or least squares. The columns are the mean estimates, mean squared error (MSE) and variances for the different estimators.

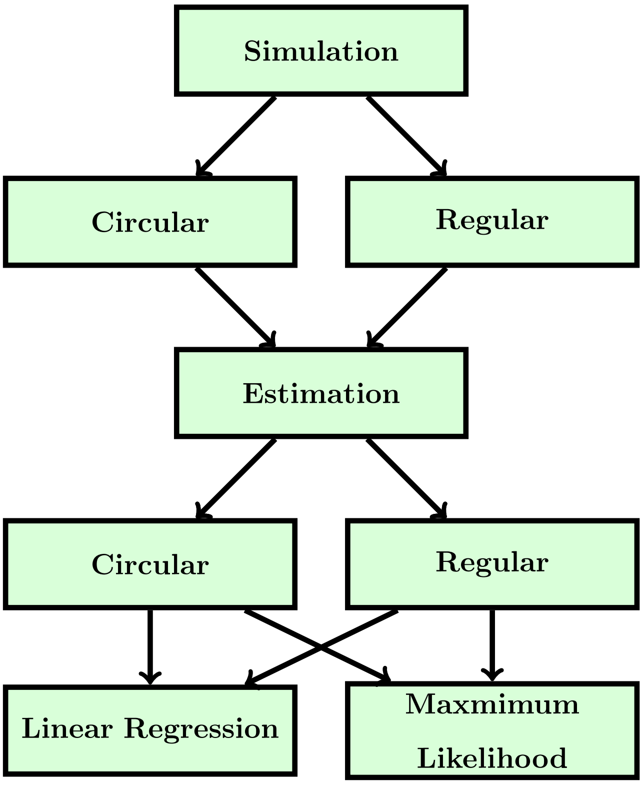


Figure 6.3: *Simulation – Estimation experiment scheme*

Table 6.1 tells us that there is not much difference between estimation approaches and methods, both for circular and regular data. Typically, maximum likelihood will be a better estimate than least squares. In this routine, the least squares estimate is the initial value for the maximum likelihood estimator, and we could view the maximum likelihood as a refinement of least

	E $\widehat{\alpha}_0$	E $\widehat{\alpha}_1$	MSE $\widehat{\alpha}_0$	MSE $\widehat{\alpha}_1$	Var $(\widehat{\alpha}_0)$	Var $(\widehat{\alpha}_1)$
Measure	10^{-2}	10^{-2}	10^{-2}	10^{-2}	10^{-5}	10^{-5}
CCML	30.18	4.82	3.21	3.20	52.73	40.10
CCLS	30.19	4.81	3.22	3.21	55.98	42.74
CRML	30.21	4.77	3.25	3.23	68.62	50.51
CRLS	30.22	4.76	3.25	3.23	72.79	54.11
RCML	30.37	4.57	3.26	3.27	57.79	43.27
RCLS	30.43	4.51	3.27	3.30	62.96	47.43
RRML	30.02	4.90	3.20	3.19	71.69	54.85
RRLS	30.10	4.82	3.21	3.23	76.42	58.68

Table 6.1: *Comparison of the different methods and approaches in 1D* STARCH(1) *with dimensions* 10×100 *and true parameters* $\alpha_0 = 0.3$ *and* $\alpha_1 = 0.05$. *Abbreviations: R is regular and C circular, ML is maximum likelihood and LS is least squares. RCML means regular simulation, circular estimation with maximum likelihood.*

squares. In all cases of this experiment, the ML estimators are closer to $\boldsymbol{\theta} = [0.3, 0.05]^T$ than the LS estimators, but the improvement does not seem to be significant. The mean squared error is lower for MLE than LSE, indicating that this is a global effect.

Circular estimates on circular data are better than regular estimation on the same data, but the opposite is true for regular data. Erroneous circular estimation misses the mark the most, but even that is not that far off. Comparing the wrongly specified circular estimation to the correct circular we have differences in mean squared error of respectively $5 \cdot 10^{-4}$ and $7 \cdot 10^{-4}$ for $\widehat{\alpha}_0^{\text{ML}}$ and $\widehat{\alpha}_1^{\text{ML}}$. This should be regarded as minor differences.

Another interesting aspect of this experiment is that the variances are smaller using the circular approach. This could be explained by the full utilization of the available data. Again, the disparities are small. We see this more clearly in higher dimension and visually in Chapter 7.

In two dimensions, the difference between approaches and methods are clearer. The reason is that we have a larger circular effect in two dimensions than in one. We do the same routine as in one dimension, only with a 5 × 5 × 100 dataset which totals up to 2,500 data points. Hence we have more data than in one dimension, but usually one need more data in higher dimension. The results are presented in Table 6.2.

A correctly specified circular model is only slightly better than the regular model on the same dataset and quite better than the correctly specified regular model according to Table 6.2. However, look at what happens to the circular estimates on a regular dataset. Both $\widehat{\alpha}_0$ and $\widehat{\alpha}_1$ completely miss their mark. It is a clear underestimation of α_1 and correspondingly overestima-

	E $\widehat{\alpha}_0$	E $\widehat{\alpha}_1$	MSE $\widehat{\alpha}_0$	MSE $\widehat{\alpha}_1$	Var ($\widehat{\alpha}_0$)	Var ($\widehat{\alpha}_1$)
Measure	10^{-2}	10^{-2}	10^{-2}	10^{-2}	10^{-5}	10^{-5}
CCML	30.47	4.92	5.67	9.33	92.67	4.42
CCLS	30.65	4.88	5.77	9.33	128.35	6.10
CRML	30.52	4.91	5.87	9.34	285.46	13.36
CRLS	30.85	4.84	6.05	9.37	379.67	17.40
RCML	36.06	3.73	6.65	9.82	116.45	5.36
RCLS	36.07	3.73	6.67	9.82	149.35	6.92
RRML	30.68	4.84	5.89	9.37	293.97	14.17
RRLS	30.86	4.80	6.00	9.40	403.09	18.90

Table 6.2: *Comparison of the different methods and approaches in 2D* STARCH(1) *with dimensions* $5 \times 5 \times 100$ *and true parameters* $\alpha_0 = 0.3$ *and* $\alpha_1 = 0.05$. *Abbreviations: R is regular and C circular, ML is maximum likelihood and LS is least squares. RCML means regular simulation, circular estimation with maximum likelihood.*

tion of α_0. This might seem alarming, but it is actually just as we should have expected. When trying to fit a circular model to non-circular data, the points furthest apart will be only weakly correlated and will therefore pull down the overall effect of α_1. The estimators of α_0 and α_1 are strongly negatively correlated (see Appendix B.1), meaning that if you decrease α_1, α_0 will increase.

We mentioned that the circular estimates have lower variance compared to regular estimates in one dimension. This is even more apparent in two dimensions. The variance is up to three times larger for regular estimates. In 2D, the loss of variables due to conditioning on the boundary is much larger. For a 5×5 grid, the spatial boundary consist of $2(5 + 5 - 2) = 16$ points out of totally 25 data points on the grid. While the circular model uses all 25 as variables, the regular model uses only 9. This can explain the larger variance.

6.3. Comparing settings

Now we want to see how the estimation routines perform under various conditions. In the previous experiment we only considered one set of parameters and dimension sizes, but now we try several settings. The conditions we vary are the parameters and temporal sample size. Since we found in the previous experiment that in one dimension almost everything works, we only consider two dimensional regular models.

In our situation, the stationarity condition is

$$\alpha_1 < \frac{1}{9} \approx 0.11, \tag{6.2}$$

since we assume nine points at one time lag influence $\boldsymbol{\sigma}_t^2$. We will therefore let $\alpha_1 = 0.11$ be our upper limit. The other parameter, α_0, is only required positive. The main goal in this situation is to see what happens when we approach the non-stationary situation. The spatial dimension is 10×10 and $\alpha_0 = 0.4$. For each set of parameters the temporal sample sizes $T = 100$, 200 and 500 are used. The three levels of α_1 used are 0.05, 0.09 and 0.11.

The resulting mean parameter estimates are presented in Table 6.3. We use the same abbreviations as in the previous experiments. Experiments 1-6 give estimates well within what could be expected. Not surprisingly, the circular estimates seem biased, due to the misspecification. As we approach the non-stationary, all estimators of α_0 break down, except RRML. Interestingly, the α_1 estimates does not seem affected, but the constant term estimate $\widehat{\alpha}_0$ explodes. In some way, the RML manage to refine the RLS estimate, while the CML do not. The reason why this happens, is not clear. It might be the lack of a finite fourth moment of $\boldsymbol{X}_t$, but this needs to be investigated further.

Experiment	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
T	100	200	500	100	200	500	100	200	500
True α_0	40.00	40.00	40.00	40.00	40.00	40.00	40.00	40.00	40.00
RCML α_0	44.04	43.94	43.90	55.91	55.58	55.35	144.28	142.65	135.64
RCLS α_0	44.00	43.82	43.87	56.43	54.53	53.80	282.76	250.66	243.21
RRML α_0	40.09	40.02	39.98	40.51	40.16	40.01	43.77	42.58	41.48
RRLS α_0	40.15	39.99	40.08	43.68	41.30	40.50	192.37	154.43	148.16
True α_1	5.00	5.00	5.00	9.00	9.00	9.00	11.00	11.00	11.00
RCML α_1	4.38	4.41	4.41	8.15	8.15	8.18	10.63	10.70	10.76
RCLS α_1	4.39	4.43	4.41	8.11	8.21	8.26	9.98	10.21	10.33
RRML α_1	4.98	5.01	5.01	8.97	8.97	8.99	10.95	10.97	10.99
RRLS α_1	4.97	5.02	4.99	8.78	8.90	8.97	10.33	10.54	10.64

Table 6.3: *Result from 9 different set ups of an 2D* STARCH(1). *Measure is* 10^{-2} *for the estimates and true parameter values. Spatial dimension is* 10×10.

Another thing worth noticing is that the estimates tend to improve as temporal sample size increase. Even though the estimates are terribly wrong for α_0 when $\alpha_1 = 0.11$, the error decrease as T goes from 100 to 200 and finally 500. At least, it helps to increase the sample size.

6.4. Student's t-distributed innovations

In some situations, the assumption that our innovations are normally distributed, might be inappropriate. We want to see what happens to our estimates, if the innovations comes from a student's t-distribution with ν degrees of freedom. For ν large, a central t-distribution approach a standard normal distribution, making this a fitting distribution for testing how sensitive the estimation methods are with respect to the assumption of standard normal innovations. The t-distribution has heavier tails, but shares many of its properties with the normal distribution.

If T has a central student's t-distribution with ν degrees of freedom, we have that

$$\mathrm{E}\, T = 0, \qquad \mathrm{Var}(T) = \frac{\nu}{\nu - 2}. \tag{6.3}$$

Let $\{\boldsymbol{T}_t\}$ be a vector process of independent student's t-distributed random variables with ν degrees of freedom. Define

$$\boldsymbol{\Psi}_t = \sqrt{\frac{\nu - 2}{\nu}} \boldsymbol{T}_t, \tag{6.4}$$

such that

$$\mathrm{E}\, \boldsymbol{\Psi}_t = \mathbf{0}, \qquad \mathrm{Var} \boldsymbol{\Psi}_t = \frac{\nu - 2}{\nu} \mathrm{Var} \boldsymbol{T}_t = \frac{\nu - 2}{\nu} \frac{\nu}{\nu - 2} \mathbb{I} = \mathbb{I}. \tag{6.5}$$

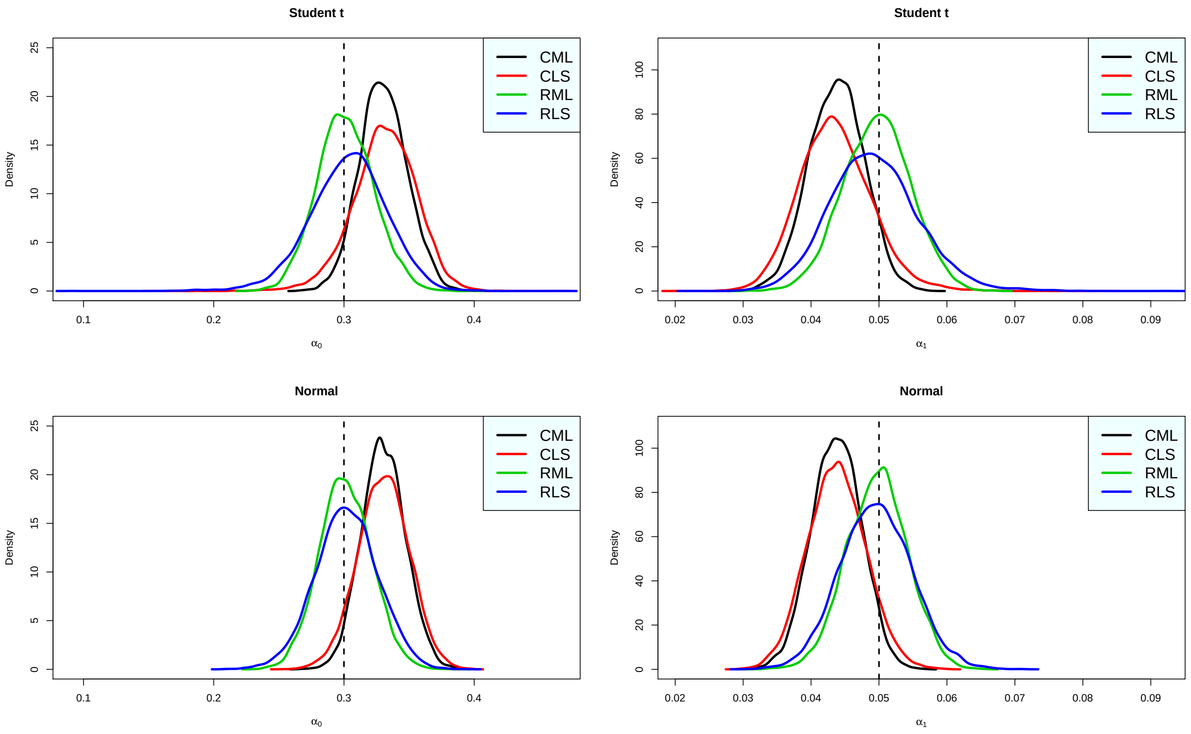


Figure 6.4: *Density plots of estimated parameters of 10,000 regular simulated 2D* STARCH(1) *processes using both student's $t(\nu = 10)$-distributed and standard Gaussian innovations. The true parameters are indicated with vertical lines. The dimensions of the datasets are* $10 \times 10 \times 100$.

In this way, $\{\mathbf{\Psi}_t\}$ is an iid $\mathrm{WN}(\mathbf{0}, \mathbb{I})$ process and fulfils the requirements on the innovations of definition 3.1.1.

We are going to simulate a STARCH(1) process using $\{\mathbf{\Psi}_t\}$ as our innovations. We also simulate a STARCH(1) using the standard normal innovation distribution for comparison. In order to get the two simulations comparable, we draw the same uniformly distributed variables in both situations and map them using the quantile functions of both student's t-distribution and normal distribution to get samples from the respective distributions. Estimates are then found using circular and regular estimation approaches and maximum likelihood and least squares methods. After 10,000 repetitions of this, we estimate the non-parametric density of the each estimator. The estimation routine still assumes normal distribution.

Figure 6.4 shows the estimated marginal density plots for all estimators, both with t- and normal distributed innovations. Since the data is regularly simulated, the circular estimates are biased, while the regular seem unbiased. The circular model systematically overestimates α_0 and underestimate α_1. Comparing across innovation distributions, the resulting densities are very similar. It does not seem to be any shift in expectation, since the maximums of the curves are found for the same α_i values $(i = 0, 1)$. However, there is a visual difference in the uncertainty. The bell curves seem more stretched out for the t-distributed innovations estimates. The estimated density lines are only visible where there are observations, and therefore the figure also show more extremes for the t-distributed innovations estimates.

Compared to the maximum likelihood estimators, the least squares estimator densities seem to be a bit wider and heavy tailed. Although we do not present it here, we have tested with lower degrees of freedom. This leads to more stretched out bell curves and heavier tailed estimator distributions.

In conclusion, this experiment illustrates that the estimators are not that sensitive to changing the innovation distribution to a student's t-distribution.

6.5. STGARCH(1,1)

The reader might argue that procedures working STARCH does not necessarily imply acceptable preformance for STGARCH. Therefore our methods need to be tested for STGARCH and here this is done using $(p, q) = (1, 1)$. There is quite a step up in complexity when going from an ARCH to a GARCH process. This extension means that the unobserved $\{\boldsymbol{\sigma}_t\}$ has to be included in state representation in order to keep the Markov property. In addition the number of parameters is increased from two to three. Therefore a GARCH process will need more data than an ARCH.

For this experiment, we only do circular estimation. Conditioning on the boundary can not be used in this situation, since it is not able to estimate the unobserved $\{\boldsymbol{\sigma}_t\}$ process on the boundary. It is however important to test if the circular model works for the more complex model STGARCH(1,1). The model we use here is given by

$$\begin{aligned} \boldsymbol{X}_t &= \operatorname{diag}\{\boldsymbol{\sigma_t}\}\,\boldsymbol{Z}_t, & \mathbb{X}_t &= \begin{bmatrix}\mathbf{1}, & \boldsymbol{\mathcal{W}}^\star \boldsymbol{X}_{t-1}^2, & \boldsymbol{\mathcal{W}}^\star \boldsymbol{\sigma}_{t-1}^2\end{bmatrix}, \\ \boldsymbol{\sigma}_t^2 &= \mathbb{X}_t \boldsymbol{\theta}, & \boldsymbol{\theta} &= \begin{bmatrix}\alpha_0, & \alpha_1, & \beta_1\end{bmatrix}^T. \end{aligned}$$

The stationarity condition for a 2D STGARCH(1,1) with the same neighbouring structure for the ARCH and GARCH terms (given by Figure 6.1) is that

$$9\alpha_1 + 9\beta_1 < 1 \quad \text{or} \quad \alpha_1 + \beta_1 < 1/9 \approx 0.11. \tag{6.6}$$

We choose the true parameters to be $\alpha_0 = 0.1$, $\alpha_1 = 0.05$ and $\beta_1 = 0.02$, which gives $\alpha_1 + \beta_1 = 0.07 < 1/9$. Hence, our simulated data is stationary. We simulate 1,000 non-circular datasets and estimate them circularly using both maximum likelihood and least squares estimation. To make the datasets non-circular, we simulate circular datasets of size $210 \times 210 \times 1000$ and keep the $10 \times 10 \times 1000$ points at the center.

A downside to STGARCH estimation in this context is that we need a user defined initial value for the parameter vector. The reason for this is that we now need to estimate the unobserved process, $\{\boldsymbol{\sigma}_t\}$, which depends on the parameters. When the process is estimated, we can optimize the likelihood function or minimize the objective function to find new parameters. These new parameters will estimate an updated version of $\{\boldsymbol{\sigma}_t\}$. We use the initial parameter vector $\widehat{\boldsymbol{\theta}}^{(0)} = [.2, .01, .01]^T$, both for the ML and LS estimation. The result is given in Table 6.4.

		α_0	α_1	β_1	Iterations
	$\boldsymbol{\theta}$ (10^{-2})	10.000	5.000	2.000	
MLE	$\mathrm{E}\,\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}$ (10^{-2})	10.597	4.402	2.352	17.91
	SD (10^{-2})	0.649	0.115	0.294	
	MSE (10^{-5})	7.769	3.711	2.104	
LSE	$\mathrm{E}\,\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}$ (10^{-2})	10.250	4.407	2.491	19.13
	SD (10^{-2})	0.747	0.139	0.336	
	MSE (10^{-5})	6.206	3.707	3.536	

Table 6.4: *Results from simulating 1000 non-circular* STGARCH(1, 1) *processes of dimensions* $10 \times 10 \times 1000$.

Comparing the two approaches we see that the mean estimates are quite satisfactory. Both seem to overestimate α_0 and β_1 and underestimate α_1. Remember that this is circular estimation of non-circular data, and some bias is expected. We have used quite large datasets for this experiment, so we suspect the bias due to the circular assumption to be relatively small. The mean squared errors of order 10^{-5} is small compared with the parameter values.

We also calculated the empirical correlation matrix of the 1,000 estimates. Table 6.5 show that there are some minor differences between the ML and LS correlation matrices, but the tendencies are the same. The GARCH coefficient estimator $\widehat{\beta}_1$ is negatively correlated with both $\widehat{\alpha}_0$ and $\widehat{\alpha}_1$, while $\widehat{\alpha}_0$ and $\widehat{\alpha}_1$ are practically uncorrelated. The negative correlation between $\widehat{\beta}_1$ and $\widehat{\alpha}_0$ must be said to be extremely high and close to minus one. These high negative correlations might be explained by (3.6). For this experiment, the empirical version of (3.6) simplifies to

$$\widehat{\sigma}_X^2 = \frac{\widehat{\alpha}_0}{1 - 9\widehat{\alpha}_1 - 9\widehat{\beta}_1}. \tag{6.7}$$

For fixed σ_X^2, an increase in the numerator must lead to an increase in the denominator to keep the fraction constant. Increasing $\widehat{\alpha}_0$ must therefore decrease $\widehat{\alpha}_1$ or $\widehat{\beta}_1$, thus explaining the strong negative correlation.

	α_0^{ML}	α_1^{ML}	β_1^{ML}	α_0^{LS}	α_1^{LS}	β_1^{LS}
α_0	1.000	0.074	-0.932	1.000	0.023	-0.922
α_1	0.074	1.000	-0.406	0.023	1.000	-0.392
β_1	-0.932	-0.406	1.000	-0.922	-0.392	1.000

Table 6.5: *Estimated correlation matrices of ML and LS estimators of* $\text{STGARCH}(1,1)$.

Chapter 7

Parametric bootstrap bias correction

As we saw in Chapter 6, when using a circular estimation technique on regular data, we get a biased estimate due to the somewhat incorrect model specification. We will in this chapter look at a method for correcting this bias. It turns out that doing a parametric bootstrap bias correction can be quite effective. We start with some general theory and then apply it to our situation.

An unbiased estimator, means that $\mathrm{E}\,\widehat{\theta} = \theta$. The bias of an estimator $\widehat{\theta}$ for θ is

$$\mathrm{bias}\{\widehat{\theta}\} = \mathrm{E}\left(\widehat{\theta} - \theta\right) = \mathrm{E}\,\widehat{\theta} - \theta.$$

Thus, an unbiased estimator has zero bias. When using circular estimators on non-circular data we get biased estimates. We are going to use resampling methods to estimate this bias and adjust our estimate accordingly.

Bootstrap was first introduced by Efron in 1979. Efron had constructed a resampling procedure that was meant to approximate the jackknife (another resampling method). He wanted to derive properties of the bootstrap to better understand the jackknife, but ended up developing a method that in some situations succeed the jackknife, according to Chernick and Labudde (2014). Efron made further developments of the bootstrap throughout the 1980s, publishing numerous articles. He also wrote a book with Tibshirani (1993) called *An Introduction to the Bootstrap*, which can be a good reference for further reading.

Bootstrap methods belong to the class of Monte Carlo methods called resampling methods. Normally, we want to be able to say something about a characteristic of a large population, such as the political opinion of the people living in Norway. It will be too costly and time consuming to call and ask everyone, therefore we only ask a random (representative) sample from the population. A resampling method treats the observed sample as the entire population, and generate random samples from the sample (resampling) to estimate population characteristics and make inference about the population (Rizzo, 2007, pp. 183). Rizzo says that the sample can be regarded as a pseudo-population with similar characteristics as the true population. The sampling distribution of a statistic can be estimated by repeatedly generating random samples from the pseudo-population.

We separate between parametric and non-parametric bootstrap. Resampling from a fully specified probability distribution is called parametric bootstrap, while for a non-parametric bootstrap the distribution is not specified at all. The parametric bootstrap is most fitting for our purpose, since we already have a parametric model. It is called a resampling method, but in practice we simulate new datasets using an estimate as the true parameter.

7.1. Bootstrap Bias Correction

Bootstrapping is here used for bias-correction. This is reasonable since the actual estimators, coming from an incorrect model, will inevitable get biased. There is no such thing as a free lunch and this applies here. The point estimate will have less bias but increased variance.

Let $\boldsymbol{\delta}$ be the bias and $\widehat{\boldsymbol{\theta}}$ the estimator of $\boldsymbol{\theta}$ based on the original sample. We then have

$$\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta} + \boldsymbol{\delta} + \boldsymbol{Z}_0. \tag{7.1}$$

Based on $\widehat{\boldsymbol{\theta}}$, the parametric bootstrap procedure simulates B datasets using $\widehat{\boldsymbol{\theta}}$ as the true parameter. Let $\{\widetilde{\boldsymbol{\theta}}_b : b = 1, \ldots, B\}$ be the set of estimates based on the bootstrap samples. Then

$$\widetilde{\boldsymbol{\theta}}_b = \widehat{\boldsymbol{\theta}} + \boldsymbol{\delta} + \boldsymbol{Z}_b, \quad b = 1 \ldots B \tag{7.2}$$

where $\{\boldsymbol{Z}_b : b = 0, \ldots, B\}$ is some zero expectation noise process. Taking the mean over all $\widetilde{\boldsymbol{\theta}}_b$ we get

$$\overline{\widetilde{\boldsymbol{\theta}}} = B^{-1} \sum_{b=1}^{B} \widetilde{\boldsymbol{\theta}}_b = \widehat{\boldsymbol{\theta}} + \boldsymbol{\delta} + \overline{\boldsymbol{Z}}. \tag{7.3}$$

Solving (7.3) for $\boldsymbol{\delta}$ using that $\overline{\boldsymbol{Z}} \to \operatorname{E} \boldsymbol{Z}_0 = \mathbf{0}$, gives an estimator of the bias, $\widehat{\boldsymbol{\delta}}$.

$$\widehat{\boldsymbol{\delta}} = \overline{\widetilde{\boldsymbol{\theta}}} - \widehat{\boldsymbol{\theta}}. \tag{7.4}$$

Compensating our original estimator with the bias estimator, gives a bias corrected estimator, $\boldsymbol{\theta}^{\star}$, expressed by

$$\boldsymbol{\theta}^{\star} = \widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\delta}} = \widehat{\boldsymbol{\theta}} - (\overline{\widetilde{\boldsymbol{\theta}}} - \widehat{\boldsymbol{\theta}}) = 2\widehat{\boldsymbol{\theta}} - \overline{\widetilde{\boldsymbol{\theta}}}. \tag{7.5}$$

We summarize how a parametric bootstrap bias correction can be carried out in Algorithm 3 below.

7.2. Example: STARCH(1)

We know that we get some bias by assuming a circular model, when the data is non-circular. In the STARCH(1) model used in Chapter 6, the estimators turned out to be strongly negatively

Algorithm 3 Parametric Bootstrap Bias Correction algorithm

```
Estimate parameters, $\widehat{\boldsymbol{\theta}}$, on the data
**for** each node $b \in \{1, \dots, B\}$ **do**
    Simulate a process with the same settings as original data and parameter vector $\hat{\boldsymbol{\theta}}$
    Estimate $\widetilde{\boldsymbol{\theta}}_b$ on the simulated data
**end for**
Compute $\overline{\widetilde{\boldsymbol{\theta}}} = B^{-1} \sum_{b=1}^{B} \widetilde{\boldsymbol{\theta}}_b$
Estimate bias by $\widehat{\delta} = \widehat{\boldsymbol{\theta}} - \overline{\widetilde{\boldsymbol{\theta}}}$
Calculate bias corrected estimate, $\boldsymbol{\theta}^{\star} = 2\hat{\boldsymbol{\theta}} - \overline{\widetilde{\boldsymbol{\theta}}}$
```

correlated (see Appendix A or B.1). The mistake of assuming circularity creates a false dependence between some points in the dataset, making the total effect of this dependence smaller and hence the approach should systematically underestimate α_1. Systematic underestimation is the same as negative bias. Due to the strong negative correlation between the parameters, we suspect that underestimation of α_1 leads to overestimation of α_0 or a positive biased estimate of α_0. We can see this systematic erroneous estimation in Figure 7.1. If the estimation was unbiased, the points in Figure 7.1 should be symmetric around zero, but we see that the distribution is skewed down to the right in the fourth quadrant. The aim of our bias correction is to move the center back to zero.

We use Algorithm 3 on a two-dimensional STARCH process (same situation as Chapter 6). We picked the true parameter vector randomly. The bootstrap replicas are simulated using the

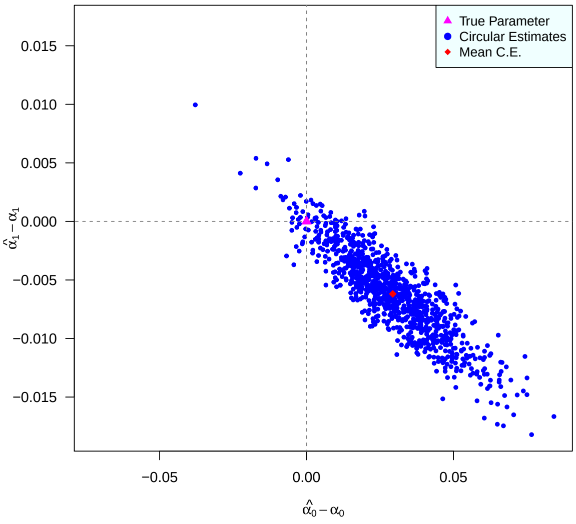


Figure 7.1: *Bias in* 1000 *simulated datasets with circular estimation.*

estimate.

We repeat the algorithm 1000 times using the same true $\boldsymbol{\theta}$ each time. For each repetition, k, we simulated an «observed» dataset with dimension $10 \times 10 \times 1000$ and estimated the parameters, $\widehat{\alpha}_{0,k}$ and $\widehat{\alpha}_{1,k}$. Then we did a parametric bootstrap by simulating $B = 200$ new data sets, using $\widehat{\boldsymbol{\alpha}}_k$ as the true parameter and estimated parameters for each of these yielding the bootstrap estimate $\widetilde{\boldsymbol{\alpha}}_{k,j}$. After all these $\widetilde{\boldsymbol{\alpha}}_{k,j}$ were computed, we calculated the mean, $\overline{\widetilde{\boldsymbol{\alpha}}}_k$ and estimated the bias $\widehat{\delta}_k = \overline{\widetilde{\boldsymbol{\alpha}}}_k - \widehat{\alpha}_k$. Using the estimated bias, we computed a bootstrap bias corrected (BBC) estimate, $\boldsymbol{\alpha}^\star_k = \widehat{\boldsymbol{\alpha}}_k - \widehat{\boldsymbol{\delta}}_k$. We also estimated parameters for each «observed» dataset using regular estimation for comparison purposes. Finally we calculated the mean of all the estimates. We calculated the mean squared error (MSE) and the empirical variance of the different estimates as well. The results are presented in Table 7.1 and Figure 7.2.

Table 7.1 shows that the circular estimates are actually close to the true values, but not impressive. They have lower variances than their competitors, but due to the biased estimates the mean squared error (MSE) is relatively large compared to the others (factor of about 3). When we do a parametric bootstrap bias correction (BBC), the estimates are improved significantly. This comes at the price of increased variance, for this particular experiment the variance increase by about 23%. The MSE however, decreases with approximately 70% for both α_0 and α_1, yielding a large improvement. A somewhat surprising result from this experiment, is that the bootstrap bias corrected estimates actually have lower variance and MSE than the regular estimates. This means that even though the mean point estimates are better for the regular approach, the estimates are more uncertain. We can see from the 95% confidence intervals for the estimates, that the true parameters are not even included in the circular estimates, but they are in the BBC

	ME (10^{-2})	VAR (10^{-5})	MSE (10^{-5})	Lower 95% CI	Upper 95% CI
Circular $\widehat{\alpha}_0$	31.934	26.210	112.256	31.616	32.251
Circular $\widehat{\alpha}_1$	4.381	1.362	5.191	4.309	4.453
BBC $\alpha^\star_0$	29.314	32.174	33.126	28.962	29.665
BBC $\alpha^\star_1$	4.935	1.679	1.719	4.855	5.015
Regular $\widehat{\alpha}_0$	29.165	35.479	35.714	28.795	29.534
Regular $\widehat{\alpha}_1$	4.962	1.899	1.912	4.877	5.047

Table 7.1: *Mean results from 1000 repeated experiments with regular simulated datasets, estimated using circular estimation, regular estimation and Bootstrap bias corrected circular estimation. ME is short for mean estimates, VAR is variance and MSE is mean squared error. The true parameters are $\alpha_0 = 0.29$ and $\alpha_1 = 0.05$ and dimensions are $10 \times 10 \times 1000$. We also estimated a 95% confidence interval for each estimator based on the aymptotic normality of the estimates.*

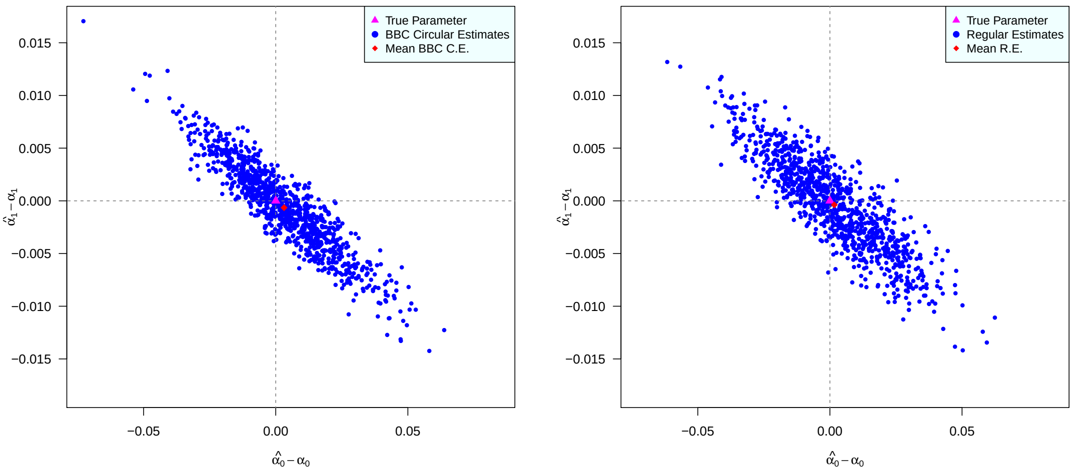


Figure 7.2: *Left: The same circular estimates as in Figure 7.1 after the parametric bootstrap bias correction. Right: Corresponding regular estimates. The axes are equal in all these three plots, making them comparable. Notice that the mean point estimate is closer for the regular approach, but the spread is also larger.*

and regular confidence intervals. In fact, the BBC interval is narrower than the regular one, due to lower variance.

Figures 7.1 and 7.2 are scatter plots of the bias from each of the 1,000 repetitions, pre- and post bias correction. The plots are made using the same axes to make them comparable. In Figure 7.1 the estimates' biases are far from centralized. After the bias correction has been performed the biases are almost perfectly centralized (left hand side of Figure 7.2). The right hand side of Figure 7.2 shows the corresponding regular estimates' biases. These are also centralized, but they are more scattered than the bias corrected. This visually confirms the higher variance of regular estimates found in Table 7.1.

Doing a bias correction is not free. It costs computing power and the variance of the estimate increase. When doing a circular estimation on non-circular data, one must deal with the bias somehow. This chapter illustrates that a parametric bootstrap bias correction is an effective way of handling the bias.

Chapter 8

Gibbs sampler and EM algorithm

Imagine that we have a circular model, perhaps over a quite large area, but we do not observe the entire process. We only observe a smaller area at the center of this circular model. The data we observe will of course not have circular properties, and hence using a circular model may be deemed fault. We would like to treat the unobserved part of the circular model as missing data, use a Gibbs sampler to simulate the missing data and the EM algorithm to estimate parameters.

At this point, we have not managed to implement this in a satisfactory manner, but we present the idea anyway and hope to continue to investigate this angle at a later stage.

8.1. Gibbs sampler

A Gibbs sampler is an algorithm from computational statistics for creating pseudo samples from multivariate distributions, by sampling from marginal densities. The method belongs to the group of Metropolis-Hastings algorithms and is much used in Bayesian statistics (Rizzo, 2007).

Let $\mathcal{S}_1$ be the space where the observations are and $\mathcal{S}_2$ the entire circular area. We call the observed data $\{\boldsymbol{X}_t\} = \{X_t(u) : u \in S_1\}$ and the unobserved data $\{\boldsymbol{Y}_t\} = \{Y_t(u) : u \in \mathcal{S}_2 \backslash \mathcal{S}_1\}$, both where $t = 1, \ldots, T$. Let $\boldsymbol{V}_t = (\boldsymbol{X}_t, \boldsymbol{Y}_t)$ be the entire circular process defined on $\mathcal{S}_2$. Figure 8.1 illustrates this notation. All the theory developed for circular STGARCH is valid for

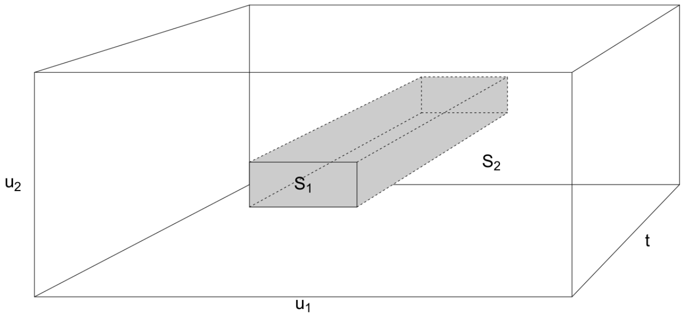


Figure 8.1: *The grey area ($\mathcal{S}_1$) illustrates where the observed sample is and the rest of the area ($\mathcal{S}_2 \backslash \mathcal{S}_1$) is where the unobserved variables are. Together they form a circular model on $\mathcal{S}_2$.*

$\{\boldsymbol{V}_t\}$. We use the common notation that $\boldsymbol{V}_{-t} = \{\boldsymbol{V}_s : s \neq t\}$. We will be using the Markov property and therefore we will assume that $\{\boldsymbol{V}_t\}$ is a STARCH(1) process for this derivation. We could develop the procedure in a more general fashion, but it will complicate calculations. As the goal is to indicate an idea, we want to keep calculations simple. Using Bayes' theorem and the Markov property of $\{\boldsymbol{V}_t\}$, we have that

$$\begin{aligned} f(\boldsymbol{V}_t|\boldsymbol{V}_{-t}) &= f(\boldsymbol{V}_t|\boldsymbol{V}_{t-1}, \boldsymbol{V}_{t+1}) \\ &= \frac{f(\boldsymbol{V}_t|\boldsymbol{V}_{t-1})f(\boldsymbol{V}_{t+1}|\boldsymbol{V}_t)}{f(\boldsymbol{V}_{t+1}|\boldsymbol{V}_{t-1})}. \end{aligned} \tag{8.1}$$

The denominator of this expression is a normalizing constant in terms of $\boldsymbol{V}_t$. $\boldsymbol{V}_t$ only contributes in the numerator and $f(\boldsymbol{V}_{t+1}|\boldsymbol{V}_t)$ and $f(\boldsymbol{V}_t|\boldsymbol{V}_{t-1})$ are consecutive terms in the likelihood we developed in Chapter 5. We also have that

$$f(\boldsymbol{Y}_t, \boldsymbol{X}_t|\boldsymbol{V}_{-t}) = f(\boldsymbol{Y}_t|\boldsymbol{X}_t, \boldsymbol{V}_{-t})f(\boldsymbol{X}_t|\boldsymbol{V}_{-t}). \tag{8.2}$$

Now, what we are really interested in, is the distribution of $\{\boldsymbol{Y}_t|\boldsymbol{X}_t, \boldsymbol{V}_{-t}\}$, for the purpose of sampling from it. If we combine the two equations (8.1) and (8.2), we get that

$$f(\boldsymbol{Y}_t|\boldsymbol{X}_t, \boldsymbol{V}_{-t}) = \frac{f(\boldsymbol{X}_t, \boldsymbol{Y}_t|\boldsymbol{V}_{-t})}{f(\boldsymbol{X}_t|\boldsymbol{V}_{-t})} = \frac{f(\boldsymbol{V}_t|\boldsymbol{V}_{t-1})f(\boldsymbol{V}_{t+1}|\boldsymbol{V}_t)}{f(\boldsymbol{X}_t|\boldsymbol{V}_{-t})f(\boldsymbol{V}_{t+1}|\boldsymbol{V}_{t-1})}, \tag{8.3}$$

where $c^{-1} = f(\boldsymbol{X}_t|\boldsymbol{V}_{-t})f(\boldsymbol{V}_{t+1}|\boldsymbol{V}_{t-1})$ is a normalizing constant. We can write this expression explicitly. Like before we assume independently normally distributed innovations with zero expectation, and hence $\boldsymbol{V}_t|\boldsymbol{V}_{t-1} \sim \text{MVN}\left(\boldsymbol{0}, \text{diag}\left\{\boldsymbol{\sigma}_t^2\right\}\right)$. To ease notation, let $\mathbb{H}_t = \text{diag}\left\{\boldsymbol{\sigma}_t^2\right\}$. We have that

$$f(\boldsymbol{V}_t|\boldsymbol{V}_{t-1}) = (2\pi)^{-n/2}|\mathbb{H}_t|^{-1/2}\exp\Big\{-\frac{1}{2}\boldsymbol{V}_t^T\mathbb{H}_t^{-1}\boldsymbol{V}_t\Big\}, \tag{8.4}$$

where n is the length of vector $\boldsymbol{V}_t$. Hence,

$$f(\boldsymbol{Y}_t|\boldsymbol{X}_t, \boldsymbol{V}_{-t}) = c(2\pi)^{-n}|\mathbb{H}_t|^{-1/2}|\mathbb{H}_{t+1}|^{-1/2}\exp\Big\{-\frac{1}{2}\big(\boldsymbol{V}_t^T\mathbb{H}_t^{-1}\boldsymbol{V}_t + \boldsymbol{V}_{t+1}^T\mathbb{H}_{t+1}^{-1}\boldsymbol{V}_{t+1}\big)\Big\} \tag{8.5}$$

It may be convenient to split up the parts of $\boldsymbol{V}_t$ into the known $\boldsymbol{X}_t$ and the variable $\boldsymbol{Y}_t$. Since these are conditionally independent and since $\mathbb{H}_t$ is a diagonal matrix, let $\mathbb{H}_{t,Y}$ correspond to diag $\left\{\boldsymbol{\sigma}_{t,Y}^2\right\}$ and $\mathbb{H}_{t,X} = \text{diag}\left\{\boldsymbol{\sigma}_{t,X}^2\right\}$. Since $\mathbb{H}_t$ is diagonal, the determinant is simply the product of the diagonal elements. Hence $|\mathbb{H}_t|^{-1/2} = |\mathbb{H}_{t,Y}|^{-1/2} \cdot |\mathbb{H}_{t,X}|^{-1/2}$. The sum of squares in the exponent is also separable. Since $\boldsymbol{V}_t$ is a $n \times 1$ vector and $\mathbb{H}_t$ is a $n \times n$ diagonal matrix, the product $\boldsymbol{V}_t^T\mathbb{H}_t^{-1}\boldsymbol{V}_t$ is 1×1. In fact, $\boldsymbol{V}_t^T\mathbb{H}_t^{-1}\boldsymbol{V}_t = \boldsymbol{X}_t^T\mathbb{H}_{t,X}^{-1}\boldsymbol{X}_t + \boldsymbol{Y}_t^T\mathbb{H}_{t,Y}^{-1}\boldsymbol{Y}_t$ and the separated version of Equation (8.5) is

$$\begin{aligned} f(\boldsymbol{Y}_t|\boldsymbol{X}_t, \boldsymbol{V}_{-t}) = c \cdot (2\pi)^{-n}|\mathbb{H}_t|^{-1/2}&|\mathbb{H}_{t+1,X}|^{-1/2}|\mathbb{H}_{t+1,Y}|^{-1/2} \\ &\times \exp\Big\{-\frac{1}{2}\big(\boldsymbol{X}_t^T\mathbb{H}_{t,X}^{-1}\boldsymbol{X}_t + \boldsymbol{Y}_t^T\mathbb{H}_{t,Y}^{-1}\boldsymbol{Y}_t \\ &+ \boldsymbol{X}_{t+1}^T\mathbb{H}_{t+1,X}^{-1}\boldsymbol{X}_{t+1} + \boldsymbol{Y}_{t+1}^T\mathbb{H}_{t+1,Y}^{-1}\boldsymbol{Y}_{t+1}\big)\Big\}. \end{aligned} \tag{8.6}$$

It is important to acknowledge that it is only the elements involving $\boldsymbol{Y}_t$ that are considered stochastic in this density. Besides from $\boldsymbol{Y}_t$ itself, this is $\mathbb{H}_{t+1,Y}$ and some elements of $\mathbb{H}_{t+1,X}$, neighbouring $\mathcal{S}_2 \backslash \mathcal{S}_1$. For convenience, we have marked these terms in blue, the rest are deemed known.

Theoretically, c is determined by

$$c^{-1} = \int_{-\infty}^{\infty} g(\boldsymbol{Y}_t | \boldsymbol{X}_t, \boldsymbol{V}_{-t}) d\boldsymbol{Y}_t, \text{ where } g(x) = c^{-1} f(x), \tag{8.7}$$

but this integral can be difficult to express analytically. The sole purpose here is to use (8.5) to create a Gibbs sample of the missing data in $\mathcal{S}_2 \backslash \mathcal{S}_1$. To be able to do this, we need to numerically estimate c. One possible way of estimating c, is by Monte Carlo integration (Rizzo, 2007, Ch. 5).

8.2. EM algorithm

The EM algorithm is a very useful algorithm for dealing with incomplete data. For more information about using the EM algorithm when dealing with missing data, see Dempster et al. (1977).

The following is inspired by unpublished work of Karlsen (2015a). We use the notation $\boldsymbol{X} = \{\boldsymbol{X}_t : t = 1, \ldots, T\}$ and $\boldsymbol{Y} = \{\boldsymbol{Y}_t : t = 1, \ldots, T\}$ to denote the respective sets of variables. The EM algorithm defines a sequence $\{\boldsymbol{\theta}^{(m)}, m \geq 0\}$, where $\boldsymbol{\theta}^{(0)}$ is an initial value. The sequence is defined by

$$\boldsymbol{\theta}^{(m+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \, \mathrm{E}_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\theta}^{(m)}} \left(\log f(\boldsymbol{X}, \boldsymbol{Y} | \boldsymbol{\theta}) \right) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \int \log f(\boldsymbol{X}, \boldsymbol{y} | \boldsymbol{\theta}) f(\boldsymbol{y} | \boldsymbol{X}, \boldsymbol{\theta}^{(m)}) d\boldsymbol{y} \tag{8.8}$$

for $m \geq 0$. The maximization problem can in many cases be formulated as

$$\nabla_{\boldsymbol{\theta}} \, \mathrm{E}_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\theta}^{(m)}} \left(\log f(\boldsymbol{X}, \boldsymbol{Y} | \boldsymbol{\theta}) \right) \Big|_{\boldsymbol{\theta}^{(m+1)}} = \boldsymbol{0}, \tag{8.9}$$

which is the same as solving

$$\int \frac{\partial \log f}{\partial \theta_i} (\boldsymbol{X}, \boldsymbol{y} | \boldsymbol{\theta}) f(\boldsymbol{y} | \boldsymbol{X}, \boldsymbol{\theta}^{(m)}) d\boldsymbol{y} = 0, \qquad i = 1, \ldots, p, \tag{8.10}$$

with respect to $\boldsymbol{\theta}$. The rightmost part we derived in the previous section. The other part is

$$\nabla_{\boldsymbol{\theta}} \log f(\boldsymbol{X}, \boldsymbol{Y}) = \nabla_{\boldsymbol{\theta}} \log f(\boldsymbol{V}_0 | \boldsymbol{\theta}) + \sum_{t=1}^{T} \nabla_{\boldsymbol{\theta}} \log f(\boldsymbol{V}_t | \boldsymbol{V}_{t-1}, \boldsymbol{\theta}) = \sum_{t=1}^{T} \nabla_{\boldsymbol{\theta}} \log f(\boldsymbol{V}_t | \boldsymbol{V}_{t-1}, \boldsymbol{\theta}) + \mathcal{O}(1),$$

where deleting the first term is relatively harmless. This type of approximation is similar to what is done in pseudo-likelihood theory.

Suppose we are able to generate a Markov process $\{\boldsymbol{Y}^{(k)} : k = 1, \ldots, K\}$ (the Gibbs sample) of dimension equal to the number of unobserved variables with stationary measure $f(\boldsymbol{Y} | \boldsymbol{X}, \boldsymbol{\theta}^{(m)})$.

Then

$$\int \frac{\partial \log f}{\partial \theta_i}(\boldsymbol{X}, \boldsymbol{y}|\boldsymbol{\theta}) f(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}^{(m)}) d\boldsymbol{y} = \lim_{K\to\infty} K^{-1} \sum_{k=1}^{K} \frac{\partial \log f}{\partial \theta_i}(\boldsymbol{X}, \boldsymbol{Y}^{(k)}|\boldsymbol{\theta}), \quad i = 1, \ldots, p, \tag{8.11}$$

where $\boldsymbol{\theta}$ is p-dimensional. By choosing K sufficiently large, we can use (8.11) to approximate the left hand side of (8.10) and solve the system of equations

$$K^{-1} \sum_{k=1}^{K} \frac{\partial \log f}{\partial \theta_i}(\boldsymbol{X}, \boldsymbol{Y}^{(k)}|\boldsymbol{\theta}) = 0, \quad i = 1, \ldots, p, \tag{8.12}$$

with respect to $\boldsymbol{\theta}$. Algorithm 4 is an EM-algorithm for estimating $\boldsymbol{\theta}$ based on this approximation.

Algorithm 4 EM-algorithm

```
Initialize $\boldsymbol{\theta}_0$
loop over $m \geq 1$
    Approximate $c$
    Create the Gibbs sample $\{\boldsymbol{Y}^{(k)} : k = 1, \ldots, K\}$, using $\boldsymbol{\theta}^{(m)}$ and $\boldsymbol{X}$
    Solve
        $K^{-1} \sum_{k=1}^{K} \frac{\partial \log f}{\partial \theta_i}\left(\boldsymbol{X}, \boldsymbol{Y}^{(k)}|\widehat{\boldsymbol{\theta}}^{(m+1)}\right) = 0, \quad i = 1, \ldots, p$
    with respect to $\widehat{\boldsymbol{\theta}}^{(m+1)}$.
end loop
```

Like we mentioned in the introduction, we have not yet made this procedure work in a satisfactory manner. We therefore only include this chapter in the thesis as an interesting idea for future academic work.

Chapter 9

Concluding Remarks

To finalize this thesis we would like to make some concluding remarks and suggest future research. In many ways, this work is not finished. The thesis is coming to an end, but there are still many open questions and we do not have all the answers. If we did, there would be nothing left to do tomorrow.

STGARCH is a spatial temporal extension of the ordinary GARCH model. It seems to be rather difficult to extend GARCH modelling to pure spatial processes, but by including the time dimension we have demonstrated that such an extension is possible. From a theoretical point of view this extension has some interest apart from applications. Moreover it is connected to MGARCH and in fact the circular version with a fixed spatial region represent a subclass of MGARCH which may have some advantages compared to a general MGARCH model.

An important part of this thesis is finding and analysing computable and reasonable estimators of the parameters in the model. We have considered both maximum likelihood and least squares types of estimators. The STGARCH model is a relatively simple model but nevertheless it is still an open question to obtain the complete likelihood in a computable way. Here we have used a circular approach which slightly modify the model and also is best suited for the situation with a fixed finite spatial region but with an infinite or increasing time horizon.

When the idea of assuming a circular model first came up, it seemed very artificial. We find it amazing how something can seem foolish, but turn out to be so theoretically useful. Without it, we can not even express the likelihood, but with it, the likelihood derivation becomes trivial. We manage to fully utilize the available data and achieve lower variance on our estimates compared to other approaches. We have only considered STGARCH processes, but the circular model has much wider potential. It could be applied to other spatio-temporal processes as well.

As an alternative to the circular approach for the likelihood we have also considered the so-called conditioning on the boundary approach. This is a modified conditional likelihood where the spatial effect of conditioning is neglected in the sense that we still assume a Markov type structure of the actual likelihood approximation method that is used. This method represents a modification of the true likelihood and may be related to composite likelihood methods (Varin et al., 2011). It is always better to test a procedure, when you have a competing one. Conditioning on the boundary got the role of the competitor in this thesis. The Monte Carlo studies in Chapter 6 indicate that it definitely is a viable approach to estimation and should not be rejected. However, theoretically, it might be more challenging than the circular approach. We suspect that when you condition on the entire boundary and $\mathcal{F}_{t-1}$, there will be some information about $\boldsymbol{X}_t$ in the boundary at time $\geq t$. We are uncertain how this will affect the likelihood at this point.

After giving praise, we should be self-critical. In Chapter 5 there are many things left undone. We should have proven that our iteration routine for estimating $\widehat{\boldsymbol{\theta}}_{\text{ML}}$ converge to a solution of the normal equations. We should also show that the solution is truly a maximum of the likelihood function. We did a simulation experiment to see if (5.16) converged to a constant, instead of proving it theoretically. We should also prove that the remainder of the Taylor approximation of the score function is $\mathcal{O}_P(N^{-1})$. These are things we hope to do in the future.

We used Comte and Lieberman (2003) to argue for consistency and asymptotic normality. The conditions for consistency of estimators are not very restrictive, but the 8^{th} order moment required for asymptotic normality is somewhat restrictive since ordinary GARCH only needs finiteness of the 4^{th} order moments. In our future work, we will search for less restrictive conditions for asymptotic normality of our estimators for the STGARCH. Being a less general model, we suspect such an ease of requirements exist. We will work on the sketched proof in Chapter 5 and make it more rigorous.

We presented the idea of using a Gibbs sampler and an EM algorithm for applying the circular model on non-circular variables in Chapter 8. This is a very interesting idea, and should be looked further into. It could potentially be a good alternative to the parametric bootstrap bias correction.

There are many ways to take this work further. Applying the procedures developed in this thesis to real-life data is something we have not done yet. This is important from a practical point of view and will of course be a part of the next step.

Bibliography

Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 192–236, 1974.

Patrick Billingsley. *Probability and measure.* Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, third edition, 1995. ISBN 0-471-00710-2. A Wiley-Interscience Publication.

Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, 31(3):307–327, 1986.

Tim Bollerslev. Glossary to ARCH (GARCH). *CREATES Research Paper*, 49, 2008.

Philippe Bougerol and Nico Picard. Stationarity of GARCH processes and of some nonnegative time series. *J. Econometrics*, 52(1-2):115–127, 1992.

Peter J Brockwell and Richard A Davis. *Introduction to time series and forecasting.* Springer Science & Business Media, 2006.

Michael R. Chernick and Robert A. Labudde. *Introduction to Bootstrap Methods with Applications to R.* Wiley, 2014.

David Clayton and Luisa Bernardinelli. Bayesian methods for mapping disease risk. *Geographical and environmental epidemiology: methods for small-area studies*, pages 205–220, 1992.

Fabienne Comte and Offer Lieberman. Asymptotic theory for multivariate GARCH processes. *Journal of Multivariate Analysis*, 84(1):61–84, 2003.

Noel Cressie and Christopher K Wikle. *Statistics for spatio-temporal data.* John Wiley & Sons, 2011.

Richard A Davis and Thomas Mikosch. Extreme value theory for GARCH processes. In *Handbook of financial time series*, pages 187–200. Springer, 2009.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

Annette J Dobson and Adrian Barnett. *An introduction to generalized linear models*. CRC press, 2008.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7 (1):1–26, 1979.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the bootstrap*. Chapman & Hall/CRC, 1993.

Robert F Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica: Journal of the Econometric Society*, pages 987–1007, 1982.

Robert F Engle and Kenneth F Kroner. Multivariate simultaneous generalized ARCH. *Econometric theory*, 11(01):122–150, 1995.

Christian Francq, Jean-Michel Zakoian, et al. Maximum likelihood estimation of pure GARCH and ARMA-GARCH processes. *Bernoulli*, 10(4):605–637, 2004.

Hans Karlsen. A partial observed model. Unpublished note, 2015a.

Hans Karlsen. SGARCH. Working paper, 17 pages, 2015b.

Kartverket. Gratis Kart og Kartdata. http://kartverket.no/kart/gratis-kartdata/, 2015. [Online; accessed December 2015].

Alexander J McNeil, Rüdiger Frey, and Paul Embrechts. *Quantitative risk management: Concepts, techniques and tools*. Princeton university press, 2005.

Daniel B. Nelson. Stationarity and persistence in the GARCH$(1,1)$ model. *Econometric Theory*, 6(3):318–334, 1990.

Nobelprize.org. The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2003. http://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/2003/, 2003. [Online; accessed 6. March 2016].

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL https://www.R-project.org/.

Maria L Rizzo. *Statistical computing with R*. CRC Press, 2007.

Chris Seidel. Plotting a table of numbers as an image using R. http://www.phaget4.org/R/image_matrix.html, 2016. [Online; accessed March 2016].

Karline Soetaert. *plot3D: Plotting multi-dimensional data.*, 2014. URL http://CRAN.R-project.org/package=plot3D. R package version 1.0-2.

SSB. Folkemengde, 1.januar 2015. https://www.ssb.no/befolkning/statistikker/folkemengde/aar/2015-02-19#content, 2015. [Online; accessed 26. October 2015].

Robert Stelzer. On the relation between the VEC and BEKK multivariate GARCH models. *Econometric Theory*, 24:1131–1136, 8 2008. ISSN 1469-4360. doi: 10.1017/S0266466608080456. URL http://journals.cambridge.org/article_S0266466608080456.

Daniel Straumann, Thomas Mikosch, et al. Quasi-maximum-likelihood estimation in conditionally heteroscedastic time series: a stochastic recurrence equations approach. *The Annals of Statistics*, 34(5):2449–2495, 2006.

Cristiano Varin, Nancy Reid, and David Firth. An overview of composite likelihood methods. *Statistica Sinica*, 21(1):5–42, 2011.

Melanie M Wall. A close look at the spatial structure implied by the CAR and SAR models. *Journal of Statistical Planning and Inference*, 121(2):311–324, 2004.

Andrew A Weiss. Asymptotic theory for ARCH models: estimation and testing. *Econometric theory*, 2(01):107–131, 1986.

Peter Whittle. On stationary processes in the plane. *Biometrika*, pages 434–449, 1954.

Diethelm Wuertz, Yohan Chalabi with contribution from Michal Miklovic, Chris Boudt, Pierre Chausse, and others. *fGarch: Rmetrics - Autoregressive Conditional Heteroskedastic Modelling*, 2013. URL http://CRAN.R-project.org/package=fGarch. R package version 3010.82.

Appendices

Appendix A

Spatial consistency

In this appendix we present the results from an experiment where a small temporal sample size ($T = 10$) was used. We estimated parameters for gradually increasing spatial sample size, increasing each dimension by 10 points for each step. We started with 10×10 and ended up with 90×90. The model we considered was the STARCH(1) model described in the introduction of Chapter 6. We let $\alpha_0 = 40 \cdot 10^{-2}$ and did the experiment twice, once with $\alpha_1 = 10^{-2}$ and once with $\alpha_1 = 10^{-1}$. The estimates are presented in tables A.1 and A.3, while the standard deviations and correlation are presented in tables A.2 and A.4.

We see that for $\alpha_1 = 0.01$, the estimates gradually improve as sample size increase (see Table A.1) and the standard deviations in Table A.2 decrease unambiguously. At the bottom of Table A.2 one can see how extremely correlated the parameters are.

Note that when $\alpha_1 = 0.1$, we are close to a non-stationary process, since the stationarity condition is $\alpha_1 < 1/9 \approx 0.11$. In Table Table A.3 you can really see the effect of being close to stationarity. The α_1 estimates start off below, but approach the correct value as sample size increase. The situation is not the same for α_0. Compared to Table A.1 the mean estimates starts off much worse, but as sample size increase they actually approach the correct value from above. An interesting aspect here, is that the RRML estimates seems almost unaffected by the approach to non-stationarity. We saw the same tendency in the experiments of section 6.3. From Table A.4 we also see declining standard deviations as sample size increase. This experiment also show high negative correlation between parameters. For the LS estimates the correlation seems to be approaching -1, which is not a good thing, while the ML estimates is approximately -0.8.

Spatial	10×10	20×20	30×30	40×40	50×50	60×60	70×70	80×80	90×90
True α_0	40.00	40.00	40.00	40.00	40.00	40.00	40.00	40.00	40.00
RCML α_0	40.89	40.45	40.14	40.08	40.15	40.02	40.08	40.13	40.03
RCLS α_0	40.81	40.54	40.17	40.08	40.16	40.05	40.08	40.13	40.02
RRML α_0	40.49	40.17	39.96	39.98	40.07	39.92	40.01	40.06	39.98
RRLS α_0	40.34	40.29	39.98	39.99	40.07	39.95	40.02	40.06	39.97
True α_1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
RCML α_1	0.80	0.92	0.99	0.99	0.95	1.00	0.97	0.97	0.98
RCLS α_1	0.82	0.90	0.98	1.00	0.95	0.99	0.97	0.97	0.99
RRML α_1	0.89	1.00	1.03	1.02	0.98	1.03	0.99	0.99	1.00
RRLS α_1	0.93	0.97	1.03	1.02	0.98	1.02	0.99	0.99	1.00

Table A.1: *Result from 9 different set ups of 2D* STARCH(1) *with* $\alpha = 0.4$ *and* $\alpha_1 = 0.01$. *Measure is* 10^{-2} *for the parameter estimates and true parameters. Temporal sample size,* T, *is* 10.

Spatial	10×10	20×20	30×30	40×40	50×50	60×60	70×70	80×80	90×90
RCML SD (α_0)	48.85	24.28	16.11	12.09	9.66	8.05	6.89	6.05	5.36
RCLS SD (α_0)	49.42	24.46	16.24	12.19	9.73	8.11	6.94	6.09	5.40
RRML SD (α_0)	60.98	26.98	17.24	12.73	10.07	8.33	7.10	6.20	5.48
RRLS SD (α_0)	62.14	27.22	17.39	12.83	10.14	8.39	7.15	6.25	5.52
RCML SD (α_1)	11.50	5.74	3.82	2.87	2.29	1.91	1.64	1.44	1.27
RCLS SD (α_1)	11.67	5.78	3.85	2.90	2.31	1.93	1.65	1.45	1.28
RRML SD (α_1)	14.42	6.39	4.09	3.02	2.39	1.98	1.69	1.47	1.30
RRLS SD (α_1)	14.76	6.45	4.13	3.05	2.41	2.00	1.70	1.49	1.31
RCML ρ	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90
RCLS ρ	-0.91	-0.91	-0.90	-0.91	-0.91	-0.91	-0.91	-0.91	-0.91
RRML ρ	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90
RRLS ρ	-0.91	-0.91	-0.90	-0.91	-0.91	-0.91	-0.91	-0.91	-0.91

Table A.2: *Standard deviations and correlations coefficient from 2D settings experiment with* STARCH(1) *with* $\alpha_1 = 0.01$. *Measure is* 10^{-3} *for the* SD (α_0) *and* SD (α_1).

Spatial	10 × 10	20 × 20	30 × 30	40 × 40	50 × 50	60 × 60	70 × 70	80 × 80	90 × 90
True α_0	40.00	40.00	40.00	40.00	40.00	40.00	40.00	40.00	40.00
RCML α_0	76.25	54.97	49.49	46.57	45.43	44.60	43.94	42.92	42.81
RCLS α_0	99.82	62.72	52.79	50.93	49.14	46.86	45.01	44.42	43.65
RRML α_0	44.45	41.27	41.68	40.87	40.56	40.28	40.24	39.91	40.15
RRLS α_0	78.68	51.11	46.42	45.94	45.29	43.96	42.20	41.73	41.41
True α_1	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00
RCML α_1	8.95	9.57	9.70	9.85	9.83	9.87	9.89	9.93	9.93
RCLS α_1	8.19	9.32	9.60	9.72	9.71	9.80	9.86	9.89	9.90
RRML α_1	9.83	9.95	9.92	10.00	9.96	9.99	10.00	10.02	10.00
RRLS α_1	8.81	9.63	9.78	9.85	9.82	9.88	9.94	9.96	9.96

Table A.3: *Result from 9 different set ups of 2D* STARCH(1) *with* $\alpha = 0.4$ *and* $\alpha_1 = 0.1$. *Measure is* 10^{-2} *for the parameter estimates and true parameters. Temporal sample size,* T, *is* 10

Spatial	10 × 10	20 × 20	30 × 30	40 × 40	50 × 50	60 × 60	70 × 70	80 × 80	90 × 90
RCML SD (α_0)	241.86	102.63	65.58	48.70	38.02	31.48	26.96	23.43	20.80
RCLS SD (α_0)	518.57	346.88	242.74	213.21	195.26	151.57	134.97	124.78	118.52
RRML SD (α_0)	266.14	105.72	67.06	49.48	38.41	31.72	27.13	23.57	20.91
RRLS SD (α_0)	662.13	387.84	262.52	225.00	204.72	156.71	139.81	129.07	121.84
RCML SD (α_1)	9.99	4.70	3.11	2.31	1.84	1.53	1.31	1.15	1.02
RCLS SD (α_1)	16.98	10.76	7.82	6.57	6.02	4.71	4.16	3.85	3.64
RRML SD (α_1)	12.07	5.10	3.28	2.41	1.90	1.57	1.34	1.17	1.04
RRLS SD (α_1)	21.97	12.13	8.48	6.95	6.32	4.88	4.31	3.98	3.75
RCML ρ	-0.82	-0.81	-0.80	-0.80	-0.80	-0.80	-0.80	-0.80	-0.80
RCLS ρ	-0.93	-0.95	-0.96	-0.97	-0.98	-0.98	-0.98	-0.98	-0.98
RRML ρ	-0.81	-0.80	-0.80	-0.80	-0.80	-0.80	-0.80	-0.80	-0.80
RRLS ρ	-0.92	-0.95	-0.96	-0.97	-0.98	-0.98	-0.98	-0.98	-0.98

Table A.4: *Standard deviations and correlations coefficient from 2D settings experiment with* STARCH(1) *with* $\alpha_1 = 0.1$. *Measure is* 10^{-3} *for the* SD (α_0) *and* SD (α_1)

Appendix B

Covariance matrix experiments

In this Appendix we will do experiments related to the covariance matrix of the ML estimators. First we will compare the formula estimate in Equation (5.14) to Monte Carlo estimates of the covariance matrix of the maximum likelihood estimators in Appendix B.1 and secondly we will test the convergence of Equation (5.16) for a STGARCH(1,1) model in Appendix B.2.

B.1. Testing of (5.14)

The approximation formula for the covariance matrix of $\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}$ is given by Equation (5.14). Let

$$h(\widehat{\boldsymbol{\theta}}, \boldsymbol{X}) = \widehat{\mathrm{Var}}(\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}) = \left(\sum_{t=1}^{N} \widetilde{\mathbb{X}}_t^T \widetilde{\mathbb{W}}_t \widetilde{\mathbb{X}}_t \right)^{-1} \tag{B.1}$$

For the STARCH process presented at the beginning of Chapter 6, the routine for checking the formula becomes the following.

```
For $k = 1, \dots, K\{$

    ⋆ Simulate $\{\boldsymbol{X}_k\}$

    ⋆ Estimate $\widehat{\boldsymbol{\theta}}_k$

    ⋆ Calculate $h(\widehat{\boldsymbol{\theta}}_k, \boldsymbol{X}_k)$ and $h(\boldsymbol{\theta}_k, \boldsymbol{X}_k)$

$\}$
Calculate $\frac{1}{K}\sum_k h(\widehat{\theta}_k, X)$ and $\frac{1}{K}\sum_k h(\theta_k, X)$
Estimate Cov $(\theta_0, \theta_1)$ based on $\{\widehat{\theta}_{0,k}, \widehat{\theta}_{1,k} : k = 1, \dots, K\}$
```

We ran this routine using dimensions 5 × 5 × 1000, 10 × 10 × 1000 and 20 × 20 × 1000. The results are presented in the tables below.

Taking the largest dimensioned dataset, we can find the correlation matrix of $\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}$, by $\operatorname{diag}\{\boldsymbol{\Sigma}\}^{-1/2}\,\boldsymbol{\Sigma}\operatorname{diag}\{\boldsymbol{\Sigma}\}^{-1/2}$, where $\operatorname{diag}\{\boldsymbol{\Sigma}\}$ is the diagonal matrix consisting of the diagonal elements of $\boldsymbol{\Sigma}$. Since the covariance matrices are almost the same, we use the Monte Carlo estimate.

$$\operatorname{diag}\{\boldsymbol{\Sigma}\}^{-1/2}\,\boldsymbol{\Sigma}\operatorname{diag}\{\boldsymbol{\Sigma}\}^{-1/2} = \begin{bmatrix} 1 & -0.896 \\ -0.896 & 1 \end{bmatrix} \tag{B.2}$$

	Formula $\boldsymbol{\alpha}$	Formula $\widehat{\boldsymbol{\alpha}}$	$\widehat{\boldsymbol{\alpha}}$
Var $(\widehat{\alpha}_0)$	107.20	107.12	105.50
Cov $(\widehat{\alpha}_1, \widehat{\alpha}_0)$	-65.79	-65.70	-64.83
Cov $(\widehat{\alpha}_0, \widehat{\alpha}_1)$	-65.79	-65.70	-64.83
Var $(\widehat{\alpha}_1)$	50.02	49.92	49.91

Table B.1: *Comparing Monte Carlo estimated covariance matrix and true covariance matrix calculated using formula (5.2). Data dimension was* $5 \times 5 \times 1000$. *Measure is* 10^{-7}.

	$h(\boldsymbol{\alpha}, \boldsymbol{X})$	$h(\widehat{\boldsymbol{\alpha}}, \boldsymbol{X})$	$\widehat{\boldsymbol{\alpha}}$
Var $(\widehat{\alpha}_0)$	26.79	26.79	23.47
Cov $(\widehat{\alpha}_1, \widehat{\alpha}_0)$	-16.42	-16.42	-14.58
Cov $(\widehat{\alpha}_0, \widehat{\alpha}_1)$	-16.42	-16.42	-14.58
Var $(\widehat{\alpha}_1)$	12.48	12.48	11.42

Table B.2: *Comparing Monte Carlo estimated covariance matrix and true covariance matrix calculated using formula (5.2). Data dimension was* $10 \times 10 \times 1000$. *Measure is* 10^{-7}.

	Formula $\boldsymbol{\alpha}$	Formula $\widehat{\boldsymbol{\alpha}}$	$\widehat{\boldsymbol{\alpha}}$
Var $(\widehat{\alpha}_0)$	6.70	6.69	6.56
Cov $(\widehat{\alpha}_1, \widehat{\alpha}_0)$	-4.11	-4.11	-3.93
Cov $(\widehat{\alpha}_0, \widehat{\alpha}_1)$	-4.11	-4.11	-3.93
Var $(\widehat{\alpha}_1)$	3.12	3.12	2.93

Table B.3: *Comparing Monte Carlo estimated covariance matrix and true covariance matrix calculated using formula (5.2). Data dimension was* $20 \times 20 \times 1000$. *Measure is* 10^{-7}.

The Tables B.1, B.2 and B.3 tells us two things. The first thing is that Equation (5.14) is a good approximation, because the results are similar. Remember that the numbers here are of order 10^{-7}, so the differences are very small. The second thing is that the variances decreases as the sample size increase. This is the consistency property of the estimators. Another important thing is that the empirical correlation between $\widehat{\alpha}_0$ and $\widehat{\alpha}_1$ is $\widehat{\rho} \approx -0.9$, which is a strong negative correlation. This can be explained by 3.6, which empirically in this situation means that

$$\widehat{\sigma}_X^2 = \frac{\widehat{\alpha}_0}{1 - 9\widehat{\alpha}_1}. \tag{B.3}$$

For fixed $\widehat{\sigma}_X^2$, an increase in $\widehat{\alpha}_0$ must lead to a decrease in $\widehat{\alpha}_1$ to keep the fraction constant.

B.2. Convergence of (5.16)

We simulated a STGARCH(1,1) of size 10 × 10 × 100.000, circularly. We then used the true parameter and estimated $\widetilde{\mathbb{X}}_t$ and $\widetilde{\mathbb{W}}_t$ for $t = 1, \ldots, T$, for various values of T. Then calculated

$$\widetilde{\mathcal{I}} = \frac{1}{10^2 \cdot T} \sum_{t=1}^{T} \widetilde{\mathbb{X}}_t^T \widetilde{\mathbb{W}}_t \widetilde{\mathbb{X}}_t. \tag{B.4}$$

This formula can be recognized as the inverse of (5.14). Due to $\mathcal{I}$ being symmetric, we have that $\mathcal{I}_{ij} = \mathcal{I}_{ji}$ for $1 \leq i, j \leq 3$. The resulting matrices are presented in Table B.4. The column names of Table B.4 reflect the following positions of the matrix

$$\mathcal{I} = \begin{bmatrix} \mathcal{I}_{11} & \mathcal{I}_{12} & \mathcal{I}_{13} \\ \mathcal{I}_{21} & \mathcal{I}_{22} & \mathcal{I}_{23} \\ \mathcal{I}_{31} & \mathcal{I}_{32} & \mathcal{I}_{33} \end{bmatrix}. \tag{B.5}$$

We see from Table B.4 that the matrix converge as T grows larger, but the convergence is not extremely fast. For this realization, the matrix converge to a constant matrix $\mathcal{I}$ as was claimed in (5.16).

	$\mathcal{I}_{11}$	$\mathcal{I}_{21}$	$\mathcal{I}_{31}$	$\mathcal{I}_{12}$	$\mathcal{I}_{22}$	$\mathcal{I}_{32}$	$\mathcal{I}_{13}$	$\mathcal{I}_{23}$	$\mathcal{I}_{33}$
T=10	16.89	55.37	39.97	55.37	223.82	134.33	39.97	134.33	132.02
T=20	14.97	47.89	43.25	47.89	188.07	139.30	43.25	139.30	149.27
T=40	13.63	45.09	44.09	45.09	184.58	147.30	44.09	147.30	157.20
T=60	13.22	44.13	44.43	44.13	182.03	149.62	44.43	149.62	159.74
T=80	13.04	43.52	44.66	43.52	179.91	150.27	44.66	150.27	161.17
T=100	12.96	43.17	44.84	43.17	177.75	150.43	44.84	150.43	161.95
T=200	12.72	42.61	45.04	42.61	175.84	151.84	45.04	151.84	163.40
T=300	12.63	42.44	45.11	42.44	175.32	152.34	45.11	152.34	163.93
T=400	12.59	42.32	45.14	42.32	175.01	152.54	45.14	152.54	164.21
T=500	12.57	42.24	45.18	42.24	174.71	152.58	45.18	152.58	164.38
T=1000	12.54	42.09	45.25	42.09	173.90	152.62	45.25	152.62	164.69
T=5000	12.49	42.00	45.27	42.00	173.99	152.92	45.27	152.92	164.96
T=10000	12.49	41.99	45.28	41.99	173.95	152.95	45.28	152.95	164.99
T=49999	12.49	41.98	45.28	41.98	173.82	152.95	45.28	152.95	165.01
T=99999	12.49	41.98	45.29	41.98	173.76	152.92	45.29	152.92	165.01

Table B.4: *Resulting information matrices for a* $10 \times 10 \times T$ STGARCH$(1,1)$ *for increasing* T*.*

Appendix C

R-Code

In this appendix, we will present some R-code used in this thesis. There are many reasons for presenting this. One motivation is to show the reader an example of how one can implement the algorithms presented in this thesis. We do not claim the code presented in this appendix optimal and the reader is welcome to suggest other ways of implementation.

C.1. Simulation algorithm

Note: The R-code written for the simulation algorithms was created before we came up with the convenient vector notation of section 4.1. Therefore the routine is formulated somewhat different from Algorithm 1, but they do the same thing.

1D STARCH

```
#Simulation in 1D STARCH
#Explanation of arguments:
# param is parameter vector of length 2, alpha0 and alpha1
# dim is sample size vector of length 2, spatial and temporal size
# burnin is numeric giving the length of the temporal burnin
# cutoff is numeric determining how much of the circular simulation
# should be truncated in each direction to form a non-circular dataset.
# If cutoff=0, the simulation is circular.
simulation_1d_arch<-function(param=NULL, dim=NULL, burnin=100, cutoff=0){

  #Check for correct input:
  if(length(dim)!=2) stop("Incorrect dimension specification")
  if(length(param)!=2) stop("Incorrect parameter specification")
  #Creating constants from vectors to ease notation
  a0<-param[1]
  a1<-param[2]
  X<-dim[1]+2*cutoff
  n<-dim[2]+burnin

  #Pre-allocating:
  data<-array(NA_real_,dim=c(X,n))
  data[,1]=rnorm(X,0,1)
  S<-array(NA_real_, dim=c(X,n))
  S[,1]<-abs(rnorm(X,0,1))
  # Circular simulation routine
    for(t in 2:n){
      a.tmp<-data[,t-1]^2
      a.tmp<-a.tmp[c(X,1:(X-1))]+a.tmp[1:X]+a.tmp[c(2:X,1)]
      S[,t]<-sqrt(a0+a1*a.tmp)
      data[,t]<-S[,t]*rnorm(X,0,1)
    }
  #Returning the simulated dataset with cutoff and burn-in removed
  return(data[(1+cutoff):(X-cutoff),(burnin+1):n])
}
```

2D STARCH

```
#Simulation of 2D STARCH-process
#Explanation of arguments:
# param is parameter vecotr of length 2, alpha0 and alpha1
# dim is sample size vector of length 3, spatial and temporal size
# burnin is numeric giving the length of the temporal burnin
# cutoff is numeric determining how much of the circular simulation
# should be truncated in each direction to form a non-circular dataset.
# If cutoff=0, the simulation is circular.

simulation_2d_arch<-function(param=NULL, dim=NULL, burnin=100, cutoff=0){
  #Check for correct input:
```

```
  if(length(dim)!=3) stop("Incorrect dimension specification")
  if(length(param)!=2) stop("Incorrect parameter specification")

  #Creating constants from vectors to ease notation
  a0<-param[1]; a1<-param[2];
  X<-dim[1]+2*cutoff; Y<-dim[2]+2*cutoff;n<-dim[3]+burnin;

  #Pre-allocating:
  data<-array(NA_real_,dim=c(X,Y,n))
  data[,,1]=rnorm(X*Y,0,1)
  S<-array(NA_real_, dim=c(X,Y,n))
  S[,,1]<-rnorm(X*Y,0,1)^2

  # Circular simulation routine
    for(t in 2:n){
      a.tmp<-data[,,t-1]^2
      a.tmp<-a.tmp[c(X,1:(X-1)),]+a.tmp[1:X,]+a.tmp[c(2:X,1),]
      a.tmp<-a.tmp[,c(Y,1:(Y-1))]+a.tmp[,1:Y]+a.tmp[,c(2:Y,1)]
      S[,,t]<-sqrt(a0+a1*a.tmp)
      data[,,t]<-S[,,t]*matrix(rnorm(X*Y,0,1),nrow=X, ncol=Y)

    }
  #Returning the simulated dataset with cutoff and burn-in removed
  return(data[(1+cutoff):(X-cutoff),(1+cutoff):(Y-cutoff),(burnin+1):n])
}
```

C.2. Estimation

Here are the algorithms presented in Chapter 5.

Least Squares estimation

The following function is dimension independent. It takes in the vectorized process and the design matrix, where the neighbourhood structure has been implemented and this matrix is specified according to the dimension. We here have no initial value in need of specification, because this routine is written for the STARCH process considered in Chapter 6.

```
LS.estimation<-function(X, mat, var_mat=FALSE){
  #X is the data, arranged as a vector
  #mat is the design array with dimension (M,2,N-1).
  #var_mat is a logic vector used for decided wether or not to return the
  #covariance matrix of the estimators.
  Y<-X^2
  d<-dim(X);N<-d[2];M<-d[1];
  #Creates a temporary array in order to use the apply function
  tmp.mat<-array(NA_real_, dim=c(M,3,N-1))
  tmp.mat[,1:2,]<-mat
  tmp.mat[,3,]<-Y[,2:N]
  v<-apply(apply(tmp.mat, 3, function(x) t(x[,1:2])%*%x[,3]),1,sum)
  design<-matrix(apply(apply(mat,3,function(x)c(t(x)%*%x)),1,sum), ncol=2, nrow=2)
  #Calculate theta
  theta<-solve(design,v)
  if(!var_mat)
    return(theta)
  else{
  #Covariance matrix calculations
    var<-2*solve(design)%*%matrix(apply(apply(mat, 3,function(x) {
      return(c(t(x)%*%diag(c((x%*%theta)^2))%*%x))}),1,sum),
      ncol=length(theta), nrow=length(theta))%*%solve(design)
    return(c(theta,var))
  }
}
```

Maximum likelihood estimation

The maximum likelihood estimation routine is also dimension independent. It lets the design matrix, *mat*, deal with the neighbourhood structure and hence the spatial dimension is included here. Typically we have set the initial parameter vector equal to the least squares estimate from the routine above, since the least squares estimation routine for a STARCH process does not require initial values.

```
#ML estimation of STARCH
max_lik<-function(init, X,mat, n.iterations=20, return_all=FALSE, var_mat=FALSE){
```

```
# init is initial parameter vector
# X vectorized data
# mat design matrix
# n.iterations is number of iterations
# return_all decides if all iterations should be returned
# var_mat decides if covariance matrix should be calculated and returned

# Initialize theta matrix
theta<-matrix(NA_real_, ncol=2, nrow=n.iterations+1)
#Intial value
theta[1,]<-init
d<-dim(X)
w<-array(NA_real_, dim=dim(mat))
#Iteration process
for(i in 1:n.iterations){
  #Calculating weights
  w<-apply(mat,3,function(x) x%*%theta[i,])
  w<-1/(2*w^2)
  sum1<-matrix(0, ncol=2, nrow=2)
  sum2<-numeric(length=2)
  for(t in 2:d[2]){
    W<-diag(w[,t-1])
    sum1<-sum1+t(mat[,,t-1])%*%W%*%mat[,,t-1]
    sum2<-sum2+t(mat[,,t-1])%*%W%*%X[,t]^2

  }
  #Calculating next theta
  theta[i+1,]<-solve(sum1,sum2)
}
#Covariance matrix
if(var_mat){
  retur_list<-list()
  if(return_all)
    retur_list$theta<-theta # returning all iterations
    else
      retur_list$theta<-theta[n.iterations+1,] # returning only last iteration
    retur_list$var<-solve(sum1) # returning covariance matrix
    return(retur_list)
  }else{
    if(return_all)
      return(theta) # returning all iterations
    else
      return(theta[n.iterations+1,]) # returning only last iteration
  }
}
```

C.3. Parametric Bootstrap Bias Correction

We also present some code for the parameteric bootstrap bias correction. The function *est* simply uses the least squares function to get an initial estimate and then the maximum likelihood estimation function to refine the estimates.

```
bootstrap.est<-function(true_param, M=100, burnin=200,
                        cutoff=100, dim=c(30,30,200),
                        ncores=32, n.iterations=30){
    #Simulating "original" dataset
    data<- simulation_2d_arch(param=true_param, dim=dim,
     burnin=burnin, cutoff = cutoff)
    param.est<-est(data, n.iterations)
    registerDoParallel(cores=ncores)
    value<-foreach(j = 1:M) %dopar%{
    # Simulating data
      data.boot<- simulation_2d_arch(param=param.est, dim=dim,
       burnin=burnin, cutoff = cutoff)
    # Estimation by maximum likelihood
      boot.est<-est(data.boot,n.iterations)
      names(boot.est)<-c("param1", "param2")
      boot.est
    }
    theta<-t(sapply(value, function(x) return(x)))
    theta_mean<-apply(theta,2,mean) # calculating mean of bootstrap estimates
    bias_corrected<-2*param.est-theta_mean # bias corrected estimate
  result<-c(param.est, bias_corrected) # vector to be returned
  names(result)<-c("est.0", "est.1","bias.corr.0","bias.corr.1")
  return(result)
}
```