Doctoral Thesis

Statistical methods for high-multivariate financial time series

Author(s):
Audrino, Francesco

Publication Date:
2002

Permanent Link:
https://doi.org/10.3929/ethz-a-004324184

Rights / License:
In Copyright - Non-Commercial Use Permitted
Statistical Methods for High-Multivariate Financial Time Series

A dissertation submitted to the
SWISS FEDERAL INSTITUTE OF TECHNOLOGY
ZURICH

for the degree of
Doctor of Mathematics

presented by
FRANCESCO AUDRINO
Dipl. Math. ETH
born April 28, 1975
citizen of Chiasso TI

accepted on the recommendation of
Prof. Dr. P. Bühlmann, examiner
Prof. Dr. P. Embrechts, co-examiner

2002
Acknowledgments

First of all, it is a great pleasure to thank my supervisor P. Bühlmann for his guidance throughout my Ph.D. studies. This piece of work would not have been possible without his advice and help. He has always been patient and has always found time to discuss problems and ideas with me.

I have profited from fruitful discussions with many people at several meetings and seminars. In particular, very special thanks go to M. Dacorogna for his helpful and motivating discussions. Furthermore, I thank P. Embrechts for being co-examiner.

Thanks go to my friends Enrico Degiorgi, Nicola Kistler, Giacomo Catenazzi, Tommaso Cereghetti, Fausta Leonardi, Carlo Rossi, Andrea Binda, Alex Juri and all my colleagues from ETH for their friendship and for the beautiful time spent outside the ETH.

Especially, I would like to thank Tatiana and my family for their love and support over the last three years.

Financial support from Olsen & Associated is gratefully acknowledged.
Seite Leer / Blank leaf
Abstract

This thesis brings together some new univariate and multivariate models and techniques to estimate volatility and risk measures for financial instruments as for example real, global defined portfolios.

In the first part, we concentrate on the univariate analysis. We propose two different volatility estimation approaches: nonparametric ARCH(1) modeling within the local likelihood framework and a new class of GARCH-type models with tree-structured multiple thresholds. The second approach relies on the idea of a binary tree where every terminal node parameterizes a (local) GARCH model for a partition cell of the predictor space. We derive supporting asymptotic results and we investigate on simulated and real data the better predictive potential and the better performance of the volatility estimates in comparison with the ones from a classical analysis (for example local regression and GARCH(1,1)).

The second part of this thesis deals with the main problem of finding computationally feasible and well performing strategies to estimate volatility for (very) high-dimensional, asynchronous multivariate financial time series. We propose a synchronization technique which takes into account the fact that information continues to flow for closed markets while others are still open. We also propose a functional gradient descent algorithm (FGD), a recent technique from the area of machine learning. Our FGD algorithm is computationally feasible in multivariate problems with dozens up to thousands of individual return series. These new methods potentially increase the predictive performance of any reasonable model such as for example constant conditional correlation GARCH-type models. Since multivariate analysis is generally important for analyzing time-changing portfolios and for better portfolio predictions (even when the portfolio weights are time-constant), synchronization and FGD are valuable techniques for a variety of problems with multivariate financial data.
Seite Leer / Blank leaf
Riassunto

In questa tesi sono riuniti nuovi modelli uni- e multi-dimensionalì, e nuove tecniche per stimare la volatilità e le misure di rischio di strumenti finanziari come per esempio portafogli globalmente definiti.

Nella prima parte della tesi mi focalizzo su modelli univariati. Due metodi differenti per stimare la volatilità sono proposti: modelli ARCH(1) non parametrici costruiti all’interno della struttura di likelihood locale, e una nuova classe di modelli GARCH con soglie multiple strutturate ad albero. Questa seconda tecnica si basa sull’idea di un albero binario, dove ogni nodo parametrizza un modello GARCH (locale) in una specifica cella dello spazio. Alcuni risultati asintotici sono dimostrati e l’attrattiva della stimata volatilità è testata su dati simulati e reali e confrontata con quella ottenuta usando metodi classici (per esempio regressione locale e GARCH(1,1)).

La seconda parte di questa tesi tratta il problema principale: trovare strategie eseguibili e con buone attrattive per stimare la volatilità di serie finanziarie non sincronizzate multi-dimensionalì per dimensioni (molto) grandi. Una tecnica di sincronizzazione che tenga conto del fatto che l’informazione continua a confluire anche in mercati chiusi mentre altri sono ancora aperti è proposta. Inoltre anche un algoritmo basato su “functional gradient descent” (FGD), una tecnica recente sviluppatisi nel campo del machine learning viene presentata. Questo FGD algoritmo è eseguibile per problemi multi-dimensionalì con dozzine fino a migliaia di serie finanziarie. Questi nuovi metodi aumentano di molto il potenziale di ogni ragionevole modello, come per esempio i modelli di tipo CCC-GARCH. Visto che un’analisi multivariata è generalmente importante per analizzare portafogli che cambiano con il passare del tempo, sincronizzazione e FGD sono due tecniche sfruttabili in una notevole quantità di problemi differenti che utilizzano dati finanziari multi-dimensionalì.
Seite Leer / Blank leaf
Contents

Acknowledgments ii
Abstract iv
Riassunto v
Introduction 1

I Univariate methods 5

1 Local Likelihood for non-parametric ARCH(1) models 7
  1.1 Abstract 7
  1.2 Introduction 7
  1.3 The local constant log-likelihood estimation 10
    1.3.1 Local likelihood estimator with Gaussian or scaled $t_0$-distributed innovations 10
    1.3.2 The convexity of $\rho(y, g_x) = -\log \left( f_{U_i} (y - g_x) \right)$ 13
  1.4 Consistency of the local likelihood estimator 15
    1.4.1 Consistency of $\hat{g}_{x:n}^N$ and $\hat{f}_{x:n}^N$ 16
    1.4.2 Consistency of $\hat{g}_{x:n}^N$ and $\hat{f}_{x:n}^N$ under pseudo-likelihood estimation 19

ix
# Contents

1.4.3 Consistency of $\hat{g}_{\lambda,n}^{1/\lambda}$ .................................. 20

1.5 Asymptotic normality of the local likelihood estimator .... 22
  1.5.1 The asymptotic variance of $\hat{g}_{\lambda,n}^{N}$ ............... 26
  1.5.2 The asymptotic variance of $\hat{g}_{\lambda,n}^{1/\lambda}$ ........... 27
  1.5.3 Asymptotic variance of the local regression estimator . 29

1.6 Additional aspects ......................................................... 32
  1.6.1 Uniform convergence of $\hat{g}_{\lambda,n}^{N}$ ............... 32
  1.6.2 The estimation of $\beta$ ........................................... 33

1.7 Some numerical results .................................................... 35
  1.7.1 Simulations ....................................................... 36
  1.7.2 Three real data examples ........................................ 37

1.8 Multiplicative ARCH models with $p > 1$ predictor variables . 38
  1.8.1 The normal case ................................................. 39
  1.8.2 Numerical results ................................................. 40

1.9 Concluding remarks ....................................................... 40

2 Tree-Structured GARCH Models ............................................ 43
  2.1 Abstract .............................................................. 43
  2.2 Introduction .......................................................... 43
  2.3 Tree-structured GARCH estimation ................................. 46
    2.3.1 Forward entering of thresholds: growing the binary tree 48
    2.3.2 Pruning the tree .............................................. 50
  2.4 Consistency .......................................................... 51
    2.4.1 Assumptions and proof of Theorem 2.1 ....................... 52
  2.5 Estimation with generalized additive models (GAM) ............ 55
  2.6 Numerical results ..................................................... 56
    2.6.1 Simulations ..................................................... 56
    2.6.2 Two real data examples ....................................... 64
    2.6.3 Summarizing numerical results ............................... 67
2.6.4 How appropriate is GARCH(1,1) for daily returns? 68
2.7 Concluding remarks 68

II Multivariate volatility estimation

3 Synchronizing Multivariate Financial Time Series 73
3.1 Abstract 73
3.2 Introduction 73
3.3 The synchronous CCC-GARCH(1,1) model 75
  3.3.1 Synchronization of the data 75
  3.3.2 The model 78
  3.3.3 Estimating the model 80
3.4 Numerical results 81
  3.4.1 Simple estimate of A and synchronization 81
  3.4.2 Estimates for the synchronous CCC-GARCH(1,1) and its performance 82
  3.4.3 Estimating the performance at the portfolio level 87
3.5 More sophisticated asynchronous models as another source for improvements? 93
  3.5.1 Flexible thresholding for the asynchronized returns 93
  3.5.2 Numerical results 96
3.6 Estimating risk measures 98
  3.6.1 The estimates 99
  3.6.2 Backtesting 100
3.7 Conclusions 100

4 Volatility Estimation with Functional Gradient Descent for Very High-Dimensional Financial Time Series 105
4.1 Abstract 105
4.2 Introduction 105
## Contents

4.3 The generic functional gradient descent algorithm ............... 106  
   4.3.1 Choice of the base learner ..................................... 109  
4.4 Univariate volatility estimation with FGD ...................... 110  
   4.4.1 Numerical results for simulated data ....................... 114  
   4.4.2 Some real example ............................................ 117  
4.5 Volatility estimation for high multivariate time series .................. 119  
   4.5.1 Base learners with variable selection .................... 124  
   4.5.2 Computational cost and parallelization ................. 125  
   4.5.3 Numerical results ............................................ 125  
4.6 Conclusions ....................................................... 130

Bibliography .......................................................... 131

List of figures ....................................................... 134

List of tables ....................................................... 136
Introduction

Returns of asset prices in a portfolio generate a multivariate time series, often in dozens or hundreds of dimensions. Denote by $P_{t,i}$ the price of asset $i$ at time $t$ and its returns by $X_{t,i} = (P_{t,i} - P_{t-1,i})/P_{t-1,i}$. The multivariate time series $\{X_{t,i}; t = 1, \ldots, n, i = 1, \ldots, d\}$ is usually assumed to be stationary.

On the one side, the modeling and prediction effort is on the multivariate (squared) volatility

$$V_t = \text{Cov}_{d \times d}(X_t|\mathcal{F}_{t-1}), \quad X_t = (X_{t,1}, \ldots, X_{t,d})^T,$$

where $\mathcal{F}_{t-1}$ denotes the information up to time $t - 1$, the $\sigma$-algebra generated by $\{X_s; s \leq t - 1\}$. Besides the genuine interest of volatility in finance, $V_t$ is a key quantity because the following model often yields a reasonable approximation,

$$X_t = \Sigma_t Z_t,$$

where $\Sigma_t \Sigma_t^T = V_t$ and $Z_t$ are i.i.d. multivariate innovations with uncorrelated components and componentwise variances equal to one. Due to the enormous complexity of $V_t$ as a function of the past $\mathcal{F}_{t-1}$ for $d$ in the hundreds, the problem of predicting such high-dimensional $V_t$ received only little attention, particularly when (nonlinear) methods are used which take cross-dependencies between the time series and auto-dependences from past values into account. Predicting truly high-dimensional volatility in (0.1) raises huge challenges in computational and modeling issues due to the well known curse of dimensionality. Previous work on multivariate volatility models has been given by Bollerslev (1990), Engle et al. (1990), Lin (1992) and Engle and Kroner (1995) in the framework of GARCH-type models, and by Harvey et al. (1994), Aguilar and West (2000) and Chib et al. (1999) within the stochastic volatility (state space model) framework. In the GARCH-type framework, only very simple models (Bollerslev, 1990) are feasible in high dimensions,
whereas with stochastic volatility models, only Chib et al. (1999) present an example with dimensionality as large as 40 which is still far lower than the degree of multivariateness we can deal with here and which often occurs in practice.

It is well known that if the main quantity of interest is a portfolio price

$$P_t = \sum_{i=1}^{d} \alpha_i P_{t,i}$$

with portfolio weights $$\{\alpha_i; i = 1, \ldots, d\}$$, rather than the behavior of the individual assets therein, a first look may suggest that the high-dimensional problem can be bypassed to a large extent by just looking at the univariate portfolio prices $$\{P_t; t = 1, \ldots, n\}$$ for predicting volatility of the portfolio returns. Proceeding in this way, a substantial information loss has typically to be paid (by using the portfolio prices rather than all the individual prices) resulting in less accurate predictions of volatility for portfolio returns. But more striking is the fact that for time-changing portfolio weights $$\alpha_i = \alpha_{t,i}$$ – which is most often the case in practice – the portfolio returns become typically non-stationary. We then have to model the highly multivariate time series of asset returns to obtain volatility predictions of time-changing portfolios. Furthermore, the problem of optimal allocation in a portfolio requires the multivariate study of all the asset returns therein.

On the other side, the time of measurement of daily financial data, typically the closing time, is often different in markets which do not have the same trading hours. For example, between US and Japan there are no common open hours and between US and Europe there is only partial overlap. Therefore, the value of real global portfolios constructed on daily data across different markets is never known at a fixed point in time and consequently the calculation of risk measures such as the Value at Risk (quantile of the Profit-and-Loss distribution of a given portfolio over a prescribed holding period) and the conditional Value at Risk or expected shortfall (the expected loss given that the loss exceeds VaR) could give misleading results. A consequence of using such asynchronous data is the fact that correlations across the assets are often small, see Burns et al. (1998). This may lead to inaccurate (estimated) risk calculations such as VaR or expected shortfall.

The aim of this thesis is threefold: to present some new univariate models for volatility estimation which can be extended in the multivariate framework, to propose a synchronization technique for real global portfolios and to solve with a new method, already known in the area of machine learning, the problem of volatility estimation for very high-dimensional financial time series.
All the models and methods proposed are tested on simulated and real data and their performance with respect to different measures is confronted with the one from the classical GARCH(1,1) model (Bollerslev, 1986) and from the standard constant conditional correlation (CCC) model (Bollerslev, 1990) in the univariate and in the multivariate setting, respectively. The resulting gains over these classical models are always relevant and sometimes considerable, depending on how the performance is measured.

Some consistency and asymptotic results for the models proposed to estimate the volatility at the univariate level are also proved in this thesis.

**Outline.** This thesis is divided into two parts and consists of three articles developed during my Ph. D. research over the last three years in collaboration with my advisor Prof. P. Bühlmann. Part 1 consists of the description and testing on simulated and real data of two new models for estimating the volatility of financial univariate return time series. Consistency and asymptotic results for these models are given. Part 2 is the main part of this thesis: methods to synchronize the data and to estimate the volatility also for very high-dimensional financial problems are presented and tested on simulated and real data examples.

In Chapter 1, we give consistency and asymptotic results for a local log-likelihood model used to solve general non-parametric ARCH(1) problems. The performance of the resulting estimates for the volatility is tested and confronted with the one from a standard local regression.

Chapter 2 is the article *Tree-Structured GARCH Models* published by the Journal of the Royal Statistical Society, Series B, 63, No. 4, 727-744. The modeling technique proposed is parametric and potentially high-dimensional: it relies on estimating thresholds by using the idea of binary tree construction for partitioning a predictor space. This model is essentially different from CART (Breiman et al., 1984). The article here present a consistency result and demonstrate the new procedure on simulated and real data.

Chapter 3 is the working paper *Synchronizing Multivariate Financial Time Series*, Research Report 97 by the Seminar für Statistik at the ETH Zürich. This paper proposes a synchronization of daily data in real global portfolios. Proceeding as in Burns et al. (1998), our general approach recognizes that even when markets are closed, the asset values may change before the market reopens. The performance of this new method is compared with the one from the classical CCC-GARCH(1,1) model introduced by Bollerslev (1990) on asynchronized data. The influence of synchronization on the calculation of risk measures is discussed. The comparison also includes an univariate approach for the portfolio returns. We collect strong empirical evidence of
the power of synchronization in a first step and that our multivariate modelling technique is superior over univariate models for the portfolio returns.

Chapter 4 is the working paper *Volatility Estimation with Functional Gradient Descent for Very High-Dimensional Financial Time Series*, Research Report 99 by the Seminar für Statistik at the ETH Zürich. This report proposes a version of functional gradient descent (FGD) (Breiman, 1999; Mason et al., 1999; Friedman et al., 2000; Friedman, 2001) adapted to become successful in the field of high-dimensional financial time series. The method produces accurate estimates and is still feasible for very high-dimensional problems (where there is virtually no other competitive alternative method).
Part I

Univariate methods
Seite Leer / Blank leaf
Chapter 1

Local Likelihood for non-parametric ARCH(1) models

1.1 Abstract

We propose a local likelihood estimation for the log-transformed ARCH(1) model in the financial field. Our nonparametric estimator is constructed within the likelihood framework for non-Gaussian observations: it is different from standard kernel regression smoothing, where the innovations are assumed to be normally distributed. We derive consistency and asymptotic normality for our estimators and conclude from simulation and real data analysis that the local likelihood estimator has better predictive potential than classical local regression.

1.2 Introduction

As a starting point, we consider the following non-parametric ARCH(1) model

\[ X_t = \sigma_t Z_t \quad (t \geq 2) \]
\[ \sigma_t^2 = f(X_{t-1}), \quad f : \mathbb{R} \rightarrow \mathbb{R}^+ , \quad (1.1) \]

where the innovations \( Z_t \) are identically independent distributed, with distribution function \( G \), zero mean, variance one and independent from \( \{X_s; s < t\} \).
Chapter 1. Local Likelihood for non-parametric ARCH(1) models

We assume the time series \( \{X_t\}_{t \geq 1} \) to be stationary (which is approximately true for return time series in a time-window of about 2 years; for more details see Mikosch and Starica, 1999) and that certain mixing conditions hold (see Section 1.4 for more details).

For our purposes, it is useful to transform the model (1.1) into regression by taking the logarithm

\[
\log (X_t^2) = \log (\sigma_t^2 Z_t^2) = \log (\sigma_t^2) + \log (Z_t^2) = \beta + \log (\sigma_t^2) + \log (Z_t^2) - \beta,
\]

where \( \beta = \mathbb{E} \left[ \log (Z_t^2) \right] \) (assumed to be negative). To simplify, from now on we will use the following notation:

\[
Y_t := \log (X_t^2), \\
U_t := \log (Z_t^2) - \beta \quad \text{and} \\
g(X_{t-1}) := \beta + \log (f(X_{t-1})) = \beta + \log (\sigma_t^2).
\]

Then, the transformed non-parametric ARCH(1) model can be rewritten as

\[
Y_t = g(X_{t-1}) + U_t,
\]

where \( \mathbb{E}[U_t] = 0 \), \( U_t \) are i.i.d. and independent from \( \{Y_s; s \leq t\} \).

The classical method used to find a non-parametric estimator for the function \( g(\cdot) \) in (1.2) is kernel regression (least squares) smoothing, as for example in Härdle and Vieu (1992), Härdle (1994) or Yang et al. (1999). In this paper, we propose an alternative, non-parametric estimator for \( g(\cdot) \) using local likelihood estimation in a similar way as for example in Loader (1999). For this purpose, the distribution of the innovations \( U_t \) in (1.2) must be derived from the one of the variables \( Z_t \) in (1.1). This is done as follows.

If the distribution of the innovations \( Z_t \) in (1.1) is known, the distribution of \( U_t \) can be calculated using the results of the following transformation theorem.

**Theorem.** (Transformation theorem)

Let \( X \) be a random variable with density \( X \sim f_X, \mathbb{P}[X \leq x] = F_X(x) \). Let \( Y = h(X), h(\cdot) \) an invertible function. Then:

\[
\mathbb{P}[Y \leq y] = \mathbb{P}[h(X) \leq y] = \mathbb{P}[X \leq h^{-1}(y)] = F_X(h^{-1}(y)).
\]

Computing the derivative, we get

\[
f_Y(y) = f_X(h^{-1}(y)) \cdot \frac{d}{dy}(h^{-1}(y)).
\]
1.2. Introduction

In the following examples, we calculate explicitly the density function of $U_t$ using the transformation theorem in the case of standard normally and scaled $t_v$, $v > 4$, distributed innovations $Z_t$ in (1.1).

**Example 1.1.** Let the innovations $Z_t$ in (1.1) be standard normally distributed, i.e. $G = \mathcal{N}(0, 1)$. Then we have that $Z_t^2 \sim \chi_1^2$ with density function given by

$$f_{Z_t^2}(z) = \begin{cases} 0 & \text{if } z < 0 \\ \frac{1}{\sqrt{2\pi}} z^{-\frac{1}{2}} e^{-\frac{z}{2}} & \text{otherwise.} \end{cases}$$

Now, applying the transformation theorem on $X = Z_t^2 \sim \chi_1^2$ and $h(x) = \log x - \beta$, we have

$$f_{U_t}(u) = f_{Z_t^2}(e^{\beta+u}) \cdot e^{\beta+u} = \frac{1}{\sqrt{2\Gamma(\frac{1}{2})}} (e^{\beta+u})^{-\frac{1}{2}} \exp \left( - \frac{e^{\beta+u}}{2} \right) e^{\beta+u} =$$

$$= c \cdot \exp \left( \frac{u}{2} - \frac{\tilde{\beta} e^u}{2} \right), \quad u \in \mathbb{R},$$

where the constants $c$ and $\tilde{\beta}$ equal

$$c = \frac{1}{\sqrt{2\Gamma(\frac{1}{2})}} \tilde{\beta}^{\frac{1}{2}} > 0 \quad \text{and}$$

$$\tilde{\beta} = \exp(\beta) = \exp(\mathbb{E}[\log(Z_t^2)]) > 0.$$ 

**Example 1.2.** Let the innovations $Z_t$ in (1.1) be scaled $t_v$ distributed, with degrees of freedom $v > 4$, i.e. $G = \sqrt{\frac{v-2}{v}} t_v$. Note that the scaling factor $\sqrt{\frac{v-2}{v}}$ is used to satisfy the condition $\text{Var}(Z_t) = 1$. Using the transformation theorem, we obtain that $Z_t^2$ has the following density

$$f_{Z_t^2}(z) = \begin{cases} 0 & \text{if } z < 0 \\ \frac{1}{\sqrt{2\pi}} \sqrt{\frac{v}{v-2}} z^{-\frac{1}{2}} (1 + \frac{z}{v-2})^{-\frac{v+1}{2}} & \text{otherwise.} \end{cases}$$

Now, applying the transformation theorem a second time, analogously to Example 1.1, we get

$$f_{U_t}(u) = c \cdot e^u \left( 1 + \frac{\tilde{\beta} e^u}{v-2} \right)^{-\frac{v+1}{2}}, \quad u \in \mathbb{R},$$

where the constant $c = \frac{1}{\sqrt{2\Gamma(\frac{1}{2})}} \tilde{\beta}^{\frac{1}{2}} > 0$, $\tilde{\beta}$ defined as in Example 1.1.
1.3 The local constant log-likelihood estimation

In this section, we want to illustrate the non-parametric strategy we use for constructing the estimate of the function \( g(\cdot) \) in the model (1.2). The local constant log-likelihood of the model (1.2) for a fitting point \( x \) looks like

\[
\mathcal{L}_x(X^n_1, g_x) = \sum_{t=2}^{n} \left( w_t(x) \rho(Y_t, g_x) \right) = \\
= \sum_{t=2}^{n} \left( W \left( \frac{x - X_{t-1}}{h_n} \right) \left( - \log(f_{U_t}(Y_t - g_x)) \right) \right),
\]

where the kernel \( W(\cdot) : \mathbb{R} \rightarrow \mathbb{R} \) is a symmetric (i.e. \( W(u) = W(-u), \ u \in \mathbb{R} \)), nonnegative bounded function satisfying

\[
u W(u) \rightarrow 0 \ \text{as} \ |u| \rightarrow +\infty, \ \int_{\mathbb{R}} W(u) \, du = 1.
\]

It typically assigns largest weights to observations close to \( x \) (for example Gaussian or tricube kernel). \( h_n \) is a global bandwidth. It is also possible to choose the bandwidth \( h_n = h_n(x) \) dependent on the fitting point \( x \), but this does not belong to the scope of this work. The sequence \( \{h_n\}_{n \in \mathbb{N}} \) is such that

\[
h_n \rightarrow n \rightarrow 0, \ \ n h_n \rightarrow n \rightarrow \infty, \ \ h_n > 0, \ \ \forall n \geq 2 \in \mathbb{N}.
\]

For more details about the more general local polynomial log-likelihood estimation, we refer the reader to the book of Loader (1999). Minimizing (1.3) with respect to the parameter \( g_x \) leads to the local constant log-likelihood estimate \( \hat{g}_{x;n} \)

\[
\hat{g}_{x;n} = \arg\min_{g_x} \sum_{t=2}^{n} \left( w_t(x) \left( - \log(f_{U_t}(Y_t - g_x)) \right) \right)
\]

for every fitting point \( x \).

1.3.1 Local likelihood estimator with Gaussian or scaled \( t_v \)-distributed innovations

In this Section, we will focus on the important case of standard normal distributed innovations \( Z_t \) in (1.1), i.e. \( G = \mathcal{N}(0, 1) \), as we do in the Examples 1.1 and 1.6. The importance of this particular case is also given from
1.3. The local constant log-likelihood estimation

the fact that the local likelihood estimator can be written in closed form. This is not the case if the innovations are scaled $t_v$-distributed, for example. The general problem of the existence and uniqueness of the local polynomial log-likelihood estimate is considered for example in Loader (1999), Theorem 4.1.

**Theorem 1.3. (Existence and uniqueness of $\widehat{g}_{x:n}$ under standard normal assumption)**

Assume that the innovations $Z_t$ in (1.1) are standard normally distributed. Then the local constant log-likelihood estimator $\widehat{g}_{x:n}$ given by (1.4) for the function $g(x)$ of the transformed ARCH(1) model (1.2) exists, is unique and equals

$$\widehat{g}_{x:n} = \beta^N + \log \left( \frac{\sum_{t=2}^n \left( W \left( \frac{x-X_{t-1}}{h_n} \right) X_t^2 \right)}{\sum_{s=2}^n W \left( \frac{x-X_{s-1}}{h_n} \right)} \right)$$

(1.5)

for every fitting point $x$, where $\beta^N = \mathbb{E}[\log(Z_1^2)] \approx -1.27$.

**Proof.** Differentiating, the estimator $\widehat{g}_{x:n}$ (1.4) must be a solution of the local log-likelihood equation

$$\sum_{t=2}^n \left( w_t(x) \rho'(Y_t, g_x) \right) = 0,$$

(1.6)

where $\rho'(y, g_x) = \frac{\partial}{\partial g_x} \rho(y, g_x)$.

Using the results of Example 1.6 of Section 1.3.2 for the local constant log-likelihood equations (1.6) we get

$$\sum_{t=2}^n \left( w_t(x) \rho'(Y_t, g_x) \right) = \sum_{t=2}^n \left( w_t(x) \frac{1}{2} \left( 1 - \beta^e(Y_t - g_x) \right) \right) = 0$$

$$\iff \quad \frac{1}{2} \sum_{t=2}^n w_t(x) = \frac{\beta^e g_x}{2} \sum_{t=2}^n \left( w_t(x) e^{Y_t} \right).$$

Thus, solving the last equation with respect to $g_x$ we obtain

$$\widehat{g}_{x:n}^N = \log \left( \beta^N \right) + \log \left( \frac{\sum_{t=2}^n \left( W \left( \frac{x-X_{t-1}}{h_n} \right) \exp(Y_t) \right)}{\sum_{s=2}^n W \left( \frac{x-X_{s-1}}{h_n} \right)} \right)$$

and (1.5) is proved. The uniqueness of this estimator follows directly from the strict convexity of $\rho(Y_t, g_x)$ in $g_x$ (see Example 1.6 of Section 1.3.2).
Remark. Note that the particular form of \( \hat{g}^{N}_{x,n} \) as
\[
\hat{g}^{N}_{x,n} = \beta^{N} + \log \left( \hat{k}(x) \right),
\]
where \( \hat{k}(x) \) is a kernel estimator for \( k(x) = \mathbb{E} \left[ e^{Y_t} \mid X_{t-1} = x \right] \), allows for uniform convergence of the estimator. The sufficient conditions on the kernel function \( W(\cdot) \), on the bandwidth \( h_n \) and on the process \( \{X_t\} \) needed for a strong uniform convergence of the kernel estimator \( \hat{k}(x) \) in a compact set \( C \) have been studied for example in Collomb (1984, 1985). Thanks to the result of Lemma 1.20, we also have uniform convergence of the estimator \( \hat{g}^{N}_{x,n} \) given by (1.5), i.e.
\[
\sup_{x \in C} \left| \hat{g}^{N}_{x,n} - g(x) \right| \xrightarrow{P} 0, \quad \text{as } n \to \infty. \tag{1.7}
\]
The conditions on the kernel function \( W(\cdot) \), on the bandwidth \( h_n \), on the process \( \{X_t\} \), and on the functions \( \psi_x(Y_t - g_x) = \rho'(Y_t, g_x) \) needed for a strong uniform convergence of general robust non-parametric time series estimators, like for example (1.5), have been investigated in Collomb and Härdle (1986). For more details, we refer the reader to Theorem 1.18.

The following Corollary 1.4 gives us the local constant log-likelihood estimator \( \hat{f}^{N}_{x,n} \) for the function \( f(x) \) in the non-parametric ARCH(1) model (1.1) with normal distributed innovations.

Corollary 1.4. Assume that the innovations \( Z_t \) in (1.1) are standard normally distributed. Then the local log-likelihood estimator \( \hat{f}^{N}_{x,n} \) for the function \( f(x) \) in the non-parametric ARCH(1) model (1.1) is given by
\[
\hat{f}^{N}_{x,n} = \frac{\sum_{t=2}^{n} \left( W\left( \frac{x-X_{t-1}}{h_n} \right) X_t^2 \right)}{\sum_{t=2}^{n} W\left( \frac{x-X_{t-1}}{h_n} \right)}.
\]

Proof. Since \( f(x) = e^{\beta} e^{g(x)} \) and \( e^{Y_t} = X_t^2 \), the result follows directly from (1.5). \( \square \)

Remark. The estimator \( \hat{f}^{N}_{x,n} \) is the same as the one that we get when we make a local regression for the quadratic ARCH(1) model
\[
X_t^2 = f(X_{t-1})Z_t^2 = f(X_{t-1}) + f(X_{t-1})(Z_t^2 - 1) = f(X_{t-1}) + \eta_t,
\]
where \( \eta_t \) is a martingale difference with \( \mathbb{E} [\eta_t] = 0 \), which follows from the definition of \( Z_t \) in (1.1). In other words, the estimator (1.8) is the popular kernel estimator in the quadratic ARCH model. What is new here is the insight
that it is a (transformed) local likelihood estimator and, as we will argue in Section 1.5, it is efficient under Gaussian innovations. The same result can be found also for more than one predictor, i.e. \( p > 1 \), and standard normally distributed innovations \( Z_t \). For more details, we refer the reader to Section 1.8.

As we have already said at the beginning of this Section, in the case of scaled \( t_v \)-distributed innovations the local likelihood estimator can not be written in a closed form. In spite of this, the following theorem proves the existence and the uniqueness of the local likelihood estimator under scaled \( t_v \)-distributed innovations.

**Theorem 1.5.** (Existence and uniqueness of \( \widehat{g}_{x;n} \) under scaled \( t_v \)-distributed innovations)

Assume that the innovations \( Z_t \) in (1.1) are scaled \( t_v \)-distributed, \( v > 4 \). Then the local constant log-likelihood estimator \( \widehat{g}_{x;n}^{t_v} \) given as the solution of the local likelihood equation (1.6) exists and is unique.

**Proof.** The result follows directly from Theorem 4.1 in Loader (1999). Note that in Example 1.7 of Section 1.3.2 we prove the convexity of \( \rho(y, g_x) \) in the case of scaled \( t_v \)-distributed innovations. \( \Box \)

### 1.3.2 The convexity of \( \rho(y, g_x) = -\log \left( f_{U_t}(y - g_x) \right) \)

We want to find here a sufficient condition to determine whether the function \( \rho(y, \cdot) \) is convex for all \( y \). Naturally, if the function \( \rho(y, \cdot) \) is twice differentiable in the second argument (which is true in many cases), it is well known that a sufficient condition for the convexity is that the second partial derivative of \( \rho(y, \cdot) \) with respect to the second argument must be positive, i.e.

\[
\frac{\partial^2}{\partial g_x^2} \rho(y, g_x) > 0.
\]

Therefore, we only have to calculate an explicit formula for the second partial derivative of \( \rho(y, g_x) = -\log \left( f_{U_t}(y - g_x) \right) \) in our particular case. The first partial derivative of \( \rho(y, g_x) \) with respect to \( g_x \) looks like

\[
\frac{\partial}{\partial g_x} \rho(y, g_x) = -\frac{1}{f_{U_t}(y - g_x)} \cdot f_{U_t}'(y - g_x),
\]

where \( f_{U_t}'(y - g_x) = \frac{\partial}{\partial g_x} f_{U_t}(y - g_x) \). Differentiating a second time with respect to \( g_x \) we obtain the following condition for the convexity

\[
\frac{\partial^2}{\partial g_x^2} \rho(y, g_x) = \frac{f_{U_t}'(y - g_x)^2}{f_{U_t}(y - g_x)^2} - \frac{f_{U_t}''(y - g_x)}{f_{U_t}(y - g_x)} > 0 \quad \forall y,
\]
where $f''_{U_t}(y - g_x) = \frac{\partial^2}{\partial g_x^2} f_{U_t}(y - g_x)$.

**Example 1.6.** We verify whether the condition (1.9) holds, and consequently whether the function $\rho(y, \cdot)$ is convex, in the case of standard normally distributed innovations $Z_t$ in (1.1). In the Example 1.1 we have already calculated the density function $f_{U_t}(-)$ in the case $G = \mathcal{N}(0, 1)$

$$f_{U_t}(y - g_x) = c \exp \left( \frac{y - g_x}{2} \left( \tilde{\beta} e^{(y-g_x)} - 1 \right) \right).$$

Thus, the first and second partial derivative of $f_{U_t}(-)$ with respect to $g_x$ are given by:

$$f'_{U_t}(y - g_x) = c \exp \left( \frac{y - g_x}{2} \right) \left( \frac{\tilde{\beta} e^{(y-g_x)} - \tilde{\beta} e^{(y-g_x)}}{2} \right) = f_{U_t}(y - g_x) \cdot \frac{1}{2} \left( \tilde{\beta} e^{(y-g_x)} - 1 \right)$$

and

$$f''_{U_t}(y - g_x) = f'_{U_t}(y - g_x) \cdot \frac{1}{2} \left( \tilde{\beta} e^{(y-g_x)} - 1 \right) - f_{U_t}(y - g_x) \cdot \frac{1}{2} \tilde{\beta} e^{(y-g_x)} = f_{U_t}(y - g_x) \cdot \left( \frac{1}{4} \left( \tilde{\beta} e^{(y-g_x)} - 1 \right)^2 - \frac{1}{2} \tilde{\beta} e^{(y-g_x)} \right).$$

Using these results in (1.9), we obtain that

$$\frac{\partial^2}{\partial g_x^2} \rho(y, g_x) = \frac{1}{4} \left( \tilde{\beta} e^{(y-g_x)} - 1 \right)^2 - \left( \frac{1}{4} \left( \tilde{\beta} e^{(y-g_x)} - 1 \right)^2 - \frac{1}{2} \tilde{\beta} e^{(y-g_x)} \right) = \frac{1}{2} \tilde{\beta} e^{(y-g_x)} > 0 \forall y$$

and hence in the case of standard normally distributed innovations $Z_t$ the function $\rho(y, \cdot)$ is convex in $g_x$.

An identical result can (in some cases easily) be found directly from the calculation of the partial derivatives of $\rho$ with respect to $g_x$.

**Example 1.7.** We calculate explicitly $\rho$ in the case of scaled $t_\nu$-distributed innovations $Z_t$. From the result of Example 1.2, we have

$$\rho(y, g_x) = - \log \left( f_{U_t}(y - g_x) \right) = - \log \left( c e^{\frac{y - g_x}{\tilde{\beta}} \left( 1 + \frac{\tilde{\beta} e^{(y-g_x)}}{\nu - 2} \right)^{-\frac{\nu + 1}{2}}} \right) = \frac{\nu + 1}{2} \log \left( 1 + \frac{\tilde{\beta} e^{(y-g_x)}}{\nu - 2} \right) - \frac{y - g_x}{2} - \log(c).$$
The first and second partial derivative of \( \rho \) with respect to \( g_x \) equal

\[
\rho'(y, g_x) = \frac{1}{2} - \frac{v+1}{2} \frac{1}{1 + \frac{v-2}{\beta e^{(y-g_x)}}} := \psi(y - g_x)
\]

and

\[
\rho''(y, g_x) = \frac{v+1}{2} \left( \frac{1}{1 + \frac{v-2}{\beta e^{(y-g_x)}}} \right)^2 \frac{v-2}{\beta e^{(y-g_x)}},
\]

respectively. Since clearly the condition (1.9) is satisfied, the convexity of \( \rho \) also holds in this case.

### 1.4 Consistency of the local likelihood estimator

In this Section, we want to prove the consistency of the normal local constant log-likelihood estimator \( \hat{g}_{\rho,n} \) given by (1.5) in the “right” case of standard normally distributed innovations \( Z_t \) and in the misspecified case of innovations \( Z_t \) with distribution function \( G \neq \mathcal{N} \), having zero mean and variance one (pseudo-likelihood estimation). Moreover, we also want to show the consistency of the local constant log-likelihood estimator \( \hat{g}_{\rho,n}^{\text{ht}} \) given as the unique solution of (1.6) in the case of scaled \( t_v \)-distributed innovations.

In order to study the consistency of our estimators, we have to consider certain dependence or mixing conditions for stochastic processes. The process \( \{X_t\}_{t \geq 1} \) satisfies the following strong mixing or \( \alpha \)-mixing condition (1.10), firstly introduced by Rosenblatt (1956):

there exists a sequence \( \alpha(n) \) of positive numbers such that \( \lim_{n \to \infty} \alpha(n) = 0 \) and for any \( A \in \mathcal{M}_{1,j}, B \in \mathcal{M}_{j+n, +\infty} \) we have

\[
|P(A \cap B) - P(A)P(B)| \leq \alpha(n),
\]  

where we denote by \( M_{a,b} \) the \( \sigma \)-algebra generated by the random variables \( \{X_t : a \leq t \leq b\} \), \( 1 \leq a \leq \infty \);

or the process satisfies the more often studied \( \phi \)-mixing or uniform mixing condition (1.11) (Billingsley, 1968)

there exist coefficients \( \phi(n) \) such that \( \lim_{n \to \infty} \phi(n) = 0 \) and the inequality

\[
|P(A \cap B) - P(A)P(B)| \leq \phi(n)P(A)
\]  

holds for any \( A \in \mathcal{M}_{1,j}, B \in \mathcal{M}_{j+n, +\infty} \).
A process which satisfies the condition (1.10) or (1.11) is called to be \( \alpha \)-mixing or \( \phi \)-mixing respectively. Note that the \( \phi \)-mixing condition is considerably stronger than the \( \alpha \)-mixing condition.

Note also that if the process \( \{X_t\}_{t \geq 1} \) is \( \alpha \)- or \( \phi \)-mixing, then also the process \( \{X_{t-1}, Y_t\}_{t \geq 2} \) is mixing with coefficients bounded by the ones from \( \{X_t\}_{t \geq 1} \) but with separation lag being smaller by one.

We assume that the mixing coefficients \( \alpha(n) \) (or \( \phi(n) \)) satisfy the following condition (A1)

\[
\sum_{n=2}^{\infty} \alpha(n)^{\frac{2}{2+\tau}} < \infty \quad \text{and} \quad \mathbb{E}[|Y_t|^{2+\tau} | X_{t-1} = x] < \infty \quad \text{for some} \ \tau > 0 \quad (A1)
\]

or

\[
\sum_{n=2}^{\infty} \phi(n)^{\frac{1}{2+\tau}} < \infty \quad \text{and} \quad \mathbb{E}[|Y_t|^{2+\tau} | X_{t-1} = x] < \infty \quad \text{for some} \ \tau > 0. \quad (A1)
\]

This condition ensures that versions of the law of large numbers and of the central limit theorem for the dependent variables \( \{X_{t-1}, Y_t\}_{t \geq 2} \) exist. Note that in the particular case where the innovations \( Z_t \) in (1.1) are normally or scaled \( t_\nu \)-distributed, (A1) holds (for more details, see for example Doukhan, 1994).

1.4.1 Consistency of \( \hat{g}_{X; n}^N \) and \( \hat{f}_{X; n}^N \)

The consistency in the "right" case of normally distributed innovations \( Z_t \) is proved in the following Theorem 1.8.

**Theorem 1.8.** (Consistency of \( \hat{g}_{X; n}^N \))

Suppose that the process \( \{X_{t-1}, Y_t\}_{t \geq 2} \) is from the model (1.1)-(1.2) with standard normal distributed innovations \( Z_t \). Then, for every fitting point \( x \) the estimator \( \hat{g}_{X; n}^N \) for the function \( g(x) \) of the transformed ARCH(1) model (1.2) given by (1.5) is consistent, i.e.

\[
\hat{g}_{X; n}^N \xrightarrow{P} g(x) = \beta^N + \log \left( f(x) \right) \quad \text{as} \ \bar{h}_n \to 0, \ nh_n \to \infty.
\]

In the proof of Theorem 1.8 we need the results of the following Proposition 1.9.

**Proposition 1.9.** For a fixed fitting point \( x \), denote by \( \gamma \left( g(x), \bar{g}(x) \right) := \mathbb{E}_g \left[ \rho(Y_t, \bar{g}(x)) \mid X_{t-1} = x \right] \) and with \( \mathcal{C} \) the compact set in \( \mathbb{R} \) (i.e. a closed interval) of all possible constant functions \( g(x) \). Let we assume that the following two conditions are satisfied:
1.4. Consistency

i) \( \gamma(g(x), g(x)) < \gamma(g(x), \bar{g}(x)) \) for all \( \bar{g}(x) \neq g(x) \in \mathcal{C} \);

ii) for all \( g, \bar{g} \): \( \mathbb{E}_g \left[ \inf_{\tau \in U(\bar{g})} \rho(Y_t, \tau) \mid X_{t-1} = x \right] \rightarrow \gamma(g(x), \bar{g}(x)) \)

for \( U(\bar{g}) \downarrow \{\bar{g}\} \), where \( U(\bar{g}) \) is a neighborhood of \( \bar{g} \).

Then:

\[ \hat{g}_{x;n} \xrightarrow{P} g(x), \]

with \( \hat{g}_{x;n} \) defined as in (1.4).

Proof. The proof can be found for example in Künsch (1997), Satz 6.6. □

The second assumption of Proposition 1.9 is satisfied if the function \( \rho(y, g_x) \) is continuous \( \forall y \) and if we have that \( \inf_{\tau \in U(\bar{g})} \rho(Y_t, \tau) \) is bounded from below. In this case, using the monotone convergence, we can exchange expectation and the limiting operation and the assumption ii) holds.

Proof of Theorem 1.8. For every fitting point \( x \), denote by \( \hat{C}_n(g_x) \) the kernel estimator

\[ \hat{C}_n(g_x) := \frac{1}{(n - 1)h_n} \sum_{t=2}^{n} \left( w_t(x) \rho(Y_t, g_x) \right). \]

Thanks to the regularity conditions assumed for the process \( \{X_{t-1}, Y_t\}_{t \geq 2} \) and to (A1), using the WLLN for dependent observations with mixing structure (see for example Doukhan, 1994) we have that

\[ \forall g_x \in \mathcal{C} : \hat{C}_n(g_x) \xrightarrow{P} \gamma(g(x), g_x) \text{ as } nh_n \rightarrow \infty, \]

where \( \gamma(\cdot, \cdot) \) is defined as in Proposition 1.9. Now, we have to show that

\[ \hat{g}^*_{x;n} = \arg\min_{g_x} \hat{C}_n(g_x) \xrightarrow{P} \arg\min_{g_x} \gamma(g(x), g_x) = g(x) \]

holds for every fitting point \( x \). This follows directly from the result of Proposition 1.9. Thus, what we need is to prove that the assumptions of Proposition 1.9 hold. Firstly, we prove that assumption i) is satisfied, which corresponds to show that \( \arg\min_{g_x} \gamma(g(x), g_x) = g(x) \). From the model (1.2), we know that

\[ g(x) = \beta^N + \log (f(x)) = \mathbb{E}_g \left[ Y_t \mid X_{t-1} = x \right]. \]

Now, from the definition of \( \gamma \) it follows that

\[ \arg\min_{g_x} \gamma(g(x), g_x) = \arg\min_{g_x} \mathbb{E}_g \left[ \rho(Y_t, g_x) \mid X_{t-1} = x \right] \]
and differentiating we obtain that the minimum is a solution of the following equation

$$\mathbb{E}_g \left[ \rho' (Y_t, g_x) \mid X_{t-1} = x \right] = 0.$$ 

In the case of standard normal distributed innovations $Z_t$, we have already calculated in Examples 1.1-1.6 that $\rho' (Y_t, g_x) = \frac{1}{2} - \tilde{\beta}_{N} e^{(Y_t - g_x)}$. Introducing the explicit form of $\rho'$ in the equation above, we obtain that

$$\arg\min g_x \gamma (g(x), g_x) = \log(\tilde{\beta}_{N}) + \log \left( \mathbb{E}_g e^{Y_t} \mid X_{t-1} = x \right).$$

Using the definition of $\tilde{\beta}$, we rewrite $\log(\tilde{\beta}_{N})$ as $\beta^N$ and considering that from the model (1.1) follows $f(x) = \mathbb{E}[X_t^2 \mid X_{t-1} = x] = \mathbb{E}_g e^{Y_t} \mid X_{t-1} = x$ we have proved assumption i) of Proposition 1.9.

To end the proof, we just have to show the continuity of the function $\rho(y, \cdot) \forall y$ and to verify if the assumption

$$\inf_{\tau \in U(g_x)} \rho(Y_t, \tau)$$

is satisfied for a neighborhood $U(g_x) \subset \mathcal{C}$ of the form $U(g_x) = (a_1, a_2)$, $a_1 < a_2$, $a_1 \neq -\infty$. This follows directly from the explicit form of $\rho(Y_t, g_x)$ as

$$\rho(Y_t, g_x) = \frac{\tilde{\beta}_{N}}{2} e^{Y_t - g_x} - \frac{Y_t - g_x}{2} - \log(c),$$

c as in Example 1.1. □

**Remark.** It is also possible using stronger assumptions and robust theory (note that the estimator $\tilde{\beta}_{N}$ must satisfy the local log-likelihood equation (1.6)) to prove the almost sure convergence of the normal local constant log-likelihood estimator (1.5) to $g(x)$, the solution of

$$\mathbb{E} \left[ \rho'(Y_t, g(x)) \mid X_{t-1} = x \right] = \int_{\mathbb{R}} \rho'(Y_t, g(x)) dF(Y_t \mid X_{t-1} = x) = 0.$$ 

See for example Boente and Fraiman (1989), Theorem 2.1.

The following Corollary 1.10 gives us the consistency of the local constant log-likelihood estimator $\hat{f}_{x:1}^{N}$ given by (1.8) for the function $f(x)$ in the non-parametric ARCH(1) model (1.1) with standard normally distributed innovations $Z_t$.

**Corollary 1.10.** (Consistency of $\hat{f}_{x:1}^{N}$)

*Under the same assumptions of Theorem 1.8, the estimator $\hat{f}_{x:1}^{N}$ given by (1.8) is consistent for every fitting point $x$.**
1.4. Consistency

Proof. Thanks to the result of Lemma 1.20 (exp is a continuous function) and since \( f(x) = e^{-\beta^N} e^{g(x)} \), the consistency of \( \hat{f}_{x:n}^N \) follows directly from Theorem 1.8.

1.4.2 Consistency of \( \hat{g}_{X:n}^N \) and \( \hat{f}_{x:n}^N \) under pseudo-likelihood estimation

We want to investigate here the consistency of the estimators \( \hat{g}_{X:n}^N \) and \( \hat{f}_{x:n}^N \) in the misspecified case of innovations \( Z_t \) as in (1.1) with distribution function \( G \neq \mathcal{N}(0, 1) \).

Theorem 1.11. (Consistency under pseudo-likelihood estimation)

Suppose that the process \{\( X_{t-1}, Y_t \)\}(\( t \geq 2 \)) is from the model (1.1)-(1.2) and let the innovations \( Z_t \) have a distribution function \( G \neq \mathcal{N}(0, 1) \). Denote by

\[
\delta := \mathbb{E}\left[\log(Z_t^2)\right] - \beta^N,
\]

where \( \beta^N \) is defined as in Theorem 1.3. Then for every fitting point \( x \):

i) \( \hat{g}_{X:n}^N \) given by (1.5) is a consistent estimator for \( g(x) - \delta \), where \( g(x) \) equals \( \mathbb{E}[Y_t | X_{t-1} = x] \) in the model (1.2);

ii) \( \hat{f}_{x:n}^N \) is a consistent estimator for \( f(x) \) in the model (1.1).

A very important consequence of the results of Theorem 1.11 is that even in the misspecified case of non-Gaussian innovations \( Z_t \) the normal estimator \( \hat{f}_{x:n}^N \) is consistent. On the other side, we see that the estimator \( \hat{g}_{X:n}^N \) is not consistent for the less interesting function \( g(x) \).

Proof of Theorem 1.11. We rewrite \( g(X_{t-1}) \) as \( \delta + g^N(X_{t-1}) \) and \( U_t \) as \( U_t^N - \delta \) in the model (1.2)

\[
Y_t = g(X_{t-1}) + U_t = \delta + g^N(X_{t-1}) + U_t^N - \delta,
\]

where \( g^N(X_{t-1}) \) and \( U_t^N \) are the variables in the misspecified case of normal distributed innovations \( Z_t \). Note that in the "right" model is

\[
\mathbb{E}[U_t] = \mathbb{E}[U_t^N - \delta] = 0, \text{ but } \delta = \mathbb{E}[U_t^N] = \mathbb{E}\left[\log(Z_t^2) - \beta^N\right] \neq 0.
\]

Assuming standard normally distributed innovations \( Z_t \), we proceed with a pseudo-likelihood estimation for the function \( g(x) = \delta + g^N(x) = \delta + \beta^N + \log(f(x)) \). As before, the optimal normal local constant log-likelihood estimator is given by (1.5). The proof of i) is analogous to the one of Theorem 1.8.
From our construction we have that \( g(x) = \delta + \beta^N + \log \left( f(x) \right) \). Solving the equation with respect to \( f(\cdot) \) we get that
\[
 f(x) = e^{-\beta^N} e^{\left( g(x) - \delta \right)}
\]
and from i) and Corollary 1.4 follows ii). \( \square \)

### 1.4.3 Consistency of \( \hat{g}_{x; n}^{f_y} \)

Since in this case we do not know the explicit form of the estimator \( \hat{g}_{x; n}^{f_y} \), we have to show the consistency in a different way as the one used for the normal constant log-likelihood estimator. As a starting point, we know that \( \hat{g}_{x; n}^{f_y} \) is the unique solution of the constant local log-likelihood equation (1.6) with \( \rho'(y, g_x) = \psi(y - g_x) : \mathbb{R} \to \mathbb{R} \) already calculated in Example 1.7 and given by
\[
\psi(y - g_x) = \frac{1}{2} - \frac{1}{2} \frac{\nu + 1}{1 + \frac{\nu - 2}{\beta e^{(y-g_x)}}}.
\]

The most important condition for consistency is that the function \( g(x) \) must satisfy
\[
E_{g_y} \left[ \psi \left( Y_t - g(x) \right) \mid X_{t-1} = x \right] = 0
\]
for every fitting point \( x \). The consistency of the estimator \( \hat{g}_{x; n}^{f_y} \) follows from the general result of Proposition 1.12.

**Proposition 1.12.** Suppose that the process \( \{X_{t-1}, Y_t\}_{t \geq 2} \) is stationary, \( \alpha \)-or \( \phi \)-mixing and that assumption (AI) holds. Let \( \psi : \mathbb{R} \to \mathbb{R} \) be continuous and \( E_{g_y} \left[ |\psi(Y_t - g_x)| \mid X_{t-1} = x \right] < \infty \) for every fitting point \( x \). If:
\[
E_{g_y} \left[ \psi(Y_t - g_x) \mid X_{t-1} = x \right] > 0 \text{ for } g_x > g(x) \text{ and } E_{g_y} \left[ \psi(Y_t - g_x) \mid X_{t-1} = x \right] < 0 \text{ for } g_x < g(x) \quad (1.12)
\]
then it exists a sequence of estimators \( \{\hat{g}_{x; n}\}_n \) which converges almost surely to \( g(x) \) and satisfies
\[
\sum_{t=2}^{n} \left( w_t(x) \psi(Y_t - \hat{g}_{x; n}) \right) = 0.
\]

**Proof.** The proof can be found in Künsch (1997), Satz 6.7. Note that the mixing assumption ensures that also for the dependent variables \( \psi(Y_t - g_x) \), a general version of the SLLN holds; for more details, see Doukhan (1994) \( \square \)
1.4. Consistency

Now, we are in position to prove the consistency of the estimator $\hat{g}^{t_v}_{x,n}$ in the case of scaled $t_v$ distributed innovations $Z_t$.

**Theorem 1.13. (Consistency of $\hat{g}^{t_v}_{x,n}$)**

Suppose that the process $\{X_{t-1}, Y_t\}_{t \geq 2}$ is from the model (1.1)-(1.2) and let the innovations $Z_t$ in (1.1) be scaled $t_v$-distributed with degrees of freedom parameter $v > 4$. Then, the local constant log-likelihood estimator $\hat{g}^{t_v}_{x,n}$ given as the unique solution of the likelihood equation (1.6) is consistent for $g(x)$ satisfying

$$E_g \left[ \psi \left( Y_t - g(x) \right) \mid X_{t-1} = x \right] = 0$$

for every fitting point $x$.

**Proof.** We have already shown that for scaled $t_v$ distributed innovations the function $\psi$ takes the values

$$\psi(u) = \frac{1}{2} - \frac{v + 1}{2} \frac{1}{1 + \frac{v-2}{\beta v e^u}}$$

and is clearly continuous for all $u \in \mathbb{R}$. It also directly follows that $\psi(u)$ is bounded for all $u \in \mathbb{R}$, since:

- For $u \to +\infty : \psi(u) \to \frac{1}{2} - \frac{v + 1}{2}$
- For $u \to -\infty : \psi(u) \to \frac{1}{2}$.

Therefore, we have that $|\psi(u)| \leq C(v) := -\left(\frac{1}{2} - \frac{v+1}{2}\right)$, for $4 < v < \infty$ and $x \in \mathbb{R}$ and consequently, thanks to the monotonicity of the integral, we get that $E_g \left[ |\psi(Y_t - g_x)| \mid X_{t-1} = x \right] < \infty$ for every fitting point $x$.

We end the proof by showing that the condition (1.12) of Proposition 1.12 is satisfied. The result follows then directly from Proposition 1.12. Define

$$\lambda(g_x) := E_g \left[ \psi(Y_t - g_x) \mid X_{t-1} = x \right] = \int_{\mathbb{R}} \psi(y - g_x) dF_{Y_t|X=x}(y).$$

Now, if $g_x > g(x)$ it follows that $y - g_x < y - g(x)$. Since $\psi(u)$ is strictly monotone, we get that $\psi(y - g_x) > \psi(y - g(x))$. According to the monotonicity of the integral, we finally have that $\lambda(g_x) > \lambda(g(x)) = 0$ by definition. On the other hand, if $g_x < g(x)$ we get analogously that $\lambda(g_x) < \lambda(g(x)) = 0$ for all $x \in \mathbb{R}$ and (1.12) holds. $\square$
1.5 Asymptotic normality of the local likelihood estimator

In this Section, we investigate the asymptotic representation of the normal and scaled $t_v$ local constant log-likelihood estimator $\hat{g}_{x,n}$ given by (1.5) and $\hat{g}^V_{x,n}$, respectively. For this purpose, we will need the following general assumptions:

**H1.** $E[p''(Y_t, g(x)) | X_{t-1} = x] \neq 0$.

**H2.** The process $\{X_{t-1}, Y_t\}_{t \geq 2}$ is a stationary $\alpha$-mixing process, with mixing coefficients $\alpha(n)$ defined analogously to (1.10), such that

$$\sum_{i=2}^{\infty} \alpha(i) \frac{r^r}{r} < \infty$$

and

$$E[|\psi(Y_t - g(x))|^{2+r} | X_{t-1} = x] < \infty$$

for some $r > 0$.

(If the process is $\phi$-mixing, an analogous condition exists for the coefficients $\phi(n)$.)

**H3.** The stationary density $d$ of the vector $X_t$ is continuous and positive at $x$.

**H4.** For all $s \geq 2$, the density $d_s(u, v)$ of $(X_t, X_{t+s})$ is bounded uniformly in $s$.

**H5.** The kernel $W(\cdot)$ is defined as in Section 1.3.

**H6.** There exists a constant $0 \leq s < \infty$ such that $h_n^{3/2} \to s$ as $n \to \infty$.

**H7.** We assume that the function $g(x)$ in the model (1.2) verifies a Lipschitz condition of order 1, i.e.

$$|g(u) - g(x)| \leq C |u - x|$$

for some $C > 0$, and that the following limit exists

$$\lim_{\epsilon \to 0} \frac{g(x + \epsilon u) - g(x)}{\epsilon} = g'(x, u).$$

**Remark.** Note that assumption H7 on the function $g(x) = \beta + \log(f(x))$ is satisfied if we assume that

$$\sup_x \left| \frac{f'(x)}{f(x)} \right| \leq C < \infty$$

implied for example by $\sup_x |f'(x)| < \infty$ and $\inf_x f(x) > 0$.

This condition holds for example for the parametric ARCH(1) model, where $f(x) = \alpha_0 + \alpha_1 x^2$, $\alpha_0 > 0$, $\alpha_1 \geq 0$. 
1.5. Asymptotic normality

H8. Consider the function $\psi : \mathbb{R} \to \mathbb{R}$ defined by

$$\psi(Y_t - g_x) = \frac{\partial}{\partial g_x} (Y_t, g_x).$$

Assume that $\psi(\cdot)$ is differentiable with finite second moment and satisfying

$$\left| \frac{\partial \psi}{\partial t_1} (Y_t - t_1) - \frac{\partial \psi}{\partial t_2} (Y_t - t_2) \right| \leq H(Y_t) |t_1 - t_2|,$$

for all $t_1, t_2 \in \mathbb{C}$, a compact set $\subset \mathbb{R}$. Moreover, analogously to H2 we assume that

$$E \left[ \left| \frac{\partial \psi}{\partial s} (Y_t - s) \right|_{s=g(x)} \right]^{2+\tau} < \infty \text{ for some } \tau > 0.$$

In the proof of the following Theorem 1.15, we will need the results of Lemma 1.14. For simplicity from now on, we will denote by $n' = n - 1$.

Lemma 1.14. Let $\{X_t, Z_t\}_{t \geq 1}$ be a stationary random process verifying H2. Denote by $F(u \mid X_t = x)$ the conditional distribution of $Z_t$ given that $X_t = x$, by $\mu_1(x) = E[Z_1 \mid X_1 = x]$ and by $\sigma_1^2(x) = E[(Z_1 - \Phi(x))^2 \mid X_1 = x]$. Let us suppose that

i) $\mu_1$ is Lipschitz and $\lim_{\epsilon \to 0} \frac{\mu_1(x + \epsilon u) - \mu_1(x)}{\epsilon} = \mu_1'(x, u)$ exists;

ii) $\sigma^2$ is continuous in a neighborhood of $x$.

For a fitting point $x$, denote by $\hat{\mu}_{1,x,n}^K = \sum_{t=2}^n (k_t(x)Z_t)$ the kernel estimator, i.e.

$$k_t(x) = \frac{K \left( \frac{X_t - x}{h_n} \right)}{\sum_{t=2}^n K \left( \frac{X_t - x}{h_n} \right)},$$

where $K(\cdot)$ is a kernel function as in Section 1.3. Then H3 to H6 imply that

$$\frac{\sqrt{n' h_n}}{h_n} (\hat{\mu}_{1,x,n}^K - \mu_1(x))$$

is asymptotically normally distributed with mean $\mu_2 = s \int_{\mathbb{R}} \mu_1'(x, u)K(u)du$ and variance $\sigma_2^2 = \sigma_1^2(x)K_0$, with $K_0 = \int_{\mathbb{R}} K^2(u)du$.

Proof. The proof can be found in Boente and Fraiman (1990), Lemma 2. □

First of all, we give the asymptotic representation of the general local constant log-likelihood estimator $\hat{g}_{x,n}$ given by (1.4), i.e. without knowing the distribution function $G$ of the innovations $Z_t$. 

\[1.5.\text{Asymptotic normality}\]
Theorem 1.15. (Asymptotic normality of $\hat{g}_{x,n}$)

Assume that the conditions H1 to H8 are satisfied and that $\hat{g}_{x,n} \xrightarrow{P} g(x)$ as $n'h_n \to \infty$. Then for every fitting point $x$

$$\sqrt{n'h_n}(\hat{g}_{x,n} - g(x)) \xrightarrow{d} \mathcal{N}(\mu, \sigma^2) \quad \text{as } n \to \infty,$$

where the asymptotic mean $\mu$ is given by

$$\mu = \frac{3}{2} \int_{\mathbb{R}} g'(x, u) W(u) du$$

and the asymptotic variance $\sigma^2$ by

$$\sigma^2 = V(\psi, G) \frac{\int_{\mathbb{R}} W^2(u) du}{d(x)},$$

where the factor $V(\psi, G)$ equals

$$V(\psi, G) = \frac{\int_{\mathbb{R}} \psi^2(u) dF_{U_t}(u)}{\left( \int_{\mathbb{R}} \psi(u) dF_{U_t}(u) \right)^2}. \quad (1.13)$$

Note that if the function $g$ is smooth, we have that $g'(x, u) = g'(x) \cdot u$ and the asymptotic mean $\mu = 0$ (since the kernel $W(\cdot)$ is assumed to be symmetric). On the other hand, the asymptotic variance $\sigma^2$ strictly depends from the factor $V(\psi, G)$ and, as we can see from the explicit calculation for the examples with $G = \mathcal{N}(0, 1)$ and $G = \sqrt{\frac{v-2}{v} t_v}$, this can lead to big differences if we confront the asymptotic variance of the local likelihood estimator with the one of other non-parametric estimators like for example classical local regression.

Proof of Theorem 1.15. As a starting point, we remind the reader of the fact that the local constant log-likelihood estimator $\hat{g}_{x,n}$ given by (1.4) must be a solution of the log-likelihood equation (1.6). Since the function $\psi(Y_t - g_x) = \rho'(Y_t, g_x)$ is assumed to be differentiable for all $Y_t$ (see H8) and since $\hat{g}_{x,n}$ should be an estimator for $g(x)$, the idea is to consider the Taylor expansion of $\hat{g}_{x,n}$ around $g(x)$. This leads to the following result:

$$0 = \frac{1}{n'h_n} \sum_{t=2}^{n} \{w_t(x) \psi(Y_t - \hat{g}_{x,n})\} = \frac{1}{n'h_n} \sum_{t=2}^{n} \{w_t(x) \psi(Y_t - g(x))\} +$$

$$+ \{\hat{g}_{x,n} - g(x)\} \frac{1}{n'h_n} \sum_{t=2}^{n} \{w_t(x) \frac{\partial \psi}{\partial \xi_n}(Y_t - \xi_n)\}.$$
where \( w_t(x) = W\left( \frac{x - X_{t-1}}{h_n} \right) \) and we have that \(|\xi_n - g(x)| \leq |\tilde{g}_{x,n} - g(x)|\). We denote by

\[
M_n := \frac{1}{n'h_n} \sum_{t=2}^{n} \left( w_t(x) \frac{\partial \psi}{\partial \xi_n} (Y_t - \xi_n) \right)
\]

and by

\[
M(g(x)) = \mathbb{E}_g \left[ \frac{\partial \psi}{\partial t} (Y_t - t) \bigg|_{t=g(x)} \right].
\]

Since the estimator \( \tilde{g}_{x,n} \) is assumed to be consistent, it follows from assumption H8 and the LLN that

\[
\left| M_n - M(g(x)) \right| \leq \left| \frac{1}{n'h_n} \sum_{t=2}^{n} \left( w_t(x)\psi'(Y_t - g(x)) \right) - M(g(x)) \right| + \\
+ \left| \tilde{g}_{x,n} - g(x) \right| \cdot \left| \frac{1}{n'h_n} \sum_{t=2}^{n} \left( w_t(x)H(Y_t) \right) \right|
\]

where the right side converges to zero in probability. Thus, we have proved that

\[
M_n \xrightarrow{p} M(g(x)) \quad \text{as } n'h_n \to \infty, \ h_n \to 0.
\]

Therefore, assuming that \( M^{-1} \) exists (see H1), we can approximate the difference \( \tilde{g}_{x,n} - g(x) \) by an arithmetic mean as follows:

\[
\tilde{g}_{x,n} - g(x) = -(M_n)^{-1} \frac{1}{n'h_n} \sum_{t=2}^{n} \left( w_t(x)\psi'(Y_t - g(x)) \right)
\]

\[
= -(M(g))^{-1} \frac{1}{n'h_n} \sum_{t=2}^{n} \left( w_t(x)\psi'(Y_t - g(x)) \right) + R_n,
\]

where \( R_n \) is a rest term such that \( \sqrt{n'h_n}R_n \xrightarrow{p} 0 \) as \( n'h_n \to \infty \).

Rewriting the last equation without using the introduced notation, we get that

\[
\tilde{g}_{x,n} - g(x) = \frac{1}{n'h_n} \sum_{t=2}^{n} \left( w_t(x)\psi(Y_t - g(x)) \right) + R_n.
\]

Now we can use the result of Lemma 1.15 for the term

\[
\frac{\sqrt{n'h_n}}{n'h_n} \sum_{t=2}^{n} \left( w_t(x)\psi(Y_t - g(x)) \right)
\]
and the dependent variables $Z_t = \psi(Y_t - g(x))$. Note that in our case

$$\mu_1(x) = \mathbb{E}\left[\psi(Y_t - g(x)) \mid X_{t-1} = x\right] = 0$$

and

$$\sigma^2_1(x) = \int_{\mathbb{R}} \psi(y - g(x))^2 dF_{U_t}(y - g(x)).$$

The result follows since by Slutsky’s Lemma the error term $R_n$ can be neglected, and since from the definition of $\mu'_1(x, u)$ we have that

$$\mu'_1(x, u) = g'(x, u) \mathbb{E}\left[\psi'(Y_t - g(x)) \mid X_{t-1} = x\right].$$

1.5.1 The asymptotic variance of $\widehat{g}^N_{x,n}$

Using the results of Theorem 1.15, we calculate the asymptotic variance of the normal local constant log-likelihood estimator $\widehat{g}^N_{x,n}$ given by (1.5). This is done in the following theorem.

**Theorem 1.16. (Asymptotic variance of $\widehat{g}^N_{x,n}$)**

Suppose that the process \{X_{t-1}, Y_t\}_t is from the model (1.1)-(1.2) and let the innovations $Z_t$ be standard normally distributed. Under the assumptions H4 to H7, the asymptotic variance of the normal local constant log-likelihood estimator $\widehat{g}^N_{x,n}$ given by (1.5) of the function $g(x)$ equals

$$\sigma^2 = 2 \cdot \int_{\mathbb{R}} W^2(u) du \frac{d(x)}{d(x)}$$

(1.14)

for every fitting point $x$, where the kernel $W(\cdot)$ is defined as in Section 1.3.

**Proof.** The result follows directly from Theorem 1.15. In the case of standard normally distributed innovations $Z_t$ we have that assumptions H2-H3 hold (see Doukhan, 1994). Moreover, we have already shown in Examples 1.1 and 1.6 that the function $\psi(Y_t - g_x) = \rho'(Y_t, g_x)$ equals

$$\psi(Y_t - g_x) = \frac{1}{2} - \frac{\beta_N}{2} e^{(Y_t - g_x)}.$$

Knowing the explicit form of $\psi$, it is easy to verify that assumption H8 is satisfied (the exponential function is Lipschitz-continuous in a compact set $C$).
1.5. Asymptotic normality

and for example for $\tau = 1$, the second condition holds. Now, we want to show that $H_1$ holds, too:

$$
\mathbb{E}
\left[
\frac{\hat{\beta}_N^N}{2} e^{(Y_t-g(x))}
| X_{t-1} = x
\right] =
\frac{\hat{\beta}_N^N}{2} \int_{-\infty}^{\infty} e^u \exp\left(-\frac{u}{2} - \frac{\hat{\beta}_N^N e^u}{2}\right) du =
= \frac{\hat{\beta}_N^N}{2} \int_{-\infty}^{\infty} e^{2u} \exp\left(-\frac{\hat{\beta}_N^N e^u}{2}\right) du =
= \frac{\hat{\beta}_N^N}{2} \int_{z=\hat{\beta}_N^N}^{\infty} \sqrt{z} \exp\left(-\frac{\hat{\beta}_N^N z}{2}\right) dz > 0.
$$

Thus, we can use the result of Theorem 1.15 to calculate the asymptotic variance of $\hat{g}_{X,n}$, taking into account also that this estimator is consistent for $g(x)$ (see Theorem 1.8). For simplicity, we denote by $k := \frac{\hat{\beta}_N^N}{2}$ in the following calculations. We have already shown that

$$
\int_{\mathbb{R}} \psi'(y-g(x)) dF_{U_t}(y-g(x)) = k \int_{0}^{\infty} e^{-kz} \sqrt{z} dz =
= k \left[ -c e^{-kz} \right]_{0}^{\infty} \frac{c}{2} \int_{0}^{\infty} e^{-kz} \sqrt{z} dz =
= \frac{c}{2} \int_{0}^{\infty} e^{-kz} \sqrt{z} dz = \frac{c}{\sqrt{\hat{\beta}_N^N}}. \frac{\sqrt{2\pi}}{2}.
$$

Using the definition of $c$ made in Example 1.1, we finally get that

$$
\int_{\mathbb{R}} \psi'(y-g(x)) dF_{U_t}(y-g(x)) = \frac{1}{2}.
$$

Analogously we obtain that

$$
\int_{\mathbb{R}} \psi^2(y-g(x)) dF_{U_t}(y-g(x)) =
= c \int_{\mathbb{R}} \left( \frac{1}{2} - \frac{\hat{\beta}_N^N}{2} e^{(Y_t-g(x))} \right)^2 \exp\left(-\frac{u}{2} - \frac{\hat{\beta}_N^N e^u}{2}\right) du = \frac{1}{2}
$$

and the result is proved.

1.5.2 The asymptotic variance of $\hat{g}_{X,n}$

We are interested here in the explicit calculation from (1.13) of the asymptotic variance of $\hat{g}_{X,n}$ in the case of scaled $t_v$, $v > 4$, distributed innovations $Z_t$. 

\[ \square \]
Theorem 1.17. (Asymptotic variance of $\hat{g}_{x,n}^{(\nu)}$)
Suppose that the process \{X_{t-1}, Y_t\}_{t \geq 2} is from the model (1.1)-(1.2) and let the innovations $Z_t$ in (1.1) be scaled $\nu$, \( \nu > 4 \), distributed. Under the assumptions H4 to H7, the asymptotic variance of the local constant log-likelihood estimator $\hat{g}_{x,n}^{(\nu)}$, given as the unique solution of (1.6) equals

$$\sigma^2 = \frac{2(\nu + 3)}{\nu} \cdot \int_{\mathbb{R}} W^2(u) du$$

for every fitting point $x$, where the kernel $W(\cdot)$ is defined as in Section 1.3.

Remark. Note that for $\nu \to \infty$, the factor $V(\psi, G) = \frac{2(\nu + 3)}{\nu}$ in (1.15) converges to 2, which is consistent with the result found in (1.14) for a standard normal distribution.

Proof of Theorem 1.17. The result follows directly from Theorem 1.15. In the case of scaled $\nu$-distributed innovations $Z_t$, we have that assumptions H2-H3 hold (see Doukhan, 1994). Moreover, we have already shown in Examples 1.2-1.7 that the function $\psi(Y_t - g_x) = \rho'(Y_t, g_x)$ equals

$$\psi(Y_t - g_x) = \frac{1}{2} - \frac{\nu + 1}{2} \frac{1}{1 + \frac{\nu - 2}{\beta^2 e^{(Y_t - g_x)}}}.$$

Like in the normal case, knowing the explicit form of $\psi$, it is easy to verify that H8 is satisfied. Now, we want to show that H1 holds, too. This is done by proving that

$$\mathbb{E} \left[ \frac{\nu + 1}{2} \left( \frac{1}{1 + \frac{\nu - 2}{\beta^2 e^{(Y_t - g_x)}}} \right)^2 \cdot \frac{\nu - 2}{\beta^2 e^{(Y_t - g_x)}} \right]_{X_{t-1} = x} =$$

$$= \frac{\nu + 1}{2} \int_{\mathbb{R}} \left( \frac{1}{1 + \frac{\nu - 2}{\beta^2 e^{u}}} \right)^2 \frac{\nu - 2}{\beta^2 e^{u}} dF_U(u) > 0.$$

Therefore we can use the result of Theorem 1.15 to calculate the asymptotic variance of $\hat{g}_{x,n}^{(\nu)}$, considering also that this estimator is consistent for $g(x)$ (see Theorem 1.13). For simplicity, we denote by $k := \frac{\nu + 1}{2} \frac{\Gamma\left(\frac{\nu + 1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\sqrt{\pi}}$ in the
1.5. Asymptotic normality

following calculations. We have already shown that

$$
\int_{\mathbb{R}} \psi'(y - g(x)) dF_{U_t}(y - g(x)) = \frac{v + 1}{2} \int_{\mathbb{R}} \left( \frac{1}{1 + \frac{v-2}{\beta v e^u}} \right)^2 \frac{v - 2}{\beta v e^u} dF_{U_t}(u)
$$

$$
= c \int_{\mathbb{R}} \left( \frac{1}{2} - \frac{v + 1}{2} \frac{1}{1 + \frac{v-2}{\beta v e^u}} \right)^2 \left( 1 + \frac{\widetilde{\beta v}}{v - 2} e^u \right)^{-\left(\frac{v+1}{2}\right)} e^u du =
$$

$$
= k \int_{0}^{\infty} z^{-\frac{3}{2}} \frac{1}{(1 + \frac{1}{z})^2} dz = k \int_{0}^{\infty} \sqrt{z} (1 + z)^{-\left(\frac{v+3}{2}\right)} dz =
$$

$$
k \frac{4\sqrt{\pi}}{\psi(\frac{\nu}{2} + 1)} = \frac{v + 1}{4} \frac{\Gamma(\frac{\nu+1}{2})\Gamma(\frac{\nu}{2} + 1)}{\Gamma(\frac{\nu}{2})\Gamma(\frac{\nu+3}{2})} = \frac{v}{2(v + 3)},
$$

because of the definition of $\Gamma(x + 1) = x\Gamma(x)$, $\forall x > 0$. Analogously we obtain that

$$
\int_{\mathbb{R}} \psi^2(u) dF_{U_t}(u) =
$$

$$
= c \int_{\mathbb{R}} \left( \frac{1}{2} - \frac{v + 1}{2} \frac{1}{1 + \frac{v-2}{\beta v e^u}} \right)^2 \left( 1 + \frac{\widetilde{\beta v}}{v - 2} e^u \right)^{-\left(\frac{v+1}{2}\right)} e^u du =
$$

$$
= \frac{1}{4} - \frac{v + 1}{4} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu+3}{2})} + \frac{3(v + 1)^2}{16} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu+5}{2})} = \frac{v}{2(v + 3)},
$$

and the result is proved. \qed

1.5.3 Asymptotic variance of the local regression estimator

We now want to compare the asymptotic variances given by (1.14) and (1.15) in the normal and in the scaled $t_\nu$, $\nu > 4$, case respectively with the one obtained from a local regression estimation. In the local regression, we have that the estimator must be a solution of the constant local log-likelihood equation (1.6) with $\rho(Y_t, g_x) = (Y_t - g_x)^2$. Note that this happens exactly in the misspecified case when the variables $U_t$ are standard normally distributed. Therefore, in this case we get that

$$
\int_{\mathbb{R}} \psi'(y - g(x)) dF_{U_t}(y - g(x)) = 2 \int_{\mathbb{R}} dF_{U_t}(y - g(x)) = 2
$$

and

$$
\int_{\mathbb{R}} \psi^2(y - g(x)) dF_{U_t}(y - g(x)) = 4 \int_{\mathbb{R}} u^2 dF_{U_t}(u) = 4 \mathbb{E}[U_t^2] = 4 \sigma_{U_t}^2.
$$
Chapter 1. Local Likelihood for non-parametric ARCH(1) models

since $\mathbb{E}[U_t] = 0$, and from the result of Theorem 1.15 follows that the asymptotic variance equals

$$
\sigma_{\text{reg}}^2 = \sigma_{U_t}^2 \cdot \int_{\mathbb{R}} K^2(u) du \frac{d(x)}{d(u)}.
$$

(1.16)

An approximation of the term $V(\psi, G) = \sigma_{U_t}^2$ in (1.16) can be easily calculated by simulating. Since the estimator $\hat{\beta}_{x:n}^{N}$ given by (1.5) is consistent for $\beta_{x,n}^N + \log(f(x))$ also under pseudo-likelihood estimation (see the results of Theorem 1.11), we simulate the values for the factor $V(\psi, G)$ in (1.13) for $G \neq \mathcal{N}(0, 1)$ and $\psi(u) = \frac{1}{2} - \frac{\beta_{x,n}^N}{2} e^u$ proceeding as follows.

1. Generate $m$ variables $W_1, \ldots, W_m$ from the distribution $G \neq \mathcal{N}(0, 1)$. For the results of the following Figure 1.1, we take $m = 5 \cdot 10^5$ and $G = \sqrt{\frac{v-2}{v}} t_v, \ v > 4$.

2. Construct variables $U_t = \log(W_t^2) - \hat{\beta}_{x,n}^N, \ t = 1, \ldots, m$, where $\hat{\beta}_{x,n}^N$ is an estimate of $\beta_{x,n}^N$.

3. Since

$$
\mathbb{E}[\psi(U_t) \mid X_{t-1} = x] \approx \frac{1}{m} \sum_{i=1}^{m} \psi(U_i)
$$

and

$$
\mathbb{E}[\psi(U_t)^2 \mid X_{t-1} = x] \approx \frac{1}{m} \sum_{i=1}^{m} \psi(U_i)^2,
$$

approximate the factor $V(\psi, G)$ in (1.13) with

$$
V(\psi, G) \approx \frac{1}{m} \sum_{i=1}^{m} \psi(U_i)^2 \left( \frac{1}{m} \sum_{i=1}^{m} \psi'(U_i) \right)^2.
$$

where $\psi(u) = \frac{1}{2} - \frac{\beta_{x,n}^N}{2} e^u$.

The behaviour of the factor $V(\psi, G)$ of the asymptotic variance of the local constant log-likelihood estimators $\hat{\beta}_{x:n}^{N}$ and $\hat{\beta}_{x:n}^{r}$, and of the local regression estimator in the case of scaled $t_v$-distributed innovations $Z_t$ with degrees of freedom $v > 4$ is illustrated in the following Figure 1.1.

As we expect, the asymptotic variance of the estimator from a local regression is bigger than the one of the normal and $t_v$ local log-likelihood estimator in the range $v \in [6, +\infty]$. The reason is that in the case of innovations
1.5. Asymptotic normality

Figure 1.1: The behaviour of the factor $V(\psi, G)$ for the three estimators introduced in this Section in dependence of the degrees of freedom factor $v$ of the innovations distribution function. Note that the right points corresponds to the value $v = \infty$, i.e. $G = \mathcal{N}(0, 1)$.

$Z_t$ with distribution $G$ approximately standard normal, as for example the scaled $t_v$-distribution with $v$ big enough, the variables $U_t$ have a distribution function very different from a normal one as assumed by a local regression estimation. We also see that the “right” estimator (i.e. the estimator $\hat{\theta}^{t_v}_{X;n}$ with the true parameter $v$) is always minimal. A main disadvantage occurring when we work with the normal local log-likelihood estimator in the misspecified case of
scaled $t_v$-distributed innovations $Z_t$ is that the asymptotic variance increases more rapidly than for example when we work with the local regression estimator. A consequence of this fact is that the local regression estimator has a smaller factor $V$ for degrees of freedom $v < 5.5$. In other words, the local regression estimator and the $t_v$ local log-likelihood estimator are more stable than the normal one under pseudo-likelihood estimation. The big improvement in the asymptotic variance of the local log-likelihood estimators over the local regression happens when the innovations are standard normally distributed (2 vs 4.94).

1.6 Additional aspects

1.6.1 Uniform convergence of $\hat{g}_{x;n}^N$

As we have already seen in the Remark at the end of Theorem 1.3, to ensure the uniform convergence of the estimator $\hat{g}_{x;n}^N$ to $g(x)$ (called for a consistent estimation of $\beta$, see Theorem 1.19) we will need stronger assumptions. In addition to the conditions of Section 1.3, we assume that the kernel function $W(\cdot)$ is positive and fulfills a Lipschitz-condition

$$\exists \delta > 0, c < \infty : |W(u) - W(v)| \leq c |u - v|^{\delta} \forall u, v \in \mathbb{R}. \quad (K)$$

We also assume that the stationary density $d$ of the stationary process $\{X_t\}_t$ is strictly positive and uniformly bounded on a compact set $S$ of $\mathbb{R}$, and that the uniform equicontinuity condition

$$\forall \epsilon \exists \delta : \sup_{x \in S} \sup_{|u-x| \leq \delta} |E[\rho'(Y_t, g(x) + s) | X_{t-1} = u] f(u) - E[\rho'(Y_t, g(x) + s) | X_{t-1} = x] f(x)| \leq \epsilon, \quad (D)$$

is satisfied for all fixed $s$.

The last assumption we make is that there is an increasing sequence $\{m_n\}_{n \in \mathbb{N}}$ of positive integers such that

$$\exists A < \infty : \frac{n \phi_{m_n}}{m_n} \leq A, 1 \leq m_n \leq n, \forall n \geq 2, \quad (M)$$

where $\phi_n$ are the mixing coefficients introduced in (1.11) (an analogous condition exists also for the coefficients $\alpha_n$ introduced in (1.10)).

Note that in our particular case we do not have to make any further assumption on the function $\psi_x : \mathbb{R} \to \mathbb{R}$ given for every fitting point $x$ by

$$\psi_x(Y_t - g_x) = \rho'(Y_t, g_x) = \frac{1}{2} - \frac{1}{2} e^{(Y_t - g_x)} , \quad g_x \in C.$$
1.6. Additional aspects

Theorem 1.18. (Uniform convergence of $\hat{g}^{N}_{x,n}$)
Suppose that the same assumptions of Theorem 1.8 hold. If the conditions (K), (D) and (M) are satisfied and if

$$\frac{nh_{n}}{m_{n} \log(n)} \to \infty \text{ as } n \to \infty$$

holds, then for all $x \in S$ and sufficiently large $n$

$$\sup_{x \in S} |\hat{g}^{N}_{x,n} - g(x)| \xrightarrow{P} 0, \text{ as } n \to \infty.$$

Proof. The proof can be found in Collomb and Härdle (1986), Theorem 1. □

1.6.2 The estimation of $\beta$

If the distribution of the innovations $Z_t$ in (1.1) is not assumed to be known, a moment estimate for the constant $\beta$ (and consequently for $\hat{\beta}$) in (1.2) can be derived as follows (see also Chapter 2). From the identity $g(X_{t-1}) = \beta + \log (f(X_{t-1}))$, we obtain that $\sigma_t^2 = f(X_{t-1}) = \exp (-\beta) \cdot \exp (g(X_{t-1}))$. Thus, from the assumption that the innovations $Z_t$ have zero mean and variance one, it follows that

$$E X^2_t = \frac{X^2_t}{\sigma^2_t} = E \left[ \frac{X^2_t}{\exp (-\beta) \cdot \exp (g(X_{t-1}))} \right] = E \left[ Z^2_t \right] = 1.$$

Consequently, the moment estimate $\hat{\beta}$ can be derived from the empirical version

$$\frac{1}{n-1} \sum_{t=2}^{n} \left( \frac{X^2_t}{\exp (-\beta) \cdot \exp (g(X_{t-1}))} \right) = 1$$

and equals

$$\hat{\beta} = -\log \left( \frac{1}{n-1} \sum_{t=2}^{n} \left( \frac{X^2_t}{\exp (\hat{g}(X_{t-1}))} \right) \right), \quad (1.17)$$

where $\hat{g}(\cdot)$ is any consistent estimate of $g(\cdot)$ and naturally we assume that $\inf_{x} e^{\hat{g}(x)} > 0$.

In the following Theorem 1.19, we show that the moment estimate $\hat{\beta}$ given by (1.17) is consistent.
Theorem 1.19. (Consistency of \( \hat{\beta} \))

Suppose that \( \{X_t\}_{(t \geq 1)} \) is stationary, \( \alpha \)-mixing with coefficients \( \alpha(n) \) as in (1.10) so that

\[
\sum_{n=2}^{\infty} \alpha(n) \frac{1}{n^\tau} < \infty \text{ and } \mathbb{E}[|X_t|^{2+\tau}] < \infty \text{ for some } \tau > 0, \tag{A1}
\]

or \( \phi \)-mixing with coefficients \( \phi(n) \) as in (1.11) so that

\[
\sum_{n=2}^{\infty} \phi(n) \frac{1}{n^\tau} < \infty \text{ and } \mathbb{E}[|X_t|^{2+\tau}] < \infty \text{ for some } \tau > 0. \tag{A1}
\]

If \( \{X_t\}_{(t \geq 1)} \) has stationary density \( d \), suppose that we have uniform convergence of the estimators \( \hat{g}(X_{t-1}) \) of \( g(X_{t-1}) \):

\[
\forall \epsilon > 0 \exists \ n_0 \in \mathbb{N} \text{ such that } \sup_{n \geq n_0} \mathbb{P}\left[ \sup_{|x| \leq M} \left| \hat{g}(x) - g(x) \right| > \epsilon \right] < \epsilon, \tag{A2}
\]

where \( M \) is chosen such that \( \mathbb{P}[|X_t| > M] < \epsilon \). Assume moreover that

\[
\inf_{x} g(x) > -\infty, \tag{A3}
\]

then the moment estimator \( \hat{\beta} \) given by (1.17) is consistent.

Remark. Note that (A2) holds under suitable conditions, see for example Section 1.6.1.

In the proof of Theorem 1.19, we need the following Lemma 1.20 (better known also as a Corollary of the continuous mapping theorem).

Lemma 1.20. Let \( r(\cdot) \) be an ordinary, real, measurable function, and \( \{X_n\} \) a sequence of random variables. If \( X_n \xrightarrow{P} a \) and if \( r \) is continuous in \( a \), then \( r(X_n) \xrightarrow{P} r(a) \).

Proof. The result follows directly from the continuous mapping theorem; see for example Billingsley (1968), Theorem 5.1. \( \Box \)

Proof of Theorem 1.19. Rewriting the quotient \( \frac{1}{\exp \left( \hat{g}(X_{t-1}) \right)} \) in (1.17) as

\[
\frac{1}{\exp \left( \hat{g}(X_{t-1}) \right)} = \frac{1}{\exp \left( g(X_{t-1}) \right)} + \left( \frac{1}{\exp \left( \hat{g}(X_{t-1}) \right)} - \frac{1}{\exp \left( g(X_{t-1}) \right)} \right)
\]
1.7. Some numerical results

and using Taylor we obtain that
\[
\frac{1}{\exp(\hat{g}(X_{t-1}))} - \frac{1}{\exp(g(X_{t-1}))} - \frac{1}{\xi^2} \left( \exp(\hat{g}(X_{t-1})) - \exp(g(X_{t-1})) \right),
\]
where the point \(\xi\) is chosen such that
\[
|\xi - \exp(g(X_{t-1}))| < |\exp(\hat{g}(X_{t-1})) - \exp(g(X_{t-1}))|.
\]
Note that assumption (A3) ensures that \(\xi\) is far enough from zero. Thus, we can write the moment estimator \(\hat{\beta} = \hat{\beta}_n\) in (1.17) as
\[
\hat{\beta}_n = -\log \left( \frac{1}{n-1} \sum_{t=2}^{n} \left( \frac{X_t^2}{\exp(g(X_{t-1}))} \right) + \Delta_n \right),
\]
where the remainder term
\[
\Delta_n = -\frac{1}{n-1} \sum_{t=2}^{n} \left( \frac{X_t^2}{\xi^2} \left( \exp(\hat{g}(X_{t-1})) - \exp(g(X_{t-1})) \right) \right) \overset{P}{\to} 0
\]
thanks to the assumptions (A1) and (A2), Lemma 1.20 and the WLLN for the mixing process \(\{X_t\}_t\) (see Doukhan, 1994). Now, rewriting \(X_t^2\) as \(Z_t^2\sigma_t^2\) and \(\exp(g(X_{t-1}))\) as \(\sigma_t^2\exp(\beta)\) in the last formula for \(\hat{\beta}_n\), we obtain
\[
\hat{\beta}_n = -\log \left( \exp(-\beta) \cdot \frac{1}{n-1} \sum_{t=2}^{n} Z_t^2 + \Delta_n \right).
\]
Thus, from the assumption \(\mathbb{E}[Z_t^2] = 1\) in (1.1), using the WLLN for the independent innovations \(Z_t^2\), we get that
\[
\hat{\beta}_n \overset{P}{\to} -\log \left( \exp(-\beta) \right) = \beta \quad \text{as} \quad n \to \infty.
\]

1.7 Some numerical results

We consider here the performance of the local constant log-likelihood estimators for simulated and real data. We compare the results with the ones from a local regression as we have done in the last section. The problem of the choice of an optimal bandwidth is not considered here (see for example Loader, 1999, or Hart and Vieu, 1990), but we always report with the use of the bandwidth that minimize our performance measure.
1.7.1 Simulations

The model that we use for simulating data is as in (1.1) with function \( f(\cdot) \) given by

\[
 f(x) = 0.2 + 0.8x^2 \cdot \exp(1 - 0.5|x|). \tag{1.18}
\]

The distribution of the innovations \( Z_t \) in (1.1) is either standard normal or scaled \( \sqrt{\frac{v}{v-2}} Z_t \sim t_v, v > 4 \), so that \( Z_t \) has variance one.

<table>
<thead>
<tr>
<th>degrees of freedom ( v )</th>
<th>OS-L2 measure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>local reg.</td>
</tr>
<tr>
<td>( \infty )</td>
<td>20.262</td>
</tr>
<tr>
<td>10</td>
<td>8.359</td>
</tr>
<tr>
<td>8</td>
<td>5.283</td>
</tr>
<tr>
<td>6</td>
<td>1.325</td>
</tr>
<tr>
<td>5</td>
<td>6.639</td>
</tr>
</tbody>
</table>

Table 1.1: The OS-L2 measure for five independent realizations of \( n = 500 \) days from the model (1.1) with \( f(\cdot) \) given by (1.18) with various degrees of freedom parameters \( v > 4 \). The relative gains over the classical local regression estimation are given between parenthesis.

For quantifying the goodness of fit, we consider the following measure:

\[
\text{OS-L}_2 = \frac{1}{n-1} \sum_{t=2}^{n} \left| \sigma_t^2 - \hat{\sigma}_{Q_{t-1};n}^2 \right|^2, \quad Q_1^n \text{ a new test set (out-sample loss),} \tag{1.19}
\]

where \( \hat{\sigma}_{Q_{t-1};n} \) uses the model estimated from the data \( X^n_t \) but evaluates it on new test data \( Q^n_1 \) that is another independent realization of the data. The out-sample OS-L2 statistic (1.19) is a measure for predictive performance. The results for five independent realizations of \( n = 500 \) days from the model (1.1) with \( f(\cdot) \) given by (1.18) with various degrees of freedom parameters \( v > 4 \) for the normal local log-likelihood estimator \( \hat{\sigma}_{x;n}^N \) given by (1.8), the \( t_v \) local log-likelihood estimator \( \hat{\sigma}_{x;n}^{t_v} = \exp \left( \hat{\sigma}_{x;n}^{t_v} - \hat{\beta} \right) \) coming out from (1.4) and the local regression estimator are reported in Table 1.1.
1.7. Some numerical results

Note that the differences in the first row (i.e. for $G = \mathcal{N}$) between the normal and the $t_\nu$ local estimator are caused by the estimation of $\hat{\beta}$ (with the method introduced in Section 1.6.2) needed by the construction of $\hat{f}_{x;n}^v$.

As we expect, the local log-likelihood estimators consistently outperform the classical local regression estimation for all degrees of freedom parameters $\nu$ considered: the relative gain in the out-sample OS-L$_2$ measure is always bigger than 60%.

1.7.2 Three real data examples

We consider three financial instruments with 1000 daily negative log-returns $X_t = \log \left( \frac{P_t}{P_{t-1}} \right)$ (in percentages): from the German DAX index between January 18, 1994 and November 17, 1997; from the US DJIA index between December 11, 1995 and November 24, 1999; and from the BMW stock price between September 23, 1992 and July 23, 1996. We consider the normal local log-likelihood estimator $\hat{f}_{x;n}^\mathcal{N}$, again in comparison with a local regression estimation, and with the parametric ARCH(1) model, i.e. the function $f(\cdot)$ in (1.1) is given by

$$f(x) = \alpha_0 + \alpha_1 x^2,$$

where $\alpha_0, \alpha_1$ are real positive parameters.

Since the OS-L$_2$ measure introduced by (1.19) can not be calculated for real data, we measure goodness of fit with the following OS-PL$_2$ (out-of-sample prediction loss) statistic:

$$OS-PL_2 = \frac{1}{n-1} \sum_{t=2}^{n} \left| \hat{f}_{Q_t-1;n} - Q_t^2 \right|^2, \quad Q_t^1 \text{ a new test set},$$

with $\hat{f}_{Q_{t-1};n}$ using the estimated model from the data $X^n_t$ but evaluating at test data $Q^n_1 = X^{2n}_{n+1}$ with $n = 500$. Note that the OS-PL$_2$ criterion (and also others) for real data allows only to discriminate between volatility forecasts whit performance different in large orders of magnitude. Otherwise, small differences could be obscured by low signal to noise ratio; see for example Section 2.6.2 for more details. The results are summarized in the following Table 1.2.

On the contrary to the results of Table 1.1, the differences between the models of Table 1.2 are very small. For this reason, we can not reach any definitive conclusion, even if the normal local log-likelihood estimator yields better forecast results than a classical local regression, also for real data.
Chapter 1. Local Likelihood for non-parametric ARCH(1) models

<table>
<thead>
<tr>
<th>financial instrument</th>
<th>OS-PL(_2) measure</th>
<th>parametric ARCH(1)</th>
<th>local regression</th>
<th>normal local likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMW share price</td>
<td>3.80</td>
<td>3.90 (+2.6%)</td>
<td>3.74 (-1.6%)</td>
<td></td>
</tr>
<tr>
<td>DAX index</td>
<td>17.79</td>
<td>18.48 (+3.9%)</td>
<td>18.25 (+2.6%)</td>
<td></td>
</tr>
<tr>
<td>DJIA index</td>
<td>8.44</td>
<td>8.35 (-1.1%)</td>
<td>8.38 (-0.7%)</td>
<td></td>
</tr>
</tbody>
</table>

Table 1.2: The OS-PL\(_2\) measure for three financial instruments: the German DAX index, the US DJIA index and the BMW share price. The relative gains over the parametric ARCH(1) model are given between parenthesis.

1.8 Multiplicative ARCH models with \( p > 1 \) predictor variables

Analogously to Section 1.2, we consider as a starting point the non-parametric ARCH(p) model

\[
X_t = \sigma_t Z_t \quad (t \geq p + 1)
\]

\[
\sigma_t^2 = f(X_{t-1}, \ldots, X_{t-p}), \quad f : \mathbb{R}^p \rightarrow \mathbb{R}^+,
\]

(1.21)

where the innovations \( Z_t \) are independent identically distributed, with distribution function \( G \), zero mean, variance one and independent from \( \{X_s; s \leq t\} \). We make the same assumptions on the process \( \{X_t\}_{(t \geq p+1)} \) as in Section 1.2.

To estimate the function \( f \), we proceed here as follows. We assume that the function \( f \) is of the form

\[
f(X_{t-1}, \ldots, X_{t-p}) = f_1(X_{t-1})f_2(X_{t-2}) \cdots f_p(X_{t-p}),
\]

(1.22)

\[
f_i : \mathbb{R} \rightarrow \mathbb{R}^+, \ i = 1 \ldots p,
\]

i.e. the product of \( p \) functions depending on one predictor variable only. Analogous to Section 1.2, it is useful to transform the model (1.21)-(1.22) logarithmically. We get

\[
Y_t = \beta + \sum_{i=1}^{p} g_i(X_{t-i}) + U_t,
\]

(1.23)

where \( Y_t, \ \beta \) and \( U_t \) are defined as in (1.2), and \( g_i(X_{t-i}) = \log (f_i(X_{t-i})) \) for \( i = 1, \ldots, p \). Note that (1.23) belongs to the class of the generalized
additive models (GAM), see Hastie and Tibshirani (1990). For this reason we can estimate the different functions $g_i$ (and consequently the functions $f_i$ and the squared volatility $f$ of (1.21)) using a backfitting algorithm.

The goal is to minimize
\[
\sum_{t=p+1}^{n} \rho(Y_t, \{g_i(X_{t-i})\}_{i=1}^{p}) = \sum_{t=p+1}^{n} \left( - \log (f_U(Y_t - \beta - \sum_{i=1}^{p} g_i(X_{t-i}))) \right).
\]

Proceeding with a constant local log-likelihood estimation we get that for a given fitting point $x = (x_1, \ldots, x_p) \in \mathcal{C} \subset \mathbb{R}^p$

\[
\hat{g}_{i:n}(x) = \arg\min_{g} \sum_{t=p+1}^{n} W\left(\frac{x_i - X_{t-i}}{h_n}\right) \cdot R_{t,i}, \ i = 1, \ldots, p, \quad (1.24)
\]

where the residuals $R_{t,i}$ are given by $R_{t,i} = \rho(Y_t - \sum_{j \neq i} \hat{g}_{j:n}(X_{t-j}) - g)$. For more details and a explicit description of the use in this particular case of the backfitting algorithm, we refer to Widmer (2001).

Now, we want to show that the Remark at the end of Corollary 1.4 is still true also for $p > 1$. This is done in the next section.

### 1.8.1 The normal case

We assume here that the innovations $Z_t$ are standard normally distributed. From the result of Theorem 1.3, we have that the local constant log-likelihood estimator $\hat{g}_{i:n}^N(x)$ for the function $g_i$, $i = 1, \ldots, p$, and every fitting point $x = (x_1, \ldots, x_p)$ is given by

\[
\hat{g}_{i:n}^N(x_i) = \log \left( \frac{\sum_{t=p+1}^{n} W\left(\frac{x_i - X_{t-i}}{h_n}\right) \exp \left( Y_t - \sum_{j \neq i} \hat{g}_j(X_{t-j}) \right)}{\sum_{s=p+1}^{n} W\left(\frac{x_i - X_{t-i}}{h_n}\right)} \right).
\]

With a back transformation, we find that the estimator $\hat{f}_{i:n}^N(x_i)$ for the original function $f_i$ equals

\[
\hat{f}_{i:n}^N(x_i) = \exp \left( \hat{g}_{i:n}^N(x_i) \right) = \frac{\sum_{t=p+1}^{n} W\left(\frac{x_i - X_{t-i}}{h_n}\right) \exp \left( Y_t - \sum_{j \neq i} \hat{g}_j(X_{t-j}) \right)}{\sum_{s=p+1}^{n} W\left(\frac{x_i - X_{t-i}}{h_n}\right)} = \frac{\sum_{t=p+1}^{n} W\left(\frac{x_i - X_{t-i}}{h_n}\right) X_i^2}{\sum_{s=p+1}^{n} W\left(\frac{x_i - X_{t-i}}{h_n}\right) \prod_{j \neq i} f_j(X_{t-j})}, \ i = 1, \ldots, p.
\]
Note that, analogous to the case where only one predictor variable is involved, multiplicative models of the form (1.21)-(1.22) with Gaussian innovations $Z_t$ can be fitted with closed form expressions.

Analogously to the remark at the end of Corollary 1.4, exactly the same result can be obtained rewriting the model (1.21) as

$$X_t^2 = \sigma_t^2 + \eta_t = f_1(X_{t-1}) f_2(X_{t-2}) \cdots f_p(X_{t-p}) + \eta_t,$$

where $\eta_t = \sigma_t^2 (Z_t^2 - 1)$, and estimating the functions $f_i$ with a local regression using a backfitting algorithm for the current residuals

$$\frac{X_t^2}{\prod_{j \neq i} \hat{f}_j(X_{t-j})}, \ i = 1, \ldots, p.$$

### 1.8.2 Numerical results

We test the GAM model with normal local log-likelihood estimation of the last section on the same real financial instruments of Section 1.7.2, in comparison with a GAM model again given by (1.23) but with local regression estimation in (1.24) (i.e. $\rho(\cdot) = (\cdot)^2$). For quantifying the goodness of fit, we consider the same OS-PL$_2$ statistic introduced by (1.20). The results for different values of $p$ are summarized in Table 1.3.

The improvements of the GAM model with normal local log-likelihood over the GAM model with local regression are bigger than the ones of Table 1.2. On the other hand, we see that using more than one predictor in the estimation does not yield a relevant improvement in the OS-PL$_2$ statistic.

### 1.9 Concluding remarks

We have proposed a non-parametric local likelihood estimator for volatility in the ARCH(1) model (1.1) which leads to more accurate predictions than classical kernel regression smoothing. We have also shown how our estimation procedure can be generalized for multiplicative ARCH models with $p > 1$ predictor variables (1.21)-(1.22).

As supporting asymptotics, we have presented consistency and asymptotic normality results for our local likelihood estimator in the general situation where the distribution function of the innovations $Z_t$ may be mis-specified.

We analyze the results of our local likelihood estimator on simulated and real return time series and we confront the performance of the volatility forecasts with the ones from a classical local regression with respect to the out-of-
Table 1.3: The OS-PL$_2$ measure for the same three financial instruments of Table 1.2. The relative gains over the GAM model with local regression are given in parenthesis.

sample loss (simulations) and out-of-sample prediction loss (real data sets). More specifically, we have the following:

- Predicting the volatility using classical local regression for the variables $Y_t = \log(X_t^2)$ versus $X_{t-1}, \ldots, X_{t-p}$ in the transformed ARCH model (1.2) and back-transforming is not a good idea.

- A better strategy is to locally regress $X_t^2$ versus $X_{t-1}, \ldots, X_{t-p}$ directly in the quadratic ARCH model. We found that this is equal to making a local likelihood estimation under Gaussian innovations in the transformed ARCH model (1.2).

- The volatility local likelihood estimator under Gaussian innovations has
closed form solution.

- For real data, we found that the local likelihood estimator gives best (or at least equally good) volatility forecasts than the local regression one.

- For simulated data, the local likelihood estimator consistently outperforms the classical local regression one; the relative gain with respect to the out-of-sample loss (OS-L2) statistic is always bigger than 60%.
Chapter 2

Tree-Structured GARCH Models

2.1 Abstract

We propose a new GARCH model with tree-structured multiple thresholds for volatility estimation in financial time series. The approach relies on the idea of a binary tree where every terminal node parameterizes a (local) GARCH model for a partition cell of the predictor space. Fitting of such trees is constructed within the likelihood framework for non-Gaussian observations: it is very different from the well-known CART procedure for regression which is based on residual sums of squares. Our strategy includes the classical GARCH model as a special case and allows one to increase model-complexity in a systematic and flexible way. We derive a consistency result and conclude by simulation and real data analysis that the new method has better predictive potential than other approaches.

2.2 Introduction

We propose a new method for estimating volatility in stationary financial time series. The real data examples of interest are daily log-returns $X_t = \log(P_t/P_{t-1})$, where $P_t$ denotes the price of an asset at day $t$. Our modeling technique is parametric and potentially high-dimensional: it relies on estimating thresholds by using the idea of binary tree construction for partitioning a predictor space.
As a starting point, consider a nonparametric GARCH(1,1) model,

\[ X_t = \sigma_t Z_t \quad (t \in \mathbb{Z}), \]
\[ \sigma_t^2 = f(X_{t-1}, \sigma_{t-1}^2), \quad f : \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}^+, \quad (2.1) \]

where \((Z_t)_{t \in \mathbb{Z}}\) is a sequence of independent identically distributed innovation variables with \(\mathbb{E}[Z_t] = 0, \ Var(Z_t) = 1\) and \(Z_t\) independent from \(\{X_s; s < t\}\). The so-called volatility \(\sigma_t\) is then defined by

\[ \sigma_t^2 = \text{Var}(X_t|\mathcal{F}_{t-1}) = f(X_{t-1}, \sigma_{t-1}^2), \]

where \(\mathcal{F}_{t-1}\) denotes the \(\sigma\)-algebra (the information) of the variables \(\{X_s; s \leq t - 1\}\). The restriction to model the squared volatility \(\sigma_t^2\) as a function of the previous values \(X_{t-1}\) and \(\sigma_{t-1}^2\) only is natural in finance. Note that it still generates a dependence of \(\sigma_t^2\) from all previous observations \(\{X_s; s < t\}\) due to the recursive definition of \(\sigma_{t-1}^2\); this is the important difference between ARCH and GARCH models. Also, the implicit assumption in (2.1) that \(\mathbb{E}[X_t|\mathcal{F}_{t-1}] = 0\) is a reasonable approximation for many financial time series: the substantial modeling effort goes into the dominant volatility, although for real data, we usually subtract first a linear AR(1) estimate for the conditional mean. The simplest but often used example for (2.1) is the classical GARCH(1,1) model (Bollerslev, 1986),

\[ f(x, \sigma^2) = \alpha_0 + \alpha_1 x^2 + \beta \sigma^2, \quad \alpha_0, \alpha_1, \beta > 0. \quad (2.2) \]

Generally, the unknown function \(f(\cdot, \cdot)\) in (2.1) may be nonlinear and even not smooth; for example, an asymmetry in financial trading with positive and negative return values from the past implies asymmetric or even discontinuous behavior for \(f(x, y)\) around \(x = 0\). Estimation of \(f(\cdot, \cdot)\) in general is very difficult due to the non-observable volatility in the second argument. An iterative nonparametric estimation procedure has been proposed in Bühlmann and McNeil (1999). Its advantage is the general form for \(f(\cdot, \cdot)\), but quite a few theoretical issues are not rigorously settled yet. Its disadvantages are mainly poor performance at edges, including the high values of volatility which are of particular interest in practice, lack of ability to deal with non-Gaussian observations and sensitivity to the choice of smoothing parameters. Our approach here is more in the spirit of a sieve approximation with parametric models for the nonparametric function \(f(\cdot, \cdot)\). The approximation builds on the following principles:

1. it includes the classical GARCH(1,1) model as a simple special case;
2.2. Introduction

(2) it uses a binary tree type selection strategy to estimate thresholds (splits) for building up an approximating multiple threshold GARCH model. The binary tree construction, where every terminal node represents a (local) three-dimensional GARCH model, is based on the likelihood in model (2.1).

Item (1) has an important link to practice: there is a relatively strong belief that the classical GARCH(1,1) model is appropriate, despite its simplicity with only three parameters. Our tree-structured nested modeling strategy allows one to verify this by using known selection techniques for nested models: as we will see, there is potential to improve upon the classical GARCH(1,1) for real data and we will quantify such gains in terms of prediction accuracy for volatility, rather than testing structural properties of \( f(\cdot, \cdot) \).

The likelihood-driven tree method mentioned in item (2) marks an essential difference to CART (Breiman et al., 1984). Underlying an approximate normality assumption for observations, CART uses residual sum of squares. For financial data, the normality assumption for observations and the corresponding techniques is not appropriate and can result in very poor performance. Our approach resembles more the general tree fitting with the deviance criterion used by Clark and Pregibon (1993). Another difference to CART (or more general versions driven by deviance) is that our tree structured scheme for conditional variance estimation employs a three-dimensional (local) GARCH model in every terminal node from the binary tree; CART uses only one location parameter per node. Finally, our tree GARCH procedure models the function \( f(\cdot, \cdot) \) in (2.1) and hence the infinite past in terms of the observations, whereas CART (or versions thereof) in autoregressive modeling deals with a \( p \)-dimensional predictor space \( (p < \infty) \) from finitely many lagged observations. Our methodology is also markedly different from autoregressive threshold models (SETAR) for conditional expectations (Tong, 1990), as we focus on the conditional variance of very non-Gaussian observations, and because we allow for non-Markovian models.

Extending the GARCH(1,1) model in (2.2) in the direction of adding potentially high parametric complexity has not yet been considered. Other versions of GARCH(1,1) with three or four parameters and a GARCH model with one or two thresholds at fixed locations (Rabemananjara and Zakoian, 1993) have been proposed. But there seems to be no systematic flexible way to build up a class of models from the classical GARCH(1,1) in (2.2) to a potentially high dimensional approximation of the general model in (2.1). The paper here deals mainly with this latter task: we describe the methodology in Section 2.3, present a consistency result in Section 2.4 and demonstrate the new procedure on simulated and real data in Section 2.6.
2.3 Tree-structured GARCH estimation

We describe here our methodology for approximating \( f(x, \theta) \) in (2.1) by a piecewise linear function. The novel part is the estimation of thresholds in \( \mathbb{R} \times \mathbb{R}^+ \), the start- and end-points of piecewise approximating functions. The working model is

\[
X_t = \phi X_{t-1} + \sigma_t(\theta) Z_t, \quad \sigma_t^2(\theta) = f_\theta(X_{t-1}, \sigma_{t-1}^2(\theta)) (t \in \mathbb{Z}),
\]  

(2.3)

with \( \sigma_t(\theta) Z_t \) as in model (2.1), but the functional form \( f(\cdot, \cdot) = f_\theta(\cdot, \cdot) \) now parameterized by a threshold function, see formula (2.4) below. We also add here a linear autoregressive term for estimating a conditional mean (being of minor importance for many financial time series). The function \( f_\theta(\cdot, \cdot) \) is parameterized as a binary tree structured GARCH(1,1). It involves a partition

\[
P = \{ \mathcal{R}_1, \ldots, \mathcal{R}_k \}, \quad \bigcup_{j=1}^k \mathcal{R}_j = \mathbb{R} \times \mathbb{R}^+, \quad \mathcal{R}_i \cap \mathcal{R}_j = \emptyset (i \neq j),
\]

for the predictor space. For every partition cell \( \mathcal{R}_j \), we employ a GARCH(1,1) model: the parametric form of the function then depends on \( \mathcal{P} \),

\[
f_\theta(x, \sigma^2) = f_\theta^p(x, \sigma^2) = \sum_{j=1}^k \left( \alpha_{0,j} + \alpha_{1,j} x^2 + \beta_{j} \sigma^2 \right) I_{(x, \sigma^2) \in \mathcal{R}_j},
\]

(2.4)

where \( \theta \) denotes the parameter set \( \{ \alpha_{0,j}, \alpha_{1,j}, \beta_j; \ j = 1, \ldots, k \} \). For \( k = 1 \), we have the classical GARCH(1,1) model from (2.2). As we will discuss in Section 2.3.1, the partition \( \mathcal{P} = \{ \mathcal{R}_1, \ldots, \mathcal{R}_k \} \) is constructed from a binary tree: every terminal node represents a rectangular partition cell \( \mathcal{R}_j \) whose edges are determined by thresholds.

Figure 2.1 represents an example of a binary tree partition of the predictor space \( \mathbb{R} \times \mathbb{R}^+ = \{(x, \sigma^2); x \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+\} \): the partition cells, corresponding to terminal nodes in the tree, are \( \mathcal{R}_1 = \{(x, \sigma^2); x \leq d_1\}, \mathcal{R}_2 = \{(x, \sigma^2); x > d_1 \text{ and } \sigma^2 \leq d_2\}, \mathcal{R}_3 = \{(x, \sigma^2); x > d_1 \text{ and } \sigma^2 > d_2\} \).

The negative log-likelihood in the working model (2.3) is

\[
-\ell(\phi, \theta; X^n_T) = -\sum_{t=2}^n \log \left[ \sigma_t^{-1}(\theta) p_Z \left( \frac{X_t - \phi X_{t-1}}{\sigma_t(\theta)} \right) \right],
\]

(2.5)

where \( p_Z(\cdot) \) denotes the density of the innovation \( Z_t \). The log-likelihood is always considered conditional on \( X_1 \) and some reasonable starting value \( \sigma_1^2(\theta) \), e.g. \( \sigma_1^2(\theta) = \text{Var}(X_1) \).

The strategy of our flexible tree-structured GARCH estimation is as follows.
2.3. Tree-structured GARCH estimation

Figure 2.1: Example of a binary tree partition $\mathcal{P} = \{\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3\}$ of predictor space $\{(x, \sigma^2); x \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+\}$.

(a) The minimizing criterion is always the negative log-likelihood in (2.5) with innovation density $p_Z(\cdot)$, either specified (e.g. standard normal) or of parametric form such as scaled Student’s $t$ with unknown degrees of freedom. The parametric form of $\sigma_t(\theta)$ in (2.5) is always of the form of a threshold model in (2.4). As we will see in Section 2.4, the model and innovation density in (2.3) do not need to be true for good approximating properties.

(b) Optimization with threshold functions in item (a) becomes an estimation and model selection problem. The former is done by maximum likelihood. For the latter, an exhaustive search is computationally pro-
hibitive and we propose a tree-structured partial search: within a data-
determined tree structure, the optimal model is estimated using the AIC
criterion.

2.3.1 Forward entering of thresholds: growing the binary

tree

Forward entering of threshold variables, using a binary tree construction, in-
duces a partition for \( \mathbb{R} \times \mathbb{R}^+ \) as follows. A first threshold \( d_1 \in \mathbb{R} \) or \( \mathbb{R}^+ \) together with a component index \( t_1 \in \{1, 2\} \) partitions

\[
\mathbb{R} \times \mathbb{R}^+ = R_{left} \cup R_{right},
\]

where \( R_{left} = \{(x, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+; (x, \sigma^2)_{t_1} \leq d_1\} \) and \( R_{right} \) analogously but with the relation ‘\( > \)’ instead. Then, one of the partition cells \( R_{left} \) or \( R_{right} \) is again partitioned with a threshold \( d_2 \) and a component index \( t_2 \) in the same fashion. We iterate this procedure: for the \( n \)th iteration step, we specify a pair \((d_m, t_m)\) and an existing partition cell for further refinement by splitting it into two cells. The refinement of an existing partition \( P^{(old)} \) is always constructed according to the following rule:

\[
P^{(old)} = \bigcup_j R_j \text{ an existing partition} \quad \rightarrow \quad \text{pick an element } R_{j^*} \in P^{(old)}
\]

\[
\rightarrow \quad \text{split } R_{j^*} = R_{j^*,left} \cup R_{j^*,right} \text{ according to a pair } (d, t) \in \mathbb{R} \times \{1, 2\}
\]

\[
\rightarrow \quad P^{(new)} = \bigcup_{j \neq j^*} R_j \bigcup (R_{j^*,left} \cup R_{j^*,right}),
\]

with \((d, t)\) describing a threshold and component index, \( R_{j^*,left} = \{(x, \sigma^2) \in R_{j^*} \subset \mathbb{R} \times \mathbb{R}^+; (x, \sigma^2)_{t_1} \leq d\} \) and analogously for \( R_{j^*,right} \) with the relation ‘\( > \)’. The whole procedure then produces a partition \( P = \{R_1, \ldots, R_k\} \) which can be described as a binary tree whose terminal nodes represent the partition cells, see also Figure 2.1. This is conceptually as in CART (Breiman et al., 1984). But as discussed below, the data-driven construction of a tree-
structured partition and estimation of parameters are very different.

Our algorithm below constructs a binary tree, corresponding to a parti-
tion of \( \mathbb{R} \times \mathbb{R}^+ \), by optimizing the reduction of a conditional negative log-
likelihood.

Step 1. Compute the negative log-likelihood from (2.5) without partition-
ing (i.e. in the partition \( P^{(0)} = \mathbb{R} \times \mathbb{R}^+ \)) with

\[
\ell_{\theta^{(0)}}(x, \sigma^2) = a_0 + \alpha_1 x^2 + \beta \sigma^2, \quad \theta^{(0)} = (\alpha_0, \alpha_1, \beta),
\]
and derive from it the maximum likelihood estimates $\hat{\phi}(0), \hat{\theta}(0)$ using a quasi-Newton method, cf. Nocedal and Wright (1999). Set $m = 0$.

**Step 2.** Increment $m$ by one. Search for the best refined partition $\mathcal{P}^{(m)}_{opt}$ by binary splitting of a cell from $\mathcal{P}^{(m-1)}_{opt}$ as described in (2.6). The details are as follows:

(I) Given $\mathcal{P}^{(m-1)}_{opt} = \{R_1, \ldots, R_m\}$, consider a new partition $\mathcal{P}^{(m)}$, where

one partition cell $R_{j^*} \in \mathcal{P}^{(m-1)}$ is split into $R_{j^*} = R_{j^*,left} \cup R_{j^*,right}$ as in (2.6). The function associated with $\mathcal{P}^{(m)}$ is

$$f_{(\theta^{(m-1)}\backslash j^*}, \theta^*)(x, \sigma^2) = \sum_{j \neq j^*} (a_{0,j} + a_{1,j}x^2 + \beta_j \sigma^2) I_{[(x, \sigma^2) \in R_j]}$$

$$+ \sum_{i \in \{j^*_{left}, j^*_right\}} (a_{0,i}^* + a_{1,i}^*x^2 + \beta_i^* \sigma^2) I_{[(x, \sigma^2) \in R_i]},$$

where

$$\theta^{(m-1)}\backslash j^* = \{a_{0,j}, a_{1,j}, \beta_j; j = 1, \ldots, m, j \neq j^*\} \in (\mathbb{R}^+)^{3(m-1)},$$

$$\theta^* = \{a_{0,i}^*, a_{1,i}^*, \beta_i^*; i \in \{j^*_{left}, j^*_right\}\} \in (\mathbb{R}^+)^6.$$  

(II) Compute the minimal negative conditional log-likelihood in the refined partition $\mathcal{P}^{(m)}$, holding the parameter vector $\hat{\theta}^{(m-1)}\backslash j^*, \hat{\theta}^{(m-1)}\backslash j^*$ fixed,

$$\min_{\hat{\theta}^*} \left( -\ell_{\mathcal{P}^{(m)}}(\hat{\phi}^{(0)}, (\hat{\theta}^{(m-1)}\backslash j^*), \theta^*; X^n_2) \right),$$

by numerical minimization over $\theta^*$ using a quasi-Newton method. Here, $-\ell_{\mathcal{P}^{(m)}}$ is as in (2.5) with the function $f_{\mathcal{P}^{(m)}}(\cdot, \cdot)$ from (2.7). For this numerical minimization over $\theta^*$, use as starting values for the parameters $\theta^*$ in both new cells $R_{j^*,left}, R_{j^*,right}$ the components of $\hat{\theta}^{(m-1)}$ corresponding to cell $R_{j^*}$.

(III) Optimize (2.8) by varying $\mathcal{P}^{(m)}$ in (I) and recomputing (II). Denote the optimal refined partition by $\mathcal{P}^{(m)}_{opt}$.

**Step 3.** Compute the maximum likelihood in partition $\mathcal{P}^{(m)}_{opt}$. Minimize with a quasi-Newton method the negative log-likelihood in (2.5) with function $f_{\mathcal{P}^{(m)}_{opt}}(\cdot, \cdot)$ from (2.4) to obtain $(\hat{\phi}^{(m)}, \hat{\theta}^{(m)})$. For that purpose, use the starting values $\phi^{(m-1)}, \theta^{(m-1)}\backslash j^*$ and the minimizer $\hat{\theta}^*$ in (2.8) which is computed in Step 2.

**Step 4.** Repeat Steps 2 and 3 until $m = M$. This yields a partition $\mathcal{P}^{(M)}_{opt}$ corresponding to a large binary tree equipped with parameter estimates $(\hat{\phi}^{(M)}, \hat{\theta}^{(M)})$.
Chapter 2. Tree-Structured GARCH Models

The value $M$, corresponding to $M + 1$ partition cells or terminal nodes in the binary tree, is pre-specified in advance such that the binary tree is sufficiently large. With financial return data, choosing $M$ around 6 is often appropriate. The search for the splitting value in Step 2 is done on a grid: we propose grid-points that are empirical $\alpha$-quantiles of the data with $\alpha = i/\text{mesh}, \ i = 1, \ldots, \text{mesh} - 1$. We typically choose $\text{mesh} = 8$ or 16.

The conditional log-likelihood in (2.8) yields a substantial computational short-cut compared to a full likelihood approach. For every given partition $\mathcal{P}^{(m)}$, the numerical nonlinear minimization in (2.8) involves only the six-dimensional parameter $\theta^*$. Since our algorithm searches over many candidate partitions $\mathcal{P}^{(m)}$ in every iteration step $m$, a relatively fast nonlinear minimization is important. Finding the best split is thus determined by maximal reduction of the negative conditional log-likelihood. The parameter estimates in Step 3 are computed from the full likelihood. For $\mathcal{P}^{(m)}$, we take advantage of the fact that the starting values specified in Step 3 are very reasonable for obtaining a reliable and fast maximum likelihood estimate in a possibly high-dimensional parameter space.

2.3.2 Pruning the tree

The binary tree, or the partition $\mathcal{P}_{\text{opt}}^{(M)}$, constructed in Section 2.3.1 is too large, or too fine, respectively. We correct by pruning: we search for a best subtree (with respect to AIC in (2.9) below), which is often computationally feasible since $M$ around 6 is usually large enough for financial time series. Denote by $\tau$ the set of all binary subtrees from $\mathcal{P}_{\text{opt}}^{(M)}$: its elements are denoted by $\mathcal{P}_i$. Note that $\tau$ is generally larger than the set $\{\mathcal{P}_{\text{opt}}^{(0)}, \mathcal{P}_{\text{opt}}^{(1)}, \ldots, \mathcal{P}_{\text{opt}}^{(M)}\}$.

For every $\mathcal{P}_i$, we compute the maximum likelihood estimates $(\hat{\phi}^{\mathcal{P}_i}, \hat{\theta}^{\mathcal{P}_i})$ with a quasi-Newton method, according to (2.5) with $f^{\mathcal{P}_i}(\cdot, \cdot)$ of the form (2.4). Note that reasonable starting values are again at hand by going backwards from $(\hat{\phi}^{(M)}, \hat{\theta}^{(M)})$ in a stage-wise manner. We then consider the penalized negative log-likelihood or AIC statistic

$$\text{AIC}(\mathcal{P}_i) = -2\ell(\hat{\phi}^{\mathcal{P}_i}, \hat{\theta}^{\mathcal{P}_i}; X_2^n) + 2(\text{dim}(\hat{\theta}^{\mathcal{P}_i}) + 1)$$

(2.9)

as a measure for predictive performance. The additional contribution 1 in the penalty term arises whenever the conditional expectation parameter $\phi$ in (2.3) is estimated. Choose the binary tree, or the partition $\mathcal{P}$, minimizing (2.9). The final tree-structured GARCH model is thus given by (2.3) with $\hat{\phi}^{\mathcal{P}}$, and $f_{\hat{\phi}^{\mathcal{P}}}(\cdot, \cdot)$ as in (2.4) based on partition $\mathcal{P}$. 
2.4 Consistency

The AIC statistic in (2.9) can be replaced by any other sensible model selection criterion. We have experimented with two other versions but found that overall performance with AIC is satisfactory.

2.4 Consistency

We give here some supporting asymptotics for a threshold model specified by (2.3) with Gaussian innovations and (2.4). Extensions to non-Gaussian innovations with suitably nice densities will be analogous. We exclude the effect of model selection and assume that the model structure is fixed, i.e. the structure of a binary tree partition $P$ in (2.4) is specified. The unknown parameters of the model are then

$$\xi = (\phi, \theta, d)$$

for the autoregressive, GARCH (within a partition cell) and threshold parameters, respectively. The volatility can thus be written as

$$\sigma^2_t(\theta, d) = f^{P_d}_\theta(X_{t-1}, \sigma^2_{t-1}(\theta, d)),$$

where $f^{P_d}_\theta$ as in (2.4) but with $P = P_d$ parameterized with unknown thresholds $d$.

Since we view the tree GARCH model only as a suitable approximation for the data generating process, we argue here that our procedure yields consistent estimates for the best parameters $\xi_0$, projected on the model specified by (2.3) with Gaussian innovations and (2.10),

$$\xi_0 = \arg\max_\xi h(\xi), \quad h(\xi) = E \left[ \log \left( \sigma_t(\theta, d)^{-1} \varphi \left( \frac{X_t - \phi X_{t-1}}{\sigma_t(\theta, d)} \right) \right) \right].$$

We assume that $\xi_0$ is unique, e.g. $h(\cdot)$ is strictly concave. If the model is true, $\xi_0$ will be the true parameter. Although $\xi_0$ is unique, the maximum likelihood estimator (based on Gaussian innovations) $\hat{\xi}$ is not: given data, the log-likelihood is constant between observations with respect to the threshold parameters $d$. For theoretical purposes, we define $\hat{d}$ as the smallest values representing a maximum likelihood estimator.

**Theorem 2.1.** Assume that the data generating process $(X_t)_{t\in\mathbb{Z}}$ is stationary. Under the regularity conditions (A1)–(A5) given in Section 2.4.1, the maximum likelihood estimator in the model specified by (2.3) with Gaussian innovations and (2.10) satisfies $\hat{\xi} \to \xi_0$ in probability, as $n \to \infty$. 
Theorem 2.1 describes consistency in mis-specified models. The true model isn't necessarily of GARCH type as in (2.1); or the true model is of the form (2.1) but with non-normal innovations \((Z_t)_{t \in \mathbb{Z}}\). The interpretation of the consistency result is then as follows: the model with parameter \(\xi_0\) is the closest element in the class of threshold models, described by (2.3) and (2.10), to the true underlying stationary process with respect to the Kullback-Leibler divergence; see for example Shibata (1997). In particular, if the fitted model is true, then the true parameters are consistently estimated.

Asymptotic normality with convergence rate \(n^{-1/2}\) for the parameters \(\phi, \theta\) usually holds due to partial differentiability with respect to \(\phi, \theta\) of the log-likelihood function; see Pollard (1984, Ch. VII). Under the realistic assumption that the data generating model is not exactly of threshold-type form, the estimated thresholds \(\hat{d}\) will generally have a different limiting distribution and, provided that the true model has a smooth distribution, its convergence rate is slower than \(n^{-1/2}\), due to a non-continuous likelihood function with respect to the split point parameters. A general relevant theory for this phenomenon is given by Kim and Pollard (1990); see also Hansen (2000).

### 2.4.1 Assumptions and proof of Theorem 2.1

We first give and discuss a set of regularity conditions for Theorem 2.1.

(A1) The data generating process \((X_t)_{t \in \mathbb{Z}}\) is stationary, \(\beta\)-mixing (cf. Yu, 1994) with \(\beta(k) \leq C k^{-\delta}\) for some \(C, \delta > 0\). Moreover, we assume that \(\mathbb{E}|X_t|^{\nu} < \infty, \nu > 0\).

(A2) Consider the parameter set \(\mathcal{L} = \{((\theta, d); \sup_{t \in \mathbb{Z}} \sigma_t(\theta, d) = O_p(1)\}\), i.e. the set where the volatility doesn't 'explode'. Denote by \(\Xi = \{\xi; (\theta, d) \in \mathcal{L}, |\phi| < 1\}\). Assume that the best projected parameter vector \(\xi_0\) is an interior point of \(\Xi\).

(A3) The function \(f_\theta^d(\cdot, \cdot)\) in (2.10) and \((X_t)_{t \in \mathbb{Z}}\) are such that \(h(\cdot)\) in (2.11) has a unique maximizer and is continuous for \(\xi \in \Xi\).

(A4) Define the truncated squared volatility as

\[
\sigma_t^{\text{trunc}(p)}(\theta, d)^2 = f_{X_{t-1}} \circ f_{X_{t-2}} \circ \ldots \circ f_{X_{t-p}}(X_{t-p}^2),
\]

\[
f_x(\tau^2) = f^d_\theta(x, \tau^2),
\]

i.e. an approximation with \(p\) (instead of infinitely many) lagged observed values for the volatility. Assume that the function \(f^d_\theta(\cdot, \cdot)\) in
2.4. Consistency

(2.10) satisfies:

\[ \sup_{t,(\theta,d) \in \mathcal{L}} |\sigma_t(\theta, d) - \sigma_t^{trunc(p)}(\theta, d)| = O_P(\gamma(p)), \]

for some \( \gamma(p) \to 0 \) (\( p \to \infty \)).

(A5) The function \( f_{\theta}^{p,d}(\cdot, \cdot) \) in (2.10) is such that

\[ \inf_{x,\sigma^2,(\theta,d) \in \mathcal{L}} f_{\theta}^{p,d}(x, \sigma^2) > 0. \]

Assumption (A1) indicates that we do not require the data being generated from a GARCH-type model as in (2.1) or (2.3) and (2.4). (A2) is related to stationarity of the model. Continuity of \( h(\cdot) \) in (A3) is implied by smoothness of the distribution of \( (X_t)_{t \in \mathbb{Z}} \). (A4) requires that the non-Markovian volatility process can be approximated by the truncated Markovian model; this holds for example, if all \( |\beta_j| < 1 \) in (2.4) and \( (X_t)_{t \in \mathbb{Z}} \) having a smooth distribution. The assumption is related to the Shannon-McMillan-Breiman Theorem. (A5) implies strict positivity of the volatility function (which is reasonable).

**Proof.** (Proof of Theorem 2.1)

For \( n \) large enough, the maximizer of the likelihood is in \( \Xi \) (otherwise, the values \( \sigma_t(\hat{\theta}, \hat{d}) \) would explode implying a small likelihood).

Denote the log-likelihood (conditioned on the first observation and using any \( \sigma_1 \neq 0 \)) by

\[ \ell_n(\xi) = n^{-1} \sum_{t=2}^{n} \log \left\{ \sigma_t^{-1} \varphi \left( \frac{X_t - \phi X_{t-1}}{\sigma_t} \right) \right\}, \]

abbreviating \( \sigma_t(\theta, d) \) by \( \sigma_t \). Analogously, define the truncated log-likelihood as

\[ \ell_n^{trunc(p)}(\xi) = n^{-1} \sum_{t=p+1}^{n} \log \left\{ (\sigma_t^{trunc(p)})^{-1} \varphi \left( \frac{X_t - \phi X_{t-1}}{\sigma_t^{trunc(p)}} \right) \right\}. \]

We will show that

\[ \sup_{\xi \in \Xi} |\ell_n(\xi) - \ell_n^{trunc(p)}(\xi)| = O_P(\gamma(p)) \]  

(2.12)

with \( \gamma(p) \) as in assumption (A4). Similarly, denoting by

\[ h^{trunc(p)}(\xi) = \mathbb{E} \left\{ \log \left\{ \frac{1}{\sigma_t^{trunc(p)}(\theta, d)} \varphi \left( \frac{X_t - \phi X_{t-1}}{\sigma_t^{trunc(p)}(\theta, d)} \right) \right\} \right\}, \]
we get for the population version,
\[ \sup_{\xi \in \Xi} |h(\xi) - h^{\text{trunc}(p)}(\xi)| = O(y(p)). \tag{2.13} \]

We give the proofs for (2.12) and (2.13) at the end. Now, the idea is to work with the truncated log-likelihood which is of the form
\[ \ell_{n}^{\text{trunc}(p)}(\xi) = n^{-1} \sum_{t=p+1}^{n} g(X_{t}, \ldots, X_{t-p+1}; \xi), \]
\[ g(X_{t}, \ldots, X_{t-p+1}; \xi) = \log \left\{ \left( \sigma_{t}^{\text{trunc}(p)} \right)^{-1} \varphi \left( \frac{X_{t} - \phi X_{t-1}}{\sigma_{t}^{\text{trunc}(p)}} \right) \right\}. \]

By the mixing property of \( (X_{t})_{t \in \mathbb{Z}} \),
\[ \sup_{\xi \in \Xi} |\ell_{n}^{\text{trunc}(p)}(\xi) - h^{\text{trunc}(p)}(\xi)| = o_{p}(1), \tag{2.14} \]
for every \( p \), see Yu (1994). More details proving (2.14) are given at the end.

By (2.12)–(2.14) it follows that
\[ \sup_{\xi \in \Xi} |\ell_{n}(\xi) - h(\xi)| = o_{p}(1). \tag{2.15} \]

Now, use a sandwich argument:
\[ \sup_{\xi \in \Xi} |\ell_{n}(\xi) - h(\xi)| \geq |h(\xi)| \geq |\ell_{n}(\hat{\xi}) - h(\hat{\xi})| \]
\[ \geq \ell_{n}(\hat{\xi}) = \ell_{n}(\hat{\xi}) - |h(\hat{\xi})| + |h(\hat{\xi})| \]
where we have used that \( \hat{\xi} \) and \( \hat{\xi} \) are the maximizers of \( h(\cdot) \) and \( \ell_{n}(\cdot) \), respectively. Due to (2.15), the left and right hand side are asymptotically equal to \( h(\hat{\xi}) \), implying that \( h(\hat{\xi}) = h(\hat{\xi} + o_{p}(1)) \). Since \( h(\cdot) \) is continuous and \( \hat{\xi} \) its unique maximizer, it follows that \( \hat{\xi} = \hat{\xi} + o_{p}(1) \), i.e. consistency.

Proof of (2.12) and (2.13): using a first order Taylor expansion,
\[ \log(\sigma^{-1} \varphi(u/\sigma)) = \log(\tau^{-1} \varphi(u/\tau)) + h(u, \tilde{\sigma})(\sigma - \tau), \]
\[ |\sigma - \tilde{\sigma}| < |\sigma - \tau|, \]
\[ h(u, \tilde{\sigma}) = -\tilde{\sigma}^{-1} + \tilde{\sigma}^{-3} u^{2}. \]

Use this with \( u = X_{t} - \phi X_{t-1}, \sigma = \sigma_{t}, \tau = \sigma_{t}^{\text{trunc}(p)} \). Assumption (A5), implying that \( \sigma_{t}, \sigma_{t}^{\text{trunc}(p)} \) are bounded away from zero, and the moment
2.5. Estimation with generalized additive models (GAM)

condition in (A1) yield that

\[ \ell_n(\xi) = n^{-1} \sum_{t=p+1}^{n} \Delta_t(\xi) (\sigma_t - \sigma_t^{\text{trunc}(p)}), \]

\[ \sup_{t,\xi \in \mathcal{Z}} \left| n^{-1} \sum_{t=p+1}^{n} \Delta_t(\xi) (\sigma_t - \sigma_t^{\text{trunc}(p)}) \right| = O_P(1), \]

with \( \Delta_t(\xi) = h(X_t - \phi X_{t-1}, \tilde{\sigma}_t(\theta, d)) \). Now use (A4): since \( \gamma(p) \) is uniform with respect to \( t \) and \( (\theta, d) \), we obtain (2.12). Formula (2.13) follows similarly.

Proof of (2.14): apply Theorem 3.4 in Yu (1994). Consider the permissible class \( \mathcal{G} = \{g(\cdot; \xi); \xi \in \mathcal{Z}\} \). An envelope function \( G(\cdot) \) can be constructed straightforwardly using (A5), and integrability of \( G(\cdot) \) is implied by the moment assumption in (A1). Furthermore, the class \( \mathcal{G} \) consists of functions which are compositions of a smooth, with a piecewise (involving indicators) quadratic function (from GARCH model in a partition cell): a good bound for the metric entropy condition then follows.

2.5 Estimation with generalized additive models (GAM)

For confrontation purposes, we consider here a different estimation technique for the volatility in (2.1). A nonparametric estimate for \( \sigma_t^2 \) can be derived as follows.

1. Estimate the conditional mean \( \mu_t = \mathbb{E}[X_t|\mathcal{F}_{t-1}] \) by an AR(1) model,

\[ \hat{\mu}_t = \hat{\phi} X_{t-1} \]

with parametric estimate \( \hat{\phi} \) obtained from least squares fitting.

2. Compute \( Y_t = \log((X_t - \hat{\mu}_t)^2) \), \( t = 2, \ldots, n \).

3. In model (2.3), we have \( Y_t \approx \beta + \log(\sigma_t^2) + (\log(Z_t^2) - \beta) \) with \( \beta = \mathbb{E}[\log(Z_t^2)] \). Let \( \gamma_t = \beta + \log(\sigma_t^2) \). Fit a GAM model with the transformed data \( Y_2, \ldots, Y_n \),

\[ \hat{\gamma}_t = \hat{h}_1(X_{t-1}) + \hat{h}_2(X_{t-2}) + \cdots + \hat{h}_k(X_{t-k}), \ t = k + 1, \ldots, n, \]

with nonparametric estimates \( \hat{h}_i(\cdot) \) obtained from a least squares backfitting algorithm with response variables \( Y_t \), cf. Hastie and Tibshirani (1990). The optimal value of \( k \) is chosen by minimizing the AIC statistic for Gaussian additive modeling of \( Y_t \).
4. Back-transform \( \delta_t = \exp(\hat{\gamma}_t) \approx \exp(\beta)\sigma_t^2 = \frac{1}{\epsilon}\sigma_t^2 \) and build \( R_t^2 = (X_t - \hat{\mu}_t)^2/\delta_t \approx c Z_t^2 \). Thus, set

\[
\hat{\epsilon} = (n)^{-1} \sum_{i=1}^{n} R_t^2.
\]

5. Then, set

\[
\hat{\sigma}^2_t = \hat{\epsilon} \delta_t.
\]

6. Iterate steps 1.-5. Thereby use weighted estimation in step 1 with weights \( w_t = \frac{1}{\delta_t} \), where \( \hat{\sigma}^2_t \) is the estimate from the previous iteration. Stop iterating by monitoring convergence of \( \hat{\sigma}^2_t \) and \( \hat{\mu}_t \).

A related technique is given in Yang et al. (1999): they don’t use the log-transform but work with squared observations and dependent, but uncorrelated innovations.

### 2.6 Numerical results

We consider here the performance of tree-structured GARCH models for simulated and real data. We compare the results with the GARCH(1,1) estimate in (2.2) and a nonparametric generalized additive model for log-transformed squared data which is described in Section 2.5. We always report with the use of \( M = 5 \) in Step 4 from Section 2.3.1 and with a grid search using mesh = 8 as described in Section 2.3.1 (except in Table 2.4 where mesh = 16 is also used): these specifications lead to good tree-structured model fits, despite their simplicity.

#### 2.6.1 Simulations

The models that we use for simulating data are as in (2.1) with various \( f(\cdot, \cdot) \). One is a threshold model

\[
f(x, \sigma^2) = \begin{cases} 
0.1 + 0.5x^2, & \text{if } x \leq d_1 = 0, \\
0.2 + 0.2x^2 + 0.75\sigma^2, & \text{if } x > d_1 = 0, \sigma^2 \leq d_2 = 0.5, \\
0.8 + 0.5\sigma^2, & \text{if } x > d_1 = 0, \sigma^2 > d_2 = 0.5,
\end{cases}
\]

(2.16)

The parameters are chosen to mimic time series of real log-returns. Also, the first threshold \( d_1 = 0 \) allows for an asymmetry natural in finance. Another
model is a classical GARCH(1,1)

\[ f(x, \sigma^2) = 0.05 + 0.1x^2 + 0.85\sigma^2. \] (2.17)

A third model is neither GARCH nor threshold GARCH

\[ f(x, \sigma^2) = (0.1 + 0.2|x| + 0.9x^2) \cdot (0.8 \exp(-1.5|x|\sigma)) + 
(0.4x^2 + 0.5\sigma^2)^{3/4}. \] (2.18)

Finally, we consider a multiplicative model for which the GAM fitting procedure, described in Section 2.5, is consistent,

\[ X_t = \sigma_t Z_t \ (t \in \mathbb{Z}), \]
\[ \log \sigma_t^2 = \log(0.4 + 0.28|X_{t-1}|^3) - 0.15X_{t-2}^2, \] (2.19)

where \((Z_t)_{t \in \mathbb{Z}}\) is as in (2.1). The distribution of innovations is either standard normal \(Z_t \sim \mathcal{N}(0, 1)\) or scaled \(t_6, \sqrt{6} \mathcal{Z}_t \sim t_6\), so that \(Z_t\) has variance one. We always take sample size \(n = 1000\): for real daily data, this would correspond to about four years and stationarity would be expected to be approximately true.

Estimation always uses the knowledge that \(\mu_t = \mathbb{E}[X_t|\mathcal{F}_{t-1}] = 0\), i.e. \(\phi = 0\) in (2.3). Figures 2.2 and 2.3 display some results from the tree-structured GARCH model, in comparison with the classical fit from the well-known GARCH(1,1) in (2.2) and with the estimated generalized additive model (GAM) described in Section 2.5.

We observe the following in Figure 2.2. The tree structured GARCH fit using \(\mathcal{N}(0, 1)\) innovations overestimates the number of thresholds: the first two thresholds are approximately correct. An improvement is given by the tree-structured GARCH fit with scaled \(t_6\)-distributed innovations: the thresholds and also the maximum likelihood estimated degrees of freedom \(\hat{v} = 5.12\) for the true \(v = 6\) are very satisfactory. The classical GARCH(1,1) model with scaled \(t_6\)-distributed innovations yields \(\hat{v} = 4.37\) and of course, has no thresholds in the volatility surface. There is no surprise that the tree structured GARCH with scaled \(t_6\)-distributed innovations is best, since the true model is of this form. But the tree GARCH estimate with \(\mathcal{N}(0, 1)\) misspecified innovations fits, as a quasi-maximum-likelihood, still reasonably well.

Figure 2.3 highlights a desirable important feature: the tree-structured GARCH with scaled \(t_6\)-distributed innovations performs very well in regions of high conditional variance. This is not the case when we look at the classical GARCH(1,1) fit; and the GAM estimate described in Section 2.5 is very poor in regions of high volatility. Note the different scales for the various parts of Figure 2.3.
Figure 2.2: Function estimation of conditional variance. Top left: true conditional variance $f(x, \sigma)$ given by (2.16), plotted against $x$ and $\sigma$. This is used in model (2.3) with scaled $t_v$-distributed innovations for a data realization (data 4 from Table 2.1) which is the basis for the other panels. Top right: estimated conditional variance from classical GARCH(1,1) model with scaled $t_v$-distributed innovations ($v$ unknown). Bottom left: estimated conditional variance from tree GARCH model with standard normal innovations. Bottom right: estimated conditional variance from tree GARCH model with scaled $t_v$-distributed innovations ($v$ unknown).
2.6. Numerical results

Figure 2.3: Errors of estimates for conditional variance. Top left: realization of volatility from model (2.3) with (2.16) and $\mathcal{N}(0,1)$-distributed innovations (data 2 from Table 2.1). Top right: errors $\hat{E}_t = \hat{\sigma}_t^2 - \sigma_t^2$ from tree GARCH fit with scaled $t_\nu$-distributed innovations ($\nu$ unknown) versus true conditional variance $\sigma_t^2$. Bottom left: errors $\hat{E}_t$ from classical GARCH(1,1) fit with scaled $t_\nu$-distributed innovations ($\nu$ unknown) versus $\sigma_t^2$. Bottom right: errors $\hat{E}_t$ from GAM estimate (as described in Section 2.5) versus $\sigma_t^2$. 
For quantifying the goodness of fit, we consider various measures:

\[
\text{IS-L}_p = \sum_{t=1}^{n} |\sigma_t^2 - \hat{\sigma}_t^2|^p, \quad p = 1, 2, \quad \text{(in-sample loss)}
\]

the AIC statistic from (2.9),

\[
\text{OS-L}_p = \sum_{t=1}^{n} |\sigma_t^2 - \hat{\sigma}_t^2(Y_{1t}^{-1})|^p, \quad p = 1, 2,
\]

\[Y_1^n\] a new test set (out-sample loss),

where for OS-L, \(\hat{\sigma}_t^2(Y_{1t}^{-1})\) uses the model estimated from the data \(X^n_1\) but evaluates it on new test data \(Y_1^n\) that is another independent realization of the data. Both out-of-sample OS-L- and AIC-statistic are measures for predictive performance. The IS- and even more the OS-L-statistics are interesting measures for our simulations, but we can’t calculate them for real data. The ratio \(\hat{\sigma}_t^2/\sigma_t^2 = 1 + (\hat{\sigma}_t^2 - \sigma_t^2)/\sigma_t^2\) would measure a relative performance: in financial applications, the regions with large \(\sigma_t^2\) are of particular interest and we therefore prefer the absolute scale which does not down-weight errors where volatility is high. Detailed results for eight independent realizations from model (2.1) with various \(f(\cdot, \cdot)\) and from model (2.19) are reported in Tables 2.1-2.3.

As an out-of-sample statistic, we view OS-L_2 as the most important measure for simulations. It gives more weight to large deviations than the OS-L_1 criterion: therefore, it is more appropriate when focusing on regions of high volatility.

The tree-structured GARCH procedure consistently outperforms the classical GARCH(1,1): for data 5 (GARCH(1,1) realization), it doesn’t fit any thresholds and coincides with the GARCH(1,1) fit. Our new tree GARCH procedure compares favorably with the GAM estimate: it is much better for data 5 (GARCH(1,1) realization), data 1–4 (threshold GARCH realizations) and even for data 6 (GAM model), but a bit worse for data 7 and 8 (realizations with (2.18)). The GAM procedure has notorious difficulties in regions of high volatility which is exploited most by its poor OS-L_2 performance. The reason for this may be that least squares (back-) fitting is applied to strongly non-Gaussian data. The classical GARCH(1,1) can exhibit huge variation in out-sample accuracy OS-L, for example a poor performance in data 2. This may be due to a very flat log-likelihood at the observed data: see Zumbach (1999) who also proposes a remedy. Interestingly, the tree-structured GARCH model exhibits much more stability here.
### Table 2.1: Goodness of fit measures for eight simulations using the classical GARCH(1,1) in (2.2). The data 1-5 and 7-8 are from model (2.1): data 1-3 with (2.16) and \( \mathcal{N}(0, 1) \) innovations, data 4 with (2.16) and scaled innovations \( \sqrt{6/\nu} Z_t \sim t_\nu \), data 5 with (2.17) and \( \mathcal{N}(0, 1) \) innovations, data 7 and 8 with (2.18) and \( \mathcal{N}(0, 1) \) innovations. Data 6 is from model (2.19) with \( \mathcal{N}(0, 1) \) innovations. The likelihood for estimation is based on standard normal innovations (data 1–3, data 5–8); for data 4, we use standard normal innovation likelihood (upper part) and scaled \( t_\nu \) innovation likelihood with \( \nu \) unknown (lower part). Out-sample performances OS-L are evaluated at two independent test-sets.
<table>
<thead>
<tr>
<th>data</th>
<th>( \hat{d}_1 ) = -0.04 in ( x )</th>
<th>( \hat{d}_2 ) = 0.77 in ( \sigma^2 )</th>
<th>( \hat{d}_3 ) = 0.56 in ( \sigma^2 )</th>
<th>AIC</th>
<th>L₁</th>
<th>L₂</th>
<th>OS-L₁</th>
<th>OS-L₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1842.2</td>
<td>102.9</td>
<td>43.5</td>
<td>110.7</td>
<td>44.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>121.0</td>
<td>61.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>data 2</td>
<td>( \hat{d}_1 ) = -0.02 in ( x )</td>
<td>( \hat{d}_2 ) = 0.47 in ( \sigma^2 )</td>
<td>( \hat{d}_3 ) = 0.72 in ( x )</td>
<td>1862.7</td>
<td>74.5</td>
<td>25.2</td>
<td>70.8</td>
<td>28.2</td>
</tr>
<tr>
<td>data 3</td>
<td>( \hat{d}_1 ) = -0.01 in ( x )</td>
<td>( \hat{d}_2 ) = 0.19 in ( x )</td>
<td>( \hat{d}_3 ) = 0.98 in ( \sigma^2 )</td>
<td>1886.0</td>
<td>101.6</td>
<td>47.1</td>
<td>86.5</td>
<td>32.5</td>
</tr>
<tr>
<td>average</td>
<td>-</td>
<td></td>
<td></td>
<td>1863.6</td>
<td>93.0</td>
<td>38.6</td>
<td>94.2</td>
<td>40.1</td>
</tr>
<tr>
<td>data 4</td>
<td>( \hat{d}_1 ) = -0.01 in ( x )</td>
<td>( \hat{d}_2 ) = 0.39 in ( \sigma^2 )</td>
<td>( \hat{d}_3 ) = 0.31 in ( x )</td>
<td>1744.0</td>
<td>114.0</td>
<td>40.5</td>
<td>131.1</td>
<td>65.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>122.5</td>
<td>55.9</td>
</tr>
<tr>
<td>data 5</td>
<td>no thresholds</td>
<td></td>
<td></td>
<td>2935.8</td>
<td>85.1</td>
<td>20.5</td>
<td>67.4</td>
<td>9.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>82.6</td>
<td>16.5</td>
</tr>
<tr>
<td>data 6</td>
<td>( \hat{d}_1 ) = -0.23 in ( x )</td>
<td>( \hat{d}_2 ) = 0.24 in ( x )</td>
<td></td>
<td>2114.0</td>
<td>74.9</td>
<td>150.2</td>
<td>66.0</td>
<td>36.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>61.0</td>
<td>20.8</td>
</tr>
<tr>
<td>data 7</td>
<td>( \hat{d}_1 ) = -0.87 in ( x )</td>
<td>( \hat{d}_2 ) = 0.91 in ( x )</td>
<td>( \hat{d}_3 ) = 1.45 in ( \sigma^2 )</td>
<td>3350.4</td>
<td>327.9</td>
<td>301.1</td>
<td>330.1</td>
<td>321.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( \hat{d}_4 ) = 1.71 in ( \sigma^2 )</td>
<td></td>
<td></td>
<td></td>
<td>354.1</td>
<td>297.1</td>
</tr>
<tr>
<td>data 8</td>
<td>( \hat{d}_1 ) = -0.49 in ( x )</td>
<td>( \hat{d}_2 ) = 0.85 in ( x )</td>
<td></td>
<td>3372.0</td>
<td>252.4</td>
<td>117.2</td>
<td>265.8</td>
<td>128.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>273.0</td>
<td>129.7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 2.2:** Estimated thresholds \( \hat{d}_i \) and goodness of fit measures for the same eight simulations of Table 2.1 from the tree-structured GARCH\((1,1)\) model with standard normal distributed innovations, except data 4 (lower part) where scaled \( t_v \)-distributed innovations, with \( v \) unknown, are used.
### Table 2.3: Goodness of fit measures for the same eight simulations from Table 2.1 using the GAM model described in Section 2.5.

<table>
<thead>
<tr>
<th></th>
<th>GAM</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L₁</td>
<td>L₂</td>
<td>OS-L₁</td>
<td>OS-L₂</td>
</tr>
<tr>
<td>data 1</td>
<td>200.0</td>
<td>138.7</td>
<td>225.3</td>
<td>152.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>227.4</td>
<td>206.6</td>
</tr>
<tr>
<td>data 2</td>
<td>208.8</td>
<td>171.3</td>
<td>197.0</td>
<td>168.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>245.7</td>
<td>496.3</td>
</tr>
<tr>
<td>data 3</td>
<td>250.1</td>
<td>377.7</td>
<td>231.0</td>
<td>348.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>282.7</td>
<td>549.1</td>
</tr>
<tr>
<td>average</td>
<td>219.7</td>
<td>229.2</td>
<td>234.8</td>
<td>320.2</td>
</tr>
<tr>
<td>data 4</td>
<td>220.0</td>
<td>428.0</td>
<td>275.0</td>
<td>459.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>281.0</td>
<td>869.4</td>
</tr>
<tr>
<td>data 5</td>
<td>549.6</td>
<td>4353.0</td>
<td>416.4</td>
<td>987.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>451.2</td>
<td>1028.2</td>
</tr>
<tr>
<td>data 6</td>
<td>100.7</td>
<td>77.1</td>
<td>95.6</td>
<td>147.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>100.9</td>
<td>82.1</td>
</tr>
<tr>
<td>data 7</td>
<td>315.8</td>
<td>388.2</td>
<td>259.9</td>
<td>152.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>347.8</td>
<td>325.6</td>
</tr>
<tr>
<td>data 8</td>
<td>227.8</td>
<td>124.6</td>
<td>229.6</td>
<td>133.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>231.5</td>
<td>136.9</td>
</tr>
</tbody>
</table>
2.6.2 Two real data examples

We consider two financial instruments with 1000 daily negative log-returns \( X_t = -100 \log(P_t/P_{t-1}) \) (in percentages): from the German DAX index between January 18, 1994 and November 17, 1997; and from the BMW stock price between September 23, 1992 and July 23, 1996. We consider the tree-structured GARCH model, again in comparison with the GARCH(1,1) from (2.2) (both with the additional model term \( \phi X_{t-1} \) from (2.3) for \( \mathbb{E}[X_t \mid \mathcal{F}_{t-1}] \)), and with the GAM model described in Section 2.5.

Figure 2.4 shows the result for the DAX index from a tree GARCH fit with \( \mathcal{N}(0, 1) \)-distributed innovations. Three thresholds are fitted in the volatility surface. Graphical diagnostics for the residuals are satisfactory, with a tendency for heavier tails than standard normal.

For these real-data examples, we measure goodness of fit with the AIC statistic from (2.9) and with in- and out-sample losses for predicting centered second moments,

\[
\text{IS-PL}_2 = \sum_{t=1}^{n} \left( \hat{\sigma}_t^2 - (X_t - \hat{\mu}_t)^2 \right)^2 \quad \text{(in-sample prediction loss)},
\]

\[
\text{OS-PL}_2 = \sum_{t=1}^{n} \left( \hat{\sigma}_t^2 (Y_{1,t}^{t-1}) - (Y_t - \hat{\mu}_t(Y_{1,t}^{t-1}))^2 \right)^2,
\]

with \( \hat{\mu}_t = \hat{\phi} X_{t-1} \) and \( \hat{\mu}_t(Y_{1,t}^{t-1}) \), \( \hat{\sigma}_t^2 (Y_{1,t}^{t-1}) \) using the estimated model from the data \( X_1^n \) but evaluating at test data \( Y_{1,n}^n = X_{2,n}^n \) with \( n = 1000 \). A straightforward calculation shows that any stationary process \( (Y_t)_{t \in \mathbb{Z}} \), having finite second moments and conditional first and second moments, can be decomposed as

\[ Y_t = \mu_t + \sigma_t U_t, \]

\[ \mu_t = \mathbb{E}[Y_t \mid Y_{t-1}, Y_{t-2}, \ldots], \quad \sigma_t^2 = \text{Var}(Y_t \mid Y_{t-1}, Y_{t-2}, \ldots) \]

with uncorrelated mean zero variables \( U_t \) with \( \mathbb{E}[U_t^2 \mid Y_{t-1}, Y_{t-2}, \ldots] = 1 \). Hence,

\[ (Y_t - \mu_t)^2 = \sigma_t^2 + V_t, \quad V_t = \sigma_t^2 (U_t^2 - 1), \]

so that \( \mathbb{E}[V_t \mid Y_{t-1}, Y_{t-2}, \ldots] = 0 \). When ignoring the estimation effect of \( \hat{\mu}_t \), the quantity \( n^{-1} \text{OS-PL}_2 \) approximates

\[ \mathbb{E}_{\mathbb{Y}}[(\hat{\sigma}_t^2 (Y_{1,t}^{t-1}) - \sigma_t^2)^2] + \mathbb{E}[V_t^2]. \]
Figure 2.4: Results for negative log-returns of the DAX index from the optimal tree-structured GARCH model with $\mathcal{N}(0, 1)$-distributed innovations. Top left: estimated function $\hat{f}(x, \sigma)$ for the conditional variance, plotted against $x$ and $\sigma$. Top right: residuals $\hat{Z}_t = (X_t - \hat{\mu}_t)/\hat{\sigma}_t$ from the tree GARCH fit versus time $t$. Bottom left and right: autocorrelation function of the absolute residuals $|\hat{Z}_t|$ and normal-plot for the residuals $\hat{Z}_t$, respectively.
The left term is our target, whereas the right term represents the noise variance in the problem. For real data, the variance $\mathbb{E}[V_t^2]$ seems often dominant and differences of various volatility estimates in terms of $\mathbb{E}_Y[(\hat{\sigma}_t^2(Y_{t-1}^i) - \sigma_t^2)^2]$ will be masked. It indicates that the evaluation of volatility forecasts for real data is difficult. Thus, the criterion OS-PL$_2$ (and also others) for real data allows only to discriminate between volatility forecasts whose performance is different in large orders of magnitude.

<table>
<thead>
<tr>
<th>tree-structured GARCH</th>
<th>AIC</th>
<th>IS-PL$_2$</th>
<th>OS-PL$_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DAX</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\hat{d}_1 = -0.35$ in $x$</td>
<td>2776.2</td>
<td>8555.1</td>
<td>20001.3</td>
</tr>
<tr>
<td>$\hat{d}_2 = 1.24$ in $\sigma^2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\hat{d}_3 = 1.66$ in $\sigma^2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\hat{d}_1 = -0.35$ in $x$</td>
<td>2764.4</td>
<td>8507.6</td>
<td>20535.6</td>
</tr>
<tr>
<td>$\hat{d}_2 = -0.89$ in $x$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\hat{d}_3 = 1.66$ in $\sigma^2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\hat{d}_4 = -0.51$ in $x$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>BMW</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\hat{d}_1 = -0.32$ in $x$</td>
<td>3155.0</td>
<td>12059.2</td>
<td>15111.6</td>
</tr>
<tr>
<td>$\hat{d}_2 = 1.11$ in $\sigma^2$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| **Table 2.4**: Estimated thresholds $\hat{d}_i$ and goodness of fit measures for negative log-returns from the DAX index and the BMW stock price. The tree GARCH model (with $\mathcal{N}(0, 1)$-distributed innovations) is fitted with mesh = 8 (DAX, upper part; BMW) and mesh = 16 (DAX, lower part). The GARCH(1,1) model is with $\mathcal{N}(0, 1)$-distributed innovations.

We view OS-PL$_2$ as the most important measure, followed by AIC: as an in-sample loss, IS-PL$_2$ is not very relevant. One of the tree-structured mod-
2.6. Numerical results

els improves upon the classical GARCH(1,1) (both with \( \mathcal{N}(0, 1) \)-distributed innovations) for the DAX index; for the BMW data, the two procedures have about the same performance. This is consistent with a common belief that classical GARCH(1,1) is better for individual prices than indices. The GAM estimate has the poorest performance in both data sets: for the BMW series, its behavior is extremely poor.

2.6.3 Summarizing numerical results

The tree-structured GARCH procedure often outperforms the estimate from the GARCH(1,1) model in (2.2) and the nonparametric GAM fit described in Section 2.5: such better performance is with respect to many goodness of fit and graphical criteria. More specifically:

(i) For real data, the tree-structured GARCH procedure was slightly best for log-returns from an index (DAX) and equally good as the classical GARCH(1,1) fit for an individual share (BMW). The GAM estimate was poorest on real data. Despite the difficulties with discriminating performances for real data (see Section 2.6.2), we found strong evidence that the GAM method is not competitive.

(ii) For simulated data, the tree-structured GARCH is clearly better than classical GARCH(1,1). In one case (data 7), the GAM fit was found best; but overall, GAM was poorer and it may be very unstable, resulting in extremely low performance.

(iii) The tree-structured GARCH estimate may be much better in the interesting regions where the true volatility is high, see Figure 2.3. The nonparametric GAM fit can be extremely poor in regions of high volatility (this problem doesn’t disappear when trying other smoothing parameters).

(iv) The tree-structured fitting procedure is sometimes slightly improved by assuming scaled \( t_\nu \)-distributed innovations \( Z_t \) for the likelihood in (2.5), provided that the underlying innovations are heavier tailed. This is weakly evident in Figure 2.4 for real data. See also Table 2.2.

(v) The AIC-statistic is an indicator for ranking out-sample performance; and pruning with AIC in (2.9) works reasonably well.

Items (i), (ii) and (iii) indicate a strong advantage of parametric, likelihood based methods over nonparametric least squares smoothing techniques such
as the GAM specification used here or multiplicative nonparametric ARCH models in Hafner (1998) or Yang et al. (1999). As pointed out in (iv), the likelihood approach can be easily modified to heavier tailed innovations inducing then even more heavy tails for the observations in the model. If performance is judged with a criterion putting emphasis on accurate prediction in high volatility regions, our tree based GARCH model seems clearly best among all alternative methods considered here.

2.6.4 How appropriate is GARCH(1,1) for daily returns?

The GARCH(1,1) model in (2.2) is very popular for analyzing daily log-returns of financial assets: it is often argued that it performs well despite that it has only three parameters describing a very low-dimensional model for sample size in the range of 1000. We quantify here the possible gains by using the flexible tree-structured GARCH model. In examples of daily log-return stock data, the new tree GARCH procedure was never found to be significantly worse than GARCH(1,1). A better performance for volatility forecasting may be masked by the high noise variance in evaluating predictions, see Section 2.6.2. In simulations, where evaluation of procedures is not obscured, we found substantial gains with tree-structured GARCH over GARCH(1,1). Particularly, the tree-structured GARCH seems to have an advantage in periods of high volatility. With real data, the first split has always been found in the x-axis (for the first lagged log-return) around zero: this is compatible with the interpretation that there is an asymmetric behavior depending on the sign of the previous log-return.

2.7 Concluding remarks

We propose a tree-structured GARCH model which is more flexible and accurate for prediction of volatility in financial time series than the well-known GARCH(1,1). The modeling strategy includes the classical GARCH(1,1) as a special case (no thresholds) and allows to increase complexity in a systematic way. On finite data, the new method compares favorably with a nonparametric technique based on additive models: especially in the interesting regions where the true volatility is moderate or large.

As supporting asymptotics, we present a consistency result about estimation of a best tree-structured GARCH model for approximating a general stationary process, see Section 2.4. More refined statistical inference is difficult due to the non-continuous nature of thresholds or trees: limiting distri-
2.7. Concluding remarks

Butions of estimated thresholds are typically non-Gaussian, see the discussion following Theorem 2.1. If the primary goal is construction of better volatility forecasts, which in turn can be used for dynamic risk management (cf. McNeil and Frey, 2000), we choose the route to optimize an information or complexity criterion rather than the somewhat inappropriate structural tool of testing for a model structure. As a simple solution, we use the AIC criterion: we have gained evidence on numerical examples that it can be used as a reasonable guideline.

Our univariate tree-structured GARCH procedure has a straightforward application in multivariate models where the conditional variance of an individual series is modeled as a function of the individual lagged values and individual lagged volatility. The multivariate cross-dependence is then modeled with cross-dependent innovations. For example, individual tree GARCH models for volatilities lead to the attractive version of the multivariate, constant conditional correlation model (Bollerslev, 1990) presented in Section 3.5.1. Another straightforward possible extension of our methodology is tree structured GARCH(\(p, q\)) modeling with \(p > 1\) or \(q > 1\). As already mentioned in Section 2.2, this may be of minor importance since the general model in (2.1) is in vogue and believed to capture the most important aspects of the underlying mechanism.
Part II

Multivariate volatility estimation
Chapter 3

Synchronizing Multivariate Financial Time Series

3.1 Abstract

Prices or returns of financial assets are most often collected in local times of the trading markets. The need to synchronize multivariate time series of financial prices or returns is motivated by the fact that information continues to flow for closed markets while others are still open. We propose here a synchronization technique which takes this into account.

Besides the nice interpretation of synchronization, the method potentially increases the predictive performance of any reasonable model. We found empirically that this was the case for the CCC-GARCH(1,1) model for a 7-dimensional time series of daily exchange rate returns. Since multivariate analysis is generally important for analyzing time-changing portfolios and for better portfolio predictions (even when portfolio weights are time-constant), synchronization is a valuable technique for a variety of problems with multivariate financial data.

3.2 Introduction

We propose here a synchronization of daily data in real global portfolios. Proceeding as in Burns et al. (1998), our general approach recognizes that even when markets are closed, the asset values may change before the market reopens. Synchronizing data involves estimates of asset values at a specified
(synchronization) time point in every day; we always use the closing time of the New York stock exchange, i.e. 4 pm local New York time, as the synchronization time point. The estimated asset values at the same synchronization time across markets are called synchronized. Having constructed synchronized data, we then advocate any reasonable multivariate model. We consider here the CCC-GARCH(1,1) model (Bollerslev, 1990) with time varying conditional variances and covariances but constant conditional correlations for synchronized data which is a different and new model for the original asynchronous data. This new model is called synchronous CCC-GARCH(1,1) and is presented in Section 3.3.

In Section 3.4, we compare empirically the performance of the results obtained with synchronized and asynchronized data, using the same multivariate CCC-GARCH(1,1) model. The analysis is for a real global portfolio with negative daily log-retuns (in percentages) \( X_{t,i} = -100 \cdot \log \frac{S_{t,i}}{S_{t-1,i}} \) of seven indices all over the world, where \( S_{t,i} \) denotes the price of the asset \( i \) at day \( t \) (time synchronized or asynchronized, respectively). In Section 3.6, we discuss the influence of synchronization on the calculation of risk measures such as VaR and expected shortfall.

The resulting gains when using synchronous data are sometimes considerable, depending on how we measure performance. We argue in Section 3.5 that these relevant gains are not due to strong model-misspecification of the volatilities in the CCC-GARCH(1,1) model with asynchronous data. More sophisticated threshold models with asynchronous data like the multivariate extension of the univariate tree-structured GARCH model of Chapter 2, which constructs a potentially high dimensional approximation of a general non-parametric CCC model, yields only marginal improvements over the standard CCC-GARCH(1,1) model. Thus, it emphasizes the power of synchronizing data in a first step.

Our empirical comparisons in Sections 3.4-3.6 also include univariate approaches for a portfolio index price

\[
P_t = \sum_i \alpha_i S_{t,i} \tag{3.1}
\]

modelled by previous \( P_{t-1}, P_{t-2}, \ldots \). Note that in the more realistic case where the weights \( \alpha_i = \alpha_{t,i} \) are depending (deterministically) on \( t \), stationarity of all \( \{S_{t,i}\} \) is not implying stationarity of the portfolio prices \( \{P_t\} \); but a multivariate analysis of \( \{S_{t,i}\}; i = 1, \cdots, M \) still yields the (conditional) distribution of \( P_t \) (given the past). It turns out that even when focusing on a time-constant portfolio index as in (3.1), multivariate approaches are better than sophisticated univariate modelling of \( P_t \) based on \( P_{t-1}, P_{t-2}, \ldots \), which
is exposed to an information loss by averaging previous individual prices.

Summarizing, we collect strong empirical evidence that:

1. the multivariate modelling approach is superior over univariate models for a portfolio index as in (3.1). In particular, we find that univariate modelling yields estimates which are too conservative.

2. synchronizing data, proposed here as a novelty, leads to further improvements in multivariate modelling and in calculating risk measures.

3.3 The synchronous CCC-GARCH(1,1) model

3.3.1 Synchronization of the data

Consider a global portfolio including stocks traded in New York and London. At the closing time of the trading in New York, the value of the portfolio should be measured with an estimate of the value of the London stocks at the closing time in New York. For example, to take the closing prices of the London stocks at one day when the US market goes down 1 percent (after London closes) for pricing the portfolio at New York closing time is highly unrealistic. It will follow that the US share of the portfolio declines today while the London share declines tomorrow. We associate with synchronization some estimates of the prices of the share traded in London at the closing time in New York (from the viewpoint of a British investor, the data could also be synchronized at the closing time in London).

We denote by $S_{t,j}$, $j = 1, \cdots M$ the continuous time price of an asset $j$. The time $t$ is here always measured as New York local time (in units of days) and $t \in \mathbb{N}$ corresponds to 4:00 pm New York local time on day $t$. For example, $S_{1,1}$ denotes the price of an asset of the NYSE at 4:00 pm New York local time on the first day. Since 4:00 pm corresponds to 9:00 pm in London and since London closes at 5:00 pm, 4 hours before New York closes, the observed closing price of an asset in London on the first day would be denoted by $S_{0.83,2}$. Generally, the observed data is taken at closing times of different markets. It has the structure

$$S_{t,j} (j = 1, \cdots, M),$$

where

$$t_j = k - c_j \ (k = 1, 2, \cdots, T; \ 0 < c_j \leq 1).$$

We always synchronize to the closing time (in New York) of asset $j = 1$ so that $c_1 = 0$ and $t_1 = 1, 2, \cdots, T$. The goal is to construct synchronized prices
with \( t = 1, 2, \ldots, T \) for all \( j \). These prices, or returns thereof, are more appropriate for many multivariate discrete time series models.

We define the synchronized prices \( S^s_{t,j} \) by

\[
\log (S^s_{t,j}) = \mathbb{E} \left[ \log (S_{t,j}) \mid \mathcal{F}_t \right],
\]

(3.2)

where \( \mathcal{F}_t = \left\{ S_{t_j,j} ; t_j \leq t, \ j = 1, \ldots, M \right\} \).

The logarithms are used to be consistent with continuously compounded returns, and \( \mathcal{F}_t \) is the complete information of all recorded prices up to time \( t \). We usually only need \( S^s_{t,j} \) for discrete \( t \in \mathbb{N} \).

Clearly, if the closing price \( S \) is observed at time \( t \in \mathbb{N} \), then its conditional expectation given \( \mathcal{F}_t \) is the observed price. This is the case for the stocks from New York. If the market closes before \( t \), then its past prices and all the other markets are potentially useful in predicting \( S \) at time \( t \).

As a simplifying but reasonable approximation, we assume

\[
\log (S^s_{t,j}) = \mathbb{E} \left[ \log (S_{t,j}) \mid \mathcal{F}_t \right] = \mathbb{E} \left[ \log (S_{t+j,1}) \mid \mathcal{F}_t \right],
\]

(3.3)

for all \( t \in \mathbb{N} \) with \( t_j \leq t \leq t_j + 1 \). The first equality holds by definition in (3.2). Thus, given the information at time \( t \), the next predicted future values are given by the log-transformed synchronized prices.

Now, denote the vector of negative log-returns (in percentages), in different markets and at various time points on day \( t \) as \( X_t \),

\[
X_t = -100 \cdot \begin{pmatrix}
\log \left( \frac{S_{t,1}}{S_{t-1,1}} \right) \\
\vdots \\
\log \left( \frac{S_{t,M}}{S_{t-1,M}} \right)
\end{pmatrix}
= -100 \cdot \left( \log (S_t) - \log (S_{t-1}) \right),
\]

(3.4)

where \( t = (t_1, t_2, \ldots, t_M) \) is a multi-index.

We define the synchronized returns as the change in the logarithms of the synchronized prices

\[
X^s_t = -100 \cdot \begin{pmatrix}
\log \left( \frac{S^s_{t,1}}{S^s_{t-1,1}} \right) \\
\vdots \\
\log \left( \frac{S^s_{t,M}}{S^s_{t-1,M}} \right)
\end{pmatrix}
= -100 \cdot \left( \log (S^s_t) - \log (S^s_{t-1}) \right), \ t \in \mathbb{N}.
\]

(3.5)
3.3. The synchronous CCC-GARCH(1,1) model

The synchronized returns are depending on unknown conditional expectations and have to be modelled (and estimated). We assume a simple "auxiliary" multivariate AR(1) model for the synchronization, given by

\[ X_t = A \cdot X_{t-1} + \epsilon_t, \]  

where the innovations \( \epsilon_t \) are i.i.d. \( \mathcal{N}(0, \Sigma) \), independent from \( \{X_s; \; s < t\} \) and \( A \) a \( M \times M \) matrix. The assumption of Gaussian innovations without any stochastic heteroscedasticity is unrealistic. But the model (3.6) is still useful for estimating the structure of synchronization, i.e. a sparse matrix \( A \) with many entries equal to zero as discussed below.

Substituting (3.3), (3.4) and (3.6) into (3.5) gives the synchronized returns as

\[
X_t^x = -100 \cdot \left( \log(S_t^x) - \log(S_{t-1}^x) \right) = \\
= -100 \cdot \left( \mathbb{E} [\log(S_{t+1}) | \mathcal{F}_t] - \mathbb{E} [\log(S_t) | \mathcal{F}_{t-1}] \right) = \\
= -100 \cdot \left( \mathbb{E} [\log(S_{t+1}) - \log(S_t) | \mathcal{F}_t] - \\
\quad - \mathbb{E} [\log(S_t) - \log(S_{t-1}) | \mathcal{F}_{t-1}] + \log \left( \frac{S_t}{S_{t-1}} \right) \right) = \\
= \mathbb{E} [X_{t+1} | \mathcal{F}_t] - \mathbb{E} [X_t | \mathcal{F}_{t-1}] + X_t = X_t + A \cdot X_{t-1},
\]

and thus

\[ X_t^x = X_t + A \cdot \left( X_t - X_{t-1} \right). \] (3.7)

Clearly, if \( A \) is the zero matrix, \( X_t^x = X_t \) and the data are already synchronized. Since the New York market data are already synchronized, the row of \( A \) corresponding to the New York stocks is a zero row.

Computing synchronized returns from (3.7) boils down to estimation of \( A \) in model (3.6). As already mentioned, \( A \) has a zero row corresponding to the New York stocks. We insist here on additional sparseness, setting other elements (to be selected from the data) to zero. We proceed as follows.

**Step 1.** Find the estimates for the \( M^2 \) parameters of the matrix \( A \) and for the matrix \( \Sigma \) using the Yule-Walker estimator. The Yule-Walker covariance relations for a multivariate AR(1) model are given by

\[ R(0) = R(-1) \cdot A' + \Sigma = R(1)' \cdot A' + \Sigma \]

\[ R(1) = R(0) \cdot A', \quad \text{where} \; R(k) = \mathbb{E}[X_{t-k} \cdot X_t']. \]
For more details about the Yule-Walker estimator, see Brockwell and Davis (1991) or Reinsel (1991).

Calculate the standard errors of the estimated elements of \( A \)

\[
\text{s.e.}(\hat{A}_{ij}) = \sqrt{\text{Var}(\hat{A}_{ij})} = \sqrt{\frac{\hat{\sigma}_{ii}}{T-1} \left( \hat{R}(0)^{-1} \right)_{jj}},
\]

where \( \hat{R}(0) = \frac{1}{T} \sum_{t=1}^{T} \left( X_t - \bar{X} \right) \cdot \left( X_t - \bar{X} \right)' \) with \( \bar{X} = \frac{1}{T} \sum_{t=1}^{T} X_t \), \( \hat{\sigma}_{ii} \) is the estimate of the \( i \)-th diagonal element of the matrix \( \Sigma \) and \( T \) is the sample size.

**Step 2.** Set \( A_{ij} = 0 \) if the t-ratio

\[
t_{ij} = \left| \frac{\hat{A}_{ij}}{\text{s.e.}(\hat{A}_{ij})} \right| \leq 1.96 \quad (5\% \text{ significance level})
\]

and \( A_{1j} = 0 \) for all \( j \) (corresponding to the New York stocks).

**Step 3.** Reestimate the sparse matrix from Step 2 using the MLE in model (3.6). This estimate is denoted by \( \hat{A}_{\text{synchr.}} \).

The synchronized returns are then estimated as

\[
X_{st} = X_t + \hat{A}_{\text{synchr.}} (X_t - X_{t-1}).
\]

An estimate \( \hat{A}_{\text{synchr.}} \) and synchronized returns for real data are presented in Sections 3.4.1-3.4.2. An even better estimate of \( A \) is presented in the following Section 3.3.3: \( A \) is estimated by maximum likelihood but in a more sophisticated model than (3.6). But the sparse structure of \( A \), i.e. the zero elements of the matrix, is always determined from Steps 1 and 2 above.

### 3.3.2 The model

We consider the standard CCC-GARCH(1,1) model, introduced by Bollerslev (1990), but for the synchronized returns:

\[
\begin{align*}
X_t &= \mu_t + \epsilon_t = \mu_t + \Sigma_t Z_t \quad (t \in \mathbb{Z}), \\
X_t &= X_t + A (X_t - X_{t-1}) = (I_M + A) X_t - A X_{t-1},
\end{align*}
\]

where we make the following assumptions:

\( (A1) \) \( (Z_t)_{t \in \mathbb{Z}} \) is a sequence of i.i.d. multivariate innovation variables with spherical distribution (e.g. the multivariate normal or the multivariate t distribution) with zero mean, covariance matrix \( \text{Cov}(Z_t) = I_M \) and \( Z_t \) independent from \( \{X_k^i, k < t\} \);
3.3. The synchronous CCC-GARCH(1,1) model

(A2) (CCC construction) \( \Sigma_t^\delta (\Sigma_t^\delta)' = H_t^\delta \) is almost surely positive definite for all \( t \), where the typical element of \( H_t^\delta \) is \( h_{ij,t}^\delta = \rho_{ij}^\delta (h_{ii,t}^\delta h_{jj,t}^\delta)^{1/2} \), for \( i, j = 1, \ldots, M \);

(A3) (GARCH(1,1) part) \( h_{ii,t}^\delta = (\sigma_{t-1,i}^\delta)^2 = \alpha_0^{(i)} + \alpha_1^{(i)} (X_{t-1,i}^s)^2 + \beta^{(i)} (\sigma_{t-1,i}^s)^2 \) with \( \alpha_0^{(i)}, \alpha_1^{(i)}, \beta^{(i)} > 0 \) for \( i = 1, \ldots, M \);

(A4) \( \mu_t^\delta = E[X_t^s | F_{t-1}] = (I_M + A) E[X_t | F_{t-1}] - A X_{t-1}, \)

\( E[X_t | F_{t-1}] = A X_{t-1} \) (as in (3.6)).

We call this the synchronous CCC-GARCH(1,1) model. Note that \( \rho_{ij}^\delta \) in (A2) equals the constant conditional correlation \( \text{Corr}(X_{t,i}^s, X_{t,j}^s | F_{t-1}) \).

Proposition 3.1.
Assume that the matrix \( (I_M + A) \) is invertible. Then, the synchronous CCC-GARCH(1,1) model (3.8) can be represented with asynchronized returns \( X_t \):

\[
X_t = A \cdot X_{t-1} + (I_M + A)^{-1} \Sigma_t^\delta Z_t,
\]

where the matrix \( \Sigma_t^\delta \) has the same CCC-GARCH(1,1) structure as defined in (3.8).

Proposition 3.1 implies that the synchronous CCC-GARCH(1,1) model is still a constant conditional correlation model (in terms of asynchronized data). Moreover, we should view it as a super-model of the classical CCC-GARCH(1,1): setting \( A = 0 \) yields the classical sub-model. Finally, we do not have to preestimate synchronized data as in Section 3.3.1, but should rather view \( A \) as a sparse parameter matrix which can be estimated by the joint maximum likelihood estimator in the synchronous CCC-GARCH(1,1) model.

Proof of Proposition 3.1. Using (3.7) and the fact that \( X_{t-1} \in F_{t-1} \), we calculate the conditional mean of the synchronized returns as

\[
\mu_t^\delta = E[X_t^s | F_{t-1}] = E[(I_M + A) \cdot X_t | F_{t-1}] - E[A \cdot X_{t-1} | F_{t-1}] =
\]

\[
= (I_M + A) \cdot E[X_t | F_{t-1}] - A \cdot X_{t-1} = (I_M + A) \cdot \mu_t - A \cdot X_{t-1} =
\]

\[
= (I_M + A) \cdot A \cdot X_{t-1} - A \cdot X_{t-1} = A^2 \cdot X_{t-1}.
\]

It follows that (3.8) is equivalent to \( X_t^s = A^2 \cdot X_{t-1} + \Sigma_t^\delta Z_t \). By (3.7), we
obtain the assertion:

\[ X_t^i = (I_M + A) \cdot X_t - A \cdot X_{t-1} = A^2 \cdot X_{t-1} + \Sigma_t^z Z_t \]
\[ \iff X_t = (I_M + A)^{-1} (A + A^2) \cdot X_{t-1} + (I_M + A)^{-1} \Sigma_t^z Z_t \]
\[ \iff X_t = A \cdot X_{t-1} + (I_M + A)^{-1} \Sigma_t^z Z_t. \]

### 3.3.3 Estimating the model

The parameters \( A, \alpha_{0}^{(j)}, \alpha_{1}^{(j)}, \beta_{ij}^{(j)}, \rho_{ij}^s, (j = 1, \ldots, M) \) in the synchronous CCC-GARCH(1,1) model can be estimated with the maximum likelihood method. We usually assume the innovations \( Z_t \) to be multivariate \( t_v \) distributed with zero mean and covariance matrix \( \text{Cov}(Z_t) = I_M \), where the degrees of freedom \( v \) have to be estimated as well, i.e. \( Z_t \sim t_v(0, I_M) \). The negative log-likelihood is then given by

\[
-l(\theta; X_t^q) = \frac{T M}{2} \log(\pi v) - T \log \left( \frac{\Gamma\left(\frac{M+v}{2}\right)}{\Gamma\left(\frac{v}{2}\right)} \right) + \\
\quad + \frac{T}{2} \log |R^s| + \sum_{t=1}^{T} \log |D_t^s| - T \log \left( |(I_M + A)| \right) + \\
\quad + \frac{M + v}{2} \sum_{t=2}^{T} \left( \log \left( 1 + \frac{(\epsilon_t^s)'(R_t^{-1} \epsilon_t^s)}{v} \right) \right),
\]

where \( H_t^s = D_t^s R^s D_t^s \) and from the CCC-construction, \( D_t^s \) a diagonal \( M \times M \) matrix with diagonal-elements \( \sigma_{i,1}^s, \ldots, \sigma_{i,M}^s \), \( R^s = [\rho_{ij}^s]_{1 \leq i, j \leq M} \) and \( \epsilon_t^s = (D_t^s)^{-1}(X_t^s - \mu_t^s) \); \( \theta \) denotes the vector of all parameters involved and \( X_t^s = X_t + A(X_t - X_{t-1}) \) as before. Also, we use a sparse structure of the matrix \( A \); it is estimated following Steps 1 and 2 of the procedure given in Section 3.3.1.

Problems could occur when the maximum likelihood estimate would produce an ill-conditioned matrix \((I_M + \hat{A})\) and we would have to compute estimated volatilities or risks which involve the inverse \((I_M + \hat{A})^{-1}\). Although this is not the case for our real example where the matrix \((I_M + \hat{A})\) was found to be well-conditioned, we want to illustrate a possible strategy to overcome this problem. We propose **shrinkage towards zero**: instead of \( \hat{A} \) we could take

\[
\hat{A}_{(k)} = k \hat{A} \quad (0 < k \leq 1).
\]
3.4 Numerical results

The shrinkage parameter could be optimized with respect to out-of-sample performances; or it could be chosen via graphical procedures. The effect of shrinkage is a much better numerical behavior but it may also increase statistical accuracy as with the famous Stein estimator (see for example Voinov and Nikulin, 1995).

3.4 Numerical results

We consider a real global portfolio of seven market indices: US Dow Jones Industrial Average (DJIA), French CAC40 Index, German Deutsche Aktien (DAX), Italian BCI General Index, Dutch CBS All-Share, UK FT-SE-A All-Share Index (FTAS) and Japanese NIKKEI 225 Average (NIK). The data are from the time period between January 17, 1990 and June 22, 1994, corresponding to 1000 days without holidays in the different countries.

We use here (negative) relative difference returns (in percentages) \( X_{t,i} = -100 \frac{S_{t,i} - S_{t-1,i}}{S_{t-1,i}} \), where \( S_{t,i} \) denotes the price of the asset \( i \) at the local closing time \( t_i \) of the day \( t \), because they are close approximations of the log-returns and because they allow for much simpler portfolio and risk computations; see Section 3.4.3 for more details on the construction of the negative portfolio returns. Nevertheless, we still synchronize such relative difference returns as in (3.7).

The aim is to support empirically the effect of synchronization and to compare the synchronous CCC-GARCH(1,1) model in (3.8) with the asynchronous classical CCC-GARCH(1,1) model. For the numerical optimization, we use a quasi-Newton method.

3.4.1 Simple estimate of \( A \) and synchronization

We examine here first the effect of synchronization from a descriptive point of view. The parsimoniously structured, estimated matrix \( A \) in (3.7) is obtained from the procedure illustrated in Section 3.3.1,

\[
\hat{A} = \begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.3481 & 0.1445 & -0.1065 & 0 & 0 & 0 & -0.0910 \\
0.3895 & 0.2038 & -0.1212 & 0 & 0 & 0 & -0.1195 \\
0.3225 & 0 & 0 & 0 & 0 & 0.2125 & -0.0628 \\
0.3615 & 0 & -0.0118 & 0 & 0 & 0 & -0.0432 \\
0.2397 & 0 & 0 & 0.0324 & 0 & 0.1060 & -0.0834 \\
0.4292 & 0.1136 & 0 & 0 & 0 & 0 & -0.0773 \\
\end{pmatrix}
\]  

(3.11)
where the variables are ordered as DJIA, CAC40, DAX, BCI, CBS, FTAS, NIK. The column with the largest coefficients corresponds to the DJIA: there is substantial predictability of all other markets from the DJIA at the previous day. Besides a major determining effect of the US market for the financial world, the observed pattern is natural since the exchange in New York closes last. There seems to be also predictability of all other markets from the Japanese returns (NIK), although in this case, the coefficients are small (and negative).

Using $\hat{A}$ from (3.11) and the synchronization formula (3.7), we obtain the synchronized returns $\tilde{X}_t^*$. The effect of synchronization in terms of empirical correlations is described in Table 3.1: synchronized data often exhibit larger instantaneous correlations between different returns from indices at the same day. The empirical correlations are typically too small for highly asynchronous markets. This is the case for example of the US and the Japanese markets: the empirical correlation between DJIA and NIK is nearly twice as much when synchronizing (0.388 vs. 0.189). Of course, there is no reason to believe that synchronization generally yields higher correlations. This result is consistent and similar to the analysis in Burns et al. (1998). We will see in Section 3.4.2, that some of the strong correlations in synchronized data are due to the simple model for the synchronization transform $A$. Nevertheless, synchronized returns will still be substantially stronger correlated as their asynchronized counterparts when using better models for estimating $A$.

### 3.4.2 Estimates for the synchronous CCC-GARCH(1,1) and its performance

The parameters are estimated by maximum likelihood as in Section 3.3.3. For quantifying the goodness of fit of the models, we consider the following statistics:

- the AIC statistic: $-2 \log\text{-likelihood} + 2 \# \text{parameters}$
- the outsample $- \log\text{-likelihood}$

$$- \log\text{-likelihood}(\tilde{X}_T^*: \hat{A}, \hat{\nu}, \{\hat{\alpha}_0^{(j)}, \hat{\alpha}_1^{(j)}, \hat{\beta}^{(j)}; j = 1, \ldots, M\}),$$

where $\tilde{X}_T^* = \tilde{X}_1, \ldots, \tilde{X}_T$ are new test data and the parameter estimates, equipped with hats, are from the training sample $X^n = X_1, \ldots, X^n$. The likelihood itself is given in (3.10). Both quantities are measures for out-sample performance: their values (low is better) indicate closeness to the true data generating model with respect to the Kullback-Leibler divergence. In our
3.4. Numerical results

analysis, we take \( n = 1000 \) and the test set values \( \tilde{X}_1^T = X_{n+1}^{n+500} \) are the next 500 consecutive observations (days between June 23, 1994 and September 9, 1996). We take \( T = 500 \) (little more than two years), because it seems a reasonable time period where the multivariate return series of the seven indices are believed to be (at least approximately) stationary.

The results obtained for the synchronous CCC-GARCH(1,1) model and the classical CCC-GARCH(1,1) model without synchronization are summarized in Table 3.2.

The synchronous CCC-GARCH(1,1) model is better with respect to both goodness of fit statistics than the asynchronous CCC-GARCH(1,1) model, although the difference is small (in order of 1 percent). Such small differences could be obscured by low signal to noise ratio. It is often useful to consider differences of performance terms and use the concept of hypothesis testing, rather than quantifying differences in terms of percentages.

We consider the differences of each term in the outsample log-likelihood,

\[ \hat{D}_t = \tilde{U}_{t; \text{sync}} - \tilde{U}_{t; \text{async}}, \quad t = 1, \ldots, T, \]

where

\[ \sum_{t=1}^{T} \tilde{U}_{t; \text{model}} = \text{outsample log-likelihood}. \]

Note that, up to a change of signs, \( \hat{D}_t \) is the difference between deviance residuals, see McCullagh and Nelder (1989). We are now testing the null hypothesis that the differences \( \hat{D}_t \) have mean zero against the alternative of mean less than zero, i.e. the estimates from the synchronous CCC-GARCH(1,1) are better than the ones from the asynchronous classical CCC-GARCH(1,1). For this purpose, we use versions of the t-test and sign-test, adapted to the case of dependent observations. The t-type test statistic is

\[ \sqrt{T} \frac{\overline{D}}{\sigma_{D; \infty}}, \quad \text{where} \quad \overline{D} = \frac{1}{T} \sum_{t=1}^{T} \hat{D}_t. \quad (3.12) \]

In (3.12), \( \sigma_{D; \infty}^2 = (2\pi) \hat{f}_\beta(0), \hat{f}_\beta(0) \) a smoothed periodogram estimate at frequency zero, based on \( \hat{D}_1, \ldots, \hat{D}_T \); see for example Brockwell and Davis (1991). The motivation for this estimate is based on the assumption that \( \{\hat{D}_t\} \) is stationary and satisfies suitable dependence conditions, e.g mixing. Then

\[ \sqrt{T} \left( \overline{D} - E[\hat{D}_t] \right) \Rightarrow \mathcal{N}(0, \sigma_{D; \infty}^2), \quad (T \to \infty), \]

\[ \sigma_{D; \infty}^2 = \sum_{k=-\infty}^{+\infty} \text{Cov}[\hat{D}_0, \hat{D}_k] = (2\pi) f_\beta(0). \quad (3.13) \]
where \( \hat{f}_D(0) \) is the spectral density at zero of \( \{\hat{D}_t\}_t \).

Thus, using (3.13) for the test statistic in (3.12)

\[
\sqrt{T} \frac{\overline{D}}{\overline{\sigma}_D;\infty} \Rightarrow \mathcal{N}(0, 1) \quad (T \to \infty)
\]

(3.14)

under the null hypothesis. The observed value for the test statistic (3.12) equals \(-1.6144\) with a corresponding \( P \)-value of 0.053 indicating weak significance that the synchronous model is better.

The version of the sign test is based on the number of negative differences

\[
\hat{W}_t = I_{\{\hat{D}_t \leq 0\}}, \quad t = 1, \ldots, T,
\]

for the null hypothesis that the negative differences \( \hat{W}_t \) have mean \( \frac{1}{2} \) against the alternative of mean bigger than \( \frac{1}{2} \). The test statistic is given by

\[
\sqrt{T} \frac{\overline{W} - \frac{1}{2}}{\overline{\sigma}_W;\infty}, \quad \text{where} \quad \overline{W} = \frac{1}{T} \sum_{t=1}^{T} \hat{W}_t
\]

(3.15)

and \( \overline{\sigma}_W;\infty \) as in (3.12) but based on \( \hat{W}_1, \ldots, \hat{W}_T \). As in the derivation of the \( t \)-type test above, we have

\[
\sqrt{T} \frac{\overline{W} - \frac{1}{2}}{\overline{\sigma}_W;\infty} \Rightarrow \mathcal{N}(0, 1) \quad (T \to \infty)
\]

(3.16)

under the null hypothesis. There are 302 negative differences (of total \( T = 500 \)) and the observed value of the test statistic is 1.698 with a corresponding \( P \)-value of 0.045. Therefore, this test leads to the rejection of the null hypothesis at the 5% significance level, implying a preference of the synchronous over the asynchronous CCC-GARCH(1,1) model.

Finally, we illustrate the resulting estimates of the synchronized returns obtained from the synchronous CCC-GARCH(1,1) model. This is somewhat different than using the matrix \( \hat{A} \) in (3.11) from the maximum likelihood estimate in the simple model (3.6). The estimated matrix \( \hat{A} \) equals

\[
\hat{A} = \begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.2210 & 0.0293 & 0.0159 & 0 & 0 & 0 & -0.0674 \\
0.2982 & 0.0871 & -0.0029 & 0 & 0 & 0 & -0.0893 \\
0.2862 & 0 & 0 & 0 & 0 & 0.1009 & -0.0126 \\
0.2465 & 0 & 0.0041 & 0 & 0 & 0 & -0.0393 \\
0.1719 & 0 & 0 & 0.0044 & 0 & 0.0240 & -0.0481 \\
0.3093 & 0.0538 & 0 & 0 & 0 & 0 & -0.0489
\end{pmatrix}
\]

(3.17)
where the variables are ordered, as before, as DJIA, CAC40, DAX, BCI, CBS, FTAS, NIK. As we can see, there are differences when comparing with the matrix given by (3.11), especially in the first column.

<table>
<thead>
<tr>
<th></th>
<th>DJIA</th>
<th>CAC40</th>
<th>DAX</th>
<th>BCI</th>
<th>CBS</th>
<th>FTAS</th>
<th>NIK</th>
</tr>
</thead>
<tbody>
<tr>
<td>DJIA</td>
<td>1</td>
<td>0.265</td>
<td>0.239</td>
<td>0.117</td>
<td>0.138</td>
<td>0.217</td>
<td>0.189</td>
</tr>
<tr>
<td>CAC40</td>
<td>0.265</td>
<td>1</td>
<td>0.757</td>
<td>0.513</td>
<td>0.769</td>
<td>0.702</td>
<td>0.316</td>
</tr>
<tr>
<td>DAX</td>
<td>0.239</td>
<td>0.757</td>
<td>1</td>
<td>0.553</td>
<td>0.760</td>
<td>0.599</td>
<td>0.322</td>
</tr>
<tr>
<td>BCI</td>
<td>0.117</td>
<td>0.513</td>
<td>0.553</td>
<td>1</td>
<td>0.560</td>
<td>0.478</td>
<td>0.254</td>
</tr>
<tr>
<td>CBS</td>
<td>0.138</td>
<td>0.769</td>
<td>0.760</td>
<td>0.560</td>
<td>1</td>
<td>0.735</td>
<td>0.298</td>
</tr>
<tr>
<td>FTAS</td>
<td>0.217</td>
<td>0.702</td>
<td>0.599</td>
<td>0.478</td>
<td>0.735</td>
<td>1</td>
<td>0.302</td>
</tr>
<tr>
<td>NIK</td>
<td>0.189</td>
<td>0.316</td>
<td>0.322</td>
<td>0.254</td>
<td>0.298</td>
<td>0.302</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>DJIA</th>
<th>CAC40</th>
<th>DAX</th>
<th>BCI</th>
<th>CBS</th>
<th>FTAS</th>
<th>NIK</th>
</tr>
</thead>
<tbody>
<tr>
<td>DJIA</td>
<td>1</td>
<td>0.398</td>
<td>0.419</td>
<td>0.286</td>
<td>0.358</td>
<td>0.342</td>
<td>0.388</td>
</tr>
<tr>
<td>CAC40</td>
<td>0.398</td>
<td>1</td>
<td>0.806</td>
<td>0.567</td>
<td>0.761</td>
<td>0.704</td>
<td>0.353</td>
</tr>
<tr>
<td>DAX</td>
<td>0.419</td>
<td>0.806</td>
<td>1</td>
<td>0.613</td>
<td>0.794</td>
<td>0.654</td>
<td>0.324</td>
</tr>
<tr>
<td>BCI</td>
<td>0.286</td>
<td>0.567</td>
<td>0.613</td>
<td>1</td>
<td>0.624</td>
<td>0.609</td>
<td>0.290</td>
</tr>
<tr>
<td>CBS</td>
<td>0.358</td>
<td>0.761</td>
<td>0.794</td>
<td>0.624</td>
<td>1</td>
<td>0.747</td>
<td>0.348</td>
</tr>
<tr>
<td>FTAS</td>
<td>0.342</td>
<td>0.704</td>
<td>0.654</td>
<td>0.609</td>
<td>0.747</td>
<td>1</td>
<td>0.308</td>
</tr>
<tr>
<td>NIK</td>
<td>0.388</td>
<td>0.353</td>
<td>0.324</td>
<td>0.290</td>
<td>0.348</td>
<td>0.308</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>DJIA</th>
<th>CAC40</th>
<th>DAX</th>
<th>BCI</th>
<th>CBS</th>
<th>FTAS</th>
<th>NIK</th>
</tr>
</thead>
<tbody>
<tr>
<td>DJIA</td>
<td>1</td>
<td>0.362</td>
<td>0.379</td>
<td>0.269</td>
<td>0.291</td>
<td>0.319</td>
<td>0.332</td>
</tr>
<tr>
<td>CAC40</td>
<td>0.362</td>
<td>1</td>
<td>0.797</td>
<td>0.556</td>
<td>0.778</td>
<td>0.703</td>
<td>0.312</td>
</tr>
<tr>
<td>DAX</td>
<td>0.379</td>
<td>0.797</td>
<td>1</td>
<td>0.584</td>
<td>0.778</td>
<td>0.619</td>
<td>0.286</td>
</tr>
<tr>
<td>BCI</td>
<td>0.269</td>
<td>0.556</td>
<td>0.584</td>
<td>1</td>
<td>0.594</td>
<td>0.541</td>
<td>0.290</td>
</tr>
<tr>
<td>CBS</td>
<td>0.291</td>
<td>0.778</td>
<td>0.778</td>
<td>0.594</td>
<td>1</td>
<td>0.739</td>
<td>0.293</td>
</tr>
<tr>
<td>FTAS</td>
<td>0.319</td>
<td>0.703</td>
<td>0.619</td>
<td>0.541</td>
<td>0.739</td>
<td>1</td>
<td>0.290</td>
</tr>
<tr>
<td>NIK</td>
<td>0.332</td>
<td>0.312</td>
<td>0.286</td>
<td>0.290</td>
<td>0.293</td>
<td>0.290</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3.1: Instantaneous empirical correlations between components of asynchronized returns $X_t$ (top), of estimated synchronized returns $\tilde{X}_t^s$ from the synchronization formula (3.7) with $\hat{A}_{synchr.}$ in (3.11) (center), and from a synchronous CCC-GARCH(1,1) fit with $\hat{A}$ in (3.17) (bottom).

Figure 3.1 indicates that synchronization has primarily an effect on returns of small or medium size. Table 3.1 shows that instantaneous correlations between asynchronized returns with the better estimate $\hat{A}$ in (3.17) are a bit
Chapter 3. Synchronizing Multivariate Financial Time Series

Figure 3.1: Asynchronized returns $X_{t,i}$ against estimated synchronized returns $\hat{X}_{t,i}^s$ from the synchronous CCC-GARCH(1,1) model (using $\hat{A}$ in (3.17)) for the French CAC40, the German DAX, the Italian BCI, the Dutch CBS, the British FTAS and the Japanese NIKKEI index.
3.4. Numerical results

weaker than with the more simple transform in (3.11) but still substantially stronger than for asynchronized data.

3.4.3 Estimating the performance at the portfolio level

We now examine the effect of synchronization for the estimation of volatility in a portfolio. Denote by $P_t$ the price of a portfolio at day $t$

$$P_t = \sum_{i=1}^{7} \alpha_i S_{t,i} \quad t = 1, \ldots, m. \quad (3.18)$$

This portfolio employs a constant asset division. For illustrative purposes, we use the data $S_{t,i}$ from before and choose $\alpha_1 = 0.4$, $\alpha_2 = \ldots = \alpha_6 = 0.08$ and $\alpha_7 = 0.2$, corresponding to the volumes of the different stock exchanges. We also translate all the prices to US dollars, using daily currencies. It is known that the (negative) portfolio returns $\Delta_t$ at day $t$ become then a linear combination of the individual (negative) asset returns $X_t$

$$\Delta_t = -100 \cdot \left( \frac{P_t - P_{t-1}}{P_{t-1}} \right) = \beta'_{t-1} X_t,$$

$$\beta_{t-1,i} = \alpha_i \frac{S_{t-1,i}}{P_{t-1}}, \quad i = 1, \ldots, 7. \quad (3.19)$$

Our general model for $\Delta_t$ is

$$\Delta_t = \mu_{t,p} + \sigma_{t,p} Z_t,$$

where $\mu_{t,p} \in \mathbb{R}$ and $\sigma_{t,p} \in \mathbb{R}^+$ are measurable function of $\mathcal{F}_{t-1}$ (see (3.2)).

We compare portfolio volatility estimates from multivariate synchronous and asynchronous CCC-GARCH(1,1) models with the ones from a classical GARCH(1,1) univariate analysis (and extensions thereof) for the portfolio returns $\Delta_t$: all give rise to different $\mu_{t,p}$ and $\sigma_{t,p}$ above. Note that in the more realistic case, the weights $\alpha_i = \alpha_{t,i}$ are depending on $t$. As a consequence, the univariate analysis of the returns $\Delta_t$ is inappropriate, because the returns of portfolio prices would typically be far from stationary. For univariate analyses based on returns $\Delta_t$, we always assume the model

$$\Delta_t = \phi \Delta_{t-1} + \sigma_{t,p} Z_t, \quad (3.20)$$

where $\sigma_{t,p}$ is a measurable function of previous returns $\Delta_{t-1}, \Delta_{t-2}, \ldots$ and i.i.d. innovations $Z_t \sim \sqrt{(v-2)/\nu} \nu_{\nu}$. The scaling factor $\sqrt{(v-2)/\nu}$ is used so that $\text{Var}(Z_t) = 1$. The univariate GARCH(1,1) specification is

$$\sigma^2_{t,p} = \alpha_0 + \alpha_1 \Delta^2_{t-1} + \beta \sigma^2_{t-1,p},$$
where $\alpha_0, \alpha_1, \beta > 0$. The negative log-likelihood, conditioned on the first observation $\Delta_1$ and some starting value $\sigma_{1,p}$, (e.g. the square root of the sample variance) is then

$$-\ell(\phi, \alpha_0, \alpha_1, \beta; \Delta_1^2) = - \sum_{t=2}^{n} \log \left( c(v)^{-1} \sigma_{t,p}^{-1} f_{tv} \left( \frac{\Delta_t - \phi \Delta_{t-1}}{c(v) \sigma_{t,p}} \right) \right),$$

$$c(v) = \left( (v - 2)/v \right)^{1/2}, \quad \text{(3.21)}$$

where $f_{tv}$ denotes the density of the univariate $t_v$ distribution. Minimizing the negative log-likelihood yields estimates $\hat{\mu}_{t,p} = \hat{\phi} \Delta_{t-1}$ and $\hat{\sigma}_{t,p}^2 = \hat{\alpha}_0 + \hat{\alpha}_1 \Delta_{t-1}^2 + \hat{\beta} \sigma_{t-1,p}^2$.

With the multivariate synchronized CCC-GARCH(1,1) model, we calculate estimates of the portfolio conditional means $\hat{\mu}_{t,p}$ and variances $\hat{\sigma}_{t,p}^2$, $t = 1, \ldots, n$ as follows. We always take the innovations $Z_t$ of the model in (3.8) or (3.9) to be multivariate $t_v$ distributed ($v$ unknown) with zero mean and covariance matrix $\text{Cov}(Z) = I_M$. Using the representation in Proposition 3.1, assuming that $(I_M + A)^{-1}$ exists, it follows that the synchronized returns $X_t$ given the information up to time $t - 1$ are multivariate $t_v$ distributed

$$X_t \mid F_{t-1} \sim t_v \left( A X_{t-1}, (I_M + A)^{-1} \Sigma_t^s \left((I_M + A)^{-1} \Sigma_t^s\right)' \right).$$

Now, using a nice property of elliptical distributions (see Fang et al., 1990) we find that the portfolio return $\Delta_t$ given the information up to time $t - 1$ is univariate $t_v$ distributed with the following mean and variance:

$$\Delta_t \mid F_{t-1} \sim t_v \left( \beta_{t-1}' A X_{t-1}, \beta_{t-1}' (I_M + A)^{-1} \Sigma_t^s \left( \beta_{t-1}' (I_M + A)^{-1} \Sigma_t^s \right)' \right),$$

where the vector of coefficients $\beta_{t-1}$ is given in (3.19). Thus, we calculate

$$\hat{\mu}_{t,p} = \beta_{t-1}' \tilde{A} X_{t-1} \quad \text{and} \quad \hat{\sigma}_{t,p}^2 = \beta_{t-1}' (I_M + \tilde{A})^{-1} \tilde{\Sigma}_t^s (\beta_{t-1}' (I_M + \tilde{A})^{-1} \tilde{\Sigma}_t^s)' ,$$

where $\tilde{A}$ and $\tilde{\Sigma}_t^s$ are the maximum likelihood estimates in the model (3.9). The estimates from the classical asynchronous CCC-GARCH(1,1) model are of the same form, but with $\tilde{A} = 0$ and $\tilde{\Sigma}_t^{synch}$.

The predicted portfolio conditional mean and variance estimates with the synchronous and the asynchronous CCC-GARCH(1,1) and the univariate GARCH(1,1) model are plotted in Figure 3.2.

The predicted volatilities are substantially larger when using the univariate approach. The differences between the two multivariate methods are (visually) much smaller. The conditional mean in the synchronized model is small (relative to the magnitude of the square root of conditional variance).
3.4. Numerical results

Figure 3.2: Predicted portfolio conditional variances $\hat{\sigma}_{t,P}^2$ (top) and means $\hat{\mu}_{t,P}$ (center and bottom) for the test data days between June 23, 1994 and September 9, 1996 (500 days) using the univariate classical GARCH(1,1) model (3.20) for the negative portfolio returns (solid line), the synchronous CCC-GARCH(1,1) model (dotted line) and the standard CCC-GARCH(1,1) model without synchronization (dashed line).
Now, we test the goodness of the residuals

\[
\hat{Z}_t = \frac{\hat{\Delta}_t - \hat{\mu}_{t; p}}{\hat{\sigma}_{t; p}}, \quad t = 1, \ldots, T.
\]  

(3.22)

in different models. Here \( \hat{\Delta}_t \) is from new test set data \( \Delta_{n+1}, \ldots, \Delta_{n+500} \) over the next 500 days; \( \hat{\mu}_{t; p} \) and \( \hat{\sigma}_{t; p} \) are from the different models, estimated with the training data \( \Delta_1, \ldots, \Delta_n \) but of course evaluated using the immediate lagged values in the test set.

We are particularly interested in the null hypothesis that the dynamics of the (negative) portfolio returns follow model (3.20) against the alternative of a misspecified model. Under the null hypothesis, the statistic \( \sqrt{T} \hat{Z} \) is standard normally distributed. The observed values, and the corresponding \( P \)-values (given in parenthesis), of the test for the synchronous CCC-GARCH(1,1), the asynchronous CCC-GARCH(1,1) and the univariate model are \(-1.862 \ (0.063), -2.201 \ (0.0277)\) and \(-1.534 \ (0.125)\), respectively. It shows that the model assumed for the portfolio returns is rejected only by the asynchronous CCC-GARCH(1,1) model.

<table>
<thead>
<tr>
<th></th>
<th>synchronous CCC-GARCH(1,1)</th>
<th>CCC-GARCH(1,1) without synchr.</th>
<th>Univariate approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC</td>
<td>19053.1</td>
<td>19136.796</td>
<td>–</td>
</tr>
<tr>
<td>Outsample log-likelihood</td>
<td>4240.686</td>
<td>4284.736</td>
<td>–</td>
</tr>
<tr>
<td>IS-PL₁</td>
<td>534.1155</td>
<td>573.9434</td>
<td>609.7303</td>
</tr>
<tr>
<td>IS-PL₂</td>
<td>1425.46</td>
<td>1732.441</td>
<td>1515.562</td>
</tr>
<tr>
<td>OS-PL₁</td>
<td>185.4544</td>
<td>195.5282</td>
<td>235.8809</td>
</tr>
<tr>
<td>OS-PL₂</td>
<td>227.7761</td>
<td>240.7812</td>
<td>243.5134</td>
</tr>
<tr>
<td>Portfolio out. log-likelihood</td>
<td>461.0515</td>
<td>470.5776</td>
<td>474.2656</td>
</tr>
</tbody>
</table>

**Table 3.2:** *Goodness of fit measures from the synchronous CCC-GARCH(1,1) model, the CCC-GARCH(1,1) model without synchronization and the univariate model defined by (3.20) for the same global portfolio already introduced in Section 3.4.*
3.4. Numerical results

For quantifying the goodness of fit of the models, we consider again the outsample log-likelihood performance

portfolio outsample log-likelihood: \(-\log-likelihood\left(\tilde{\Delta}_1^T;\tilde{\mu}_t,\tilde{\sigma}_t,\hat{v}\right)\),

where, as in (3.22), \(\tilde{\Delta}_1^T = \Delta_{n+500}^{n+T}\) are new test set data and \(\tilde{\mu}_t,\tilde{\sigma}_t,\hat{v}\) are estimated from the training data. The log-likelihood is from (3.21). In addition, we also consider the following in- and out-sample prediction loss statistics:

\[
\text{IS-PL}_i = \sum_{t=1}^{n} |\tilde{\sigma}_{t,P}^2 - (\Delta_t - \tilde{\mu}_t,P)^2 |^i, \quad i = 1, 2,
\]

\[
\text{OS-PL}_i = \sum_{t=1}^{T} |\tilde{\sigma}_{t,P}^2 - (\tilde{\Delta}_t - \tilde{\mu}_t,P)^2 |^i, \quad i = 1, 2.
\]

The OS-PL statistics and the portfolio outsample log-likelihood are measures for predictive performance. For this reason, we consider them as more important than the in-sample IS-PL statistics. These statistics are given in Table 3.2 for the synchronous and asynchronous CCC-GARCH(1,1) model and for the univariate GARCH(1,1) model.

Analogously to the results obtained in the Section 3.4.2, we also find at the portfolio level that the synchronous CCC-GARCH(1,1) model is better (with respect to all goodness of fit measures) than the asynchronous classical CCC-GARCH(1,1) model and than the univariate approach. We can also see that the improvement in this case is more relevant than at the multivariate level, in particular with respect to the OS-PL performances (5-10% less). Also, multivariate modelling shows substantial advantages over the univariate GARCH(1,1) analysis, even without using synchronization.

As mentioned in Section 3.4.2, more impressive gains may be masked by a low signal to noise ratio. Similar t- and sign-type tests to the ones in (3.12)-(3.14) and (3.15)-(3.16) can be done for the goodness of fit statistics at the portfolio level. The results for the same global portfolio introduced in Section 3.4 are summarized in Table 3.3.

As we can see from Table 3.3, the t-type test never yields significant differences between the models. This may be just a fact of low power due to non-Gaussian observations. On the other hand, the sign-type test which is robust against deviations from Gaussianity yields very significant results: the synchronous is better than the asynchronous CCC-GARCH(1,1) model, and multivariate is better than univariate modelling.
Table 3.3: Testing differences of performance on the portfolio level. The values of test statistics and corresponding P-values (between parenthesis) are given for t-type test (analogously to (3.12)-(3.14)) (top) and sign-type test (analogously to (3.15)-(3.16)) (bottom).
3.5 More sophisticated asynchronous models as another source for improvements?

We question here, whether the improvements with the synchronous CCC-GARCH(1,1) model could also be achieved or even surpassed by more sophisticated models for volatilities or conditional means. Of course, the synchronous CCC-GARCH(1,1) model is also a more complex model for asynchronous data (see Proposition 3.1), but motivated from the view of synchronization with a simple linear transform.

The first extension of the asynchronous classical GARCH(1,1) model is to include a conditional mean term

\[ X_t = \mu_t + \Sigma_t \mathbf{Z}_t \quad (t \in \mathbb{Z}), \]
\[ \mu_t = B \cdot X_{t-1}, \quad B = \text{diag}(b_1, \ldots, b_M), \quad (3.23) \]

where \( \mathbf{Z}_t \) and \( \Sigma_t \) are exactly as in the model (3.8).

We then further extend the model in (3.23) to more complex, potentially high dimensional parameterizations for approximating more general volatility functions:

\[ X_t \text{ as in (3.23) with } \Sigma_t \text{ as in (3.8) but } \]
\[ h_{i,t} = \sigma_{i,t}^2 = f_i(X_{t-1,i}, \sigma_{t-1,i}^2), \quad (3.24) \]

where \( f_i(\cdot, \cdot) \) piecewise quadratic in the first and linear in the second argument. Fitting such “best” piecewise functions (even if we assume that the squared volatility \( \sigma_{i,t}^2 \) of the asset \( i \) depends only on the previous values \( X_{t-1,i} \) and \( \sigma_{t-1,i}^2 \) of the same asset, without considering cross-terms) is computationally prohibitive. A tree-structured (sub-optimal) technique for univariate series has been already proposed in Chapter 2. The same idea can also be used for every individual series in the multivariate set-up and is proposed in the next Section.

3.5.1 Flexible thresholding for the asynchronized returns

In this Section we propose a new class of multivariate CCC models with tree structured multiple thresholds for the volatility estimation of the asynchronized negative multivariate return time series of different financial instruments and, as a consequence, for the volatility estimation of the negative returns of full global portfolios constructed combining these instruments. We follow a multivariate generalisation of the methodology introduced in Chapter 2.
Consider as a reference model (3.24). The simplest example for (3.24) is to model the individual squared volatilities $\sigma_{i,t}^2$ for each asset $i$ with the classical GARCH(1,1) model as in (3.8). Now, to add complexity (meaning here essentially thresholds) in the functions $f_i(\cdot, \cdot)$, we follow the procedure given in the next Algorithm based on the same methodology already proposed in Chapter 2 for the construction of the single function $f(\cdot, \cdot)$ in the univariate tree structured GARCH models. This means that the Algorithm is always based on the likelihood in the working model (3.24), where the functional forms $f_i(\cdot, \cdot)$ are parameterized by thresholds functions $f_{i,\theta_i}(\cdot, \cdot)$ which change in the fitting procedure. Note that the optimal parameters of the single functions $f_i(\cdot, \cdot)$, $i = 1, \ldots, M$ can take very different values and also the number and place of thresholds can be substantially different for each asset.

The estimation of the optimal number and value of the parameters of the functions $f_i(\cdot, \cdot)$, $i = 1, \ldots, M$ in (3.24) is made according to the following Algorithm. We call the resulting model Tree Structured CCC-GARCH Model.

### Tree Structured CCC-GARCH Algorithm

Denote by $X_{t,i}$ the available negative returns (in percentages) of the financial instrument $i$, $i = 1, \ldots, M$ and by $X_t$ the $M$-dimensional vector of the individual available negative returns on day $t$, $t = 1, \ldots, n$.

**Step 1.** Make an univariate analysis and find the optimal univariate tree structured GARCH model for each individual negative return time series $X_{t,i}$, following the estimation procedure proposed in Chapter 2. Denote the resulting individual thresholds by

$$d_{1,i}, \ldots, d_{k_i,i} \text{ with } k_i = \# \text{ optimal thresholds for the asset } i, \ i = 1, \ldots, M.$$  \hspace{1cm} (3.25)

Note that both thresholds for returns and squared volatilities are considered in (3.25).

**Step 2.** Estimate the optimal parameters for the CCC-GARCH(1,1) with a vector AR(1) term for the conditional means (3.24), maximizing the log-likelihood function. Denote the optimal parameters for the means and the squared volatilities by $\hat{\theta}_{\text{start}}$ and the optimal constant conditional correlations matrix $\Gamma$ by $\hat{\Gamma}_{\text{start}}$.

**Step 3.** Introduce the first threshold in (3.24) in the following way. Find which threshold between the optimal individual thresholds $d_{1,i}$ $\forall i$ with $k_i > 0$ gives
3.5. More sophisticated asynchronous models

the best improvement in the maximization of the log-likelihood function for the new constructed model. Take as starting parameters for the maximization a vector based on the parameters $\theta_{\text{start}}$ for the conditional means and squared volatilities and the constant conditional correlations matrix $\Gamma_{\text{start}}$ estimated in Step 2. Denote the founded first optimal threshold for the model (3.24) by $\hat{d}_1$, where

$$\hat{d}_1 = \hat{d}_{1,i_1}, \text{ with } i_1 \in \{1, \ldots, M\}.$$  

Note that the functions $f_i(\cdot, \cdot)$ in the new model are all of a GARCH(1,1)-type except for one. In fact for $i = i_1$, the addition of the threshold $\hat{d}_1$ involves a partition for the predictor space of $f_{i_1}(\cdot, \cdot)$ in two cells; for every partition cell we employ a GARCH(1,1) model. Estimate the optimal parameters $\hat{\theta}_1$ and $\hat{\Gamma}_1$ for the new model with $\hat{d}_1$ and calculate the AIC statistic. Note that there are three parameters more in $\theta$ every threshold we introduce. Stop the Algorithm, if there is no improvement in the AIC statistic.

Step 4. Introduce the second threshold in the optimal model from Step 3 proceeding as before, i.e. choosing the threshold between $d_{1,i}$ for $i \neq i_1$ with $k_i > 0$ and $d_{2,i_1}$, if $k_{i_1} > 1$ with the best improvement in maximizing the log-likelihood function for the new constructed model resulting from the addition of this second threshold to the optimal one from Step 3. Take as starting parameters for the maximization a vector based on $\hat{\theta}_1$ and the matrix $\hat{\Gamma}_1$. Denote the optimal second threshold by $\hat{d}_2$, where

$$\hat{d}_2 = \hat{d}_{1,i_2}, \text{ with } i_2 \in \{1, \ldots, M\} \setminus \{i_1\}, \text{ or } \hat{d}_2 = d_{2,i_1}.$$  

Estimate the optimal parameters $\hat{\theta}_2$ and $\hat{\Gamma}_2$ for the model (3.24) with thresholds $\hat{d}_1$ and $\hat{d}_2$. Stop the Algorithm, if there is no improvement in the AIC statistic. Set $j = 2$.

Step 5. Set $j = j + 1$. Introduce the $j$-th threshold as before (choosing the set of the possible candidates for the $j$-th threshold in the same way as by Step 3-4), estimate the optimal parameters $\hat{\theta}_j$ and $\hat{\Gamma}_j$ for (3.24) with thresholds $\hat{d}_1, \ldots, \hat{d}_j$ and calculate the AIC statistic of the new model. Stop the Algorithm, if there is no improvement in the AIC statistic.

Step 6. Iterate Step 5 until whether there is no improvement in the AIC statistic or no other individual optimal threshold is left. Set $j_{\text{opt}} = j$. 

---


Note that in the Steps from 3 to 6 of the proposed Algorithm 1, we have in the worst case to maximize the log-likelihood function $M$ times to find the optimal threshold. Clearly, if $M$ and the number of already introduced thresholds is large, we have to do quite extensive work and the estimation procedure could be very slow. This procedure can be enormously simplified giving to each individual optimal threshold in (3.25) a "relative importance", meaning in this case the improvement in the maximization of the individual log-likelihood functions. In this way, we obtain a rank for the thresholds and introducing the thresholds following this rank we have to make only one estimation by Step. We call the resulting model Tree structured CCC-GARCH model with ranked thresholds.

3.5.2 Numerical results

The individual optimal thresholds (3.25) and the corresponding value of the individual negative log-likelihood functions for the model with this new threshold are summarized in Table 3.4 for the same seven indices of Section 3.4; the results are very similar to the ones proposed in Chapter 2.

With the help of Table 3.4, we can construct the rank of the optimal individual thresholds needed for the estimation of the tree structured CCC-GARCH model with ranked thresholds (the relative improvement in minimizing the individual negative log-likelihood is given between parenthesis):

\[
\begin{align*}
&\hat{d}_1;\text{DIA} \ (25.127) , \ \hat{d}_1;\text{NIK} \ (19.037) , \ \hat{d}_1;\text{DAX} \ (15.213) , \ \hat{d}_1;\text{CAC40} \ (14.002) , \\
&\hat{d}_1;\text{CBS} \ (9.840) , \ \hat{d}_2;\text{CBS} \ (10.465) , \ \hat{d}_2;\text{DAX} \ (9.646) , \ \hat{d}_1;\text{FTAS} \ (7.859) , \\
&\hat{d}_2;\text{FTAS} \ (7.099) , \ \hat{d}_1;\text{BGI} \ (6.689) , \ \hat{d}_2;\text{DIA} \ (6.596) , \ \hat{d}_2;\text{BCI} \ (5.748) , \\
&\hat{d}_3;\text{FTAS} \ (5.254) , \ \hat{d}_2;\text{CAC40} \ (3.895) . 
\end{align*}
\]  

3.26

Note that although in some cases (for example for the FTAS index) there is a better improvement in minimizing the individual negative log-likelihood function introducing the second threshold than introducing the first one, clearly we can not position in the rank the second threshold in front of the first.

The estimated Tree Structured CCC-GARCH model has one threshold more than the one with ranked thresholds. This is due to the fact that the set of the next possible thresholds to introduce consists in the model with ranked thresholds of only one element (i.e. the next threshold coming in the rank) and if we have no improvement in the AIC statistic we stop the estimation procedure. On the other hand, the number of threshold’s candidates using the Tree structured CCC-GARCH Algorithm is in the most cases bigger than one.
### Table 3.4: Individual optimal thresholds and corresponding improvement in the minimization of the individual negative log-likelihood functions for the seven indices introduced in Section 3.4.

Table 3.5 displays the goodness of fit measures for the models in (3.23) and (3.24) constructed with the Tree Structured CCC-GARCH Algorithm with and without ranked thresholds proposed in the last Section. The main implication is that more sophisticated models (not being of synchronization type) do not show worthwhile improvements. Hence, synchronization seems to play a more substantial effect than trying to improve the model-dynamics as in (3.23) or (3.24).
3.6 Estimating risk measures

We consider here the effect of volatility estimates for conditional (dynamical) risk measures for negative returns $\Delta_t$ from a portfolio as in (3.19)-(3.20) given the information $F_{t-1}$ from previous prices. The most popular risk measures are the Value at Risk (VaR) given $F_{t-1}$

$$\delta_q^f = \inf\{\delta \in \mathbb{R} : F_{\Delta_t | F_{t-1}}(\delta) \geq q\}, \quad 0 < q < 1,$$

where $F_{\Delta_t | F_{t-1}}(\cdot)$ denotes the cumulative distribution function of $\Delta_t$ given $F_{t-1}$. This is the quantile of the predictive distribution of the negative portfolio return over the next day. Another measure of risk is the expected shortfall given $F_{t-1}$

$$S_q^f = E[\Delta_t \mid \Delta_t > \delta_q^f, F_{t-1}], \quad 0 < q < 1.$$  

We typically choose $q \in \{0.90, 0.95, 0.99\}$ (note that we consider negative returns). Unlike VaR, the expected shortfall is coherent and satisfies a sub-
additivity requirement, see Artzner et al. (1999). The illustrations are for the real global portfolio already introduced in Section 3.4.3.

### 3.6.1 The estimates

We assume that the dynamics of the negative asynchronized portfolio returns $\Delta_t$ ($t \in \mathbb{Z}$) are given by (3.20). Since

$$ F_{\Delta_t | \mathcal{F}_{t-1}}(\delta) = P \left[ \mu_{t,p} + \sigma_{t,p} Z_t \leq \delta \mid \mathcal{F}_{t-1} \right] = F_Z \left( \frac{\delta - \mu_{t,p}}{\sigma_{t,p}} \right), $$

the risk measures can then be written as

$$ \delta^t_q = \mu_{t,p} + \sigma_{t,p} z_q, \quad 0 < q < 1 \text{ and } $$

$$ S^t_q = \mu_{t,p} + \sigma_{t,p} E[Z \mid Z > z_q], \quad 0 < q < 1, $$

where $z_q$ is the $q$-th quantile of $F_Z(\cdot)$ which by assumption does not depend on time $t$.

Now, we construct estimates for the VaR and for the expected shortfall following two different methods. We can use Extreme Value Theory (EVT) to model the tails of $F_Z(\cdot)$ and we find the following estimates $\hat{\delta}^t_q$ for the VaR measures with the help of the peaks over threshold (POT) method:

$$ \hat{\delta}^t_q = \hat{\mu}_{t,p} + \hat{\sigma}_{t,p} \left( v + \frac{\hat{\beta}}{\xi} \left( \frac{n}{N_v} (1 - q)^{-\xi} - 1 \right) \right), $$

where $v$ is a threshold for residuals $\hat{Z}_t = \frac{\Delta_t - \hat{\mu}_{t,p}}{\hat{\sigma}_{t,p}}$ high enough that a generalized Pareto distribution assumption for the excesses over $v$ may be tenable, $\hat{\beta}$ and $\hat{\xi}$ are the maximum likelihood estimates based on the residuals $\hat{Z}_t$ exceeding $v$, and $N_v = \text{card}\{t : 1, \ldots, n, \hat{Z}_t > v\}$ denotes the number of exceedances of $v$ by $\hat{Z}_1, \ldots, \hat{Z}_n$. For a detailed description, see Embrechts et al. (1997).

Another approach relies on the assumption that the distribution of innovations $Z_t$ in (3.20) is scaled $t_v$, i.e. $Z_t \sim \sqrt{(v - 2)/v} t_v$. An estimate of the expected shortfall is then given by

$$ \hat{S}^t_q = \hat{\mu}_{t,p} + \hat{\sigma}_{t,p} \sqrt{\frac{v - 2}{v}} \left( \frac{1}{1-q} \frac{c}{\bar{v}} \left( 1 + \frac{(\bar{z}_q)^2}{\bar{v}} \right)^{1/2} \right), $$

where the constant $c$ equals $\Gamma\left(\frac{1}{2}(v + 1)\right)/\Gamma\left(\frac{3}{2}v\right)(\Gamma(\frac{3}{2}v)^{-1/2}$, $\bar{z}_q = \frac{\bar{z}_q}{\sqrt{(v - 2)/v}}$ and $\bar{v}$ the maximum likelihood estimate from the multivariate or univariate models as before.
We only show here estimates of the conditional expected shortfall, for illustrative purposes. The results, using the multivariate synchronous and asynchronous CCC-GARCH(1,1) and the univariate GARCH(1,1) model are shown in Figure 3.3 (note that we consider negative returns).

The multivariate models produce more progressive estimates; and the synchronous model exhibits more small scale fluctuations. The results from Section 3.4 support that the multivariate models are better, indicating that the more progressive behavior is appropriate. The better performance of the synchronized model in Section 3.4 also suggests that the small scale movements of the corresponding estimate of the expected shortfall, which wiggles more, is a good feature.

3.6.2 Backtesting

Backtesting the results for the expected shortfall is generally very difficult since a tail phenomenon is involved. As a descriptive tool, rather than a formal test, we show boxplots of residuals

$$
\tilde{R}_t = \frac{\Delta_t - \tilde{S}_t}{\tilde{\sigma}_t, p} I_{\{\Delta_t > \tilde{S}_t\}}.
$$

in Figure 3.4.

Under our model assumptions (3.20) and ignoring estimation effects, we can easily show that these (theoretical) residuals are an i.i.d. sequence with expected value zero. Figure 3.4 yields then additional evidence that the estimate from the univariate model is too conservative and it favors the more progressive behavior of the multivariate models.

3.7 Conclusions

The need to synchronize multivariate financial time series is strongly motivated by the fact that information continues to flow for closed markets while others are still open. Besides the nice interpretative structure of synchronization, we found empirically that it improved the predictive performance of the CCC-GARCH(1,1) model for a 7-dimensional time series of daily exchange rate returns. The predictive gain with synchronization seems also much more prominent than extending the GARCH to a more complex model for approximating general functions of volatility.

For analyzing returns of a univariate portfolio, multivariate analysis of the individual returns still plays a key role. If the portfolio weights change over
Figure 3.3: The expected shortfall for the portfolio returns $\tilde{\Delta}_t$ for the test data days beginning on June 23, 1994 and ending on September 9, 1996. The estimates $S_q^f$ for $q = 0.95$ (top) and $q = 0.99$ (bottom) are obtained using the univariate GARCH(1,1) model (solid line), the synchronous CCC-GARCH(1,1) model (dotted line) and the standard CCC-GARCH(1,1) model without synchronization (dashed line).
Figure 3.4: Boxplots of residuals $\tilde{R}_t$ in (3.27) for test data with $q = 0.95$ (top) and $q = 0.90$ (bottom) from the synchronous CCC-GARCH(1,1) model (left), the standard CCC-GARCH(1,1) model without synchronization (center) and the classical univariate GARCH(1,1) model (right). The number of violations $\Delta_t > \delta^t_q$ is given between parenthesis.
time, which is most often the case in practice, only the multivariate frame-
work allows to treat such a time-changing portfolio, assuming that the indi-
vidual returns are stationary. But even for a time-constant portfolio, there is a
substantial gain when using a multivariate CCC-GARCH(1,1) model for indi-
vidual returns instead of a univariate GARCH(1,1) model for portfolio returns.
It emphasizes the importance of the multivariate approach in general and the
wide range of applications where synchronization can be very valuable.
Seite Leer / Blank leaf
Chapter 4

Volatility Estimation with Functional Gradient Descent for Very High-Dimensional Financial Time Series

4.1 Abstract

We propose a functional gradient descent algorithm (FGD) for estimating volatility and conditional covariances (given the past) for very high-dimensional financial time series of asset price returns. FGD is a kind of hybrid of nonparametric statistical function estimation and numerical optimization. Our FGD algorithm is computationally feasible in multivariate problems with dozens up to thousands of individual return series. Moreover, we demonstrate on some synthetic and real data-sets with dimensions up to 100, that it yields significantly, much better predictions than more classical approaches such as a constant conditional correlation GARCH-type model. Since our FGD algorithm is constructed from a generic algorithm, the technique can be easily adapted to other problems of learning in very high dimensions.

4.2 Introduction

We propose here a version of functional gradient descent (FGD), a recent technique from the area of machine learning (Breiman, 1999; Mason et al., 1999;
Chapter 4. Volatility Estimation with FGD

Friedman et al., 2000; Friedman, 2001). Our FGD method is based on the likelihood-framework of a-priori very general GARCH-type models. FGD has mainly influenced the thinking around so-called Boosting (Freund and Shapire (1996)) which is a machine learning technique for the classification problem (predicting labels or classes from many explanatory variables). Some important modifications of more standard FGD algorithms are necessary to make the approach successful in the very different field here of high-dimensional financial time series.

It is well known that the multivariate approach is needed in many areas such as risk management and portfolio analysis. For example, the study of a portfolio

\[ P_t = \sum_{i=1}^{d} \alpha_{t,i} P_{t,i} \]

with time-changing weights requires the multivariate approach. More generally, the aim is often to estimate the conditional distribution of a “pay-out” function

\[ \psi_t(P_{t,1}, \ldots, P_{t,d}) \]

given the information \( F_{t-1} \), where \( \psi_t \) could also be nonlinear. To do so, it suffices to estimate the distribution of the multivariate returns time series \( X_t \) given \( F_{t-1} \); and this boils mainly down to estimating the volatility matrix \( V_t \) when assuming a model as in (0.2) (estimating the innovation distribution of \( Z_t \) is then easily done via estimated residuals).

Mainly for conceptual purposes, we present in Section 4.4 FGD for the estimation of volatility in univariate time series. The method produces accurate results and is better than classical GARCH prediction. But as argued above, the merit is mainly for very high-dimensional problems where there is virtually no other competitive alternative method; this will be discussed in Section 4.5. Numerical results are illustrated on real and simulated data.

4.3 The generic functional gradient descent algorithm

We are presenting here the main idea of functional gradient descent (FGD) in the framework of general regression. Consider data \( (Y_1, X_1), \ldots, (Y_n, X_n) \), where \( Y_i \) is the response and \( X_i \) the explanatory variable. For simplicity,
4.3. The generic functional gradient descent algorithm

we assume here that $Y_i$ is univariate (e.g. taking values in $\mathbb{R}$) and $X_i$ is $p$-dimensional (e.g. taking values in $\mathbb{R}^p$).

The aim is to estimate a function $F(x)$ such as $F(x) = \mathbb{E}[Y|X = x]$, $F(x) = \text{Var}(Y|X = x)$ or $F(x) = \mathbb{P}[Y = 1|X = x]$ if $Y \in \{0, 1\}$ is binary. The function $F(\cdot)$ can often be represented as the minimizer (in function space) of an expected loss function $\lambda(y, f)$,

$$F(\cdot) = \arg\min_f \mathbb{E}[\lambda(Y, F(X))]$$

As an example, $F(x) = \mathbb{E}[Y|X = x]$ can be represented as the minimizer of such an expected loss with $\lambda(y, f) = (y - f)^2$. The functional gradient descent estimate of $F(\cdot)$ is then constructed from a constrained minimization of the empirical risk

$$\hat{F}_M(\cdot) = \sum_{m=0}^{M} \hat{w}_m \hat{f}_m(\cdot).$$

The constraints require that the solution $\hat{F}(\cdot)$ is an additive expansion of “simple estimates”,

The “simple estimates” are given from a base learner $\delta$ where

$$\delta_X(U)(x) \text{ denotes the predicted value at } x \in \mathbb{R}^p \text{ from the base learner } \delta,$$

using the response vector $U = (U_1, \ldots, U_n) \in \mathbb{R}^n$ and explanatory variables $X = X_1, \ldots, X_n \in \mathbb{R}^p$. Typically, $\delta_X(U)(x)$ is an estimate of $\mathbb{E}[U_1|X_1 = x]$. It is often constructed from (constrained) least squares fitting

$$\delta_X(U)(\cdot) = \arg\min_f \sum_{i=1}^{n} (U_i - f(X_i))^2.$$

It could be the fit from a base learner such as a regression tree, a projection pursuit or a neural net.

The constraints in the additive expansion in (4.1) are automatically built in when proceeding with the following generic FGD algorithm, cf. Friedman (2001).

---

Generic functional gradient descent (FGD)

Step 1 (initialization). Specify the starting function $\hat{F}_0(\cdot)$ and set $m = 1$. 
Step 2 (projection of gradient to base learner). Compute the negative gradient

\[ U_i = -\frac{\partial \lambda(Y_i, F)}{\partial F} \bigg|_{F = \hat{F}_{m-1}(X_i)}, \quad i = 1, \ldots, n, \]
evaluated at the previous estimate \( \hat{F}_{m-1}(\cdot) \) and the data points. Then, fit the negative gradient vector with a base learner \( \delta \)

\[ \hat{f}_m(\cdot) = \delta_x(U)(\cdot). \]
The vector \( (\hat{f}_m(X_1), \ldots, \hat{f}_m(X_n))^T \) can be viewed as a kind of projection of the negative gradient to the base learner.

Step 3 (line search). Perform a one-dimensional optimization for the step-length when updating \( \hat{F}_{m-1} \) with \( \hat{f}_m \),

\[ \hat{w}_m = \arg\min_w \sum_{i=1}^n \lambda(Y_i, \hat{F}_{m-1}(X_i) + w \hat{f}_m(X_i)). \]

Up-date

\[ \hat{F}_m(\cdot) = \hat{F}_{m-1}(\cdot) + \hat{w}_m \hat{f}_m(\cdot). \]

Step 4 (iteration). Increase \( m \) by one and iterate Steps 2 and 3 until stopping with \( m = M \). This produces the FGD estimate

\[ \hat{F}_M(\cdot) = \hat{F}_0(\cdot) + \sum_{m=1}^M \hat{w}_m \hat{f}_m(\cdot). \]
The stopping value \( M \) is chosen to optimize a measure for out-sample prediction.

Remark 1. Initialization in Step 1 was so far believed to be of negligible importance: an initial function often proposed is \( \hat{F}_0 \equiv \bar{Y}_n \). But we will see in Section 4.4, from an empirical point of view, that initialization does play an important role in the application of volatility estimation.

Remark 2. The line search in Step 3 guarantees that the empirical risk is monotonely decreasing with every iteration.

Remark 3. Stopping in Step 4 is important. Typically, the algorithm would fit the data perfectly as iterations tend to infinity.
Numerical optimization has advanced to faster converging algorithms than steepest gradient methods. However, in our setting the slow "convergence" (note that we do not iterate until convergence) is very helpful as a device for regularization.

The name "functional gradient" can be justified as follows. We are asking for the "direction" \( f \) (in function space) such that \( \Lambda(\hat{F}_{m-1} + \varepsilon f) \) most rapidly decreases, for small value of \( \varepsilon \), where \( \Lambda(F) = n^{-1} \sum_{i=1}^{n} \lambda(Y_i, F(X_i)) \). Now, viewing \( \Lambda \) as a functional on \( \text{lin}(\mathcal{F}) \), the set of linear combinations of functions in a suitable class of base learners \( \mathcal{F} \), the desired direction is the negative functional derivative \(-d\Lambda(F, \cdot)\), where

\[
\frac{d\Lambda(F, x)}{dx} = \lim_{\varepsilon \searrow 0} \frac{\Lambda(F + \varepsilon \mathbb{I}_{[x]}) - \Lambda(F)}{\varepsilon}, \quad x \in \mathbb{R}^p,
\]

where \( \mathbb{I}_{[x]} \) denotes the indicator function. We are restricted to choose \( f \in \mathcal{F} \) and cannot choose \( f \) as \(-d\Lambda(F, x)\). Instead, we search for an \( f \) minimizing \( \| -d\Lambda(F) - f \|^2 \), where \( \|g\|^2 = n^{-1} \sum_{i=1}^{n} g(X_i)^2 \). This is equivalent to fit the negative gradient vector \((U_1, \ldots, U_n)^T\) in Step 2 with the base learner \( \delta \) fitted by least squares producing \( \hat{f}_m \in \mathcal{F} \).

Why should we use FGD at all? In very high-dimensional settings, particularly in connection with tree-structured base learners (see Section 4.3.1), it is a feasible method aiming to improve the starting function. Bühlmann and Yu (2001) prove a rate-optimality result for FGD with the quadratic loss function \( \lambda \) in one-dimensional function estimation of unknown smoothness. Thus, in such low-dimensional toy problems, FGD achieves some optimality criterion asymptotically. But the theory for very high-dimensional settings or other loss functions \( \lambda \) is not rigorously developed so far. Good empirical performance of FGD, particularly in the area of classification with boosting (Freund and Shapire, 1996; Friedman et al., 2000; Friedman, 2001), and the asymptotic results in univariate function estimation support evidence that FGD is a reasonable strategy. These arguments are further strengthened by the fact that there are not many other methods whose computational costs remains manageable in very high-dimensional, large scale problems.

### 4.3.1 Choice of the base learner

The base learner in Step 2 of the functional gradient descent algorithm, producing the additive terms \( \hat{f}_m(\cdot) \), obviously determines the FGD estimate \( \hat{F}_M \). The base learner should be "weak", i.e. not too complex or high-dimensional, so that FGD would not immediately produce an overfitted estimate with the first iteration. By adding further additive terms with every iteration, we in-
crease complexity or dimensionality of the FGD estimate \( \hat{F}_M \). However, this increase in complexity is not linear and performing further iterations typically changes complexity only by "small" amounts; for example, it is shown for certain cases to become exponentially diminishing as FGD iterations grow (Bühlmann and Yu, 2001). As usual, a bias-complexity trade-off is present: the complexity of the base learner and the number of FGD iterations determine the trade-off in a somewhat unusual way, see also Bühlmann and Yu (2001).

Often, decision trees are used as base learners. Particularly in high dimensions, they have the ability to do variable selection by choosing just a few of the explanatory variables for prediction. We will consider here decision trees but also projection pursuit (Friedman and Stuetzle, 1981) as base learners. It is often desirable to make a base learner sufficiently "weak", i.e. of sufficiently low complexity. A simple but effective solution to achieve this is via shrinkage towards zero: the up-date \( \hat{w}_m \hat{f}_m(\cdot) \) in Step 3 of the FGD algorithms is replaced by

\[
\nu \hat{w}_m \hat{f}_m(\cdot), \ 0 < \nu \leq 1.
\]  

Obviously, this reduces the variance (a complexity measure) by the factor \( \nu^2 \).

### 4.4 Univariate volatility estimation with FGD

Having a univariate time series of observed prices \( P_0, P_1, \ldots, P_n \) of an asset, we consider their returns \( X_t = \frac{P_t - P_{t-1}}{P_{t-1}} \), or \( X_t = \log(P_t/P_{t-1}) \) alternatively, and assume stationarity at least in a suitable time-window. The equi-distant time-spacing \( \Delta t = t - (t - 1) = 1 \) is often one business day. Our working model for the time series of such price returns is

\[
X_t = \sigma_t Z_t, \quad \sigma_t^2 = F(X_{t-1}, X_{t-2}, \ldots),
\]  

where the innovations \( Z_t \) are i.i.d. with \( \mathbb{E}[Z_t] = 0, \ \text{Var}(Z_t) = 1 \) and \( Z_t \) independent from \( \{X_s; s < t\} \); the function \( F : \mathbb{R}^\infty \rightarrow \mathbb{R}^+ \) is assumed a priori to be very general. It is later the FGD estimate which constrains \( F(\cdot) \) to be of a more particular, but often still general form.

For FGD, we assume that \( F : \mathbb{R}^p \rightarrow \mathbb{R}^+ \) with \( p \) finite; but we allow the starting function \( \hat{F}_0(\cdot) \) to depend on the whole past of the time series. We choose the loss-function \( \lambda(\cdot, \cdot) \) from the maximum-likelihood framework with innovations \( Z_t \sim \mathcal{N}(0, 1) \):

\[
\lambda(y, f) = -\log(f^{-1/2} \varphi(\sqrt{f}^{-1/2} y)) = \frac{1}{2} \left( \log(f) + \frac{y^2}{f} + \log(2\pi) \right).
\]
4.4. Univariate volatility estimation with FGD

(If course, we could drop the log(2π) term which would then represent a simpler, equivalent loss function). This, because with Gaussian innovations in (4.3), the negative log-likelihood (conditional on the first \( p \) values) is

\[
-\sum_{t=p+1}^{n} \log \left( F(X_{t-p}^{-1})^{-1/2} \varphi(X_t F(X_{t-p}^{-1})^{-1/2}) \right),
\]

where \( X_{t-p}^{-1} = X_{t-1}, \ldots, X_{t-p} \). The partial derivative of the loss-function is

\[
\frac{\partial \lambda(y, f)}{\partial f} = (f^{-1} - y^2 f^{-2})/2
\]

and the FGD algorithm from Section 4.3 can now be used.

Naturally, we can assume other, different from the standard normal, distribution functions for the innovations \( Z_t \) in (4.3) that better replicate the behaviour of return time series. For example, when we assume scaled \( t_\eta \)-distributed innovations \( Z_t \), the loss function equals

\[
\lambda(y, f) = \frac{1}{2} \log(f^{\eta-2}) - \log(c) + \frac{\eta + 1}{2} \log \left( 1 + \frac{y^2}{\eta} \right),
\]

where the constant \( c = \frac{\Gamma \left( \frac{1}{2} (\eta + 1) \right)}{\Gamma \left( \frac{1}{2} \right)} \left( \frac{1}{\eta \pi} \right)^{\frac{1}{2}} \) and \( \frac{\eta - 2}{\eta} \) is the scaling factor such that \( Z_t \) have unit variance. The problem using scaled \( t_\eta \) distribution function for the innovations is that in the FGD procedure also the degrees of freedom \( \eta \) have to be contemporarily estimated. If the parameter \( \eta \) is in every iteration step updated, the loss function can be very unstable from step to step and this can lead to a wrong minimization of the empirical risk. A possible solution could be to fix the degrees of freedom parameter \( \eta \) (for example the optimal degrees of freedom from a GARCH(1,1) fit). For our simulations and real data examples, we always assume the innovations \( Z_t \) in (4.3) to be standard normally distributed.

As a starting function, we propose to use the fit from a GARCH(1,1) model (Bollerslev, 1986)

\[
\hat{F}_0(X_{t-1}, X_{t-2}, \ldots) = \hat{\alpha}_0 + \hat{\alpha}_1 X_{t-1}^2 + \hat{\beta} \hat{F}_0(X_{t-2}, X_{t-3}, \ldots),
\]

with parameters estimates \( \hat{\alpha}_0, \hat{\alpha}_1, \hat{\beta} \) from parametric maximum-likelihood in the GARCH(1,1) model with Gaussian innovations.

Summarizing, the FGD algorithm for univariate volatility estimation then looks as follows.
FGD for univariate volatility

Step 1 (initialization). Choose the starting function $\hat{F}_0(\cdot)$ from (4.4) and denote by $\hat{F}_0(t) = \hat{F}_0(X_1^{t-1})$. Set $m = 1$.

Step 2 (projection of gradient to base learner). Compute the negative gradient

$$U_t = (X_t^2 \hat{F}_{m-1}(t)^{-2} - \hat{F}_{m-1}(t)^{-1})/2, \ t = p + 1, \ldots, n$$

Then, fit the negative gradient vector with a base learner, using always the first $p$ time-lagged predictor variables (i.e. $X_{t-p}^t$ is the predictor for $U_t$)

$$\hat{F}_m(\cdot) = \delta_X(U)(\cdot).$$

Steps 3 and 4. As in the generic FGD algorithm, generally with shrinkage as in (4.2).

The important issue is that the starting function $\hat{F}_0(\cdot)$ matters a lot for obtaining good volatility estimates. We illustrate this in Figure 4.1 by one simulation from the model (4.5) with sample size $n = 1000$. The out-sample OS-L2 losses with $n_{out} = 1000$ (see Section 4.4.1) are 174.5 (FGD with tree learners and constant starting function $\hat{F}_0(x) = S^2$ being the empirical overall variance), 94.6 (classical GARCH(1,1) prediction) and 66.9 (FGD with tree learners and starting function from a GARCH(1,1) fit).

Also, it is advisable to allow for shrinkage in Step 3. Regarding the base learner $\delta$, we have considered regression trees (Breiman et al., 1984) and projection pursuit regression (Friedman and Stuetzle, 1981). With regression trees, the FGD algorithm for univariate volatility estimation can be further modified to achieve marginal improvements taking advantage of the fact that the learner is a partitioning method which allows a more (statistically) efficient implementation for fitting with respect to the best improvement of the loss function $\lambda$.

FGD with tree learners

Modified steps 2 and 3. Given a negative gradient vector $U$, fit a regression tree to $U$ by least squares. This produces a partition $\{R_1, \ldots, R_k\}$ of the predictor space $\mathbb{R}^p$

$$\bigcup_{j=1}^k R_j = \mathbb{R}^p, \ R_i \cap R_j = \emptyset (i \neq j).$$
4.4. Univariate volatility estimation with FGD

Figure 4.1: Outsample errors $\hat{\sigma}_i^2 - \sigma_i^2$ versus true $\sigma_i^2$ in simulated model (4.5). Top left: FGD with trees using constant starting function $\tilde{F}_0(x) = S^2$ (empirical marginal variance); Top right: GARCH(1,1) prediction; Bottom: FGD with trees using the GARCH(1,1) fit as starting function.

(I.e., the partition is such that $\sum_{i=1}^n (U_i - \sum_{j=1}^k \hat{\beta}_j 1_i(X_i \in S_j))^2$ is minimal). Then, proceed with line searches for all $k$ partition cells,

$$\hat{y}_{m,j} = \arg\min_{\gamma} \sum_{i; X_i \in S_j} \lambda(Y_i, \tilde{F}_{m-1}(X_i) + \gamma 1_{X_i \in S_j}), \quad j = 1, \ldots, k.$$
Finally, update
\[ \hat{F}_m(x) = \hat{F}_{m-1}(x) + \sum_{j=1}^{k} \hat{\gamma}_{m,j} \mathbb{I}_{[x \in \mathcal{R}_j]}. \]

This is the same generic tree algorithm as in Friedman (2001).

With a \(S\)-term projection pursuit regression, the fitted learner takes the form
\[ \sum_{j=1}^{S} (\hat{\beta}_j \cdot \hat{h}_j(\hat{\alpha}_j^\prime x)), \]
with \(\hat{h}_j(\cdot)\) a (kernel) smoother, \(j = 1, \ldots, S\). The estimates are in this case constructed in a stepwise iterative manner (see also Friedman and Stuetzle, 1981), where in the \(k\)-th iteration \((k = 1, \ldots, S)\) we have:

\[
\begin{align*}
\hat{\alpha}_k, \hat{\beta}_k &= \arg\min_{\alpha, \beta} \sum_{i=1}^{n} \left( U_i - \sum_{j=1}^{k-1} \left( \hat{\beta}_j \hat{h}_j(\hat{\alpha}_j^\prime X_i) \right) - \beta \hat{f}_{k;\alpha}(\alpha^\prime X_i) \right)^2, \\
\hat{h}_{k;\alpha}(x) &= \arg\min_{\alpha} \sum_{i=1}^{n} \left( U_i - \sum_{j=1}^{k-1} \left( \hat{\beta}_j \hat{h}_j(\hat{\alpha}_j^\prime X_i) \right) - \alpha \right)^2 K\left( \frac{x - \alpha^\prime X_i}{g} \right),
\end{align*}
\]

where \(K(\cdot)\) is a kernel function and \(g\) a bandwidth. The optimal parameters \(\hat{\alpha}_j\) and \(\hat{\beta}_j\), \(j = 1, \ldots, k - 1\) are in each iteration backfitted and updated.

### 4.4.1 Numerical results for simulated data

We simulate from the following model,
\[
X_t = \sigma_t Z_t, \quad \sigma_t^2 = F(X_{t-1}, \sigma_{t-1}^2),
\]
\[
F(x, \sigma^2) = (0.1 + 0.2 |x| + 0.9x^2) \cdot (0.8 \exp(-1.5 |x| |\sigma|)) + (0.4x^2 + 0.5\sigma^2)^{3/4}, \tag{4.5}
\]
where \(Z_t \sim \mathcal{N}(0, 1)\) is as in (4.3).
4.4. Univariate volatility estimation with FGD

For quantifying the goodness of fit, we consider various measures:

\[
\text{out-sample negative log-likelihood: } \sum_{t=1}^{n_{out}} \lambda(Y_t, \hat{F}(Y_{t-1}^{t-1})), \tag{4.6}
\]

\[
\text{IS-L}_2 = \sum_{t=1}^{n} \sigma_t^2 - \hat{F}(X_{t-1})^2 \text{ (in-sample loss)}, \tag{4.7}
\]

\[
\text{OS-L}_2 = \sum_{t=1}^{n_{out}} \sigma_t^2 - \hat{F}(Y_{t-1}^{t-1})^2 \text{ (out-sample loss)}, \tag{4.8}
\]

where \(Y_1, \ldots, Y_{n_{out}}\) are new test observations, independent from but with the same distribution as the data \(X_1, \ldots, X_n\). Both, the IS- and OS-L2 statistics are interesting measures in simulations, but we can’t evaluate them for real data. The out-sample negative log-likelihood is a more generally applicable measure for out-sample performance.

<table>
<thead>
<tr>
<th>Model</th>
<th>Performance measure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OS -log-likelihood</td>
</tr>
<tr>
<td>GARCH(1,1)</td>
<td>1656.363</td>
</tr>
<tr>
<td>FGD with tree using (p = 1, v = 0.1, L = 3)</td>
<td>1653.434</td>
</tr>
<tr>
<td>FGD with tree using (p = 2, v = 0.1, L = 5)</td>
<td>1653.534</td>
</tr>
<tr>
<td>FGD with PPR using (p = 2, v = 0.1, S = 3)</td>
<td>1654.296</td>
</tr>
<tr>
<td>FGD with PPR using (p = 2, v = 0.05, S = 2)</td>
<td>1654.672</td>
</tr>
</tbody>
</table>

Table 4.1: Goodness of fit measures (on average) for fifty simulations from model (4.5) with sample size \(n = 1000\). Out-sample performances OS-L2 and negative log-likelihood as in (4.6)-(4.8) are evaluated with test-set of size \(n_{out} = 1000\). Notation: number of lagged values (\(p\)), shrinkage factor (\(v\)), number of terminal nodes (\(L\)) and number of ridge functions (\(S\)).

Table 4.1 shows the result for 50 independent realizations from model (4.5). Sample size is \(n = 1000\) and test-set size is \(n_{out} = 1000\). For this case, FGD with trees as in Section 4.4 is better than FGD with projection pursuit, and both FGD techniques outperform the predictions from classical
GARCH(1,1). The differences in the out-sample log-likelihood are small despite that the actual differences in volatility are substantial. This phenomenon is well known and occurs because the out-sample log-likelihood measures quality for predicting future returns and not future volatilities; the former is much more noisy than the latter. Thus, similar out-sample log-likelihoods (or other prediction losses) with different methods are not implying that the methods are similar in terms of the differences between estimated and true volatility, as observed in Table 4.1.

**Figure 4.2:** 300 days of predicted conditional variance estimates from the GARCH(1,1) model (dotted line) and the FGD with tree learners with optimal parameters \( v, p \) and \( L \) (dashed line), superimposed on the simulated values \( \hat{\sigma}_t^2 \) (solid line). The data are simulated from the model (4.5).

Figure 4.2 shows how well the FGD estimates replicate the “wizzly” be-
4.4. Univariate volatility estimation with FGD

haviour of the simulated conditional variances.

4.4.2 Some real example

We consider three stock indices with 1500 daily negative log-returns: from the German DAX, the British FTAS and the Japanese NIKKEI index between January 21, 1993 and September 13, 1999. We use the first 1000 time points for training (fitting) and the remaining 500 for out-sample testing (evaluation). We compare the results with the ones from a classical GARCH(1,1) fit.

<table>
<thead>
<tr>
<th>Data</th>
<th>Model</th>
<th>Performance measure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Outsample log-lik.</td>
</tr>
<tr>
<td>DAX</td>
<td>GARCH(1,1)</td>
<td>1079.95</td>
</tr>
<tr>
<td></td>
<td>FGD with tree ( p = 5, \ v = 0.5, \ L = 30 )</td>
<td>1071.99</td>
</tr>
<tr>
<td></td>
<td>FGD with PPR ( p = 5, \ v = 0.1, \ S = 1 )</td>
<td>1073.78</td>
</tr>
<tr>
<td>FTAS</td>
<td>GARCH(1,1)</td>
<td>878.817</td>
</tr>
<tr>
<td></td>
<td>FGD with tree ( p = 5, \ v = 1, \ L = 2 )</td>
<td>872.737</td>
</tr>
<tr>
<td></td>
<td>FGD with PPR ( p = 5, \ v = 0.5, \ S = 1 )</td>
<td>871.766</td>
</tr>
<tr>
<td>NIK</td>
<td>GARCH(1,1)</td>
<td>1003.04</td>
</tr>
<tr>
<td></td>
<td>FGD with tree ( p = 5, \ v = 1, \ L = 2 )</td>
<td>995.364</td>
</tr>
<tr>
<td></td>
<td>FGD with PPR ( p = 2, \ v = 0.5, \ S = 1 )</td>
<td>998.465</td>
</tr>
</tbody>
</table>

Table 4.2: Goodness of fit measures for negative returns from the German DAX, the British FTAS and the Japanese NIKKEI index between January 21, 1993 and September 13, 1999. The models are fitted assuming \( \mathcal{N}(0, 1) \)-distributed innovations.
### Chapter 4. Volatility Estimation with FGD

<table>
<thead>
<tr>
<th>Model 1 - Model 2</th>
<th>Performance measure</th>
<th>Out. log-likelihood for</th>
<th>OS-PL$_2$ for</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>DAX</td>
<td>FTAS</td>
</tr>
<tr>
<td>FGD with tree - GARCH(1,1)</td>
<td>$-0.62$</td>
<td>$-1.02$</td>
<td>$-1.77$</td>
</tr>
<tr>
<td></td>
<td>(0.27)</td>
<td>(0.15)</td>
<td>(0.038)</td>
</tr>
<tr>
<td>FGD with PPR - GARCH(1,1)</td>
<td>$-1.17$</td>
<td>$-1.06$</td>
<td>$-1.43$</td>
</tr>
<tr>
<td></td>
<td>(0.12)</td>
<td>(0.14)</td>
<td>(0.076)</td>
</tr>
<tr>
<td>FGD with tree - FGD with PPR</td>
<td>$-0.18$</td>
<td>$0.09$</td>
<td>$-0.55$</td>
</tr>
<tr>
<td></td>
<td>(0.43)</td>
<td>(0.46)</td>
<td>(0.29)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model 1 - Model 2</th>
<th>Performance measure</th>
<th>Out. log-likelihood for</th>
<th>OS-PL$_2$ for</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>DAX</td>
<td>FTAS</td>
</tr>
<tr>
<td>FGD with tree - GARCH(1,1)</td>
<td>$-0.21$</td>
<td>$0.95$</td>
<td>$1.66$</td>
</tr>
<tr>
<td></td>
<td>(0.42)</td>
<td>(0.17)</td>
<td>(0.048)</td>
</tr>
<tr>
<td>FGD with PPR - GARCH(1,1)</td>
<td>$1.68$</td>
<td>$0.34$</td>
<td>$0.10$</td>
</tr>
<tr>
<td></td>
<td>(0.046)</td>
<td>(0.37)</td>
<td>(0.46)</td>
</tr>
<tr>
<td>FGD with tree - FGD with PPR</td>
<td>$-1.49$</td>
<td>$0.64$</td>
<td>$0.27$</td>
</tr>
<tr>
<td></td>
<td>(0.07)</td>
<td>(0.26)</td>
<td>(0.40)</td>
</tr>
</tbody>
</table>

**Table 4.3:** Testing differences of performance for the negative returns of the German DAX, the British FTAS and the Japanese NIKKEI index. The values of test statistics and corresponding *P*-values (between parenthesis) are given for *t*-type test (top) and sign-type test (bottom).

We consider two goodness of fit measures for such real data. One is the negative out-sample log-likelihood in (4.6). Alternatively, we also look at the out-sample prediction $L_2$ losses,

$$OS-PL_2 = \sum_{t=1001}^{1500} |X_t^2 - \hat{F}(t)|^2.$$  \hspace{1cm} (4.9)

The IS-PL$_2$ is defined analogously using the training sample only. The results are given in Table 4.2 and are very similar to the ones obtained in the simula-
4.5. Volatility estimation for high multivariate time series

Estimations with respect to the out-sample log-likelihood measure (improvement over the estimates from a GARCH(1,1) fit of about 1-3%).

For real data, it is harder to detect differences because appropriate prediction measures like the out-sample negative log-likelihood is affected by a large noise component, as mentioned above (see also Section 2.6.2). We will show here and in Section 4.5.3 how testing can be used to judge whether out-sample log-likelihoods (or other losses) are significantly different, even when they appear to be similar.

The out-sample criterion is here generally denoted by

\[ \sum_{t=1}^{n_{out}} L_t, \]

where \( L_t \) is the out-sample loss for out-sample prediction at time \( t \) depending on a technique or model, i.e. \( L_t = L_{t,\text{model}} \). The differences between two models with respect to out-sample performance is then

\[ \Delta(\text{model}_1, \text{model}_2) = \sum_{t=1}^{n_{out}} D_t, \quad D_t = L_{t,\text{model}_1} - L_{t,\text{model}_2}. \]

We aim to test whether \( E[D_t] = 0 \) against one- (or two-) sided alternatives. We consider the same versions of the \( t \)- and sign-test, adapted to the case of dependent observations \( D_t \), already introduced in Section 3.4.2. The results are summarized in Table 4.3 and shows that in one of six for the \( t \)-type and two of six cases for the sign-type test, respectively, the nullhypothesis is rejected at the 5% level and the optimal FGD estimates are better than the GARCH(1,1) ones. In no occasion the nullhypothesis is rejected at the 5% level for differences between FGD estimates with tree or PPR as weak learners; there is not a big difference, even if FGD with tree yields in the most cases to better results.

4.5 Volatility estimation for high multivariate time series

In the multivariate set-up, we have time series of asset prices

\[ \{ P_{t,i}; t = 0, 1, \ldots, n, \ i = 1, \ldots, d \}. \]

Their returns are defined as

\[ X_{t,i} = (P_{t,i} - P_{t-1,i})/P_{t-1,i}, \ t = 1, \ldots, n. \]
As mentioned already in Section 4.2, the challenging problem is prediction of the multivariate volatility matrix

$$V_t = \text{Cov}_{d \times d}(X_t | \mathcal{F}_{t-1}), \ X_t = (X_{t,1}, \ldots, X_{t,d})^T$$

in dimensions in the hundreds. FGD becomes a powerful strategy to construct computable and good predictions for $V_t$.

We assume stationarity (at least within a suitable time-window). Our working model is a generalization of the constant conditional correlation (CCC) GARCH model (Bollerslev, 1990), already introduced in Section 3.3.2,

$$X_t = \Sigma_t Z_t, \quad (4.10)$$

where we assume the following:

(A1) (innovations) $\{Z_t\}_{t \in \mathbb{Z}}$ is a sequence of i.i.d. multivariate innovations with spherical distribution (e.g. multivariate normal) having mean zero and covariance matrix $\text{Cov}(Z_t) = I_d$. Moreover, $Z_t$ is independent from $\mathcal{F}_{t-1} = \{X_s; s \leq t - 1\}$.

(A2) (CCC construction) The conditional covariance matrix

$$V_t = \text{Cov}(X_t | \mathcal{F}_{t-1}) = \Sigma_t \Sigma_t^T$$

is almost surely positive definite for all $t$. The typical element of $V_t$ is $v_{t,ij} = \rho_{ij}(v_{t,ii}v_{t,jj})^{1/2}$ ($i, j = 1, \ldots, d$). The parameter $\rho_{ij} = \text{Corr}(X_{t,i}, X_{t,j} | \mathcal{F}_{t-1})$ equals the constant conditional correlation and hence $-1 \leq \rho_{ij} \leq 1$, $\rho_{ii} = 1$.

(A3) (functional form) The conditional variances are of the form

$$v_{t,ii} = \sigma_{t,i}^2 = \text{Var}(X_{t,i} | \mathcal{F}_{t-1}) = F_i([X_{t-j,k}; j = 1, 2, \ldots, k = 1, \ldots, d])$$

where $F_i$ takes values in $\mathbb{R}^+$. Note that (A2) can be represented in matrix form as

$$V_t = \Sigma_t \Sigma_t^T = D_t R D_t, \quad D_t = \text{diag}(\sigma_{t,1}, \ldots, \sigma_{t,d}), \quad R = [\rho_{ij}]_{i,j=1}^d.$$

For estimating the functions $F_i(\cdot)$ in (A3), we propose FGD and restrict $F_i(\cdot) : \mathbb{R}^{pd} \rightarrow \mathbb{R}^+$ with $p$ finite, i.e. involving the first $p$ lagged multivariate
4.5. Volatility estimation for high multivariate time series

Observations. Estimation of the correlations can be easily done via empirical moments of residuals.

To proceed with a FGD technique, we first specify a suitable loss function. Assuming multivariate normality of the innovations \( Z_t \), the negative log-likelihood (conditional on the first \( p \) variables) is

\[
-n \sum_{t=p+1}^{n} \log \left( (2\pi)^{-d/2} \text{det}(V_t)^{-1/2} \exp(-X_t^T V_t^{-1} X_t/2) \right) =
\]

\[
- \sum_{t=p+1}^{n} \left( \log(\text{det}(D_t)) + \frac{1}{2} (D_t^{-1} X_t)^T R^{-1}(D_t^{-1} X_t) \right) +
\]

\[+ n'd \log(2\pi)/2 + n' \log(\text{det}(R))/2\]

where \( D_t \) is diagonal with elements \( \sqrt{F_i(X_{t-p}^{-1})} \) and \( n' = n - p \). This motivates the following loss function

\[
\lambda_R(Y, f) = \log(\text{det}(D(f))) + \frac{1}{2} (D(f)^{-1} Y)^T R^{-1}(D(f)^{-1} Y) +
\]

\[+ \frac{1}{2} \log(\text{det}(R)) + \frac{d}{2} \log(2\pi), \quad D(f) = \text{diag}(f_1, \ldots, f_d).\] (4.11)

(Since the terms \( d \log(2\pi)/2 \) and \( \log(\text{det}(R))/2 \) are constants, they could be dropped). As pointed out with the subscript, the loss function depends on the unknown correlation matrix \( R \). Our FGD algorithm will be constructed iteratively by estimating \( R \) and using the loss function with the estimated \( R \) to get an estimate for all \( F_i \)'s.

Having a (previous) estimate \( \hat{F} = (\hat{F}_1, \ldots, \hat{F}_d) \), we construct the following estimate for the correlation matrix \( R \). Build the residuals

\[
\hat{e}_{t,i} = X_{t,i}/\hat{F}_i(X_{t-1}, \ldots)^{1/2}, \quad t = p + 1, \ldots, n
\]

and define

\[
\hat{R} = (n - p)^{-1} \sum_{t=p+1}^{n} \hat{e}_t \hat{e}_t^T, \quad \hat{e}_t = (\hat{e}_{t,1}, \ldots, \hat{e}_{t,d})^T.\] (4.12)

The partial derivatives of the loss function are

\[
\frac{\partial \lambda_R(Y, f)}{\partial f_i} = (f_i - \sum_{j=1}^{d} \frac{y_i y_j y_j}{\sqrt{f_i^3/2} \sqrt{f_j^3/2}})/2, \quad i = 1, \ldots, d,\] (4.13)
Chapter 4. Volatility Estimation with FGD

where $[y_{i,j}]_{i,j=1}^{d} = R^{-1}$. This will be used when computing negative gradients (see Step 2 in the generic FGD algorithm) for every component $i = 1, \ldots, d$.

As a starting function, we propose to use the fit from a CCC-GARCH(1,1) model (Bollerslev, 1990) which is of the form (4.10) with (A3) specified to

$$F_i(X_{t-1}, X_{t-2}, \ldots) = \alpha_0, + \alpha_{1,i} X_{t-1,i}^2 + \beta_{0,i} \sigma_{t-1,i}^2, \quad i = 1, \ldots, d. \quad (4.14)$$

For $d$ large, the estimates are constructed with maximum likelihood from the $d$ individual series. This ignores the more general correlation structure in $R$, causing some statistical decrease in efficiency, but gaining the advantage that the individual estimates are computable (in parallel) in very high dimensions $d$.

The FGD algorithm for multivariate volatility looks as follows.

---

**FGD for multivariate volatility**

**Step 1 (initialization).** Choose the starting function $\hat{F}_{i, 0}(\cdot)$ and denote by

$$\hat{F}_{i, 0}(t) = \hat{F}_{i, 0}(X_{t-1}, X_{t-2}, \ldots), \quad i = 1, \ldots, d.$$

Compute $\hat{R}_0$ as in (4.12) using $\hat{F}_0$. Set $m = 1$.

**Step 2i (projection of component gradients to base learner).** Compute the negative gradient

$$U_{t,i} = - \frac{\partial \hat{R}_{m-1}(X_t, F)}{\partial F_i} \bigg|_{F = \hat{F}_{m-1}(t)}, \quad t = p + 1, \ldots, n.$$

This is explicitly given in (4.13). Then, fit the negative gradient vector $U_i = (U_{p+1,i}, \ldots, U_{n,i})^T$ with a base learner, using always the first $p$ time-lagged predictor variables (i.e. $X_{t-p}$ is the predictor for $U_{t,i}$)

$$\hat{f}_{m,i}(\cdot) = g_X(U_i)(\cdot).$$

**Step 3i (line search).** Perform a one-dimensional optimization for the step-length,

$$\hat{w}_{m,i} = \text{argmin} \sum_{t=p+1}^{n} \lambda \hat{R}_{m-1}(X_t, \hat{F}_{m-1}(t)) + w \hat{f}_{m,i}(X_{t-p}^{t-1}).$$

---
4.5. Volatility estimation for high multivariate time series

\((\hat{F}_{m-1}(t) + w_{m,i}(\cdot)\) is defined as the function which is constructed by adding in the \(i\)th component only). This can be expressed more explicitly by using (4.11).

**Step 4 (up-date).** Select the best component as

\[ i^*_m = \arg\min_i \sum_{t=p+1}^{n} \lambda_{\hat{F}_{m-1}}(X_t, \hat{F}_{m-1}(t) + \hat{w}_{m,i} \hat{f}_{m,i}(X_{t-p})). \]

Up-date

\[ \hat{F}_m(\cdot) = \hat{F}_{m-1}(\cdot) + \hat{w}_{m,i^*_m} \hat{f}_{m,i^*_m}(\cdot). \]

Then, compute the new estimate \( \hat{F}_m \) according to (4.12) using \( \hat{F}_m \).

**Step 5 (iteration).** Increase \( m \) by one and iterate Steps 2–4 until stopping with \( m = M \). This produces the FGD estimate

\[ \hat{F}_M(\cdot) = \hat{F}_0(\cdot) + \sum_{m=1}^{M} \hat{w}_{m,i^*_m} \hat{f}_{m,i^*_m}(\cdot). \]

As in the generic algorithm, the stopping value \( M \) is chosen to optimize a measure for out-sample prediction.

Note that shrinkage as in (4.2) is often useful in Steps 2–4. As in the univariate case, the starting function \( \hat{F}_0(\cdot) \) matters a lot for obtaining good volatility estimates.

A crucial difference to the multivariate (multi-class) FGD algorithm from Friedman et al. (2000), who propose to cycle through the dimensions in a systematic way one after the other, is that our construction is with candidate components in Steps 2; and 3; and choosing the component \( i^* \) in Step 4 which brings the most substantial improvement ("the steepest direction") in a single FGD iteration. Cycling through in a systematic way forces to add complexity of the FGD estimate for every component: but this isn't realistic if one time series is "of simpler structure" than others. We illustrate this for one realization of a 3-dimensional model as in (4.10): the volatilities are given by \( F_1 \) and \( F_2 \) from (4.17) below with fixed parameters \( \alpha_1 = 0.1, \alpha_2 = 0.5, \alpha_3 = 0.2, \alpha_4 = 0.75, \alpha_5 = 0.5 \) and \( F_3 \) from (4.18) below with fixed parameters \( \alpha_1 = 0.1, \alpha_2 = 0.9, \alpha_3 = -1.5, \alpha_4 = 0.5 \).
Table 4.4: Goodness of fit measures for one three-dimensional realization of size \( n = 1000 \) from model (4.10) with \( \mathbf{Z}_t \sim \mathcal{N}_3(0, I) \) and individual conditional variances defined by (4.17)-(4.18). Out-sample negative log-likelihood as in (4.15) and individual OS-L2 as in (4.8) are evaluated using test-sets of size \( n_{out} = 1000 \).

Table 4.4 impressively demonstrates that choosing the best component \( i^* \) brings substantial improvements for the third series. The goodness of fit criteria used are here the following. The out-sample negative log-likelihood function (the out-sample loss \( X \)) is

\[
\sum_{t=n+1}^{n+n_{out}} \lambda_R(\mathbf{X}_t, \hat{\mathbf{F}}(t))
\]

where the estimates, denoted here by a "\( \hat{\cdot} \)", are based on the training data \( \mathbf{X}_1, \ldots, \mathbf{X}_n \). The other criteria are the univariate in- and out-sample \( L_2 \)-losses from (4.7) and (4.8).

### 4.5.1 Base learners with variable selection

Regarding the base learner \( \delta \), we have considered regression trees (Breiman et al., 1984). Fitting regression trees for the components in our multivariate FGD (Step 2) can be modified as in the univariate setting described in Section 4.4. In very high dimensions, it is essential to use a base learner which selects only a few variables from a huge predictor space. Decision trees have this property: when having \( L \) terminal nodes, the decision tree base learner selects at most
4.5. Volatility estimation for high multivariate time series

$L - 1$ different explanatory variables. In combination with FGD, we would then add to the starting function $\tilde{F}_0$ an additive correction involving at most $M(L - 1)$ different variables which may be much lower than $dp$ which is the number of the $p$ lagged predictor variables in every of the $d$ time series.

4.5.2 Computational cost and parallelization

The computational complexity of our multivariate FGD algorithm (without initialization) is

$$Md \cdot \text{complexity}(\delta) + Md \cdot \text{complexity(line search)} + M \cdot \text{complexity}(\tilde{R})$$

The complexity of the estimate $\tilde{R}$ in (4.12) is quadratic in the dimension but the computational cost of this moment estimator is not substantial relative to the other tasks. The numerical line search has to be done $Md$ times which can contribute substantially to computing time when $d$ is in the hundreds, $M$ up to 100 (which is most often big enough) and $n$ around 1000 (which is at the upper range, due to possible non-stationarity, when using daily financial data). Likewise, the base learner has to be fitted $Md$ times. When $d$ gets large, the increase of the factor $Md$ is compounded by the fact that fitting the base learner becomes more costly. For example with decision trees, the computational complexity grows linearly in $d$. For $d$ in the range of 100, $M$ up to 100 and $n$ about 1000, the total complexity is still feasible when using decision trees.

When $d$ is in the order of 1000, the simple implementation of the FGD algorithm becomes quickly computationally expensive. Fortunately, it is very easy to implement a parallel version. Searching the best component $i^*_m$ in Steps 2 and 3 requires consideration of all components $i = 1, \ldots, d$: this can be parallelized immediately, reducing the computational cost a lot. Furthermore, when using trees as base learner in high dimensions, their fitting, which requires visiting all components $i = 1, \ldots, d$ could also be substantially parallelized. While this is a bit more sophisticated, parallelization of the Steps 2 and 3 is immediate. Such relatively simple parallelizations make our FGD algorithm feasible in dimensions $d$ in the thousands.

4.5.3 Numerical results

100-dimensional simulated data

We simulate a 100-dimensional series of sample size $n = 1000$ from model (4.10) with $Z_t \sim \mathcal{N}_{100}(0, I)$ and various volatility functions $F_j$. One such
function is the classical GARCH(1,1) volatility

$$\sigma_{t,i}^2 = F_i(X_{t-1,i}, \sigma_{t-1,i}^2)$$

where

$$F_i(x, \sigma^2) = \alpha_0 + \alpha_1 x^2 + \beta \sigma^2,$$

$$\alpha_0 \sim \text{Unif}([0, 0.2]), \; \alpha_1 \sim \text{Unif}([0.05, 0.15]), \; \beta \sim \text{Unif}([0.8, 0.9])$$

and $\alpha_0, \alpha_1, \beta$ mutually independent. Another function is from a threshold model

$$\sigma_{t,i}^2 = F_i(X_{t-1,i}, \sigma_{t-1,i}^2)$$

where

$$F_i(x, \sigma^2) = \begin{cases} 
\alpha_1 + \alpha_2 x^2, & \text{if } x \leq d_1 = 0 \\
0.2 + \alpha_3 x^2 + \alpha_4 \sigma^2, & \text{if } x > d_1 = 0, \; \sigma^2 \leq d_2 = 0.5 \\
0.8 + \alpha_5 \sigma^2, & \text{if } x > d_1 = 0, \; \sigma^2 > d_2 = 0.5 
\end{cases}$$

(4.17)

where

$$\alpha_1 \sim \text{Unif}([0, 0.3]), \; \alpha_2 \sim \text{Unif}([0.4, 0.6]), \; \alpha_3 \sim \text{Unif}([0.1, 0.3]),$$

$$\alpha_4 \sim \text{Unif}([0.6, 0.8]), \; \alpha_5 \sim \text{Unif}([0.4, 0.6])$$

($\alpha_1, \ldots, \alpha_5$ mutually independent). A third and a fourth function, in which we also allow for one cross-terms, are

$$\sigma_{t,i}^2 = F_i(X_{t-1,i}, X_{t-1,j}, \sigma_{t-1,i}^2)$$

where

$$F_i(x, y, \sigma^2) = (\alpha_1 + 0.2 |y| + \alpha_2 x^2) \cdot (0.8 \exp(\alpha_3 |x| |\sigma|)) +$$

$$+ (0.4 x^2 + \alpha_4 \sigma^2)^{3/4},$$

(4.18)

where

$$\alpha_1 \sim \text{Unif}([0.05, 0.15]), \; \alpha_2 \sim \text{Unif}([0.8, 0.95]),$$

$$\alpha_3 \sim \text{Unif}([-1.6, -1.4]), \; \alpha_4 \sim \text{Unif}([0.4, 0.6])$$

($\alpha_1, \ldots, \alpha_4$ mutually independent), and

$$\sigma_{t,i}^2 = F_i(X_{t-1,i}, X_{t-1,j}, \sigma_{t-1,i}^2)$$

where

$$F_i(x, y, \sigma^2) = (0.1 + \alpha_1 |y|^3) \cdot \exp(\alpha_2 x^2) + \alpha_3 (\sigma^2)^{3/4},$$

(4.19)

where

$$\alpha_1 \sim \text{Unif}([0.1, 0.2]), \; \alpha_2 \sim \text{Unif}([-0.1, 0]), \; \alpha_3 \sim \text{Unif}([0.8, 0.9])$$
4.5. Volatility estimation for high multivariate time series

$(\alpha_1, \alpha_2, \alpha_3$ mutually independent), where the component $j \in \{1, \ldots, d\} \setminus i$ is chosen randomly. Each of the volatility functions $F_i$ is randomly chosen with probability $1/4$, independent of each other. Note that also the coefficients in these functions are randomly chosen. The constant conditional correlation matrix $R$ is chosen to mimic the one of real log-returns. This model is “fairly close” to a CCC-GARCH(1,1) model since half of the volatility functions involve only auto-dependence (no dependence on a cross-series), a quarter of them actually being linear GARCH-type, and the other half involve only one other cross-series.

The results are displayed in Table 4.5 and Figure 4.3 (test set size is $n_{out} = 1000$).

<table>
<thead>
<tr>
<th>Model</th>
<th>Performance measure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OS – log-lik.</td>
</tr>
<tr>
<td>GARCH(1,1)</td>
<td>121636.8</td>
</tr>
<tr>
<td>FGD with tree</td>
<td>119763.0</td>
</tr>
<tr>
<td>using $p = 2$, $\nu = 0.5$, $L = 5$</td>
<td>(1.5%)</td>
</tr>
</tbody>
</table>

Table 4.5: Goodness of fit measures for one 100-dimensional realization of size $n = 1000$ as described above. The measures are defined as in (4.15), as the average over all $d = 100$ individual series as in (4.8) and the extremal relative individual gains with respect to (4.8); test-set size is $n_{out} = 1000$ and relative gains with FGD are given in parentheses.

We observe only a small gain of FGD with trees over the CCC-GARCH(1,1) prediction with respect to the negative log-likelihood, because the signal to noise level is low, but a more substantial gain when looking at individual out-sample $L_2$-losses as in (4.8) where the observation noise is not present. Of course, we cannot expect to learn in all $d = 100$ components with sample size $n = 1000$ (or here in all 75 components which are not of linear GARCH(1,1)-type). On average, the OS-$L_2$ gain over all $d = 100$ components is 8.3%: this gain will generally decrease when dimension $d$ increases and keeping sample size $n$ fixed. However, we also see from Figure 4.3, that FGD mainly improves (in absolute terms) at those components where the CCC-GARCH(1,1) predictions are poorest. This is consistent with the intuition that functional gradient descent improves the “hardest” cases. Regarding the relative gains with FGD: they range from -12.1 to 27.2% and quite many large relative gains are real-
Figure 4.3: Individual OS-L2 measures as in (4.8) for a 100-dimensional realization of size $n = 1000$ as described above. First three panels: performance of CCC-GARCH(1,1) on x-axis and of FGD with trees on y-axis with the line indicating equal performance, on different scales. Lower right panel: relative gains with FGD with trees of individual OS-L2 measures (see (4.8) on y axis versus performance of CCC-GARCH(1,1) on x-axis.
4.5. Volatility estimation for high multivariate time series

ized where the CCC-GARCH predictions are fairly low (see lower right panel of Figure 4.3).

A seven-dimensional real data example

We consider here the same seven-dimensional real global portfolio introduced in Section 3.4. The dimensionality \( d = 7 \) is "mid-range" allowing for a representation of the results which is easy to survey. FGD will also turn out to be useful in such orders of dimensions. The data comprises 1500 daily returns from the US DJIA, the French CAC40, the German DAX, the Italian BCI, the Dutch CBS, the British FTAS and the Japanese NIKKEI index, during the period January 31, 1990 until September 9, 1996: we use the first 1000 time points for training (fitting) and the remaining 500 for out-sample testing (evaluation).

We consider two goodness of fit measures for such real data. One is the negative out-sample log-likelihood in (4.15). Alternatively, we also look at the individual in- and out-sample prediction \( L_2 \) losses given by (4.9). The results are summarized in Table 4.6.

<table>
<thead>
<tr>
<th>Performance measure</th>
<th>CCC-GARCH(1,1)</th>
<th>FGD with trees</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OS -log-lik.</td>
<td>IS-PL(_2)</td>
</tr>
<tr>
<td>global</td>
<td>5614.13</td>
<td>5547.46</td>
</tr>
<tr>
<td>DJIA</td>
<td>1082.89</td>
<td>5874.58</td>
</tr>
<tr>
<td>CAC</td>
<td>11191.4</td>
<td>21915.1</td>
</tr>
<tr>
<td>DAX</td>
<td>10382.6</td>
<td>41318.4</td>
</tr>
<tr>
<td>BCI</td>
<td>8753.99</td>
<td>36145.6</td>
</tr>
<tr>
<td>CBS</td>
<td>5725.22</td>
<td>25191.9</td>
</tr>
<tr>
<td>FTAS</td>
<td>2109.78</td>
<td>5434.23</td>
</tr>
<tr>
<td>NIK</td>
<td>45610.1</td>
<td>25137.2</td>
</tr>
</tbody>
</table>

Table 4.6: Goodness of fit measures for a seven-dimensional real data example. The optimal parameters in FGD with trees are \( p = 3 \) lagged values, shrinkage \( v = 0.5 \) and \( L = 3 \) terminal nodes.

With real data, differences in the out-sample negative log-likelihood or
OS-PL\textsubscript{2} can be small between different methods; see also at the end of Section 4.4.1. Also in this case, we consider versions of the \textit{t-} and \textit{sign-test}, adapted to the case of dependent observations \( D_t \) (see Section 3.4.2). We apply both tests for comparing the FGD with trees and the CCC-GARCH predictions. The alternative is one-sided where FGD with trees has lower out-sample loss. The results are given in Table 4.7. Thus, the small difference of 1.2\% in the out-sample log-likelihoods of FGD and CCC-GARCH(1,1) from Table 4.6 turns out to be highly significant.

### 4.6 Conclusions

We have presented an FGD algorithm which is a technique for estimation of the conditional covariance matrix in (0.1). It is computationally feasible in multivariate problems with several hundreds up to thousands of return series. The algorithm is constructed from a generic algorithm: hence, other FGD algorithms can be derived aiming to learn other, typically very high-dimensional, problems.

We have demonstrated on some data-sets (synthetic and real) that our FGD algorithm significantly outperforms the predictions from the classical CCC-GARCH(1,1) model (Bollerslev, 1990). The latter model has generated the starting functions in our FGD and hence, it is not so surprising that we could observe improvements.

This seems generally the attractive feature of FGD. We choose a reasonable model for generating (estimating) the starting functions in FGD, and then we try – often successfully – to improve the initial basis model with a couple FGD iterations. According to the heuristics of a steepest functional gradient, the improvements with FGD are mainly expected at those components where the initial basis model performs poorly, as we demonstrated on a 100-dimensional data-set. Therefore, FGD is not necessarily restricted to the framework of constant conditional correlation (CCC).

### Table 4.7: Test-statistics and \( P \)-values (in parentheses) for one-sided testing for differences in out-sample negative log-likelihood of the FGD with trees and the CCC-GARCH predictions. Negative test-statistics favor the FGD method.

<table>
<thead>
<tr>
<th>\textit{t-type test}</th>
<th>\textit{sign-type test}</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2.77 (0.003)</td>
<td>-2.76 (0.003)</td>
</tr>
</tbody>
</table>


Bibliography


List of Figures

1.1 The behaviour of the factor $V(\psi, G)$. ................................. 31

2.1 Example of a binary tree partition $P = \{R_1, R_2, R_3\}$ of predictor space \{(x, \sigma^2); x \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+\}. ................................. 47

2.2 Function estimation of conditional variance from both classical GARCH(1,1) and Tree-structured GARCH model. ................................. 58

2.3 Errors of estimates for conditional variance from both classical GARCH(1,1) and Tree-structured GARCH fit. ................................. 59

2.4 Results for negative log-returns of the DAX index from the optimal tree-structured GARCH model with $\mathcal{N}(0, 1)$-distributed innovations. ................................. 65

3.1 Asynchronized returns $X_{t,i}$ against estimated synchronized returns $\hat{X}_{t,i}^s$ from the synchronous CCC-GARCH(1,1) model. ................................. 86

3.2 Predicted portfolio conditional variances $\tilde{\sigma}_t^2$ and means $\tilde{\mu}_t, p$ for the test data days. ................................. 89

3.3 The expected shortfall for the portfolio returns $\tilde{\Delta}_t$ for the test data days. ................................. 101

3.4 Boxplots of residuals $\hat{R}_t$ in (3.27) for test data with $q = 0.95$ and $q = 0.90$. ................................. 102

4.1 Outsample errors $\hat{\sigma}_t^2 - \sigma_t^2$ versus true $\sigma_t^2$ in simulated model (4.5). ................................. 113

4.2 300 days of predicted conditional variance estimates from the GARCH(1,1) model and the FGD with trees. ................................. 116
4.3 Individual OS-$L_2$ measures for a 100-dimensional realization of size $n = 1000$. .......................... 128
# List of Tables

1.1 The OS-L$_2$ measure for five independent realizations from the model (1.1) with \( f(\cdot) \) given by (1.18). ...................... 36

1.2 The OS-PL$_2$ measure for the German DAX index, the US DJIA index and the BMW share price from a local regression and a normal local likelihood fit. .......................... 38

1.3 The OS-PL$_2$ measure for the same three financial instruments of Table 1.2 from a GAM with local regression and a GAM with normal local likelihood fit. ...................... 41

2.1 Goodness of fit measures for eight simulations using the classical GARCH(1,1) in (2.2). ................................. 61

2.2 Estimated thresholds \( \hat{d}_i \) and goodness of fit measures for the same eight simulations of Table 2.1 from the tree-structured GARCH(1,1) model. ................................. 62

2.3 Goodness of fit measures for the same eight simulations from Table 2.1 using the GAM model described in Section 2.5. . . . . 63

2.4 Estimated thresholds \( \hat{d}_i \) and goodness of fit measures for the negative log-returns from the DAX index and the BMW stock price. ................................. 66

3.1 Instantaneous empirical correlations between components of both asynchronized and synchronized returns. ............. 85

3.2 Goodness of fit from the synchronous CCC-GARCH(1,1), the classical CCC-GARCH(1,1) and the univariate model defined by (3.20). ................................. 90

3.3 Testing differences of performance on the portfolio level. . . . . 92
List of Tables

3.4 Individual optimal thresholds for the seven indices introduced in Section 3.4. ................................................. 97

3.5 Goodness of fit for the real global portfolio introduced in Section 3.4 using more sophisticated asynchronous models. .... 98

4.1 Goodness of fit measures (on average) for fifty simulations from model (4.5) with sample size n = 1000. ......................... 115

4.2 Goodness of fit measures for negative returns from the German DAX, the British FTAS and the Japanese NIKKEI index. 117

4.3 Testing differences of performance for the negative returns of the same three financial instruments of Table 4.2. ................. 118

4.4 Goodness of fit measures for one three-dimensional realization of size n = 1000 from model (4.10) with $\mathbf{Z}_t \sim \mathcal{N}_3(0, I)$ and individual conditional variances given by (4.17)-(4.18) . . . . 124

4.5 Goodness of fit measures for one 100-dimensional realization of size n = 1000. .................................................. 127

4.6 Goodness of fit measures for a seven-dimensional real data example. ................................................................. 129

4.7 Test-statistics and $P$-values for one-sided testing for differences in out-sample negative log-likelihood of the FGD with trees and the CCC-GARCH predictions. ......................... 130
Curriculum Vitae

Personal Data

Name Francesco Audrino
Date of Birth 28.04.1975
Nationality Swiss

Education

1981-1986 Scuola elementare Chiasso
1986-1990 Scuola media Chiasso
1990-1994 Liceo cantonale Mendrisio
   Maturità tipo B

1994-1999 Studies in Mathematics, ETH Zürich
1999 Diploma thesis under the supervision of
   Prof. P. Embrechts and Prof. A. McNeil
   Topic: Extremes in Financial Time Series:
   Measuring Risk for Bivariate Portfolios

1999-2002 Teaching assistant and Ph.D. student at
   the ETH Zürich under the supervision of
   Prof. P. Bühlmann
   Topic: Statistical Methods for
   High-Multivariate Financial Time Series