TECHNISCHE UNIVERSITÄT WIEN

VIENNA UNIVERSITY OF TECHNOLOGY

D I P L O M A R B E I T

Multivariate Generalized Autoregressive Conditional Heteroscedasticity: Theory and Application

ausgeführt am Institut für

Wirtschaftsmathematik

der Technischen Universität Wien

unter Anleitung von

Ao.Univ.Prof. Dr. Wolfgang Scherrer

durch

Gregor Kastner

Grillparzerstraße 18

2380 Perchtoldsdorf

Wien, im August 2006

Mein besonderer Dank gilt Prof. Wolfgang Scherrer für die Betreuung dieser Diplomarbeit und für die kritische und genaue Durchsicht der Ausarbeitung.

Weiters bedanke ich mich bei Prof. Peter Bühlmann und Prof. Alexander McNeil, welche aufgrund der Vorlesungen "Zeitreihenanalyse" sowie "Empirical Methods for Finance" an der ETH Zürich mein Interesse am Thema geweckt haben.

Nicht zuletzt möchte ich mich bei meiner Familie bedanken. Für die vielfältigen Arten der Unterstützung, die ich durch jedes einzelne Mitglied mein Leben lang erfahren durfte.

Außerdem danke an:

Das einzigartige  *Haus 18.*
Robert Schöftner, für Inspiration und Motivation.
Michael Petz, für außerordentlich aufmerksames Lesen.
Heidrun, für ein unendlich großes Maß an Liebe.

Summary

Time-varying volatility modeling for univariate asset returns is a well investigated topic in time series analysis, including the prominent 2003 Nobel Prize winning ARCH model by Robert F. Engle.

This thesis provides an introduction into general ideas and tools of time series modeling with a special focus on ARCH and generalized ARCH models. Theoretical properties of this model class as well as its fitting to data are discussed. For formulations of generalized ARCH in higher dimensions, several approaches - Constant Conditional Correlation, Dynamic Conditional Correlation, (Diagonal) Vector GARCH and BEKK models - are presented and analyzed. Because the curse of dimensionality plays a major role in practical applicability of MGARCH, main attention is placed on models with reasonable numbers of parameters. These are fitted to the Austrian and German stock market indices ATX and DAX making use of S-PLUS [18] with the module finmetrics [19]. Finally, results for empirical prediction performance among the different models are compared by means of both in- and out-sample measurement.

Contents

Chapter 1

Fundamentals

In this chapter the basic ideas of multivariate time series are presented. The approach is similar to McNeil et al. [20], including supplements of Tsay [26], Zivot and Wang [28], and Brockwell and Davis [8, 9]. Basic common probability theory will be assumed to be known and is covered in many textbooks such as Elstrodt [11] or Bauer [2] (German) and Durret [10] or Williams [27] (English).

1.1 Basic Definitions

A multivariate time series model for multiple risk factors is a multivariate stochastic process $(\boldsymbol{X}_t)_{t\in\mathbb{Z}} = (X_{t,1}, \ldots, X_{t,n})'_{t\in\mathbb{Z}}$. In other words, it is a family of random vectors, indexed by the integers and defined on some suitable probability space $(\Omega, \mathcal{F}, P)$.

1.1.1 Moments of a multivariate time series

Definition 1.1 (Moments of a multivariate time series)**.** We define the *mean function* $\boldsymbol{\mu}(t)$ and the *covariance matrix function* $\Gamma(t,s)$ of $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$ by

$$\begin{aligned}\boldsymbol{\mu}(t) &= \mathbb{E}[\boldsymbol{X}_t], \qquad t\in\mathbb{Z},\\ \Gamma(t,s) &= \mathbb{E}[(\boldsymbol{X}_t-\boldsymbol{\mu}(t))(\boldsymbol{X}_s-\boldsymbol{\mu}(s))'], \qquad t,s\in\mathbb{Z}.\end{aligned}$$

Obviously, these do not need to exist in $\mathbb{R}$.

Remark. It is interesting to note that $\Gamma(t,s)$ is in general not symmetric (i.e. $\Gamma(t,s)\neq\Gamma(s,t)$), which is in contrast to the univariate case. Lagged values of one of the component series can be more strongly correlated with future values of another component series than vice versa. This property, when observed in empirical data, is known as a *lead-lag* effect. Nevertheless, due to the fact that

$$\Gamma(t,s)_{ij}=\operatorname{cov}(X_{t,i},X_{s,j})=\operatorname{cov}(X_{s,j},X_{t,i})=\Gamma(s,t)_{ji} \tag{1.1}$$

it is clear that $\Gamma(t,s)=\Gamma(s,t)'$.

1.1.2 Concepts of stationarity

Similar to univariate time series analysis, several ideas of stationarity may be considered. The underlying idea is that $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$ and time-shifted $(\boldsymbol{X}_{t+h})_{t\in\mathbb{Z}}$ for any $h\in\mathbb{Z}$ should share statistical properties. One possible way of doing this is to demand that all arbitrarily lagged finite-dimensional samples are distributionally identical. Formalized, this results in the following definition:

Definition 1.2 (Strict stationarity)**.** The multivariate time series $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$ is called *strictly stationary*, if

$$(\boldsymbol{X}'_{t_1},\ldots,\boldsymbol{X}'_{t_n})\overset{d}{=}(\boldsymbol{X}'_{t_1+h},\ldots,\boldsymbol{X}'_{t_n+h})$$

holds for all $t_1,\ldots,t_n,h\in\mathbb{Z}$ and for all $n\in\mathbb{N}$.

Strict stationarity is usually hard to verify in practice and also a very strong assumption. Alternatively, there exists another possible concept of stationarity, which only involves first and second moments, demanding that they are time-invariant.

Definition 1.3 (Covariance stationarity)**.** The time series $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$ is called *covariance* stationary (or *weakly* or *second-order* stationary) if the first two moments exist and satisfy

$$\begin{aligned} \boldsymbol{\mu}(t) &= \boldsymbol{\mu}, \qquad t \in \mathbb{Z}, \\ \Gamma(t,s) &= \Gamma(t+h, s+h), \qquad t,s,h \in \mathbb{Z}. \end{aligned}$$

Remark. Assuming the covariance matrix is finite, covariance stationarity is a necessary condition for strict stationarity and is therefore often referred to as weak stationarity. Clearly the contrary is not true, and it is also possible to define infinite-variance processes which are strictly stationary but not covariance stationary (McNeil et al. [20]).

1.1.3 Correlation in stationary multivariate time series

By definition of covariance stationarity one can easily observe that for covariance stationary time series $\Gamma(t-s,0) = \Gamma(t,s)$ holds for all $s,t \in \mathbb{Z}$. This simply means that the covariance between $\boldsymbol{X}_t$ and $\boldsymbol{X}_s$ only depends on their temporal separation $t-s$, also known as the *lag*. For covariance stationary time series this enables to write the covariance matrix function as a function of one variable, i.e.

$$\Gamma(h) := \Gamma(h,0), \qquad h \in \mathbb{Z}.$$

Remark. As noted above, the sign of the lag is, in contrast to the univariate case, of importance and it is *not* sufficient to choose $h \in \mathbb{N}_0$.

Noting that $\Gamma(0) = \mathrm{cov}(\boldsymbol{X}_t)$ for all $t \in \mathbb{Z}$, it is now possible to define the correlation matrix function of a multivariate covariance stationary process.

Definition 1.4 (Correlation matrix function)**.** Let $\Delta \in \mathbb{R}^{d\times d}$ be a diagonal matrix containing the standard deviations of the component series, in other words $\Delta = \text{diag}(\sqrt{\Gamma(0)_{11}}, \ldots, \sqrt{\Gamma(0)_{dd}})$. Then the correlation matrix function $P(h)$ of a covariance stationary multivariate time series $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$ is

$$P(h) := \Delta^{-1}\Gamma(h)\Delta^{-1}, \qquad h \in \mathbb{Z}.$$

Remarks.

1. The diagonal elementes $P(h)_{ii}$ of this matrix-valued function give the autocorrelation function (i.e. the lagged correlation with itself) of the ith one-dimensional component series $(X_{t,i})_{t\in\mathbb{Z}}$.
2. The off-diagonal entries give so-called cross correlations between different component series at different times.
3. It follows directly from (1.1) that $P(h) = P(-h)'$, but just like $\Gamma(h)$ $P(h)$ need not be symmetric.

1.1.4 Noise concepts

Before constructing more interesting classes of time series models, simple *multivariate white noise* processes are defined. They are acting as building blocks for later models.

Definition 1.5 (Multivariate white noise)**.** $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$ is multivariate white noise if it is covariance stationary and its correlation matrix function is given by

$$P(h) = \begin{cases} P & h = 0, \\ 0 & h \neq 0, \end{cases}$$

for some positive-definite correlation matrix P. Centered to have mean zero with covariance matrix $\Sigma = \text{cov}(\boldsymbol{X}_t)$, it will be denoted $\text{WN}(\boldsymbol{0}, \Sigma)$.

Remark. It is clear by definition that a $\text{WN}(\boldsymbol{0}, \Sigma)$-process has no cross correlation between component series, except for contemporaneous cross correlation at lag zero. A simple example for such a process is a series of iid random vectors with finite covariance matrix, itself known as *strict white noise*.

Definition 1.6 (Multivariate strict white noise)**.** $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$ is called multivariate strict white noise if it is a series of iid random vectors with finite covariance matrix. Centered to have mean zero and covariance matrix Σ it will be denoted $\mathrm{SWN}(\mathbf{0}, \Sigma)$.

Assuming that $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$ is adapted to some filtration $(\mathcal{F}_t)_{t\in\mathbb{Z}}$, it is possible to introduce the *martingale difference noise concept* for multivariate time series. $(\mathcal{F}_t)_{t\in\mathbb{Z}}$ will typically be the so-called natural filtration, defined by $\mathcal{F}_t := \sigma\{(\boldsymbol{X}_s)_{s\leq t}\}$. It is usually interpreted as the information available up to time t.

Definition 1.7 (Multivariate martingale difference)**.** $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$ is said to have the multivariate martingale difference property with respect to the filtration $(\mathcal{F}_t)_{t\in\mathbb{Z}}$ if

1. $\mathbb{E}[|\boldsymbol{X}_t|] < \infty$ and
2. $\mathbb{E}[\boldsymbol{X}_t|\mathcal{F}_{t-1}] = \mathbf{0}$

hold for all $t \in \mathbb{Z}$.

Remark. Because of

$$\mathbb{E}[\boldsymbol{X}_t] = \mathbb{E}[\mathbb{E}[\boldsymbol{X}_t|\mathcal{F}_{t-1}]] = \mathbf{0} \qquad \forall t \in \mathbb{Z},$$

the unconditional mean of such a process is also zero. If $\mathrm{cov}(\boldsymbol{X}_t) = \mathbb{E}[\boldsymbol{X}_t\boldsymbol{X}_t']$ exists for all t, then if $t < s$,

$$\Gamma(t,s) = \mathbb{E}[\boldsymbol{X}_t\boldsymbol{X}_s'] = \mathbb{E}[\mathbb{E}[\boldsymbol{X}_t\boldsymbol{X}_s'|\mathcal{F}_{s-1}]] = \mathbb{E}[\boldsymbol{X}_t\mathbb{E}[\boldsymbol{X}_s'|\mathcal{F}_{s-1}]] = \mathbb{E}[\boldsymbol{X}_t\mathbf{0}'] = 0$$

holds. Similarly, if $t > s$, we have

$$\Gamma(t,s) = \mathbb{E}[\boldsymbol{X}_t\boldsymbol{X}_s'] = \mathbb{E}[\mathbb{E}[\boldsymbol{X}_t\boldsymbol{X}_s'|\mathcal{F}_{t-1}]] = \mathbb{E}[\mathbb{E}[\boldsymbol{X}_t|\mathcal{F}_{t-1}]\boldsymbol{X}_s'] = \mathbb{E}[\mathbf{0}\boldsymbol{X}_s'] = 0,$$

resulting in the fact that the covariance matrix function satisfies $\Gamma(t,s) = 0$ for $t \neq s$. If additionally it is constant for all $t = s$ then a process with the multivariate martingale difference property is also a multivariate white noise process.

1.1.5 Example

In order to apply some of the above ideas, a small example taken from Brockwell and Davis [8] is now being analyzed.

Example 1.1 (Brockwell and Davis [8])**.** Consider the bivariate covariance stationary time series $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$ defined by

$$\boldsymbol{X}_t = \begin{bmatrix} X_{t,1} \\ X_{t,2} \end{bmatrix} = \begin{bmatrix} Z_t \\ Z_t + .75Z_{t-10} \end{bmatrix}, \tag{1.2}$$

with $(Z_t)_{t\in\mathbb{Z}}$ being univariate WN$(0,1)$. Clearly $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$ has mean zero, i.e.

$$\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{X}_t] = \mathbb{E}\begin{bmatrix} Z_t \\ Z_t + .75Z_{t-10} \end{bmatrix} = \mathbb{E}\begin{bmatrix} Z_t \\ Z_t \end{bmatrix} + .75\mathbb{E}\begin{bmatrix} 0 \\ Z_{t-10} \end{bmatrix} = \mathbf{0},$$

and is covariance stationary with

$$\Gamma(0) = \mathbb{E}[\boldsymbol{X}_t\boldsymbol{X}_t'] = \mathbb{E}\begin{bmatrix} Z_t^2 & Z_t^2 + .75Z_tZ_{t-10} \\ Z_t^2 + .75Z_tZ_{t-10} & (Z_t + .75Z_{t-10})^2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 1.56 \end{bmatrix},$$

$$\Gamma(-10) = \begin{bmatrix} 0 & .75 \\ 0 & .75 \end{bmatrix}, \qquad \Gamma(10) = \begin{bmatrix} 0 & 0 \\ .75 & .75 \end{bmatrix},$$

and $\Gamma(h) = 0$ for $h \in \mathbb{Z} \setminus \{-10, 0, 10\}$. Recalling that the correlation matrix function $P(h) = \Delta^{-1}\Gamma(h)\Delta^{-1}$ for $\Delta = \operatorname{diag}(\sqrt{\Gamma(0)_{11}}, \ldots, \sqrt{\Gamma(0)_{dd}})$, $P(h)$ can easily be calculated by inserting:

$$P(h) = \begin{bmatrix} \frac{\Gamma(h)_{11}}{\Gamma(0)_{11}} & \frac{\Gamma(h)_{12}}{\sqrt{\Gamma(0)_{11}\Gamma(0)_{22}}} \\ \frac{\Gamma(h)_{21}}{\sqrt{\Gamma(0)_{11}\Gamma(0)_{22}}} & \frac{\Gamma(h)_{22}}{\Gamma(0)_{22}} \end{bmatrix},$$

yielding

$$P(-10) = \begin{bmatrix} 0 & .6 \\ 0 & .48 \end{bmatrix}, \qquad P(0) = \begin{bmatrix} 1 & .8 \\ .8 & 1 \end{bmatrix}, \qquad P(10) = \begin{bmatrix} 0 & 0 \\ .6 & .48 \end{bmatrix},$$

with $P(h) = 0$ elsewhere.

1.2 Analysis in the Time Domain

1.2.1 Sample covariance and correlation

Taking the practical viewpoint, it is now assumed that an arbitrary random sample $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ from a covariance stationary multivariate time series model $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$ is given. It is therefore necessary to construct empirical estimators of the covariance matrix function (and the correlation matrix function) from this given random sample. A straightforward and widely accepted estimator, the *sample covariance matrix function* $\widehat{\Gamma}(h)$, is calculated according to:

Definition 1.8 (Sample covariance and correlation matrix function)**.**

$$\widehat{\Gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-h}(\boldsymbol{X}_{t+h} - \overline{\boldsymbol{X}})(\boldsymbol{X}_t - \overline{\boldsymbol{X}})', \qquad 0 \leq h < n,$$

with $\overline{\boldsymbol{X}} = \frac{1}{n}\sum_{t=1}^{n}\boldsymbol{X}_t$ being the *sample mean*, a well known estimator for $\boldsymbol{\mu}$. From that, the *sample correlation matrix function* $P(h)$ can be estimated by

$$\widehat{P}(h) = \widehat{\Delta}^{-1}\widehat{\Gamma}(h)\widehat{\Delta}^{-1}, \qquad 0 \leq h < n,$$

with $\widehat{\Delta}$ standing for the $d \times d$ diagonal matrix of the sample standard deviations of the component series, i.e.

$$\widehat{\Delta} = \operatorname{diag}\left(\sqrt{\widehat{\Gamma}(0)_{11}}, \ldots, \sqrt{\widehat{\Gamma}(0)_{dd}}\right).$$

Remark. 1. In order for the sample covariance and correlation matrix to be positive definite, the sample size n must be greater than the number of component time series d (Zivot and Wang [28]).

2. Asymptotic properties of the sample correlation matrix function $\widehat{P}(h)$ have been investigated under various assumptions (see for example Fuller [15] for details). The estimate is consistent, but biased in a finite sample. For asset return series, the finite sample distribution of $\widehat{P}(h)$ is rather complicated partly because of the presence of conditional heteroscedasticity and high

kurtosis. Proper bootstrap resampling methods are recommended if the finite sample distribution is needed (Tsay [26]).

1.2.2 The cross correlogram

The information in the sample correlation matrix function is generally displayed in the *cross correlogram*, which is a $d \times d$ matrix of plots. The ith diagonal plot in this graphic display is the correlogram of the ith component series given by $\{(h, \widehat{P}(h)_{ii}) : h = 0, 1, 2, \ldots\}$. For the off-diagonal plots containing the estimates of *cross correlation* there are various possible presentations. Here the convention adopted for instance by S-PLUS [18] will be used: for $i < j$ the set $\{(h, \widehat{P}(h)_{ij}) : h = 0, 1, 2, \ldots\}$ is plotted, for $i > j$ $\{(-h, \widehat{P}(h)_{ij}) : h = 0, 1, 2, \ldots\}$. See Example 1.2, especially Figure 1.2 and 1.3 for details and interpretation.

Remark. Observing cross correlograms, one will usually find (dotted) horizontal lines centered around the time axis. These are the 95% Gaussian confidence bands at $(-1.96/\sqrt{n}, 1.96/\sqrt{n})$, which are only *asymptotically* correct, under the (striking) assumption that the underlying process is a white noise process. Even though they are hardly ever theoretically justified in practice, they are often used as a rough guidance for the eye (in order to determine whether correlation at a certain lag is significantly different from zero). Nevertheless, they should not be relied upon too heavily to draw conclusions, especially if the number of observations is small or the underlying process is not a white noise process (McNeil et al. [20]).

1.2.3 Example

In order to understand better how the above ideas may work in practice, a small time domain analysis is now being conducted on simulated data.

Example 1.2. Returning to Example 1.1, consider again the bivariate covariance stationary time series $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$ defined by (1.2),

$$\boldsymbol{X}_t = \begin{bmatrix} X_{t,1} \\ X_{t,2} \end{bmatrix} = \begin{bmatrix} Z_t \\ Z_t + .75Z_{t-10} \end{bmatrix}.$$

Instead of a theoretical analysis, the time series is now being simulated in S-PLUS [18] and the module finmetrics [19], as well as in R [21], making use of the library `tseries` [25]. Two different distribution functions are used for simulating the white noise $(\boldsymbol{Z}_t)_{t\in\mathbb{Z}}$. Firstly, it is chosen to be independently normally distributed with mean zero and variance one, i.e. $Z_t \sim N(0,1)$ iid. Secondly, it is chosen to be independently t distributed, centered around zero, with degrees of freedom ν equaling three, and appropriately scaled to have variance one, i.e. $Z_t \sim t(3,0,\frac{1}{3})$. A possible realization with sample size $n = 250$ is shown in Figure 1.1.

By simply looking at the plots (and assuming for a while that one does not know the generating equations beforehand), it comes to mind that stationarity may be a plausible assumption. Shocks and peaks seem to recenter quickly, and no trend is visible. There is some evidence that the two component series might be correlated, at each point in time as well as with some lag. This becomes especially clear when looking at the path with t innovations (lower plots): The large negative shock taking place simultaneously for both component time series at $t = 76$ seems to be repeated by component series two at $t = 86$. Similar effects, even though not as striking, can be observed at several other points in time.

Despite the fact that it may be possible to understand some properties of the time series being analyzed directly by observing the standard time series plot such as Figure 1.1, it is always advisable to look at the cross correlogram. Concerning our concrete example, it is given in Figure 1.2 for the time series with Gaussian innovations and in Figure 1.3 for the time series with t-innovations and provides graphically all necessary information about the correlation function in an easy to read fashion.

Looking at Figure 1.2, the picture in row one and column two (i.e. the upper right picture) shows estimated correlations between $X_{t+h,1}$ (component series one at time $t+h$) and $X_{t,2}$ (component series two at time t) for $h \geq 0$. Clearly these estimates are small and lie mostly within the confidence band with the obvious exception of the correlation estimate for values with lag zero $\widehat{P}(0)_{12} \approx 0.8137$.[1]

[1]Since the generating equations are known, we may compare this value with the true correlation $P(0)_{12} = P(0)_{21} = 0.8$, calculated in Example 1.1.

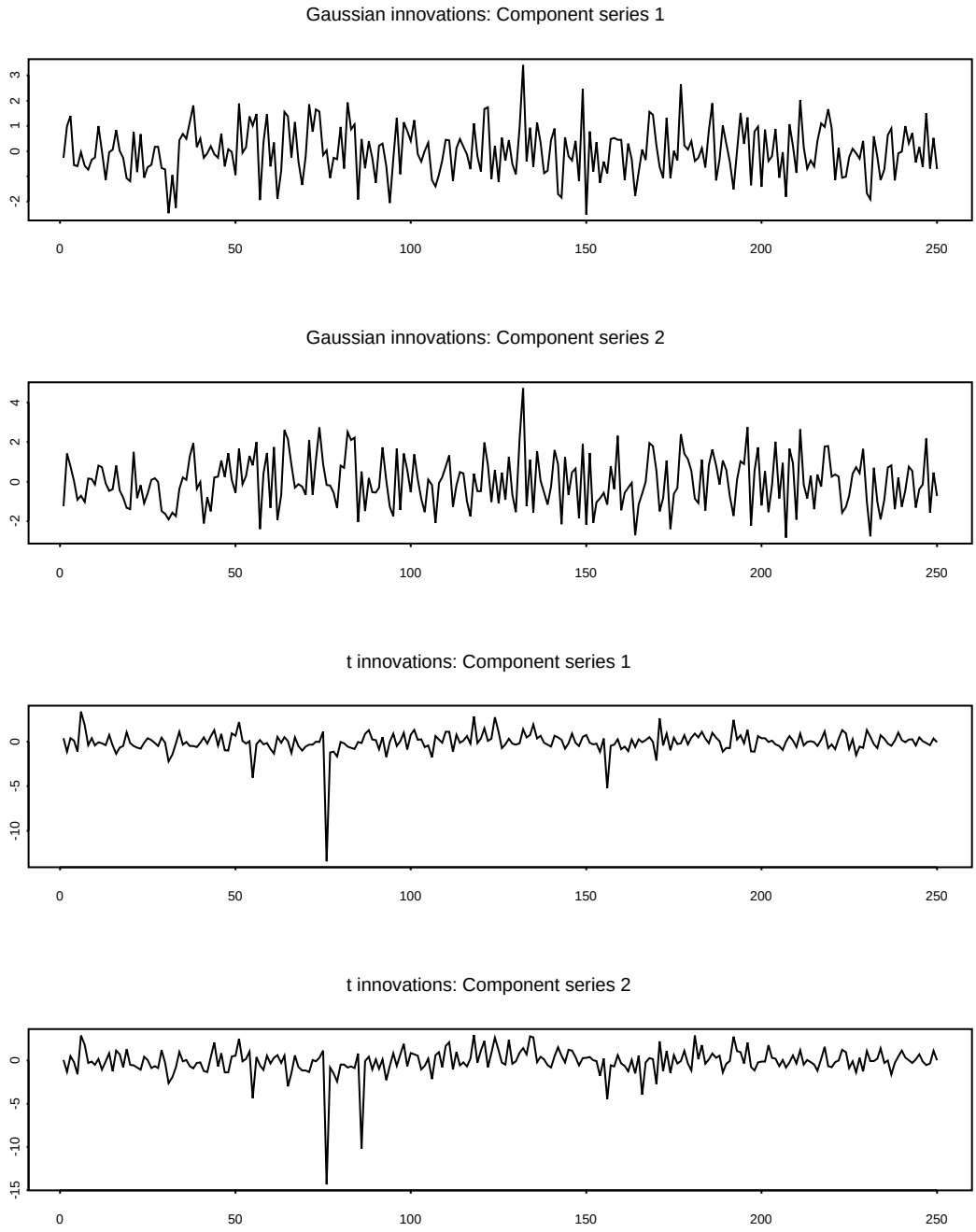

Figure 1.1: Two paths of the time series defined by (1.2). The upper two plots show a realization with iid Gaussian innovations, the lower ones show a realization with iid t innovations (df $\nu = 3$), as described in Example 1.2.

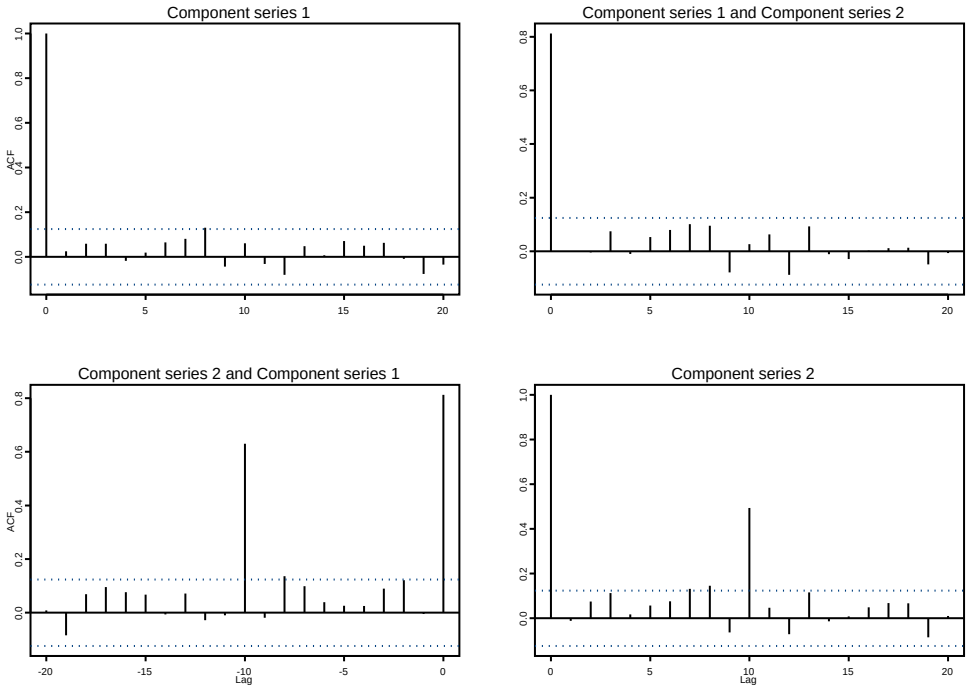

Figure 1.2: Cross correlogram of the realization of the bivariate time series defined by (1.2) with Gaussian innovations, see Example 1.2 for commentary.

The picture in row two and column one (the lower left picture) again shows estimated correlations between $X_{t+h,1}$ and $X_{t,2}$, but contrary to the above now for $h \leq 0$. Lag zero correlation $\widehat{P}(0)_{12}$ is plotted again and is exactly the same as above. Values for $h < 0$ are all within the confidence bounds except for the value $\widehat{P}(-10)_{12} \approx 0.6223$, which is significantly different from zero.[2]

The diagonal pictures in Figure 1.2 display the autocorrelations of each component series and correspond to the *correlograms* known from univariate time series analysis. One can observe the trivial peaks $\widehat{P}(0)_{ii} = 1$ for $i \in \{1, 2\}$ in both component series, but apart from that only one really significant nonzero value $\widehat{P}(10)_{22} \approx 0.4631$.[3]

[2]The true correlation at lag $h = -10$ amounts to $P(-10)_{12} = P(10)_{21} = 0.6$.
[3]The true autocorrelation at lag $h = 10$ is $P(10)_{22} = 0.48$.

Summarizing these results, we can conclude that we were able to successfully estimate auto- and intercorrelation from the simulated time series. The lead-lag effect of component series one $X_{t,1}$ onto component series two $X_{t,2}$ was preserved and could clearly be seen in the cross correlogram. S-PLUS [18] code used for simulating and generating the plots can be found in the appendix.

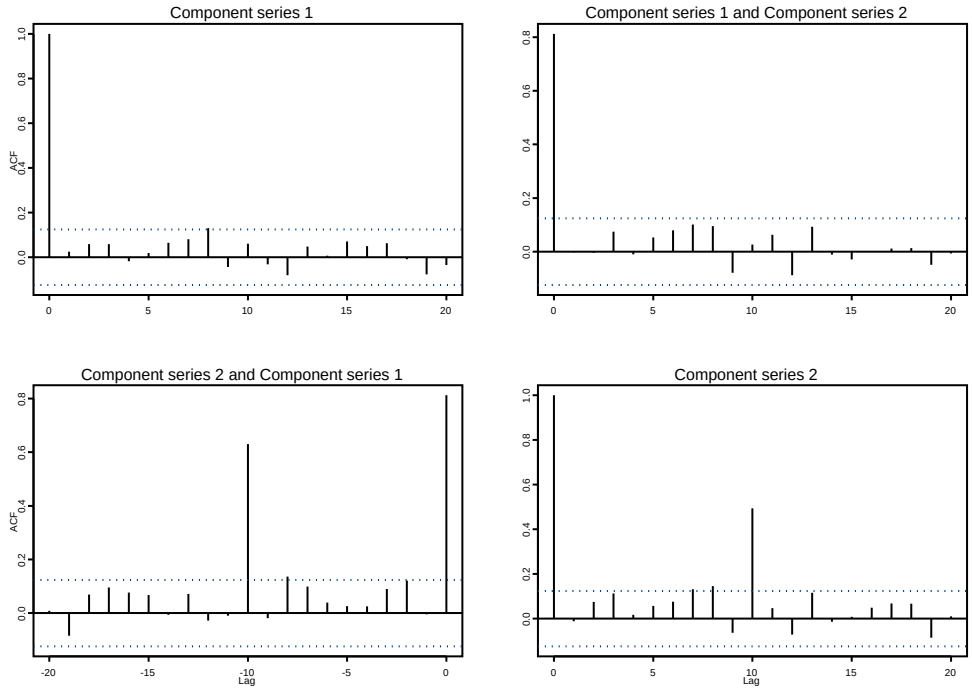

Figure 1.3: Cross correlogram of the realization of the bivariate time series defined by (1.2) with standard t innovations as plotted in Figure 1.1, see Example 1.2 for commentary.

The same analysis as above was conducted with the second time series realization, where the innovations Z_t were t distributed. The cross correlogram is displayed in Figure 1.3 and does qualitatively not differ much from the above. The estimated values are:

$$\widehat{P}(-10) \approx \begin{bmatrix} 0 & .6302 \\ 0 & .4941 \end{bmatrix}, \qquad \widehat{P}(0) \approx \begin{bmatrix} 1 & .8121 \\ .8121 & 1 \end{bmatrix},$$

with $\widehat{P}(h) \approx 0$ for $h \in \mathbb{Z} \setminus \{-10, 0, 10\}$.

1.3 Vector ARMA Processes

The main goal of this section is to provide a brief excursion into multivariate ARMA models, since they are a traditional and well-investigated tool for modeling time series in various fields. For daily financial data capturing multivariate ARMA effects is much less important than capturing multivariate volatility effects (and dynamic correlation effects) through multivariate GARCH modeling,[4] but for longer period returns the more traditional ARMA processes become increasingly useful. In econometrics literature they are more commonly known as vector ARMA or VARMA processes.

1.3.1 The VARMA model

Definition 1.9. Let $(\boldsymbol{\varepsilon}_t)_{t\in\mathbb{Z}}$ be WN$(\mathbf{0}, \Sigma_{\boldsymbol{\varepsilon}})$. The process $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$ is a zero-mean VARMA(p, q) process if it is a covariance stationary process satisfying difference equations of the form

$$\boldsymbol{X}_t - \Phi_1 \boldsymbol{X}_{t-1} - \ldots - \Phi_p \boldsymbol{X}_{t-p} = \boldsymbol{\varepsilon}_t + \Theta_1 \boldsymbol{\varepsilon}_{t-1} + \ldots + \Theta_q \boldsymbol{\varepsilon}_{t-q} \tag{1.3}$$

for all $t \in \mathbb{Z}$ and fixed parameter matrices Φ_i and Θ_j in $\mathbb{R}^{d\times d}$. $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$ is a VARMA(p, q) process with mean $\boldsymbol{\mu}$ if the centered series $(\boldsymbol{X}_t - \boldsymbol{\mu})_{t\in\mathbb{Z}}$ is a zero-mean VARMA(p, q) process.

Remarks.

1. VARMA stands for Vector Auto-Regressive Moving Average. The left hand side of (1.3) is usually referred to as the (V)AR part, the right hand side as the (V)MA part.

2. Equation (1.3) is often written in the more compact form

$$\Phi(B)\boldsymbol{X}_t = \Theta(B)\boldsymbol{\varepsilon}_t, \qquad (\boldsymbol{\varepsilon}_t)_{t\in\mathbb{Z}} \sim \text{WN}(\mathbf{0}, \Sigma_{\boldsymbol{\varepsilon}}), \tag{1.4}$$

[4] Multivariate GARCH modeling will be the main topic in Chapter 3.

where $\Phi(z) := I - \Phi_1 z - \ldots - \Phi_p z^p$ and $\Theta(z) := I + \Theta_1 z + \ldots + \Theta_q z^q$ are matrix-valued polynomials, I is the $d \times d$ identity matrix and B denotes the backward shift operator, defined by $B^k \boldsymbol{X}_t = \boldsymbol{X}_{t-k}$ for $k \in \mathbb{N}$.

3. For practical applications only *causal* VARMA processes are considered, which are, loosely speaking, processes which depend only on the past. This idea is being formalized in Definition 1.10.

Definition 1.10 (Causality)**.** A process $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$ is a *causal* process, if it admits a representation of the form

$$\boldsymbol{X}_t - \boldsymbol{\mu} = \sum_{i=0}^{\infty} \Psi_i \boldsymbol{\varepsilon}_{t-i}, \tag{1.5}$$

where $(\boldsymbol{\varepsilon}_t)_{t\in\mathbb{Z}}$ is $\mathrm{WN}(\mathbf{0}, \Sigma_\epsilon)$ and $(\Psi_i)_{i\in\mathbb{N}_0}$ is a sequence of matrices in $\mathbb{R}^{d\times d}$ whose components are absolutely summable, i.e.

$$\sum_{i=0}^{\infty} |\Psi_{i,jk}| < \infty \tag{1.6}$$

holds for any j and k in $\{1, \ldots, d\}$.

Remark. The so-called absolute summability condition (1.6) is a technical condition which ensures that $\mathbb{E}|\boldsymbol{X}_t| < \infty$. This guarantees that the infinite sum in (1.5) converges absolutely, almost surely, meaning that both $\sum_{i=0}^{\infty} |\Psi_i||\boldsymbol{\varepsilon}_{t-i}|$ and $\sum_{i=0}^{\infty} \Psi_i \boldsymbol{\varepsilon}_{t-i}$ are finite with probability one.[5] See Brockwell and Davis [8] for details.

1.3.2 Some properties of VARMA processes

One striking reason of restricting attention to causal models is the fact that causality implies covariance stationarity, and first and second moments may easily be calculated.

[5]The absolute value of matrices and vectors is to be taken component-wise and not to be confused with the determinant.

Proposition 1.1. *Any causal process is covariance stationary with* $\mathbb{E}[\boldsymbol{X}_t] = \boldsymbol{\mu}$. *For* $h \geq 0$ *the covariance matrix function is given by*

$$\Gamma(h) = \sum_{i=0}^{\infty} \Psi_{i+h} \Sigma_{\boldsymbol{\varepsilon}} \Psi_i'.$$

Proof. Considering the absolute summability condition (1.6) and linearity of expectation, it easily follows that

$$\mathbb{E}[\boldsymbol{X}_t] = \mathbb{E}[\boldsymbol{\mu} + \sum_{i=0}^{\infty} \Psi_i \boldsymbol{\varepsilon}_{t-i}] \overset{(1.6)}{=} \mathbb{E}[\boldsymbol{\mu}] + \sum_{i=0}^{\infty} \Psi_i \mathbb{E}[\boldsymbol{\varepsilon}_{t-i}] = \boldsymbol{\mu}.$$

For $h \geq 0$ the covariance matrix function is given by

$$\Gamma(t+h, t) = \mathbb{E}[(\boldsymbol{X}_{t+h} - \boldsymbol{\mu})(\boldsymbol{X}_t - \boldsymbol{\mu})'] = \mathbb{E}\Big[\sum_{i=0}^{\infty} \Psi_i \boldsymbol{\varepsilon}_{t+h-i} \sum_{j=0}^{\infty} \boldsymbol{\varepsilon}_{t-j}' \Psi_j' \Big],$$

which can, again by interchanging sums and expectations according to (1.6), be rewritten as

$$\sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \Psi_i \mathbb{E}[\boldsymbol{\varepsilon}_{t+h-i} \boldsymbol{\varepsilon}_{t-j}'] \Psi_j' = \sum_{i=h}^{\infty} \Psi_i \mathbb{E}[\boldsymbol{\varepsilon}_{t+h-i} \boldsymbol{\varepsilon}_{t+h-i}'] \Psi_{i-h}' = \sum_{i=h}^{\infty} \Psi_i \Sigma_{\boldsymbol{\varepsilon}} \Psi_{i-h}',$$

applying that $\mathbb{E}[\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_j'] \neq 0 \iff i = j$. □

The requirement that a process satisfying the VARMA-equations (1.3) (or equivalently (1.4)) be causal imposes conditions on the values that the parameter matrices Φ_i (in particular) and Θ_j may take. The theory is well-investigated, one important result being the *causality criterion*.

Proposition 1.2 (Causality criterion)**.** *If for all* $z \in \mathbb{C}$ *with* $|z| \leq 1$,

$$\det \Phi(z) \neq 0, \tag{1.7}$$

holds, then (1.4) has exactly one stationary solution. This solution is causal, i.e.

$$\boldsymbol{X}_t = \sum_{j=0}^{\infty} \Psi_j \boldsymbol{\varepsilon}_{t-j}.$$

The matrices Ψ_j are determined uniquely by

$$\Psi(z) := \sum_{j=0}^{\infty} \Psi_j z^j = \Phi^{-1}(z)\Theta(z), \qquad |z| \leq 1. \tag{1.8}$$

Proof. The proof of this proposition makes use of the existence of $\Phi^{-1}(z)$ for $|z| < 1 + \varepsilon$, which is assured by condition (1.7). Therefore it has a power series expansion with nice properties, which is then used for the desired representation. For a detailed proof the reader is referred to Brockwell and Davis [8], Theorem 11.3.1. □

Remark. The matrices Ψ_j from Proposition 1.2 can easily be found recursively from the equations

$$\begin{aligned} \Psi_0 &= I, \\ \Psi_j &= \sum_{i=1}^{j} \Phi_i \Psi_{j-i} + \Theta_j, \qquad j = 1, 2, \ldots, \end{aligned}$$

where $\Theta_j = 0$ for $j > q$, and $\Phi_i = 0$ for $i > p$. These equations are established by comparing coefficients of z^j in the power series identities (1.8) after multiplying through by $\Phi(z)$.

1.3.3 Examples and problems

In many cases the full generality of VARMA models is not required, and VAR or VMA models suffice to explain the data. One widely used, and rather simple model is the VAR(1) model, analyzed in Example 1.3.

Example 1.3 (VAR(1) process)**.** The first-order VAR process satisfies the set of vector difference equations

$$\boldsymbol{X}_t = \Phi \boldsymbol{X}_{t-1} + \boldsymbol{\varepsilon}_t, \tag{1.9}$$

for all $t \in \mathbb{Z}$. According to Theorem 1.2 it is possible to find a causal process satisfying (1.5) and (1.6) that is a solution of (1.9) if all eigenvalues of the matrix

Φ are less than one in absolute value. The causal process

$$\boldsymbol{X}_t = \sum_{i=0}^{\infty} \Phi^i \boldsymbol{\varepsilon}_{t-i},$$

calculated directly or recursively as explained above, is then the unique solution. This solution can be thought of as an infinite order vector moving average process, a so-called VMA(∞) process. The covariance matrix function of this process follows from Proposition 1.1 and is given by

$$\Gamma(h) = \sum_{i=0}^{\infty} \Phi^{i+h} \Sigma_{\boldsymbol{\varepsilon}} (\Phi^i)' = \Phi^h \sum_{i=0}^{\infty} \Phi^i \Sigma_{\boldsymbol{\varepsilon}} (\Phi^i)' = \Phi^h \Gamma(0), \qquad h \in \mathbb{N}_0.$$

Full VARMA models are less common than models from the VAR subfamily in practice, one reason being that identifiability problems arise when estimating parameters. For example, we can have situations where the first order VARMA(1,1) model $\boldsymbol{X}_t - \Phi \boldsymbol{X}_{t-1} = \boldsymbol{\varepsilon}_t + \Theta \boldsymbol{\varepsilon}_{t-1}$ can be rewritten as $\boldsymbol{X}_t - \Phi^* \boldsymbol{X}_{t-1} = \boldsymbol{\varepsilon}_t + \Theta^* \boldsymbol{\varepsilon}_{t-1}$ for completely different parameter matrices Ψ^* and Θ^*, as shown in Example 1.4. Such an identifiability problem is serious, because, without proper constraints, the likelihood function of a vector ARMA(1,1) model for the data may not have a unique maximum.

Example 1.4 (Tsay [26])**.** Consider the two-dimensional VARMA(1,1) model defined by

$$\boldsymbol{X}_t - \begin{bmatrix} .8 & -2 \\ 0 & 0 \end{bmatrix} \boldsymbol{X}_{t-1} = \boldsymbol{\varepsilon}_t + \begin{bmatrix} .5 & 0 \\ 0 & 0 \end{bmatrix} \boldsymbol{\varepsilon}_{t-1}.$$

This model is identical to the VARMA(1,1) model

$$\boldsymbol{X}_t - \begin{bmatrix} .8 & -2+a \\ 0 & b \end{bmatrix} \boldsymbol{X}_{t-1} = \boldsymbol{\varepsilon}_t + \begin{bmatrix} .5 & a \\ 0 & b \end{bmatrix} \boldsymbol{\varepsilon}_{t-1}$$

for any nonzero a and b. In this particular instance, the equivalence occurs because we have $X_{t,2} = \varepsilon_{t,2}$ in both the first and the second model. The effects on the parameters a and b on the system cancel out between AR and MA parts of the second model.

Another problem that arises when estimating VARMA models is the estimation procedure itself. In order to obtain the maximum likelihood estimate, the likelihood function has to be maximized numerically. This is computationally expensive and local maxima may be encountered.

Chapter 2

Univariate Models for Changing Volatility

The most important models for daily risk factor return series are addressed in this section. Definitions of (univariate) ARCH and GARCH models are given, and some of their mathematical properties are discussed. Throughout the chapter examples of simulated time series are presented in order to illustrate (some of) the ideas. Structure and notation again follows McNeil et al. [20], with some minor exceptions.

2.1 ARCH Processes

2.1.1 Definition and basic properties

Definition 2.1. Let $(Z_t)_{t\in\mathbb{Z}}$ be univariate SWN(0,1). The process $(X_t)_{t\in\mathbb{Z}}$ is an ARCH(p) process if it is strictly stationary and if it satisfies for all $t \in \mathbb{Z}$ and some

strictly positive-valued process $(\sigma_t)_{t\in\mathbb{Z}}$ equations of the form

$$X_t = \sigma_t Z_t, \tag{2.1}$$

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2, \tag{2.2}$$

where $\alpha_0 > 0$ and $\alpha_i \geq 0$ for $i = 1, \ldots, p$. Each Z_t shall be called an *innovation* and interpreted as such.

Remarks. 1. Let $\mathcal{F}_t = \sigma\{(X_s)_{s\leq t}\}$ again denote the σ-algebra representing the history of the process up to time t. Clearly construction (2.2) ensures that σ_t is measurable with respect to $\mathcal{F}_{t-1}$. Provided that $\mathbb{E}[|X_t|] < \infty$ and applying the independence of Z_t and $\mathcal{F}_{t-1}$, one may easily calculate that

$$\mathbb{E}[X_t|\mathcal{F}_{t-1}] = \mathbb{E}[\sigma_t Z_t|\mathcal{F}_{t-1}] = \sigma_t \mathbb{E}[Z_t|\mathcal{F}_{t-1}] = \sigma_t \mathbb{E}[Z_t] = 0, \tag{2.3}$$

meaning that the ARCH process has the martingale difference property with respect to the natural filtration $(\mathcal{F}_t)_{t\in\mathbb{Z}}$.

2. If one further assumes that $\mathbb{E}[X_t^2]$ exists so that the process $(X_t)_{t\in\mathbb{Z}}$ is a covariance stationary white noise[1], one can also calculate that

$$\mathrm{var}(X_t|\mathcal{F}_{t-1}) = \mathbb{E}[\sigma_t^2 Z_t^2|\mathcal{F}_{t-1}] = \sigma_t^2 \mathrm{var}(Z_t) = \sigma_t^2.$$

Thus the model has the interesting property that its conditional standard deviation σ_t, or *volatility*, is a continually changing function of the previous squared values of the process. If one or more of the past $|X_{t-1}|, \ldots, |X_{t-p}|$ are particularly large, then X_t is effectively drawn from a distribution with large variance, and may itself be large; in this way the model generates volatility clusters. The name ARCH (autoregressive conditionally heteroscedastic) refers to this structure: the model is

- *autoregressive*, since X_t clearly depends on previous X_{t-i}, and
- *conditionally heteroscedastic*, since the conditional standard deviation changes continually over time.

[1] A condition therefore may be formulated in terms of restrictions on the parameters α_i (namely that $\sum_{i=1}^{p} \alpha_i < 1$) and will be given in Proposition 2.4.

3. Note that the independence of Z_t and $\mathcal{F}_{t-1}$ which was assumed above is only true for causal ARCH process, i.e. the equations (2.1) and (2.2) must have a solution of the form $X_t = f(Z_t, Z_{t-1}, \ldots)$ for some f so that Z_t is independent of previous values of the process. Therefore, in practical applications only causal solutions are considered (and in fact the causality-requirement is often included in the definition itself).

4. The distribution of the innovations $(Z_t)_{t\in\mathbb{Z}}$ can in principle be any zero-mean, unit-variance distribution. For statistical fitting purposes one may or may not choose to actually specify the distribution, depending on whether a maximum-likelihood (ML), quasi-maximum-likelihood (QML) or nonparametric fitting method is implemented.[2] According to McNeil et al. [20], the most common choices for ML are standard normal innovations, i.e. $Z_t \sim N(0,1)$, or t innovations scaled in such a way that the variance is one, i.e. $Z_t \sim t(\nu, 0, \frac{\nu-2}{\nu})$.[3]

2.1.2 Stationarity aspects of ARCH(1)

In this section properties of the ARCH(1) model are analyzed. They extend to the whole ARCH class and, later on, to GARCH models. Nevertheless, these properties are pointed out in the simplest case.

Some intuitive discussion

Focusing on the question which condition an ARCH(1) process needs to satisfy in order to be stationary, it is important to note that especially in the context of ARCH models there has to be carefully distinguished between covariance stationarity and strict stationarity, since it is possible that there exist strictly stationary ARCH(1) processes with infinite variance (which are of course not covariance stationary).

[2] A more extensive discussion of fitting the models will be topic of Section 2.3.
[3] See equations (2.18) and (2.20) in Section 2.3.2 for the forms of the densities.

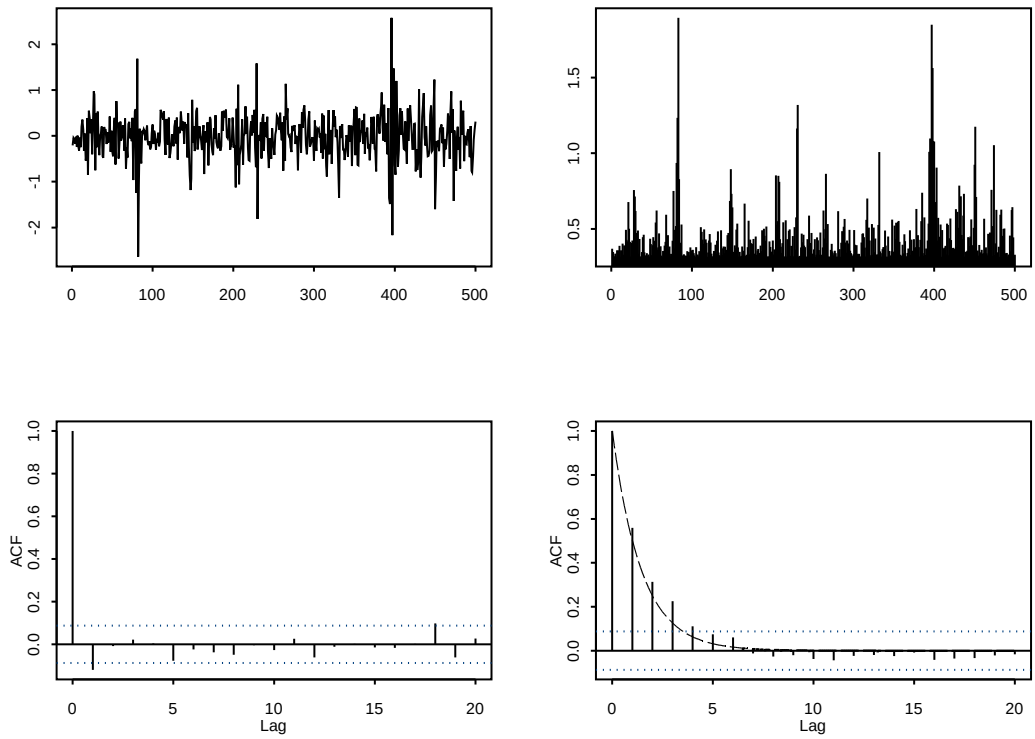

Figure 2.1: ARCH(1) process with simulated Gaussian innovations and parameters $\alpha_0 = 0.1$ and $\alpha_1 = 0.5$. The upper left picture shows the realization of the process itself, the upper right the realization of the volatility σ_t. The lower two pictures display the correlograms of the raw values X_t and squared values X_t^2. The process is covariance stationary with variance $1/5$ and finite fourth moment (since $\alpha_1 < 1/\sqrt{3} \approx 0.577$) and the squared values follow an AR(1) process. The true form of the ACF of the squared values is a dashed line in the correlogram. Details and explanation to the above ideas will be given in the following sections.

Using $X_t^2 = \sigma_t^2 Z_t^2$ and (2.2) in the case $p = 1$, the squared ARCH(1) process may be written as

$$X_t^2 = \alpha_0 Z_t^2 + \alpha_1 X_{t-1}^2 Z_t^2. \tag{2.4}$$

Taking expectation on both sides and noting the independence of X_{t-1} and Z_t, the above equation (2.4) transforms to

$$\mathbb{E}[X_t^2] = \alpha_0 + \alpha_1 \mathbb{E}[X_{t-1}^2],$$

since $\mathbb{E}[Z_t^2] = 1$ by Definition 2.1. Assuming covariance stationarity, $\mathbb{E}[X_t^2] = \mathbb{E}[X_{t-1}^2] := \sigma_x^2 < \infty$ holds and the last line simplifies to

$$\sigma_X^2 = \alpha_0 + \alpha_1 \sigma_X^2.$$

This equation may now easily be solved for σ_X^2, obtaining the variance of a covariance stationary ARCH(1) process as an elementary function of the parameters α_0 and α_1:

$$\sigma_X^2 = \frac{\alpha_0}{1 - \alpha_1}.$$

Simply because the variance σ_X^2 has to be positive and α_0 is strictly greater than zero by definition, $\alpha_1 < 1$ is a necessary condition for the ARCH(1) model to be covariance stationary. Moreover, it is also a sufficient condition, which may be seen by applying parts of the theory of *stochastic recurrence equations* (SREs).

Strict stationarity of ARCH(1)

Equation (2.4) is one particular SRE of the form

$$Y_t = A_t Y_{t-1} + B_t, \tag{2.5}$$

where A_t and B_t are each iid series of random coefficients (in this special case $A_t = \alpha_1 Z_t^2$ and $B_t = \alpha_0 Z_t^2$). In order to be able to answer the question whether there exists a stationary solution of (2.4), the following theorem for stochastic recurrence equations may be used.

Theorem 2.1 (Brandt [7])**.** *Given a stochastic recurrence equation of type (2.5) with the coefficients satisfying*

$$\mathbb{E}[\log(|A_t|)] < 0, \qquad \mathbb{E}[\max\{0, \log(|B_t|)\}] < \infty, \tag{2.6}$$

then equation (2.5) has a unique stationary solution given by

$$Y_t = \sum_{i=0}^{\infty} B_{t-i} \prod_{j=0}^{i-1} A_{t-j},$$

where the sum on the right hand side converges absolutely, almost surely.[4] *This solution is strictly stationary.*

For the detailed proof of this theorem, the reader is referred to eg. Brandt [7]. Nevertheless, some intuition for these conditions and the form of the solution is given by iterating equation (2.5) k times to obtain

$$\begin{aligned} Y_t &= A_t(A_{t-1}Y_{t-2} + B_{t-1}) + B_t \\ &= B_t + \sum_{i=1}^{k} B_{t-i} \prod_{j=0}^{i-1} A_{t-j} + Y_{t-k-1} \prod_{i=0}^{k} A_{t-i}. \end{aligned}$$

The conditions (2.6) ensure that the middle term on the right converges absolutely and the final term disappears. In particular note that

$$\frac{1}{k+1} \sum_{i=0}^{k} \log(|A_{t-i}|) \xrightarrow{a.s.} \mathbb{E}[\log(|A_t|)] < 0$$

by the strong law of large numbers. So

$$\prod_{i=0}^{k} |A_{t-i}| = \exp\left(\sum_{i=0}^{k} \log(|A_{t-i}|) \right) \xrightarrow{a.s.} 0,$$

which shows the importance of the condition that $\mathbb{E}[\log|A_t|] < 0$. The solution (2.7) is a strictly stationary process because it is a function of iid variables $(A_s, B_s)_{s \leq t}$, and the $\mathbb{E}[\log|A_t|] < 0$ condition turns out to be the key to the strict stationarity of ARCH and GARCH models.[5]

Applying Theorem 2.1 to the ARCH(1) model, then the squared process in (2.4) is a stochastic recurrence equation with $Y_t = X_t^2$, $A_t = \alpha_1 Z_t^2$ and $B_t = \alpha_0 Z_t^2$. Therefore, the conditions translate into

$$\mathbb{E}[\max\{0, \log(\alpha_0 Z_t^2)\}] < \infty, \qquad \mathbb{E}[\log(\alpha_1 Z_t^2)] < 0.$$

[4] The empty product appearing in $\prod_{j=0}^{i-1} A_{t-j}$ for $i = 0$ is set to be one.

[5] More precisely, the above condition is in fact the key to a strictly stationary solution of the ARCH equations (2.1) and (2.2) (or GARCH equations (2.11) and (2.12), respectively), which is a requirement for a (G)ARCH process by definition.

The first condition is automatically satisfied by definition. Hence, the second condition is the condition for a strictly stationary solution of the ARCH(1) equations (2.1) and (2.2) (which makes a process satisfying these equations an ARCH(1) process as defined). It may be shown that it is in fact a necessary *and* sufficient condition for strict stationarity; see Bougerol and Picard [6]. Assuming this condition fullfilled, $(X_t^2)_{t\in\mathbb{Z}}$ takes the form

$$X_t^2 = \alpha_0 \sum_{i=0}^{\infty} \alpha_1^i \prod_{j=0}^{i} Z_{t-j}^2. \tag{2.7}$$

It may easily be shown by simulation[6] that if the $(Z_t)_{t\in\mathbb{Z}}$ follow the standard normal distribution, then the condition for a strictly stationary solution is approximately $\alpha_1 \lesssim 3.56$. Maybe somewhat surprisingly, if the $(Z_t)_{t\in\mathbb{Z}}$ are scaled t distributed with four degrees of freedom and variance one, the condition is $\alpha_1 \lesssim 5.43$; choosing three degrees of freedom raises the bound to $\alpha_1 \lesssim 7.39$.

Covariance stationarity of ARCH(1)

Obviously, strict stationarity depends on the distribution of the $(Z_t)_{t\in\mathbb{Z}}$, but covariance stationarity does not. The following theorem verifies that $\alpha_1 < 1$ is a necessary and also sufficient condition for covariance stationarity, without having to specify any distributional assumptions.

Theorem 2.2. *An ARCH(1) process as in Definition 2.1 has a covariance stationary solution if and only if $\alpha_1 < 1$. Furthermore, the variance of the process is given by $\alpha_0/(1-\alpha_1)$.*

Proof. $\Longrightarrow$: This direction has already been proved.

$\Longleftarrow$: Assuming that $\alpha_1 < 1$, then by Jensen's inequality[7] for concave functions,

$$\mathbb{E}[\log(\alpha_1 Z_t^2)] \leq \log \mathbb{E}[\alpha_1 Z_t^2] = \log(\alpha_1) < 0$$

[6]The straightforward simulation was performed in S-PLUS [18], producing the same result as stated in McNeil et al. [20] and Shepard [24].

[7]See for example Elstrodt [11] for details on this famous and often useful theorem.

holds. Theorem 2.1 may be applied and taking expectations on both sides of (2.7) results in

$$\mathbb{E}[X_t^2] = \alpha_0 \mathbb{E}\left[\sum_{i=0}^{\infty} \alpha_1^i \prod_{j=0}^{i} Z_{t-j}^2\right] = \alpha_0 \sum_{i=0}^{\infty} \alpha_1^i = \frac{\alpha_0}{1-\alpha_1},$$

yielding the desired result. □

Thus, the ARCH(1) process $(X_t)_{t\in\mathbb{Z}}$ with $\alpha_1 < 1$ is a martingale difference sequence with finite and constant second moment and therefore a white noise process according to Definition 1.5.

Example 2.1. In Figure 2.2 four time series plots are displayed. Each of the time series is generated by ARCH(1) equations using the same innovations. The parameter $\alpha_0 = 0.01$ is fixed for all four, and α_1 varies in each series. The top time series is both strictly and covariance stationary ($\alpha_1 = 0.99$), whereas the middle two are not covariance stationary but strictly stationary ($\alpha_1 = 2$ and $\alpha_1 = 3$, respectively). In the bottom picture a non-stationary (explosive) process with $\alpha_1 = 4$ is displayed.

Note that the lower two pictures use a special logarithmic y-axis, where all values less than one in modulus are set to be zero. More concretely, this means that

$$y_{\log} = \operatorname{sgn}(y) \cdot \log\left(\max\{1, |y|\}\right)$$

is being plotted on the y-axis.

2.1.3 Higher moments of ARCH(1)

It is clear from (2.7) that the distribution of $(X_t)_{t\in\mathbb{Z}}$ in an ARCH(1) model bears a complicated relationship to the distribution of $(Z_t)_{t\in\mathbb{Z}}$. Even if the innovations are Gaussian, the stationary distribution of the time series is not Gaussian, but rather a *leptokurtic*[8] distribution with more slowly decaying tails. The distribution

[8] A random variable is loosely speaking called leptokurtic if its probability density curve has fatter tails and a higher peak at the mean than the normal distribution.

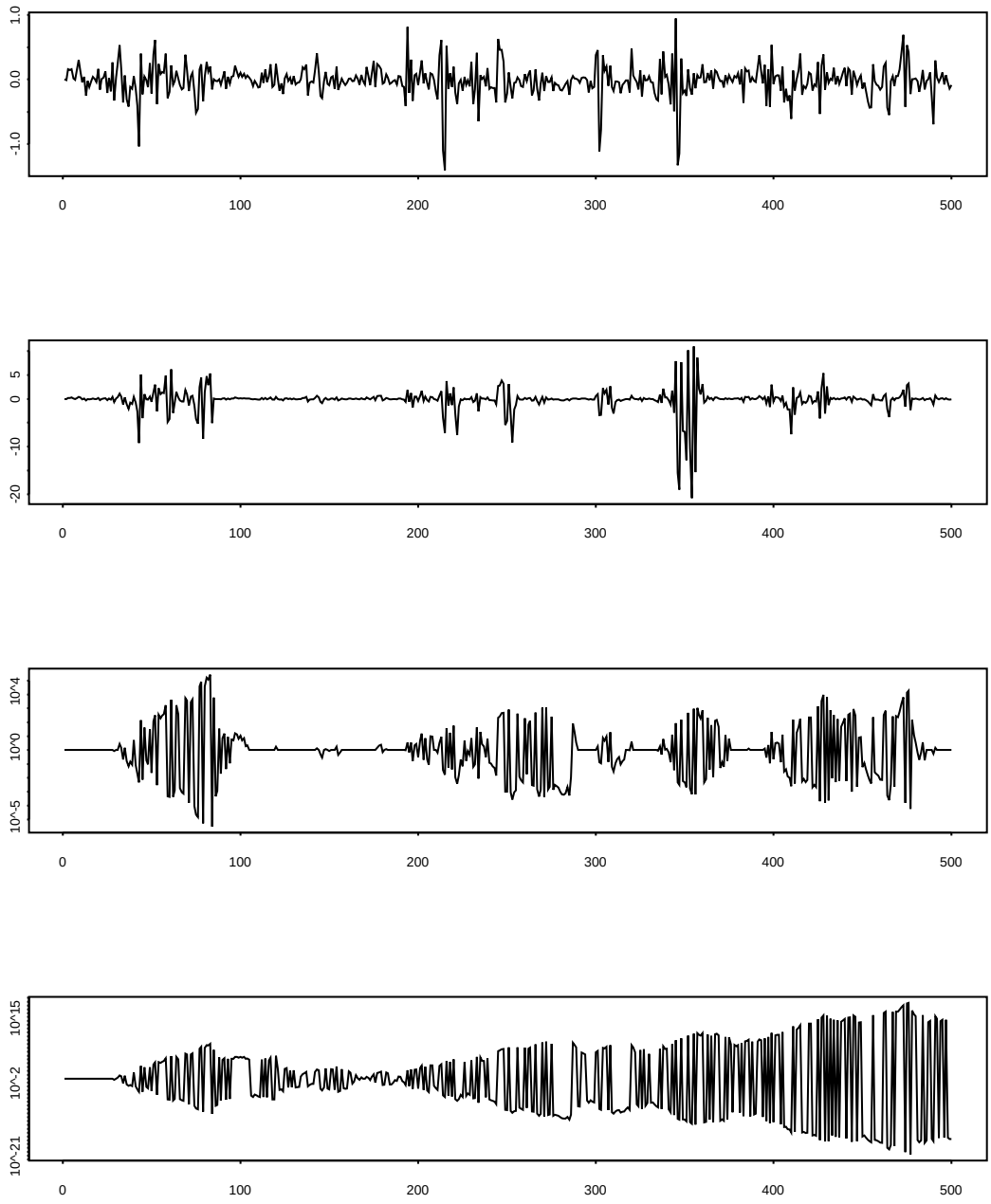

Figure 2.2: Four time series generated by ARCH(1) equations using fixed α_0 but varying α_1. From top to bottom, $\alpha_1 = 0.99$, $\alpha_1 = 2$, $\alpha_1 = 3$ and $\alpha_1 = 4$, respectively. Please refer to Example 2.1 for detailed explanation.

of X_t would essentially be known if the distribution of σ_t was known, which has no simple form. Nevertheless, a lot is known about the behaviour of its tails.

Theorem 2.3 (similar to Engle [12])**.** *For $m \geq 1$, the strictly stationary ARCH(1) process has finite moments of order $2m$ if and only if*

$$\mathbb{E}[Z_t^{2m}] < \infty \qquad \textit{and} \qquad \alpha_1 < \mathbb{E}[Z_t^{2m}]^{-1/m}.$$

Proof. We rewrite (2.7) in the form

$$X_t^2 = Z_t^2 \sum_{i=0}^{\infty} Y_{t,i},$$

for positive random variables $Y_{t,i} = \alpha_0 \alpha_1^i \prod_{j=1}^{i} Z_{t-j}^2$ for $i \geq 1$ and $Y_{t,0} = \alpha_0$. For $m \geq 1$ and an arbitrary large $n \in \mathbb{N}^+$ the following inequalities hold:

$$\sum_{k=0}^{n} \mathbb{E}[Y_{t,k}^m] \leq \mathbb{E}\left[\left(\sum_{k=0}^{n} Y_{t,k}\right)^m\right] \leq \left(\sum_{k=0}^{n} \mathbb{E}[Y_{t,k}^m]^{1/m}\right)^m.$$

The first inequality is true due to linearity of expectation and elementary calculus, and the second one is known as *Minkowski's inequality.*[9] Since

$$\mathbb{E}[X_t^{2m}] = \mathbb{E}[Z_t^{2m}]\mathbb{E}\left[\left(\sum_{i=0}^{\infty} Y_{t,i}\right)^m\right],$$

it follows that

$$\mathbb{E}[Z_t^{2m}] \sum_{i=0}^{\infty} \mathbb{E}[Y_{t,i}^m] \leq \mathbb{E}[X_t^{2m}] \leq \mathbb{E}[Z_t^{2m}]\left(\sum_{i=0}^{\infty} \mathbb{E}[Y_{t,i}^m]^{1/m}\right)^m.$$

By observing that $\mathbb{E}[Y_{t,i}^m] = \alpha_0^m(\alpha_1^m \mathbb{E}[Z_t^{2m}])^i$ it may now easily be deduced that the above sums converge if and only if $\alpha_1^m \mathbb{E}[Z_t^{2m}] < 1$ and $\mathbb{E}[Z_t^{2m}] < \infty$, which proofs the theorem. □

Remark. For a finite fourth moment (meaning that $m = 2$), the condition required amounts to $\alpha_1 < 1/\sqrt{3}$ in the case of Gaussian innovations and $\alpha_1 < 1/\sqrt{6}$ in the case of t_6 innovations; for t_4 innovations the fourth moment is undefined.

[9] Details and proof may again be found in Elstrodt [11].

Assuming the existence of a finite moment it is easy to calculate its value as well as that of the kurtosis of the process. Squaring both sides of (2.4) and taking expectations of both sides results in

$$\mathbb{E}[X_t^4] = \alpha_0^2\mathbb{E}[Z_t^4] + \alpha_1^2\mathbb{E}[Z_t^4]\mathbb{E}[X_t^4] + \frac{2\alpha_0^2\alpha_1\mathbb{E}[Z_t^4]}{1-\alpha_1}.$$

Solved for $\mathbb{E}[X_t^4]$ the above amounts to

$$\mathbb{E}[X_t^4] = \frac{\alpha_0^2\mathbb{E}[Z_t^4](1+\alpha_1)}{(1-\alpha_1)(1-\alpha_1^2\mathbb{E}[Z_t^4])}.$$

The kurtosis of the stationary distribution κ_X can then easily be calculated to be

$$\kappa_X = \frac{\mathbb{E}[X_t^4]}{\mathbb{E}[X_t^2]^2} = \frac{\kappa_Z(1-\alpha_1^2)}{1-\alpha_1^2\kappa_Z},$$

where $\kappa_Z = \mathbb{E}[Z_t^4]$ denotes the kurtosis of the innovations. Clearly when $\kappa_Z > 1$, the kurtosis of the stationary distribution κ_X is inflated in comparison with that of the innovation distribution; for Gaussian or scaled t innovations κ_X is greater than 3, so the stationary distribution is leptokurtic.

2.1.4 Parallels of ARCH(1) and AR(1)

When observing the serial dependence structure of the squared time series $(X_t^2)_{t\in\mathbb{Z}}$ in the case of covariance stationarity ($\alpha_1 < 1$), one can see a close resemblance to an univariate AR(1) process. We write the squared process as

$$X_t^2 = \sigma_t^2 Z_t^2 = \sigma_t^2 + \sigma_t^2(Z_t^2 - 1), \tag{2.8}$$

and setting $T_t = \sigma_t^2(Z_t^2 - 1)$ we note that $(T_t)_{t\in\mathbb{Z}}$ forms a martingale difference series, since $\mathbb{E}[|T_t|] < \infty$ and $\mathbb{E}[T_t|\mathcal{F}_{t-1}] = \sigma_t^2\mathbb{E}[Z_t^2 - 1] = 0$. By plugging in according to Definition 2.1, X_t^2 may further be rewritten as

$$X_t^2 = \alpha_0 + \alpha_1 X_{t-1}^2 + T_t, \tag{2.9}$$

which now closely resembles an AR(1) process for X_t^2, except that T_t is not necessarily a white noise process. In order to assure this property, one needs to restrict their attention to processes where $\mathbb{E}[X_t^4]$ is finite. T_t then has a finite and constant second moment and is a white noise process. Under this assumption, $(X_t^2)_{t\in\mathbb{Z}}$ is an AR(1) in mean process according to Definition 1.9 of the form

$$(X_t^2 - \mu) - \alpha_1(X_{t-1}^2 - \mu) = T_t, \qquad \mu = \frac{\alpha_0}{1-\alpha_1}.$$

It is a well known fact[10] that the autocorrelation function[11] of an AR(1) process is $\rho(h) = \alpha_1^{|h|}$ for $h \in \mathbb{Z}$. For an example of an ARCH(1) process with finite fourth moment whose squared values follow an AR(1) process please see Figure 2.1.

2.1.5 ARCH(p)

This section briefly concentrates on ARCH models of higher order ($p > 1$). These models allow more flexibility when modeling the correlation structure.

Reconsidering the definition of ARCH models with higher order, an ARCH(p) process $(X_t)_{t\in\mathbb{Z}}$ is given by

$$X_t = \sigma_t Z_t, \qquad \sigma_t = \sqrt{\alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2}, \qquad t \in \mathbb{Z}, \tag{2.10}$$

where $\alpha_0 > 0$, $\alpha_1, \ldots, \alpha_p \geq 0$ and $Z_t \sim \text{SWN}(0,1)$.

The basic idea of these models is to increase the order of the autoregressive polynomial. Properties of the ARCH(p) models are generalizations of the ARCH(1) model, such as the following theorem.

Theorem 2.4 (Engle [12])**.** *The ARCH(p) process is covariance stationary if and*

[10] May be verified in almost every book about time series covering fundamental theory, see for example Brockwell and Davis [8, 9], Fuller [15] or Tsay [26].

[11] The correlation matrix function $P(h)$ as introduced in Definition 1.4 may easily be applied to the univariate case, in which it is usually referred to as *autocorrelation function* (or simply ACF), and denoted by $\rho(h) = P(h)$ for dimension $d = 1$. The lag h is usually chosen to be in $\mathbb{N}_0$, since the ACF is symmetric, i.e. $\rho(-h) = \rho(h)$.

only if

$$\sum_{i=1}^{p} \alpha_i < 1.$$

The variance is then given by

$$\mathbb{E}[X_t^2] = \frac{\alpha_0}{1 - \sum_{i=1}^{p} \alpha_i}.$$

By defining $T_t = \sigma_t^2(Z_t^2 - 1)$ the ARCH(p) may analogously to the ARCH(1) process be represented in the form

$$X_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2 + T_t.$$

Once more T_t is a white noise process assuming that 4^{th} moments of the time series $(X_t)_{t\in\mathbb{Z}}$ are finite. Hence, the squared ARCH(p) process $(X_t^2)_{t\in\mathbb{Z}}$ has an AR(p) representation.

Models with high order p have to be used in applications quite often, since the influence on volatility significantly depends on (more than one) past values. This raises the problem of estimating a large number of parameters with restrictions, which is a disadvantage of high order ARCH(p). These restrictions split up in two conditions:

- non-negativity conditions (by definition) and
- stationarity conditions (e.g. $\sum_{i=1}^{p} \alpha_i < 1$).

The estimation of a large number of parameters is numerically very laborious, if, for example, efficient estimation methods such as maximum likelihood are used (Franke, Härdle and Hafner [14]). In order to reduce the computational burden but also include a large number of past values the conditional variance may be parameterized as

$$\sigma_t^2 = \alpha_0 + \alpha_1 \sum_{i=1}^{p} w_i X_{t-i}^2,$$

where the weights w_i, $i = 1, \ldots, q$ are given by

$$w_i = \frac{2(q+1-i)}{q(q+1)}.$$

The weights decline linearly and are constructed such that $\sum_{i=1}^{p} w_i = 1$, whereas yet only two parameters have to be estimated (Bera and Higgins [3]).

2.2 GARCH Processes

2.2.1 Definition and basic properties

Definition 2.2. Let $(Z_t)_{t\in\mathbb{Z}}$ be univariate SWN(0,1). The process $(X_t)_{t\in\mathbb{Z}}$ is a GARCH(p, q) process if it is strictly stationary and if it satisfies for all $t \in \mathbb{Z}$ and some strictly positive-valued process $(\sigma_t)_{t\in\mathbb{Z}}$ the equations

$$X_t = \sigma_t Z_t, \tag{2.11}$$

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2, \tag{2.12}$$

where

- $\alpha_0 > 0, \alpha_i \geq 0$ for $i = 1, \ldots, p$,
- $\beta_j \geq 0$ for $j = 1, \ldots, q$.

Remarks. 1. GARCH process are *generalized* ARCH processes in the sense that the squared volatility σ_t^2 is allowed to depend not only on the previous squared values of the process, but also on the previous squared volatilities themselves.

2. In practice low-order GARCH models are most widely used, and focus here will be placed on the GARCH(1,1) model. In this model periods of high volatility tend to be *persistent*, since $|X_t|$ has a chance of being large if *either* $|X_{t-1}|$ is large or $|\sigma_{t-1}|$ is large. Of course the same effect can be achieved

in GARCH models of high order, but lower-order GARCH models achieve this effect more parsimoniously. A simulated realization of a GARCH(1,1) process with Gaussian innovations and its volatility are shown in Figure 2.3. In comparison with the ARCH(1) model in Figure 2.1 it is clear that the volatility persists longer at higher levels before decaying to lower levels.

2.2.2 Stationarity aspects of GARCH(1,1)

It follows from (2.12) that for a GARCH(1,1) model we have

$$\sigma_t^2 = \alpha_0 + (\alpha_1 Z_{t-1}^2 + \beta_1)\sigma_{t-1}^2, \tag{2.13}$$

which is again a SRE of the form $Y_t = A_t Y_{t-1} + B_t$ as in (2.5). This time it is a SRE for $Y_t = \sigma_t^2$ rather than X_t^2, but its analysis follows easily from the ARCH(1) case: The condition $\mathbb{E}[\log |A_t|] < 0$ for a strictly stationary solution of (2.5) translates to $\mathbb{E}[\log (\alpha_1 Z_t^2 + \beta_1)] < 0$ for (2.13) and the general solution (2.7) becomes

$$\sigma_t^2 = \alpha_0 + \alpha_0 \sum_{i=1}^{\infty} \prod_{j=1}^{i} (\alpha_1 Z_{t-j}^2 + \beta_1).$$

If $(\sigma_t^2)_{t\in\mathbb{Z}}$ is a strictly stationary process then so is $(X_t)_{t\in\mathbb{Z}}$, since $X_t = \sigma_t Z_t$ and $(Z_t)_{t\in\mathbb{Z}}$ is strict white noise. The solution of the GARCH(1,1) defining equations is then

$$X_t = Z_t \sqrt{\alpha_0 \left(1 + \sum_{i=1}^{\infty} \prod_{j=1}^{i} (\alpha_1 Z_{t-j}^2 + \beta_1)\right)}. \tag{2.14}$$

Again $\mathbb{E}[\log(\alpha_1 Z_t^2 + \beta_1)] < 0$ is not only a sufficient but also necessary condition for strict stationarity of GARCH(1,1), and using the previous result the condition for covariance stationarity may be deduced.

Theorem 2.5. *The GARCH(1,1) process is a covariance-stationary white noise process if and only if* $\alpha_1 + \beta_1 < 1$. *The variance is then given by* $\frac{\alpha_0}{1-\alpha_1-\beta_1}$.

Proof. Theorem 2.5 may be proved analogously to Theorem 2.2 by making use of (2.14). □

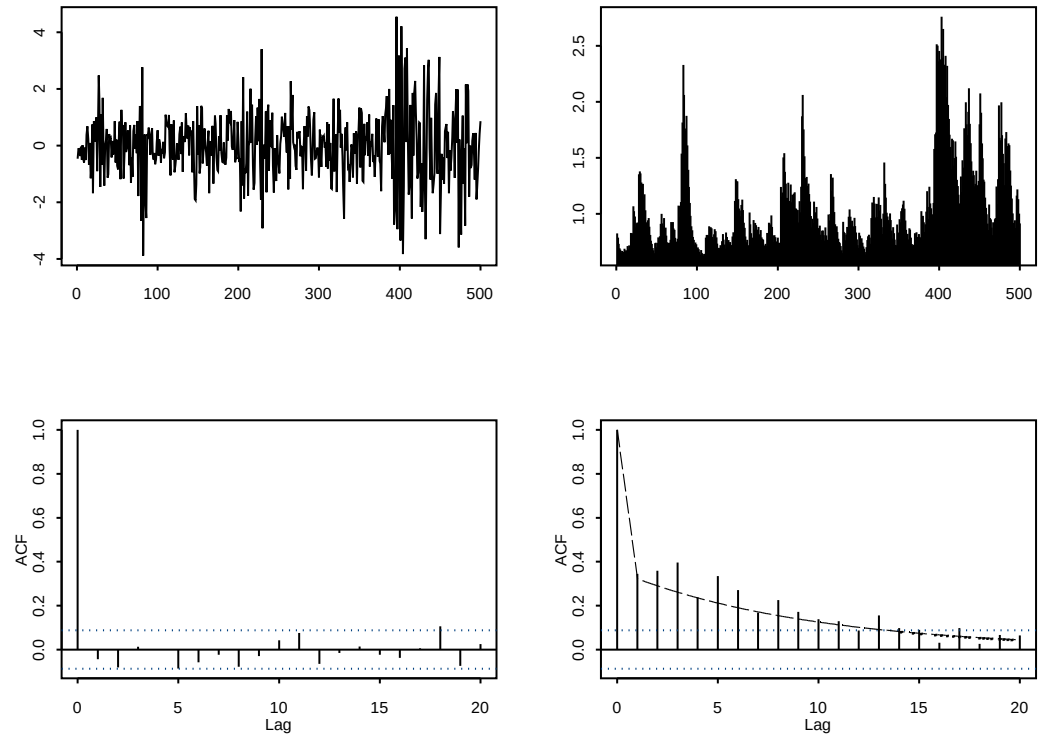

Figure 2.3: GARCH(1,1) process with simulated Gaussian innovations and parameters $\alpha_0 = 0.1, \alpha_1 = 0.2, \beta_1 = 0.7$. In the upper left picture the realization of the process itself is displayed, accompanied by the evolution of the volatility σ_t to its right. The lower two pictures show the correlograms of the raw and squared values of the time series. The process is covariance stationary according to Theorem 2.5 with unit variance and a finite fourth moment. Thus its squared values follow an ARMA(1,1) process, as derived in the following sections. The true form of the ACF of the squared values is indicated by the dashed line in the correlogram.

2.2.3 Higher moments of GARCH(1,1)

Using a similar approach to that of Theorem 2.3, representation (2.14) may be used to derive conditions for the existence of higher moments of a covariance stationary GARCH(1,1) process. For existence of a fourth moment, a necessary and sufficient condition is that $\mathbb{E}[(\alpha_1 Z_t^2 + \beta_1)^2] < 1$, which may, using elementary calculus and the fact that $\mathbb{E}[Z_t^2] = 1$, alternatively be written as

$$(\alpha_1 + \beta_1)^2 < 1 - (\kappa_Z - 1)\alpha_1^2.$$

As before, κ_Z again denotes the kurtosis of Z_t.

Assuming this to be true, fourth moment and kurtosis of X_t can again easily be calculated by squaring both sides of (2.13) and taking expectations to obtain

$$\mathbb{E}[\sigma_t^4] = \alpha_0^2 + 2\alpha_0(\alpha_1 + \beta_1)\mathbb{E}[\sigma_t^2] + (\alpha_1^2\kappa_Z + 2\alpha_1\beta_1 + \beta_1^2)\mathbb{E}[\sigma_t^4].$$

Recalling that $\mathbb{E}[\sigma_t^2] = \mathbb{E}[X_t^2] = \frac{\alpha_0}{1-\alpha_1-\beta_1}$, the above equation may be solved for $\mathbb{E}[X_t^4] = \kappa_Z \mathbb{E}[\sigma_t^4]$, obtaining

$$\mathbb{E}[X_t^4] = \frac{\alpha_0^2\kappa_Z(1 + \alpha_1 + \beta_1)}{(1 - \alpha_1 - \beta_1)(1 - \alpha_1^2\kappa_Z - 2\alpha_1\beta_1 - \beta_1^2)},$$

from which it follows that

$$\kappa_X = \frac{\kappa_Z(1 - (\alpha_1 + \beta_1)^2)}{(1 - (\alpha_1 + \beta_1)^2 - (\kappa_Z - 1)\alpha_1^2)}. \tag{2.15}$$

Equation (2.15) is written in a form which makes clear that the kurtosis of X_t is greater than that of Z_t whenever $\kappa_Z > 1$, such as for Gaussian and scaled t innovations. The kurtosis of the GARCH(1,1) model in Figure 2.3 is 5.18.

If the innovations are assumed to be standard normally distributed, there even exists a formula for calculating all higher moments. Just like Engle [12] derived a condition for $2r^{th}$ moment existence of the ARCH(1) process in 1982, Bollerslev was able to extend this theorem to the more general GARCH(1,1) class four years

later:[12]

Theorem 2.6 (Bollerslev [5])**.** *A necessary and sufficient condition for existence of the $2r^{th}$ moment for a GARCH(1,1) process with standard normal innovations is*

$$\mu(\alpha_1, \beta_1, r) = \sum_{j=0}^{r} \binom{r}{j} a_j \alpha_1^j \beta_1^{r-j} < 1,$$

where $a_0 = 1$, and $a_j = \prod_{i=1}^{j}(2j-1)$ for $j \in \mathbb{N}$. The $2r^{th}$ moment can then be expressed by the recursive formula

$$\mathbb{E}[X_t^{2r}] = \frac{a_r}{1 - \mu(\alpha_1, \beta_1, r)} \left[\sum_{n=0}^{r-1} a_n^{-1} \mathbb{E}[X_t^{2n}] \alpha_0^{r-n} \binom{r}{r-n} \mu(\alpha_1, \beta_1, n) \right].$$

Summarizing the above results for a GARCH(1,1) process X_t yields:

- The skewness of X_t equals zero if it exists and innovations themselves are not skew (true for Gaussian and t innovations).[13]
- X_t and X_{t-k} are uncorrelated for $k > 0$ (which is obviously not true for the squared values X_t^2 and X_{t-k}^2).
- The distribution of X_t is leptokurtic for both Gaussian and t innovations. More generally speaking, kurtosis of X_t is inflated in comparison with that of Z_t whenever $\kappa_Z > 1$.
- Just like ARCH(p) processes have an AR(p) representation for the squared series X_t^2, the squared GARCH(1,1) model generalizes to have a ARMA(1,1) representation (see Section 2.2.4).

Remark. Virtually all of the above features generalize unmodified to the more general GARCH(p, q) class.

[12] Please refer to Bollerslev [5] for a proof to this theorem.

[13] Can be seen analogously to (2.3): Provided that $\mathbb{E}[|X_t^3|] < \infty$,

$$\mathbb{E}[X_t^3] = \mathbb{E}[\mathbb{E}[X_t^3|\mathcal{F}_{t-1}]] = \mathbb{E}[\mathbb{E}[\sigma_t^3 Z_t^3|\mathcal{F}_{t-1}]] = \sigma_t^3 \mathbb{E}[\mathbb{E}[Z_t^3|\mathcal{F}_{t-1}]] = \sigma_t^3 \mathbb{E}[Z_t^3] = 0.$$

2.2.4 GARCH(p, q)

Higher-order GARCH models have the same general empirical behavior as the above discussed GARCH(1,1), but their mathematical analysis becomes more tedious. Nevertheless, similar to the case of ARCH processes, not only empirical but as well mathematical properties of the GARCH(1,1) models generalize to GARCH(p, q) models for $p > 1$ or $q > 1$.

Stationarity aspects of GARCH(p, q)

In higher-order models, the condition for a strictly stationary solution of the GARCH-equations (2.11) and (2.12) has been derived by Bougerol and Picard [6] in 1992, but is rather complicated. For covariance stationarity, the necessary and sufficient condition has been derived by Bollerslev in 1986 and is stated in the following theorem.

Theorem 2.7 (Bollerslev [5])**.** *The GARCH(p, q) process according to Definition 2.2 is covariance stationary if and only if*

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

The variance of the process is then given by

$$\mathbb{V}[X_t] = \mathbb{E}[X_t^2] = \frac{\alpha_0}{1 - \sum_{i=1}^{p} \alpha_i - \sum_{j=1}^{q} \beta_j}.$$

Remark. As can be seen, the condition for covariance stationarity is similar to that of the ARCH(p) case in Theorem 2.4. Not surprisingly, the GARCH(p, q) version of the theorem additionally includes the β_j terms for $j = 1, \ldots, q$ next to the familiar α_i terms for $i = 1, \ldots, p$.

Parallels of GARCH(p,q) and ARMA(p,q) process

As done for the ARCH process (see Section 2.1.4), the covariance stationary GARCH(p,q) process may be written as

$$X_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2 + T_t,$$

where T_t is a martingale difference process given by $T_t = \sigma_t^2(Z_t^2 - 1)$. Since $\sigma_{t-1}^2 = X_{t-1}^2 - T_{t-1}$, the above formula can be rewritten as

$$X_t^2 = \alpha_0 + \sum_{i=1}^{m} (\alpha_i + \beta_i) X_{t-i}^2 - \sum_{j=1}^{q} \beta_j T_{t-j} + T_t \tag{2.16}$$

by setting $m = \max(p,q)$ and

$$\begin{aligned} \alpha_i = 0 \quad \text{for} \quad i &= p+1, \ldots, q \quad \text{if} \quad q > p, \\ \beta_j = 0 \quad \text{for} \quad j &= q+1, \ldots, p \quad \text{if} \quad p > q. \end{aligned}$$

Equation (2.16) now begins to resemble an ARMA(m,q) process for X_t^2. If we further assume that $\mathbb{E}[X_t^4] < \infty$ and recall that X_t is covariance stationary (and therefore $\sum_{i=1}^{m} (\alpha_i + \beta_i) < 1$), we may rewrite (2.16) according to Definition 1.9,

$$(X_t^2 - \mu) - \sum_{i=1}^{m} \phi_i (X_{t-i}^2 - \mu) = T_t + \sum_{j=1}^{q} \theta_j T_{t-j}.$$

Now it can be observed that X_t^2 is formally an ARMA(m,q) process with parameters $\phi_i = \alpha_i + \beta_i$, $\theta_j = -\beta_j$ and mean

$$\mu = \frac{\alpha_0}{1 - \sum_{i=1}^{p} \alpha_i - \sum_{j=1}^{q} \beta_i}.$$

Remark. As before, ARMA theory may now be used to calculate (for example) the true ACF of the squared series X_t^2. This has been done in Figure 2.3, which shows a realization of a GARCH(1,1) process with finite fourth moment whose squared values follow an ARMA(1,1) process.

2.3 Fitting Univariate GARCH Models

A very versatile and most widely used approach to fitting GARCH models to data is by maximum likelihood, which will be explained in detail in this section. The topics will be the building of the (log-)likelihood, the finding of parameter estimates and the asymptotic behavior of the estimates. Some remarks about model checking will be added.

2.3.1 Building the likelihood

In order to avoid unnecessary confusion, attention is restricted to the cases of ARCH(1) and GARCH(1,1) models. The fitting of higher-order ARCH(p) and GARCH(p, q) models easily follows.

Starting from $n + 1$ data values $X_0, X_1, \ldots, X_n$, the likelihood function L can be written as the joint density of these observations, i.e.

$$L = f_{X_0, X_1, \ldots, X_n}(x_0, x_1, \ldots, x_n),$$

where f denotes the density function. Recalling that the joint density may be rewritten in terms of the product of conditional densities, the likelihood may equivalently be expressed as

$$L = f_{X_0}(x_0) \prod_{t=1}^{n} f_{X_t|X_{t-1},\ldots,X_0}(x_t|x_{t-1}, \ldots, x_0). \tag{2.17}$$

However, the marginal density $f_{X_0}(x_0)$ is not known in a tractable closed form for ARCH and GARCH models and this poses a problem for basing a likelihood on (2.17). The workaround for this problem in pactice is to use the *conditional likelihood* conditioned on X_0 instead. Doing so yields

$$\begin{aligned} L_c &= f_{X_1,\ldots,X_n|X_0}(x_1, \ldots, x_n|x_0) \\ &= \prod_{t=1}^{n} f_{X_t|X_{t-1},\ldots,X_0}(x_t|x_{t-1}, \ldots, x_0). \end{aligned}$$

In the case of the ARCH(1) model, the conditional densities simplify to

$$f_{X_t|X_{t-1},\ldots,X_0}(x_t|x_{t-1},\ldots,x_0) = f_{X_t|X_{t-1}}(x_t|x_{t-1}),$$

since the ARCH(1) process is first order *Markovian*. The conditional distribution of X_t may easily be calculated to be

$$\mathbb{P}[X_t \leq x_t|X_{t-1}] \;=\; \mathbb{P}\left[Z_t \leq \frac{x_t}{\sigma_t}|X_{t-1}\right] = G_Z\left(\frac{x_t}{\sigma_t}\right),$$

for $t = 1,\ldots,n$. $G_Z(\cdot)$ stands for the distribution function of the innovations $(Z_t)_{t\in\mathbb{Z}}$ and $\sigma_t = \sqrt{\alpha_0 + \alpha_1 X_{t-1}^2}$. Differentiating yields the conditional density for $X_t|X_{t-1}$,

$$f_{X_t|X_{t-1}}(x_t|x_{t-1}) = \frac{1}{\sigma_t} g_Z\left(\frac{x_t}{\sigma_t}\right),$$

with $g_Z(\cdot)$ denoting the density of the innovations. It follows that the conditional likelihood takes the tractable form

$$L_c(\alpha_0,\alpha_1;\boldsymbol{X}) = \prod_{t=1}^{n} \frac{1}{\sigma_t} g_Z\left(\frac{X_t}{\sigma_t}\right),$$

where $\boldsymbol{X} = (X_1,\ldots,X_n)'$. An analogous method works for ARCH(p) processes by conditioning on the first p values.

In the GARCH(1,1) model σ_t does not only depend on X_{t-1}, but also on σ_{t-1}. Therefore, the joint density of $X_1,\ldots,X_n$ conditional on realized values of both X_0 and σ_0 is constructed:

$$f_{X_1,\ldots,X_n|X_0,\sigma_0}(x_1,\ldots,x_n|x_0,\sigma_0) = \prod_{t=1}^{n} f_{X_t|X_{t-1},\ldots,X_0,\sigma_0}(x_t|x_{t-1},\ldots,x_0,\sigma_0).$$

As above, this term can be calculated using the density of the innovations by

$$f_{X_t|X_{t-1},\ldots,X_0,\sigma_0}(x_t|x_{t-1},\ldots,x_0,\sigma_0) = \frac{1}{\sigma_t} g_Z\left(\frac{x_t}{\sigma_t}\right),$$

with σ_t explained recursively by $\sigma_t = \sqrt{\alpha_0 + \alpha_1 X_{t-1}^2 + \beta_1 \sigma_{t-1}^2}$. Considering these

reasonings, the conditional likelihood can now be defined to be

$$L_c(\alpha_0, \alpha_1, \beta_1; \boldsymbol{X}) = \prod_{t=1}^{n} \frac{1}{\sigma_t} g_Z\left(\frac{X_t}{\sigma_t}\right).$$

The problem remains that the value of σ_0^2 is not actually observed. This is usually solved rather pragmatically by imputing a starting value such as the sample variance of $\boldsymbol{X}$ or even simply zero.[14]

In the GARCH(p, q) case a larger sample size is needed. Assuming that $n + p$ data values $X_{-p+1}, \ldots, X_0, \ldots, X_n$ are given, the conditional likelihood is built conditional on the observed values $X_{-p+1}, \ldots, X_0$ and the unobserved variables $\sigma_{-q+1}, \ldots, \sigma_0$. Again, for these unobserved values $\sigma_{-q+1}, \ldots, \sigma_0$ starting values are imputed.

2.3.2 Finding parameter estimates

Appropriate parameter estimates are obtained by maximizing the conditional likelihood function. By doing so it is equivalent to maximize its logarithm (being a monotone function), which is easier to cope with (Tsay [26]). Hence, denoting the set of parameters by $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)'$, the conditional log-likelihood is given by

$$\log(L_c(\boldsymbol{\theta}; \boldsymbol{X})) = \sum_{t=1}^{n} l_t(\boldsymbol{\theta}),$$

where $l_t(\boldsymbol{\theta}) = \log(f_{X_t|X_{t-1},\ldots,X_0}(x_t|x_{t-1}, \ldots, x_0))$.

In order to find estimates for the set of parameters $\boldsymbol{\theta}$, the negative log-likelihood is minimized with respect to $\boldsymbol{\theta}$. Alternatively, the root of equations $\frac{\partial}{\partial \boldsymbol{\theta}} \sum_{t=1}^{n} l_t(\boldsymbol{\theta}) = 0$ is found. These are the so-called *score equations* and are typically solved using numerical optimization procedures, such as (modified) Newton-Raphson-type methods. Concretely, negative log-likelihood can be minimized using the BHHH[15]

[14] According to McNeil et al. [20], the issue of conditioning on starting values is of relatively minor importance when datasets are large. The influence of these initial values on the final parameter estimates is small, which can be verified by experimentation.

[15] BHHH is an acronym for Berndt, Hall, Hall and Hausman [4].

algorithm as Bollerslev [5] proposed when introducing GARCH(p, q) models.

Example 2.2 (Normal innovations)**.** Assuming that the sequence of innovations $(Z_t)_{t\in\mathbb{Z}}$ is a Gaussian strict white noise with variance one, g_Z is given by

$$g_Z(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right), \tag{2.18}$$

which yields by elementary calculus that the negative conditional log likelihood function turns out to be

$$\begin{aligned} -\log L_c &= \sum_{t=1}^{n} \left(\log(\sqrt{2\pi}) + \log(\sigma_t) + \frac{1}{2}\frac{X_t^2}{\sigma_t^2}\right) \\ &= \frac{n}{2}\log(2\pi) + \frac{1}{2}\sum_{t=1}^{n}\log(\sigma_t^2) + \frac{1}{2}\sum_{t=1}^{n}\frac{X_t^2}{\sigma_t^2}. \end{aligned} \tag{2.19}$$

Example 2.3 (Student t innovations)**.** Obviously, the same basic approach can be used with non-Gaussian distributions, for example a standardized Student-t distribution. Let Y be Student-t distributed with $\nu > 2$ degrees of freedom. Then $\mathbb{V}(Y) = \frac{\nu}{\nu-2}$ for $\nu > 2$, and the innovations are chosen to be $Z_t = \sqrt{\frac{\nu}{\nu-2}}Y$. The probability density function of Z_t is then given by

$$g_Z(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{(\nu-2)\pi}}\left(1 + \frac{x^2}{\nu-2}\right)^{-\frac{\nu+1}{2}}, \tag{2.20}$$

where $\nu > 2$ and $\Gamma(a)$ is the Gamma function, i.e., $\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}dx$. Using the above density, the negative conditional log likelihood function of X_t can analogously be calculated to be

$$\begin{aligned} -\log(L_c) &= n\log\left(\frac{\Gamma\left(\frac{\nu}{2}\right)\sqrt{(\nu-2)\pi}}{\Gamma\left(\frac{\nu+1}{2}\right)}\right) \\ &+ \frac{1}{2}\sum_{t=1}^{n}\log(\sigma_t^2) \\ &+ \frac{\nu+1}{2}\sum_{t=1}^{n}\log\left(1 + \frac{X_t^2}{(\nu-2)\sigma_t^2}\right). \end{aligned}$$

Remark. Following the above ideas, other admissible distributions for Z_t can

easily be employed. Among these, which have been used with ARCH-related models belong to the Normal-Poisson mixture distribution, the power exponential distribution and the generalized exponential distribution (Hamilton [16]).

2.3.3 Quasi-maximum likelihood estimation

What if the fitted model is misspecified with respect to the distribution of the innovations? In other words, what if the model has been fitted to e.g. Gaussian innovations, but although the dynamic form of the model is correct the choice of innovations is wrong? Under this misspecification the model fitting procedure is known as *quasi-maximum* or *pseudo-maximum likelihood estimation* (QML). Essentially the negative Gaussian conditional log-likelihood specified by (2.19) is treated as an objective function to be minimized rather than a proper likelihood. Intuitively, this may still give reasonable parameter estimates - which turns out to be the case under appropriate assumptions about the true innovation distribution (McNeil et al. [20]).

Let $\widehat{\boldsymbol{\theta}}_n$ be the estimate that minimizes the negative Gaussian conditional log-likelihood (2.19) and $\boldsymbol{\theta}$ be the true value. Then even when Z_t is non-Gaussian and $\mathbb{E}[Z_t^4] < \infty$ it follows that

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} N_{p+q+1}(\mathbf{0}, I(\boldsymbol{\theta})^{-1} J(\boldsymbol{\theta}) I(\boldsymbol{\theta})^{-1}), \tag{2.21}$$

where

$$I(\boldsymbol{\theta}) = -\mathbb{E}\left[\frac{\partial^2 l_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right],$$

often said to have *outer product* form, and

$$J(\boldsymbol{\theta}) = \mathbb{E}\left[\frac{\partial l_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial l_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right],$$

referred to as *Hessian* form. Note that the expectation in $I(\boldsymbol{\theta})$ and $J(\boldsymbol{\theta})$ is taken with respect to the true model (not the misspecified Gaussian model). In general $I(\boldsymbol{\theta}) \neq J(\boldsymbol{\theta})$ unless the true model really has Gaussian innovations, i.e. the model is not misspecified.

The asymptotic covariance matrix appearing in equation (2.21) is estimated by $\overline{I}(\widehat{\boldsymbol{\theta}})^{-1}\overline{J}(\widehat{\boldsymbol{\theta}})\overline{I}(\widehat{\boldsymbol{\theta}})^{-1}$, where I is approximated by the observed information matrix

$$\overline{I}(\boldsymbol{\theta}) = -\frac{1}{n}\sum_{t=1}^{n}\frac{\partial^2 l_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'},$$

and J is approximated by

$$\overline{J}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{t=1}^{n}\frac{\partial l_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial l_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}.$$

2.3.4 Model checking

For a (correctly specified) GARCH(p,q) model, the standardized shocks or residuals z_t defined by

$$\begin{aligned} z_t &= \frac{X_t}{\widehat{\sigma}_t}, \\ \widehat{\sigma}_t^2 &= \widehat{\alpha}_0 + \sum_{i=1}^{p}\widehat{\alpha}_i X_{t-i}^2 + \sum_{j=1}^{q}\beta_j\widehat{\sigma}_{t-j}^2, \end{aligned} \tag{2.22}$$

should behave like strict white noise. This can be graphically investigated with correlograms of raw and absolute residual values (and other transformations of the raw values such as the logarithm or polynomials). Additionally, formal tests such as portmanteau SWN tests can be applied to raw and transformed values. However, due to the construction of (2.22), some initial values are needed. Usually, the starting values of X_t are set to zero and those of $\widehat{\sigma}_t$ are either set to be the sample variance or also zero. Since the first few values of z_t will strongly be influenced by these imputed starting values, they might be ignored in later analysis.

Assuming that the SWN hypothesis is sustainable (i.e. the dynamics have been captured satisfactorily), the validity of the distribution used in the ML fitting can also be investigated using QQ-plots and goodness-of-fit tests for the choice of the innovation distribution. If residuals do not behave like standard normal observations, other choices might be considered. Alternatively, if the Gaussian likelihood does a reasonable job of estimating dynamics, the QML philosophy can

be adopted and asymptotic standard errors can be estimated using the ideas of Section 2.3.3.

Chapter 3

Multivariate Models for Changing Volatility

In this chapter multivariate GARCH (often referred to as MGARCH) models are introduced. Analogously to univariate GARCH models, they are of particular use in finance, especially when modeling daily risk factor return series, because they model directly the evolution of volatilities. Several different models (all commonly summarized as MGARCH models) will be discussed, all of which can to a certain extent be viewed as generalizations of the univariate models discussed in Chapter 2.

3.1 General Structure of MGARCH Models

3.1.1 Definition

Definition 3.1. Let $(\boldsymbol{Z}_t)_{t\in\mathbb{Z}}$ be SWN$(\mathbf{0}, I_d)$. The process $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$ is said to be a multivariate GARCH process if it is strictly stationary and satisfies equations of the form

$$\boldsymbol{X}_t = \Sigma_t^{1/2}\boldsymbol{Z}_t, \qquad t \in \mathbb{Z}, \tag{3.1}$$

where $\Sigma_t^{1/2} \in \mathbb{R}^{d\times d}$ is the Cholesky factor of a positive-definite matrix Σ_t that is measurable with respect to $\mathcal{F}_{t-1} = \sigma\{(\boldsymbol{X}_s)_{s\leq t-1}\}$, the history of the process up to time $t-1$. Each $\boldsymbol{Z}_t$ shall again be called an *innovation* and interpreted as such.

Remark. Using different types of "square roots" for Σ_t other than the Cholesky factor (such as the root derived from symmetric decomposition) is possible and effects the construction of residuals when fitting the model in practice.

3.1.2 Conditional moments

It is easily calculated that a process of type (3.1) has the multivariate martingale difference property

$$\mathbb{E}[\boldsymbol{X}_t|\mathcal{F}_{t-1}] = \mathbb{E}[\Sigma_t^{1/2}\boldsymbol{Z}_t|\mathcal{F}_{t-1}] = \Sigma_t^{1/2}\mathbb{E}[\boldsymbol{Z}_t] = \boldsymbol{0}, \tag{3.2}$$

from which again it follows that the unconditional expectation $\mathbb{E}[\boldsymbol{X}_t]$ is also zero. Assuming covariance stationarity, this implies that the process is multivariate white noise. Because of

$$\operatorname{cov}(\boldsymbol{X}_t|\mathcal{F}_{t-1}) \overset{(3.2)}{=} \mathbb{E}[\boldsymbol{X}_t\boldsymbol{X}_t'|\mathcal{F}_{t-1}] = \Sigma_t^{1/2}\mathbb{E}[\boldsymbol{Z}_t\boldsymbol{Z}_t'](\Sigma_t^{1/2})' = \Sigma_t^{1/2}I_d(\Sigma_t^{1/2})' = \Sigma_t, \tag{3.3}$$

Σ_t is the *conditional covariance matrix*. Again it is possible to decompose Σ_t involving a diagonal matrix,

$$\Sigma_t = \Delta_t P_t \Delta_t, \qquad \Delta_t = \operatorname{diag}(\sigma_{t,1}, \ldots, \sigma_{t,d}), \tag{3.4}$$

where the Δ_t is known as the *volatility matrix* and contains the volatilities for the component series $(\boldsymbol{X}_{t,k})_{t\in\mathbb{Z}}$ for $k = 1, \ldots, d$ and P_t (symmetric with unit diagonal elements) is known as the *conditional correlation matrix*. When building multivariate GARCH models, the dependence of Σ_t (or equivalently of Δ_t and P_t) on the past is to be specified in a way that Σ_t always remains symmetric and positive-definite.[1]

[1] A covariance matrix must of course be symmetric and positive semidefinite and in practice we restrict our attention to the positive definite case (which facilitates fitting since the conditional distribution of $\boldsymbol{X}_t|\mathcal{F}_{t-1}$ never has a singular covariance matrix). Compare McNeil et al. [20].

3.1.3 Unconditional moments

Of course it is also possible to analyze unconditional expectation and covariance, as shall be done in the following Theorem.

Theorem 3.1 (Unconditional moments)**.** *The unconditional covariance matrix* Σ *of a process of type (3.1) is given by* $\mathbb{E}[\Sigma_t]$ *and the unconditional correlation matrix* P *has* (i,j)*th element*

$$P_{ij} = \frac{\mathbb{E}[P_{t,ij}\sigma_{t,i}\sigma_{t,j}]}{\sqrt{\mathbb{E}[\sigma_{t,i}^2]\mathbb{E}[\sigma_{t,j}^2]}}.$$

Proof. Σ may again easily be calculated using basic rules for conditional expectation and previous results:

$$\Sigma = \operatorname{cov}(\boldsymbol{X}_t) \overset{(3.2)}{=} \mathbb{E}[\boldsymbol{X}_t\boldsymbol{X}_t'] = \mathbb{E}[\mathbb{E}[\boldsymbol{X}_t\boldsymbol{X}_t'|\mathcal{F}_{t-1}]] \overset{(3.3)}{=} \mathbb{E}[\Sigma_t].$$

It also holds that

$$P_{ij} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}} = \frac{\mathbb{E}[\Sigma_{t,ij}]}{\sqrt{\mathbb{E}[\Sigma_{t,ii}]\mathbb{E}[\Sigma_{t,jj}]}} = \frac{\mathbb{E}[\sigma_{t,i}\sigma_{t,j}P_{t,ij}]}{\sqrt{\mathbb{E}[\sigma_{t,i}^2]\mathbb{E}[\sigma_{t,j}^2]}},$$

since by definition

$$\Sigma_t = \begin{pmatrix} \sigma_{t,1}^2 & \sigma_{t,1}\sigma_{t,2}P_{t,12} & \dots & \sigma_{t,1}\sigma_{t,d}P_{t,1d} \\ \sigma_{t,1}\sigma_{t,2}P_{t,12} & \sigma_{t,2}^2 & \dots & \sigma_{t,2}\sigma_{t,d}P_{t,2d} \\ \vdots & \vdots & & \vdots \\ \sigma_{t,1}\sigma_{t,d}P_{t,1d} & \sigma_{t,2}\sigma_{t,d}P_{t,2d} & \dots & \sigma_{t,d}^2 \end{pmatrix}. \qquad \square$$

Remark. Clearly P is in general not simply the expectation of the conditional correlation matrix P_t.

3.1.4 Innovations

In practical work the innovations $\boldsymbol{Z}_t$ are generally taken to be either from a multivariate Gaussian distribution $\boldsymbol{Z}_t \sim N_d(\boldsymbol{0}, I_d)$ or, more realistically for daily returns

because of the heavier tails, a spherical multivariate t distribution appropriately scaled to have covariance matrix I_d, i.e. $\boldsymbol{Z}_t \sim t_d(\nu, \mathbf{0}, \frac{\nu-2}{\nu} I_d)$. Of course any other distribution with mean zero and covariance matrix I_d is permissible as well.

3.2 Models for Conditional Correlation

In this section emphasis will be placed on models which focus on specifying the conditional correlation matrix P_t while allowing volatilities to be described by univariate GARCH models. To begin with, a popular and rather parsimonious model is described, where P_t is assumed to be constant for all t.

3.2.1 Constant conditional correlation (CCC)

Definition and basic properties

Definition 3.2. The process $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$ is a CCC-GARCH process if it is a process with the general structure given in Definition 3.1 such that the conditional covariance matrix is of the form $\Sigma_t = \Delta_t P_c \Delta_t$ for a constant and positive definite correlation matrix P_c and the components $\sigma_{t,k}$ of the diagonal volatility matrix Δ_t satisfy

$$\sigma_{t,k}^2 = \alpha_{k,0} + \sum_{i=1}^{p_k} \alpha_{k,i} X_{t-i,k}^2 + \sum_{j=1}^{q_k} \beta_{k,j} \sigma_{t-j,k}^2, \qquad k \in \{1, \dots, d\}, \tag{3.5}$$

where

- $\alpha_{k,0} > 0$,
- $\alpha_{k,i} \geq 0$ for $i = 1, \dots, p_k$,
- $\beta_{k,j} \geq 0$ for $j = 1, \dots, q_k$.

Remark. The CCC-GARCH specification represents a simple way of combining univariate GARCH processes. This may be seen by observing that in a

CCC-GARCH model observations and innovations are connected by equations $\boldsymbol{X}_t = \Delta_t P_c^{1/2} \boldsymbol{Z}_t$, which may be rewritten as $\boldsymbol{X}_t = \Delta_t \boldsymbol{Y}_t$ for a SWN$(\mathbf{0}, P_c)$ process $(\boldsymbol{Y}_t)_{t\in\mathbb{Z}}$. Clearly the component processes are univariate GARCH.[2]

Theorem 3.2. *The CCC-GARCH model is well defined in the sense that Σ_t is almost surely positive definite for all t. Moreover it is covariance stationary if and only if*

$$\sum_{i=1}^{p_k} \alpha_{k,i} + \sum_{j=1}^{q_k} \beta_{k,j} < 1$$

holds for each $k \in \{1, \dots, d\}$.

Proof. For a vector $\mathbf{v} \neq \mathbf{0}$ in $\mathbb{R}^d$ we have

$$\mathbf{v}'\Sigma_t\mathbf{v} = (\Delta_t\mathbf{v})' P_c (\Delta_t\mathbf{v}) > 0,$$

since P_c is positive definite and the strict positivity of the individual volatility processes ensures that $\Delta_t\mathbf{v} \neq \mathbf{0}$ for all t.

If $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$ is covariance stationary then each component series $(X_{t,k})_{t\in\mathbb{Z}}$ is a covariance stationary univariate GARCH process for which a necessary and sufficient condition is $\sum_{i=1}^{p_k} \alpha_{k,i} + \sum_{j=1}^{q_k} \beta_{k,j} < 1$ by Theorem 2.7. Conversely, if the component series are covariance stationary, then

$$\operatorname{cov}(\boldsymbol{X}_t)_{ij} = \mathbb{E}[\Sigma_{t,ij}] = P_{ij}\mathbb{E}[\sigma_{t,i}\sigma_{t,j}] < \infty, \qquad 1 \leq i, j \leq d,$$

and $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$ is a multivariate martingale difference with finite, non-time-dependent second moments; in other words a covariance stationary white noise. □

The CCC model is often a useful starting point before going on to fitting more complex models. Applied to financial data, it gives an adequate performance in some empirical settings, but it is generally considered that the constancy of conditional correlations in this model is an unrealistic feature and that the impact

[2] A univariate GARCH process may now be seen as a special case of a MGARCH process as described in Definition 3.1 with dimension d equaling one and $\Sigma_t^{1/2} = \sigma_t = \sigma_{t,1} \in \mathbb{R}$ satisfying equation (3.5).

of news on financial markets requires models that allow a dynamic evolution of conditional correlation as well as a dynamic evolution of volatilities. A further criticism of the model (which applies in fact to the majority of MGARCH specifications) is the fact that the individual volatility dynamics (3.5) do not allow for the possibility that large (asset) returns in one component series at a particular time point can contribute to increased volatility of another component time series at future time points.[3]

The de-volatized process

It turns out to be a fruitful approach (both for understanding and fitting the CCC model) to introduce the notion of a *de-volatized* process.

Definition 3.3. For any multivariate time series process $\boldsymbol{X}_t$ the de-volatized process $\boldsymbol{Y}_t$ is defined to be

$$\boldsymbol{Y}_t = \Delta_t^{-1} \boldsymbol{X}_t,$$

where Δ_t is, as usual, the diagonal matrix of volatilities of the component time series as in equation (3.4).

Remark. In the case of a CCC model it is easily seen that the de-volatized process $(\boldsymbol{Y}_t)_{t\in\mathbb{Z}}$ is a SWN$(\mathbf{0}, P_c)$ process.

Stepwise estimation

The above structure suggests a three-stage fitting method when dealing with empirical data. The process of de-volatizing, in this case steps one and two, is sometimes also referred to as *pre-whitening*.

1. The individual volatility processes for the component series are fixed by fitting univariate GARCH processes or, more generally, any univariate model for changing volatility.[4]

[3] Compare also Zivot and Wang [28] for further practical aspects and criticism.

[4] Well-known examples for extension of GARCH models are GARCH with leverage, threshold GARCH or exponential GARCH.

2. The de-volatized process is estimated by $\widehat{\boldsymbol{Y}}_t = \widehat{\Delta}_t^{-1}\boldsymbol{X}_t$, where $\widehat{\Delta}_t^{-1}$ is the estimate of Δ_t^{-1} found as explained above.

If the CCC-GARCH model assumption is adequate, then the $\widehat{\boldsymbol{Y}}_t$ data should behave like a realization form a SWN$(\mathbf{0}, P_c)$ process. This can be checked empirically by investigating the correlogram and cross-correlogram applied to raw and absolute values (and other standard tools used to investigate independence of random variables).

3. Assuming the adequacy of the model, the conditional correlation matrix P_c can finally be estimated from $\widehat{\boldsymbol{Y}}_t$ by the sample correlation matrix $\widehat{P}_c$ as noted in Definition 1.8 or other adequate correlation matrix estimates.

A special case of CCC-GARCH which shall be called a *pure diagonal model* occurs when $P_c = I_d$. A covariance stationary model of this kind is clearly multivariate white noise where the contemporaneous components $X_{t,i}$ and $X_{t,j}$ are also uncorrelated for $i \neq j$. Whether they are also independent or not depends on further assumption about the driving SWN$(\mathbf{0}, I_d)$ process: if the innovations have independent components (e.g. they are multivariate Gaussian), then the component series are independent; in other cases (e.g. $\boldsymbol{Z}_t \sim t_d(\nu, \mathbf{0}, \frac{\nu-2}{\nu} I_d)$) the component processes are dependent.

3.2.2 Dynamic conditional correlation (DCC)

Definition and basic properties

This model generalizes the CCC model to allow conditional correlations to evolve dynamically according to a relatively parsimonious scheme (which will make it practically useful). It can again be seen as a combination of univariate GARCH models.

Definition 3.4. The process $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$ is a DCC-GARCH process if it is a process with the general structure given in Definition 3.1 where the volatilities comprising

$\Delta_t = \text{diag}(\sigma_{t,1}, \sigma_{t,2}, \ldots, \sigma_{t,d})$ follow univariate GARCH specifications as in (3.5), i.e.

$$\sigma_{t,k}^2 = \alpha_{k,0} + \sum_{i=1}^{p_k} \alpha_{k,i} X_{t-i,k}^2 + \sum_{j=1}^{q_k} \beta_{k,j} \sigma_{t-j,k}^2, \qquad k \in \{1, \ldots, d\},$$

with the above restrictions on $\alpha_{k,i}$ and $\beta_{k,j}$. The conditional correlation matrices P_t satisfy the equations

$$Q_t = c\overline{Q} + \sum_{i=1}^{p} \alpha_i \boldsymbol{Y}_{t-i} \boldsymbol{Y}'_{t-i} + \sum_{j=1}^{q} \beta_j Q_{t-j}, \tag{3.6}$$

$$P_t = (\Delta_t^{(Q)})^{-1} Q_t (\Delta_t^{(Q)})^{-1}, \qquad t \in \mathbb{Z}, \tag{3.7}$$

where

- $\overline{Q}$ is a positive-definite covariance matrix,
- $\Delta_t^{(Q)}$ is a diagonal matrix containing the square root of the diagonal entries of Q_t, i.e. $\Delta_t^{(Q)} = \text{diag}(\sqrt{Q_{t,11}}, \sqrt{Q_{t,22}}, \ldots, \sqrt{Q_{t,dd}})$,
- $\boldsymbol{Y}_t$ denotes the de-volatized process of $\boldsymbol{X}_t$, i.e. $\boldsymbol{Y}_t = \Delta_t^{-1} \boldsymbol{X}_t$,
- $c = 1 - \sum_{i=1}^{p} \alpha_i - \sum_{j=1}^{q} \beta_j$,
- the coefficients satisfy $\alpha_i \geq 0$, $\beta_j \geq 0$, $c > 0$.

Remark. The process $(Q_t)_{t \in \mathbb{Z}}$ as defined in (3.6) is a process of matrices which are not themselves correlation matrices (although as explained below the model can be parameterized in such a way that they are correlation matrices in expectation). Hence we need to use the operation (3.7) to obtain correlation matrices. In order to see that these dynamic equations preserve the positive definiteness of P_t it suffices to check that (3.6) preserves the positive definiteness of Q_t. If $Q_{t-q}, Q_{t-q+1}, \ldots, Q_{t-1}$ are positive definite, then, for a vector $\mathbf{v} \neq \mathbf{0}$ in $\mathbb{R}^d$, we have

$$\mathbf{v}' Q_t \mathbf{v} = c\mathbf{v}' \overline{Q} \mathbf{v} + \sum_{i=1}^{p} \alpha_i \mathbf{v}' \boldsymbol{Y}_{t-i} \boldsymbol{Y}'_{t-i} \mathbf{v} + \sum_{j=1}^{q} \beta_j \mathbf{v}' Q_{t-j} \mathbf{v} > 0, \tag{3.8}$$

since the first term is strictly positive by definition and the second and third terms are non-negative.

The de-volatized process of DCC

Again the idea of de-volatization turns out to be a useful tool for analyzing the model. As can be seen in the definition, the de-volatized process $(\boldsymbol{Y}_t)_{t\in\mathbb{Z}}$ is the driving force of the dynamics of correlation.

Theorem 3.3 (McNeil et al. [20])**.** *For any covariance stationary multivariate GARCH process the de-volatized process $(\boldsymbol{Y}_t)_{t\in\mathbb{Z}}$ as introduced in Definition 3.4 is a zero-mean white noise with covariance and correlation matrix $\mathbb{E}[P_t]$.*

Proof. By observing that $\boldsymbol{Y}_t = \Delta_t^{-1}\boldsymbol{X}_t = P_t^{1/2}\boldsymbol{Z}_t$ and conditioning on $\mathcal{F}_{t-1}$, it can be seen that the de-volatized process has conditional mean zero (i.e. the martingale difference property) and conditional covariance matrix

$$\operatorname{cov}(\boldsymbol{Y}_t|\mathcal{F}_{t-1}) = \mathbb{E}[\boldsymbol{Y}_t\boldsymbol{Y}_t'|\mathcal{F}_{t-1}] = P_t^{1/2}\mathbb{E}[\boldsymbol{Z}_t\boldsymbol{Z}_t'](P_t^{1/2})' = P_t,$$

resulting in $\operatorname{cov}(\boldsymbol{Y}_t) = \rho(\boldsymbol{Y}_t) = \mathbb{E}[P_t]$. A martingale difference with a finite, non-time-dependent second moment forms a white noise process, what proves the theorem. □

If now the further condition that $\overline{Q} = \operatorname{cov}(\boldsymbol{Y}_t) = \mathbb{E}[P_t]$ is imposed in (3.6), meaning that $\overline{Q}$ should equal the unconditional covariance/correlation matrix of the de-volatized process, and attention is restricted to the covariance stationary case, it holds true that $\mathbb{E}[Q_t] = \mathbb{E}[P_t] = \overline{Q}$. This can be verified by direct calculation: Taking expectation on both sides of (3.6) yields

$$\mathbb{E}[Q_t] = \left(1 - \sum_{i=1}^{p}\alpha_i - \sum_{j=1}^{q}\beta_j\right)\overline{Q} + \sum_{i=1}^{p}\alpha_i\overline{Q} + \sum_{j=1}^{q}\beta_j\mathbb{E}[Q_t],$$

what can easily be simplified to give the desired result. Under this condition, the correlation matrix $\overline{Q}$ of the de-volatized process is the expectation of both Q_t and P_t.

Stepwise estimation

Similar to the CCC model, the estimation of a DCC-GARCH model suggested by these calculations might be conducted in the following steps:

1. Fit univariate GARCH-type models to the component series to estimate the volatility matrix Δ_t.
2. Form an estimated realization of the de-volatized process by calculating $\widehat{\boldsymbol{Y}}_t = \widehat{\Delta}_t^{-1}\boldsymbol{X}_t$.
3. Estimate $\overline{Q}$ by taking the sample correlation matrix of the de-volatized data (or again some other (robust) estimator of correlation).
4. Estimate the remaining parameters α_i and β_j in equation (3.6) by fitting a model with structure $\boldsymbol{Y}_t = P_t^{1/2}\boldsymbol{Z}_t$ to the de-volatized data. This can be achieved as a special case of the methodology for fitting general multivariate GARCH models, explained in Section 3.5. Note that in a first order model ($p = q = 1$) there will only be two remaining parameters to estimate.

3.3 Models for Conditional Covariance

Models presented in this section specify explicitly a dynamic structure for the conditional covariance matrix Σ_t in (3.1). In contrast to the previous section, these models are not designed for multiple-stage estimation based on univariate GARCH estimation procedures.

3.3.1 Some preliminaries

Before turning to the models themselves, a simple but useful tool for restructuring symmetric matrices is introduced. It will later simplify the notation of the VEC model.

Definition 3.5 (vech operator)**.** Let $\text{sym}(\mathbb{R}^{d\times d})$ be the space of real-valued symmetric matrices with dimension d and $A \in \text{sym}(\mathbb{R}^{d\times d})$ with elements a_{ij}. The *vector half operator* (vech) is a map

$$\begin{aligned} \text{sym}(\mathbb{R}^{d\times d}) &\longrightarrow \mathbb{R}^{\frac{d(d+1)}{2}} \\ A &\longmapsto \text{vech}(A) \end{aligned}$$

defined by

$$\text{vech}(A) = (a_{11}, a_{21}, \ldots, a_{d1}, a_{22}, \ldots, a_{d2}, a_{33}, \ldots, a_{dd})'.$$

In other words, vech stacks the columns of the lower triangle of A in a single column vector of length $d(d+1)/2$.

Remark. Due to the symmetry of A it is clear that $\text{vech}(A)$ contains the same information as A itself, rearranging the important components and leaving out unnecessary ones.

3.3.2 The general vector GARCH model (VEC)

A very general vector GARCH model - the VEC model - is extremely rich in parameters and therefore of limited practical use. Nevertheless, it provides a general framework for possible various restrictions on parameter matrices (e.g. diagonal VEC model). Alongside the BEKK model of Section 3.3.4, it is discussed in Engle and Kroner [13] with full mathematical rigour.

Definition 3.6 (VEC Model)**.** The process $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$ is a VEC process if it has the general structure given in Definition 3.1 and the dynamic of the conditional covariance matrix Σ_t is given by

$$\text{vech}(\Sigma_t) = \boldsymbol{a}_0 + \sum_{i=1}^{p} \overline{A}_i \text{vech}(\boldsymbol{X}_{t-i}\boldsymbol{X}'_{t-i}) + \sum_{j=1}^{q} \overline{B}_j \text{vech}(\Sigma_{t-j}), \tag{3.9}$$

for a vector $\boldsymbol{a}_0 \in \mathbb{R}^{\frac{d(d+1)}{2}}$ and matrices $\overline{A}_i, \overline{B}_i \in \mathbb{R}^{\frac{d(d+1)}{2}\times\frac{d(d+1)}{2}}$.

Remarks. 1. By using the vector half operator, equation (3.9) should be understood as specifying the dynamics for the lower-triangular portion of the conditional covariance matrix; the remaining elements of the matrix are determined by symmetry.

2. In (3.9), Σ_t is not automatically positive definite and can therefore not be interpreted as a covariance matrix. For the model to be well defined in that sense, further restrictions have to be imposed onto the parameters. A sufficient condition may easily be found by observing that Σ_t is positive definite if

   - $\text{vech}^{-1}(\boldsymbol{a}_0)$ is positive definite,
   - $\text{vech}^{-1}(\overline{A}_i \text{vech}(mm'))$ and $\text{vech}^{-1}(\overline{B}_j \text{vech}(M))$ are positive semidefinite for all applicable i and j, $m \in \mathbb{R}^d$ and M positive semidefinite.

   However, this condition is practically unverifiable.

3. In this very general form the model has $(p+q)c^2 + c$ parameters, where $c = d(d+1)/2$. It is not very hard to predict that this amounts to enormous numbers of parameters for higher dimensions (even for $p = q = 1$). Please consult Section 3.4 with Table 3.1 and Figure 3.1 for a small overview.

4. Very similar to the univariate case, conditions have been found to ensure the covariance stationarity of the process. These may be formulated in terms of restrictions onto the parameters matrices by

$$\max |\text{eig}(\sum_{i=1}^{p} \overline{A}_i + \sum_{j=1}^{q} \overline{B}_j)| < 1, \tag{3.10}$$

   where $\text{eig}(A)$ denotes the eigenvalues of A. For details and proof please consult e.g. Proposition 2.7 in Engle and Kroner [13].

3.3.3 The diagonal vector GARCH model (DVEC)

Definition and basic properties

One very common simplification of the VEC model has been to restrict attention to cases when $\overline{A}_i$ and $\overline{B}_i$ in Definition 3.6 are diagonal matrices, which is then called a *diagonal VEC* or *DVEC* model. This special case can be written very elegantly in terms of element-wise matrix multiplication.[5] The following representation is obtained, which is equivalent to a well defined general VEC model as in (3.9) for diagonal $\overline{A}_i$ and $\overline{B}_i$.

$$\Sigma_t = A_0 + \sum_{i=1}^{p} A_i \odot (\boldsymbol{X}_{t-i}\boldsymbol{X}'_{t-i}) + \sum_{j=1}^{q} B_j \odot \Sigma_{t-j}, \tag{3.11}$$

where

- $A_0 \in \mathbb{R}^{d\times d}$ is symmetric and positive definite,
- $A_i, B_j \in \mathbb{R}^{d\times d}$ are symmetric and positive semidefinite for $i = 1, \ldots, p$ and $j = 1, \ldots, q$.

Remark. This representation emphasizes beautifully the similarities to the univariate GARCH equation (2.12) for σ_t in terms of both restrictions onto the parameters as well as general structure.

Due to the fact that the parameter matrices A_i, B_i are chosen to be positive (semi)definite, Σ_t is automatically positive definite.

Theorem 3.4. *In the DVEC model (3.11) the conditional covariance matrix Σ_t is positive definite for all t.*

Proof. The above statement can be seen by observing that for $\mathbf{v} \neq \mathbf{0}$ in $\mathbb{R}^d$ it holds

[5] The element-by-element matrix multiplication is often referred to as the *Hadamard product*, defined for two matrices of the same size, and will be denoted $\odot$.

true that

$$\begin{aligned}\mathbf{v}'Q_t\mathbf{v} &= \mathbf{v}'A_0\mathbf{v} + \sum_{i=1}^{p}\mathbf{v}'(A_i \odot \boldsymbol{X}_{t-i}\boldsymbol{X}'_{t-i})\mathbf{v} + \sum_{j=1}^{q}\mathbf{v}'(B_j \odot \Sigma_{t-j})\mathbf{v} \\ &= \mathbf{v}'A_0\mathbf{v} + \sum_{i=1}^{p}(\mathbf{v} \odot \boldsymbol{X}_{\mathbf{t-i}})'A_i(\mathbf{v} \odot \boldsymbol{X}_{\mathbf{t-i}}) + \sum_{j=1}^{q}\mathbf{v}'(B_j \odot \Sigma_{t-j})\mathbf{v} \\ &> 0,\end{aligned}$$

since the first term is positive and the second term is nonnegative by definition. Because the Hadamard product of two positive semidefinite matrices is also positive semidefinite, the third term is also nonnegative. □

Due to the fact that the DVEC model is simply a special VEC model with diagonal $\overline{A}_i$ and $\overline{B}_j$, condition (3.10) for covariance stationary of a general VEC model simplifies to

$$\max_{1\leq m\leq n\leq d}|\sum_{i=1}^{p}a_{i,mn} + \sum_{j=1}^{q}b_{j,mn}| < 1,$$

where $a_{i,mn}$ and $b_{j,mn}$ denote the (m,n)th element of A_i and B_j, respectively.

The structure of DVEC

Example 3.1 (Analysis in two dimensions)**.** In order to gain an understanding of the dynamics of (3.11), consider the simplest possible DVEC model: a bivariate model with $p = q = 1$. The structural equations simplify to the three scalar terms[6]

$$\begin{aligned}\sigma_{t,11} &= a_{0,11} + a_{1,11}X^2_{t-1,1} + b_{1,11}\sigma_{t-1,11}, \\ \sigma_{t,22} &= a_{0,22} + a_{1,22}X^2_{t-1,2} + b_{1,22}\sigma_{t-1,22}, \\ \sigma_{t,12} &= a_{0,12} + a_{1,12}X_{t-1,1}X_{t-1,2} + b_{1,12}\sigma_{t-1,12}.\end{aligned} \tag{3.12}$$

By looking at the first two equations, it can be observed that the squared volatilities of the two component series ($\sigma_{t,11}$ and $\sigma_{t,22}$) follow univariate GARCH updating

[6]Please note that $\sigma_{t,ij}$ now denotes the element of Σ_t in row i and column j. The volatility of the ith component series $\sigma^2_{t,i}$ therefore corresponds to $\sigma_{t,ii}$.

patterns. The conditional covariance $\sigma_{t,12}$ in equation (3.12) has a similar structure driven by the products of the lagged values $X_{t-1,1}X_{t-1,2}$.

As for the CCC and DCC models, the volatility of a single component series is only driven by large lagged values of that series and cannot directly be affected by large lagged values in another series (no lead-lag effect). The more general (but unfortunately virtually always overparameterized) general VEC model would allow this feature.

Some possible parameter restrictions

Since it is required by definition that A_0 should be positive definite and $A_1, \ldots, A_p$ and $B_1, \ldots, B_q$ should all be positive semidefinite, Attanasio [1] suggested to parameterize the model in terms of lower-triangular Cholesky factor matrices $A_0^{1/2}, A_i^{1/2}, B_j^{1/2}$ satisfying

$$A_0 = A_0^{1/2}(A_0^{1/2})', \quad A_i = A_i^{1/2}(A_i^{1/2})', \quad B_j = B_j^{1/2}(B_j^{1/2})'. \tag{3.13}$$

Because $A_1, \ldots, A_p$ and $B_1, \ldots, B_q$ should only be positive *semi*definite, a much simpler parametrization such as

$$A_0 = A_0^{1/2}(A_0^{1/2})', \quad A_i = \boldsymbol{a}_i\boldsymbol{a}_i', \quad B_j = \boldsymbol{b}_j\boldsymbol{b}_j' \tag{3.14}$$

might be considered ($\boldsymbol{a}_i$ and $\boldsymbol{b}_j$ are vectors in $\mathbb{R}^d$). An even cruder model satisfying the requirement of positive-definiteness of Σ_t would be

$$A_0 = A_0^{1/2}(A_0^{1/2})', \quad A_i = a_i I_d, \quad B_j = b_j I_d. \tag{3.15}$$

Here a_i and b_j are simply nonnegative constants. Clearly it is possible to combine the specifications in equations (3.13), (3.14) and (3.15) in various ways for adequate parameter reduction.

3.3.4 The BEKK model

Definition and basic properties

The BEKK[7] model family discussed in this section has the great advantage that the positive-definiteness of Σ_t is ensured by construction without the need of further conditions. Together with the VEC model of section 3.3.2 it was introduced and analyzed by Engle and Kroner [13] in 1995.

Definition 3.7. The process $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$ is a BEKK process if it has the general structure given in Definition 3.1 and the dynamic of the conditional covariance matrix Σ_t is given by the equations

$$\Sigma_t = A_0 + \sum_{i=1}^{p} A_i \boldsymbol{X}_{t-i} \boldsymbol{X}'_{t-i} A'_i + \sum_{j=1}^{q} B_j \Sigma_{t-j} B'_j, \tag{3.16}$$

where

- $A_0 \in \mathbb{R}^{d\times d}$ is symmetric and positive definite,
- $A_i, B_j \in \mathbb{R}^{d\times d}$ for $i = 1, \ldots, p$ and $j = 1, \ldots, q$.

Theorem 3.5. *In the BEKK model (3.16) the conditional covariance matrix Σ_t is positive definite for all t.*

Proof. May be seen analogously to (3.8): For any vector $\boldsymbol{u} \neq \boldsymbol{0}$ in $\mathbb{R}^d$ it holds true that

$$\boldsymbol{u}'\Sigma_t\boldsymbol{u} = \boldsymbol{u}'A_0\boldsymbol{u} + \sum_{i=1}^{p} (\boldsymbol{u}'A_i\boldsymbol{X}_{t-i})^2 + \sum_{j=1}^{q} (B'_j\boldsymbol{u})'\Sigma_{t-j}(B'_j\boldsymbol{u}) > 0,$$

because the first term is positive by definition and the second and third terms are non-negative. □

[7] The acronym BEKK comes from joint work on multivariate ARCH models by Yoshi Baba, Rob Engle, Dennis Kraft and Ken Kroner.

The structure of BEKK

Example 3.2 (Analysis in two dimensions)**.** To gain insight into the BEKK model it is again useful to analyze the bivariate special case of order (1,1) and to consider the dynamics that are implied by (3.16) while comparing to those of the DVEC model given in (3.12).

$$\begin{aligned}\sigma_{t,11} &= a_{0,11} + a_{1,11}^2 X_{t-1,1}^2 + 2a_{1,11}a_{1,12}X_{t-1,1}X_{t-1,2} + a_{1,12}^2 X_{t-1,2}^2 \\ &+ b_{1,11}^2\sigma_{t-1,11} + 2b_{1,11}b_{1,12}\sigma_{t-1,12} + b_{1,12}^2\sigma_{t-1,22}. \end{aligned} \tag{3.17}$$

$$\begin{aligned}\sigma_{t,22} &= a_{0,22} + a_{1,22}^2 X_{t-1,2}^2 + 2a_{1,22}a_{1,21}X_{t-1,1}X_{t-1,2} + a_{1,21}^2 X_{t-1,1}^2 \\ &+ b_{1,22}^2\sigma_{t-1,22} + 2b_{1,22}b_{1,21}\sigma_{t-1,12} + b_{1,21}^2\sigma_{t-1,11}. \end{aligned} \tag{3.18}$$

By studying equation (3.17) it may be observed that we now have a model where a large lagged value of the second component $X_{t-1,2}$ can influence the volatility of the first series $\sigma_{t,11}$ through the parameter $a_{1,12}$. This contrasts with the bivariate DVEC model analyzed in Example 3.1, where this effect is not possible. Similarly, a large lagged second component conditional variance $\sigma_{t-1,22}$ may now effect the present first component conditional variance $\sigma_{t,11}$. Clearly, the same effects can be seen for the volatility of the second series $\sigma_{t,22}$ in (3.18).

Focusing on the conditional covariance $\sigma_{t,12} = \sigma_{t,21}$, the following structure for the off-diagonal elements in Σ_t appears:

$$\begin{aligned}\sigma_{t,12} &= a_{0,12} + a_{1,11}a_{1,21}X_{t-1,1}^2 + a_{1,22}a_{1,12}X_{t-1,2}^2 \\ &+ (a_{1,11}a_{1,22} + a_{1,12}a_{1,21})X_{t-1,1}X_{t-1,2} \\ &+ b_{1,11}b_{1,21}\sigma_{t-1,11} + b_{1,22}b_{1,12}\sigma_{t-1,22} \\ &+ (b_{1,11}b_{1,22} + b_{1,12}b_{1,21})\sigma_{t-1,12}. \end{aligned} \tag{3.19}$$

Here it can be observed that the conditional covariance of the two component series is now not only effected by lagged conditional covariances $\sigma_{t-1,12}$ and component series products $X_{t-1,1}X_{t-1,2}$ as in (3.12), but also by lagged variances and lagged values of the (squared) component processes themselves. To eliminate all crossover

effects in the conditional variance equations of the BEKK model in (3.17) and (3.18), the off-diagonal terms $a_{1,12}, a_{1,21}, b_{1,12}$ and $b_{1,21}$ would have to be set to zero and the parameters governing the individual volatilities would also govern the conditional covariance $\sigma_{t,12}$ in (3.19).

A more general BEKK class

The above example particularly shows that already a low order and low dimensional BEKK model inhibits rather complicated and nested covariance matrix structure. Nevertheless, the original definition of the BEKK class given by Engle and Kroner [13] is even broader than the previously defined. It is given by

$$\Sigma_t = A_0 A_0' + \sum_{k=1}^{K} \sum_{i=1}^{p} A_{k,i} \boldsymbol{X}_{t-i} \boldsymbol{X}_{t-i}' A_{k,i}' + \sum_{k=1}^{K} \sum_{j=1}^{q} B_{k,j} \Sigma_{t-j} B_{k,j}', \tag{3.20}$$

where $d(d+1)/2 \geq K \geq 1$ and the choice of K determines the richness of the model. This model class is of largely theoretical interest and is probably too complex for practical applications; even for $K = 1$ it is difficult to fit in higher dimensions. For this class of models, conditions have been found which assure that a given VEC model can be cast as a BEKK model (the contrary is always true), and it is also flexible enough to include all DVEC models.[8]

3.4 Model Comparison

In Table 3.1 and Figure 3.1 the numbers of parameters in the presented models are summarized. It can be seen that the general VEC model becomes completely unfeasible in higher dimensions, and BEKK and general DVEC models are also of limited use as dimension grows. For higher dimensional modeling, the remaining models are the most practically useful.

[8] More information on this topic can be found in Engle and Kroner [13] and the recent works by Scherrer and Ribarits [22], where it is for example shown that BEKK models as in (3.20) are as general as VEC models in the bivariate case, whereas in higher dimensions VEC models allow for more flexibility.

Model	Parameter Count	2	5	10	20
VEC	$\frac{d(d+1)}{2}(1+(p+q)\frac{d(d+1)}{2})$	21	465	6105	88410
BEKK ($K=1$)	$\frac{d(d+1)}{2}+d^2(p+q)$	11	65	255	1010
DVEC1 as in (3.13)	$\frac{d(d+1)}{2}(1+p+q)$	9	45	165	630
DCC	$\frac{d(d+1)}{2}+(d+1)(p+q)$	9	27	77	252
CCC	$\frac{d(d+1)}{2}+d(p+q)$	7	25	75	250
DVEC2 as in (3.14)	$\frac{d(d+1)}{2}+d(p+q)$	7	25	75	250
DVEC3 as in (3.15)	$\frac{d(d+1)}{2}+p+q$	5	17	57	212

Table 3.1: Summary of numbers of parameters in various multivariate GARCH models. Second column gives general formula and final columns give numbers for dimensions 2, 5, 10 and 20 when $p = q = 1$. In CCC and DCC it is assumed that all component series volatilities have p ARCH and q GARCH terms (i.e. $p_k = p$ and $q_k = q$ for $k = 1, \ldots, d$) in Definitions 3.2 and 3.4. Additional parameters in the innovation distribution are not considered.

3.5 Fitting Multivariate GARCH Models

In the previous sections notes have already been given on fitting some models in stages. In the high-dimensional applications this may in fact be the only feasible strategy. However, when confronted with return series of modest dimension, one can attempt to fit multivariate GARCH models by maximizing an appropriate likelihood with respect to all parameters in a single step. This procedure follows directly from the ideas developed in Section 2.3.

3.5.1 Building the likelihood

The method of building a likelihood for any multivariate GARCH model of type $\boldsymbol{X}_t = \Sigma_t^{1/2}\boldsymbol{Z}_t$ is completely analogous to the univariate case. Again, for the sake of simplicity, a first order model shall be considered and the data is assumed to be labeled $\boldsymbol{X}_0, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. For the same reasons as before, the "regular" likelihood approach as in (2.17) fails because of the unknown density $f_{\boldsymbol{X}_0}(\boldsymbol{x}_0)$. A conditional likelihood approach is based on the conditional joint density of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ given

$\boldsymbol{X}_0$ and an initial value for the conditional covariance matrix Σ_0, which is written

$$f_{\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n|\boldsymbol{X}_0,\Sigma_0}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n|\boldsymbol{x}_0,\Sigma_0)=\prod_{t=1}^{n} f_{\boldsymbol{X}_t|\boldsymbol{X}_{t-1},\ldots,\boldsymbol{X}_0,\Sigma_0}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1},\ldots,\boldsymbol{x}_0,\Sigma_0).$$

If $g(\boldsymbol{z})$ denotes the multivariate innovation density of $\boldsymbol{Z}_t$, the conditional densities can be expressed in terms of the innovation density and Σ_t:

$$f_{\boldsymbol{X}_t|\boldsymbol{X}_{t-1},\ldots,\boldsymbol{X}_0,\Sigma_0}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1},\ldots,\boldsymbol{x}_0,\Sigma_0)=|\Sigma_t|^{-1/2}g(\Sigma_t^{-1/2}\boldsymbol{X}_t),$$

where Σ_t is a matrix-valued function of $\boldsymbol{x}_{t-1},\ldots,\boldsymbol{x}_0$ and Σ_0. Most common choices of $g(\boldsymbol{z})$ are in the so called spherical family, where $g(\boldsymbol{z})=h(\boldsymbol{z}'\boldsymbol{z})$ for some function h of a scalar variable known as the density generator (McNeil et al. [20]). This yields a conditional likelihood of the form

$$L_c(\boldsymbol{\theta})=\prod_{t=1}^{n}|\Sigma_t|^{-1/2}h(\boldsymbol{X}_t'\Sigma_t^{-1}\boldsymbol{X}_t), \tag{3.21}$$

where all parameters appearing in the volatility equation and the innovation distribution are collected in $\boldsymbol{\theta}$. Equation (3.21) may of course be adapted to include a constant mean term or one of vector autoregressive structure (as introduced in Section 1.3).

Of course, the above ideas may easily be extended in order to combine the modeling of conditional expectation and conditional variance. Considering for instance VARMA(p_1,q_1)-BEKK(p_2,q_2), the likelihood would look like

$$L_c=\prod_{t=1}^{n}|\Sigma_t|^{-1/2}g(\Sigma_t^{-1/2}(\boldsymbol{X}_t-\boldsymbol{\mu}_t)),$$

with

$$\boldsymbol{\mu}_t=\boldsymbol{\mu}+\sum_{i=1}^{p_1}\Phi_i(\boldsymbol{X}_{t-i}-\boldsymbol{\mu})+\sum_{j=1}^{q_1}\Theta_j(\boldsymbol{X}_{t-j}-\boldsymbol{\mu}_{t-j})$$

and

$$\Sigma_t=A_0+\sum_{i=1}^{p}A_i(\boldsymbol{X}_{t-i}-\boldsymbol{\mu}_t)(\boldsymbol{X}_{t-i}-\boldsymbol{\mu}_t)'A_i'+\sum_{j=1}^{q_1}B_j\Sigma_{t-j}B_j'.$$

3.5.2 Finding parameter estimates

In order to evaluate the likelihood (3.21), a starting value is required for Σ_0. It is typically set to equal the sample covariance matrix (or other estimators for covariance).[9] For the numerical maximization of L_c, common choices are again modified Newton-Raphson procedures, in particular that of Berndt, Hall, Hall and Hausman [4].

3.5.3 Model checking

Standardized residuals $\boldsymbol{z}_t$ are calculated according to

$$\boldsymbol{z}_t = \widehat{\Sigma}_t^{-1/2} \boldsymbol{X}_t, \qquad t = 1, \ldots, n,$$

and should by assumption behave like a realization of a SWN$(\mathbf{0}, I_d)$ process. This can be investigated with the usual univariate procedures described in Section 2.3.4 (such as correlograms, correlograms of transformed values and portmanteau tests) applied to the component series of the residuals. Also, there should be no evidence of cross correlations at any lags for both the raw and absolute residuals in the cross correlogram.

Model selection is often performed by comparing different information criteria (like Akaike AIC or Bayes BIC). Is is important to note, however, that there is not yet much literature on theoretical aspects of the use of AIC or BIC in an univariate GARCH context, and certainly even less in a multivariate one. They are defined by

$$\begin{aligned} \mathrm{AIC}(p) &= -2\log(L_c) + 2p, \\ \mathrm{BIC}(p) &= -2\log(L_c) + p\log(n), \end{aligned}$$

where p is the number of parameters and n the length of the time series considered (S-PLUS Reference Manual [17]).

[9] The same remarks as in the univariate case apply, namely that the issue of conditioning on starting values is of relatively minor importance when datasets are large.

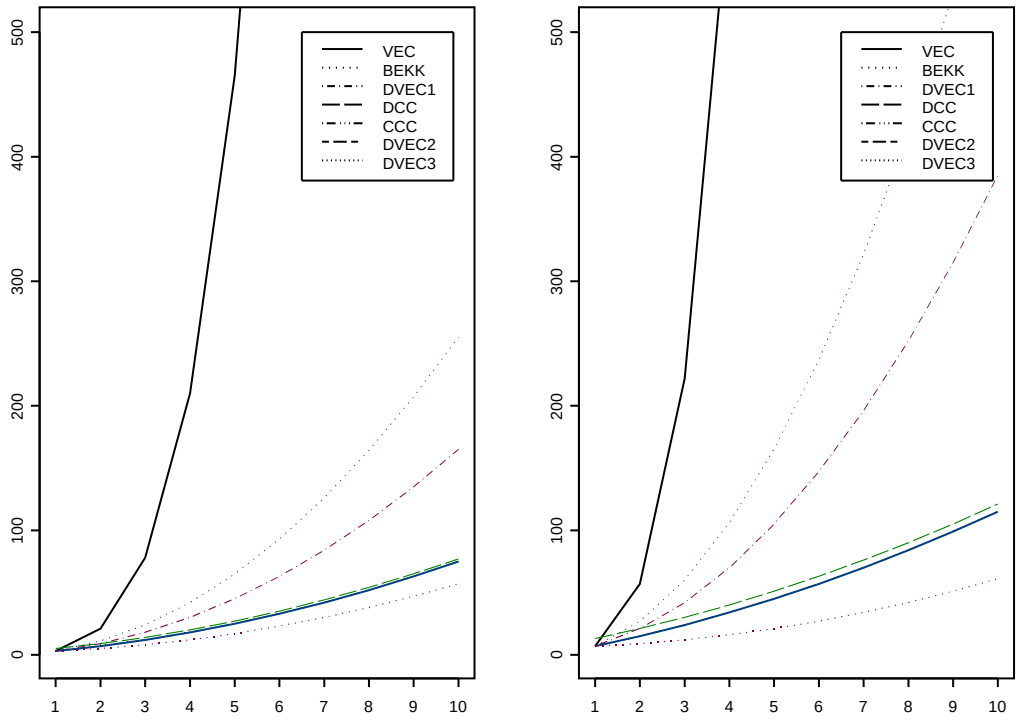

Figure 3.1: Evolution of the numbers of parameters in various multivariate GARCH models with increasing dimension d. On the left picture $p = q = 1$ are fixed, on the right $p = q = 3$. The terms DVEC1, DVEC2 and DVEC3 refer to the special DVEC parameterizations given in (3.13), (3.14) and (3.15), respectively. In CCC and DCC it is assumed that all component series volatilities have p ARCH and q GARCH terms (i.e. $p_k = p$ and $q_k = q$ for $k = 1, \ldots, d$) in Definitions 3.2 and 3.4. Additional parameters in the innovation distribution are not considered.

Chapter 4

Multivariate GARCH in Practice

Focus of this chapter will be put on fitting several different MGARCH models to data, with some attention to modeling the mean term by means of autoregressive structure. The fitting itself is performed with S-PLUS [18] making use of the module finmetrics [19]. Some sample code can be found in the appendix.

4.1 The Data

For the following analysis the ATX[1] and the German DAX[2] time series are considered. The time span ranges from November 11, 1992 to April 28, 2006 and data is collected daily (on workdays). The very few days where only one stock market was in operation were ignored in order to obtain a syncronized time series, resulting in a dataset with a total of 3321 observations. In Figure 4.1 the two (obviously non-stationary) time series are displayed graphically.

A very common way of handling a non-stationary time series $(\boldsymbol{N}_t)_{t\in\mathbb{Z}}$ is to look at

[1] ATX (abbreviated from Austrian Traded Index) is the most important stock index at the Vienna stock exchange currently consisting of stocks of the 21 largest quoted companies in Austria, such as Erste Bank Group (19,24%), OMV (17,39%) or Telekom Austria (12,97%).

[2] DAX (abbreviated from Deutscher Aktienindex) is a stock market index consisting of the 30 major German companies trading on the Frankfurt Stock Exchange. Examples are Siemens (12,44%), Deutsche Telekom (10,42%), E.ON (8,89%), Deutsche Bank (8,86%), Daimler-Chrysler (8,64%).

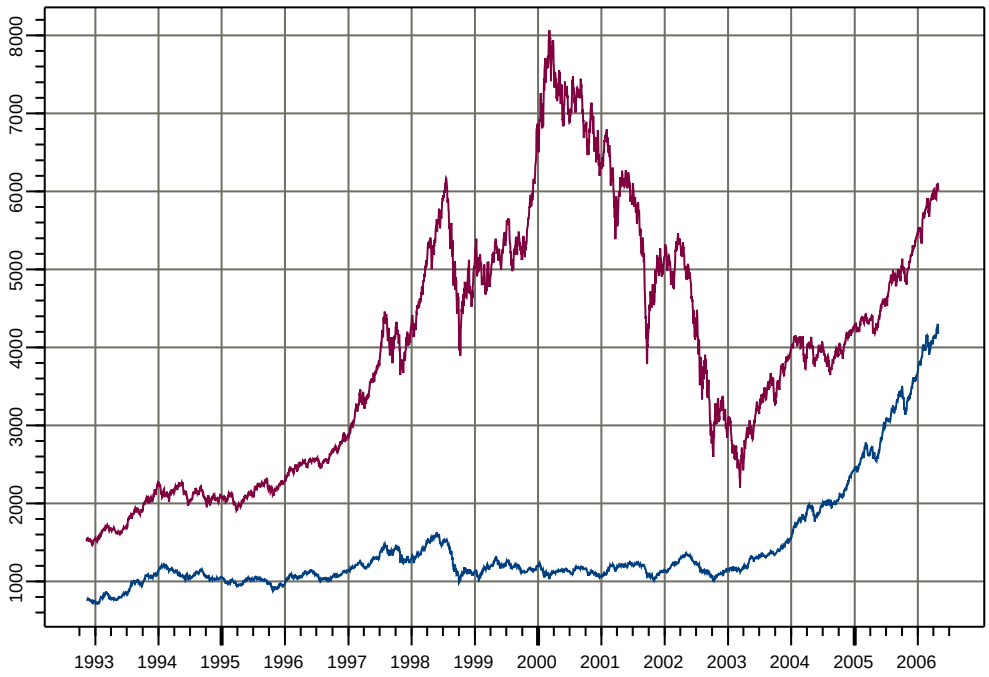

Figure 4.1: ATX (lower time series) and DAX (upper time series).

first differences $\Delta \boldsymbol{N}_t$ defined by

$$\Delta \boldsymbol{N}_t = \boldsymbol{N}_t - \boldsymbol{N}_{t-1},$$

which often yields stationary time series. Another, especially in finance literature even more widespread method is to consider *continuously compounded one period returns*[3] or simply *log-returns*, denoted by $\overline{\Delta}\boldsymbol{N}_t$ and defined to be

$$\boldsymbol{X}_t := \overline{\Delta}\boldsymbol{N}_t = \log\left(\frac{\boldsymbol{N}_t}{\boldsymbol{N}_{t-1}}\right) = \Delta(\log \boldsymbol{N}_t), \tag{4.1}$$

where $\log(\cdot)$ is the natural logarithm function. The log-returns of the ATX and

[3]Terminology taken from Zivot and Wang [28]. The name refers to the economic interpretation of these returns making $\overline{\Delta}\boldsymbol{N}_t$ the continuously compounded growth rate between periods $t-1$ and t.

DAX are shown in Figure 4.2 and some descriptive statistics of the data are presented in Table 4.1.

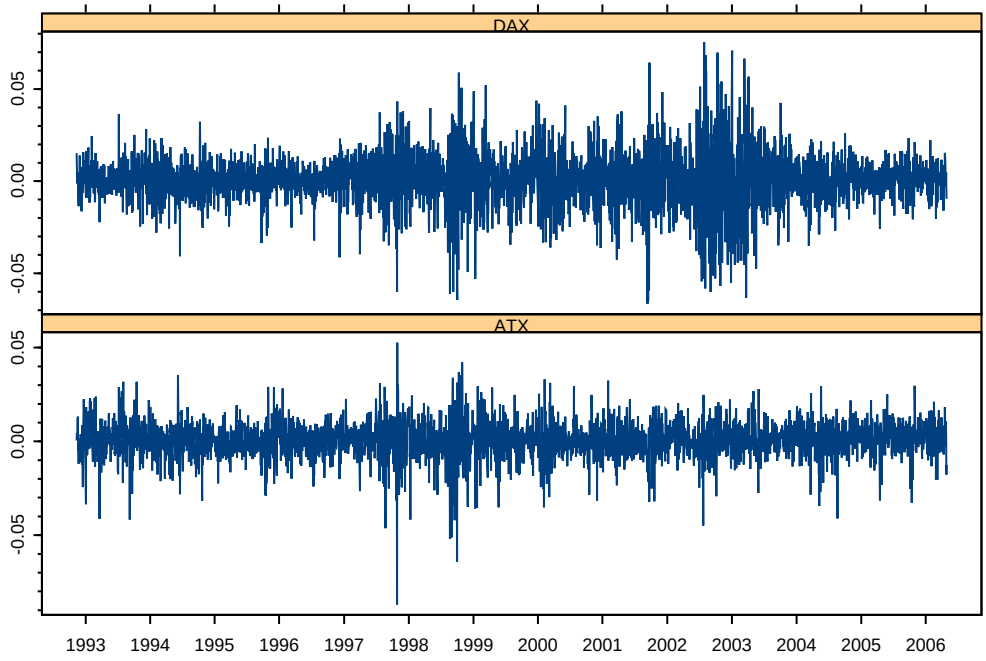

Figure 4.2: Log-returns of ATX (bottom) and DAX (top), as defined in (4.1).

	ATX	DAX
Mean	0.000514	0.000416
Median	0.000702	0.000999
Minimum	-0.086995	-0.066523
Maximum	0.052623	0.075527
Standard Deviation	0.010027	0.014649
Skewness	-0.644807	-0.153466
Excess Kurtosis	4.411840	2.929709

Table 4.1: Descriptive statistics of the log-returns.

In order to gain some insight into the correlation structure of the time series

$\boldsymbol{X}_t$, the cross-correlogram[4] of the raw and the absolute values are displayed in Figure 4.3. While the raw cross-correlogram on the top of Figure 4.3 exhibits hardly any autocorrelation (apart from ATX at lag one) and only little cross-correlation, the cross correlogram of the squared values (bottom) shows that both auto- and cross-correlation are significantly different from zero, even for large lags.

4.2 Modeling the Mean

Even though little auto- and cross-correlation of the raw time series can be spotted, a mean-equation model is implemented by means of diagonal vector autoregression.[5] Information criteria such as AIC or BIC displayed in Table 4.2 suggest that a diagonal VAR model of order one adequately captures the dynamics of the data.

lag	likelihood	AIC	BIC
1	**20309.52**	**-40607.03**	**-40570.39**
2	20303.47	-40586.94	-40525.87
3	20299.37	-40570.73	-40485.24
4	20300.55	-40565.10	-40455.18
5	20296.60	-40549.19	-40414.86
6	20297.63	-40543.25	-40384.50
7	20293.28	-40526.55	-40343.38
8	20293.19	-40518.38	-40310.80
9	20287.78	-40499.57	-40267.58
10	20285.00	-40485.99	-40229.60

Table 4.2: Likelihood, AIC and BIC for VAR(p) models, $p = 1, \ldots, 10$.

The cross correlogram of the residuals can be found in Figure 4.4, and it may be observed that some auto- and cross-correlation of the raw values has been successfully removed. Some significant cross-correlation, however, still persists at lags 4 and 15 - higher order diagonal VAR or even full VAR models would be needed in order to fully remove these. As expected, the correlation structure of the absolute values has not changed noticeably; they still show major correlation.

[4]The cross-correlogram is introduced and explained in Section 1.2.2

[5]In a diagonal VAR(p) model Φ_i is required to be diagonal for $i = 1, \ldots, p$.

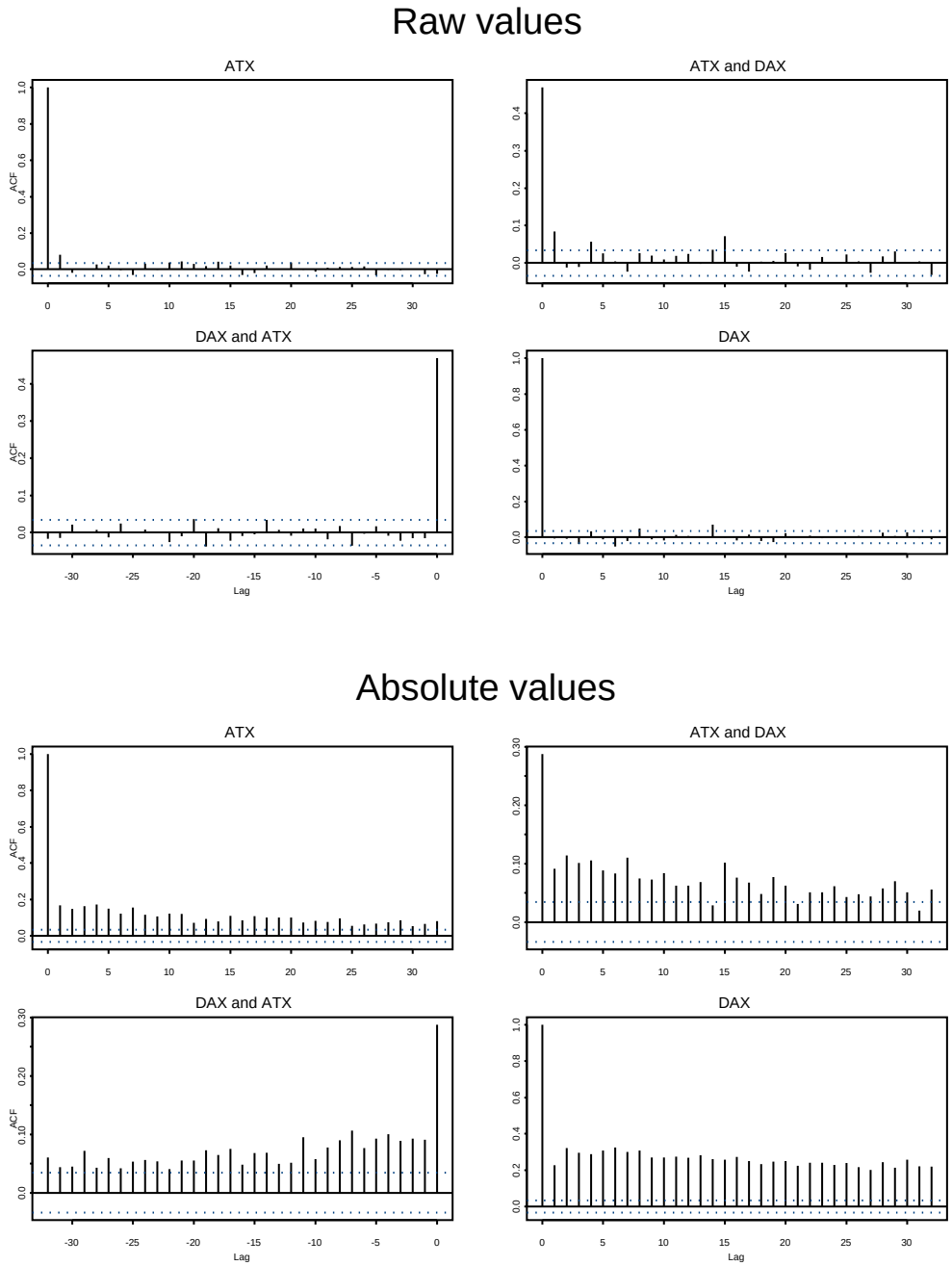

Figure 4.3: Cross Correlograms of ATX and DAX. Top pictures show auto- and cross-correlation of the raw values of log-returns, bottom pictures show auto- and cross-correlation of the absolute values.

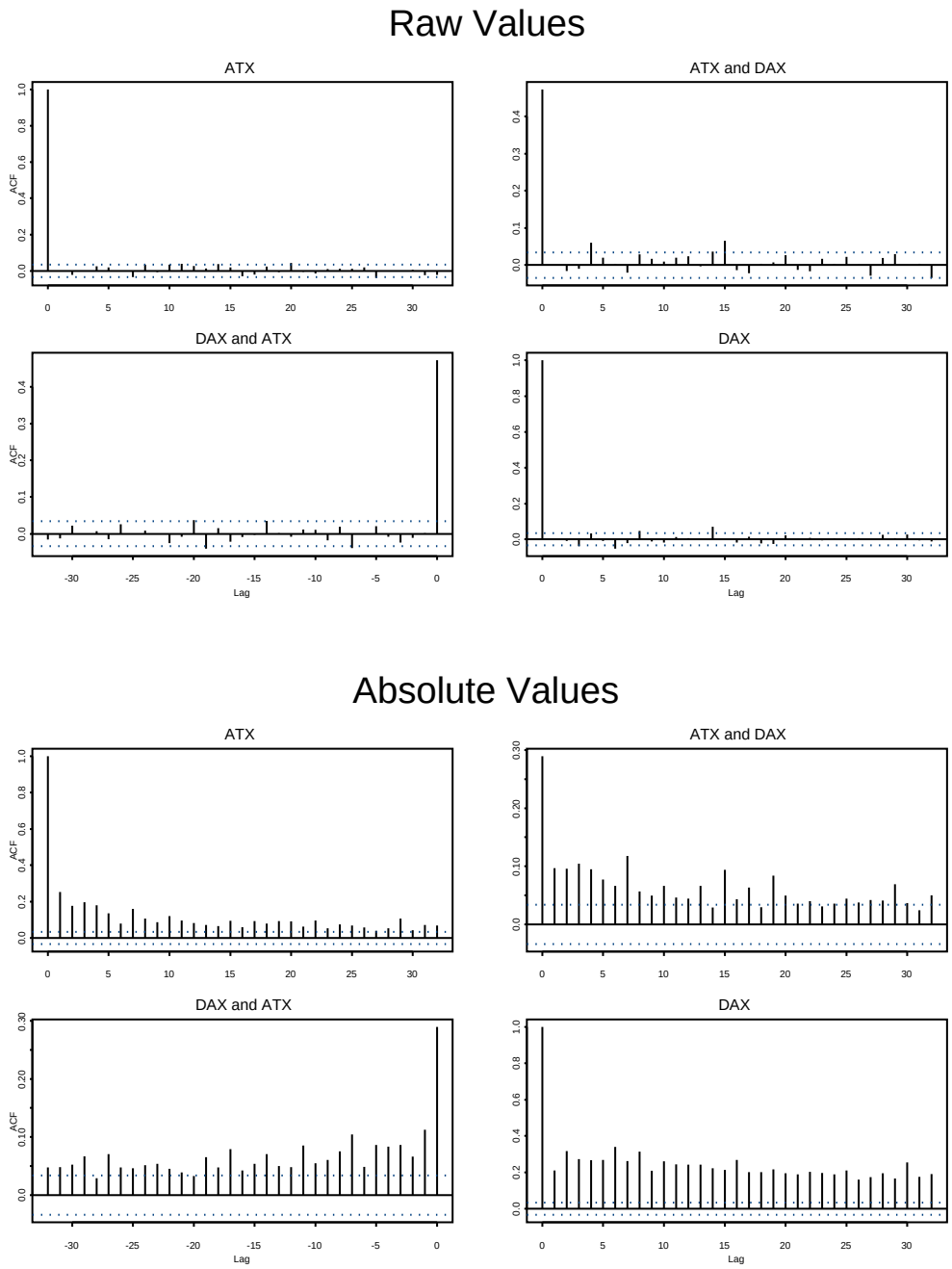

Figure 4.4: Cross correlograms of raw and absolute residuals after fitting an AR(1) model to the log returns.

For future investigations, the approach chosen is to model conditional correlation with different MGARCH models and a diagonal VAR(1) mean term. Higher order (full) VAR models will not be considered for reasons of parameter parsimony. The results will then be compared in Section 4.7.

4.3 CCC Approach

The first model fitted to the data will be a simplest possible combination of two univariate GARCH models with assumed constant conditional correlation. With a decent number of ARCH and GARCH terms, this model is rather parsimonious. Nevertheless, it provides adequate fitting quality.

4.3.1 Lag-length selection

When fitting a CCC-model as discussed in Section 3.2.1, for each of the two component series the number of ARCH coefficients p_1 and p_2 as well as the number of GARCH coefficients q_1 and q_2 appearing in equation (3.5) have to be fixed. In practice they are often set to one without further investigation (since that choice has shown reasonable results). Also, a comparison of AIC or BIC numbers may be performed. Even though this approach is not fully theoretically justified, it may be a helpful guidance. For the sake of simplicity, we set $p_1 = p_2$ and $q_1 = q_2$, i.e. the univariate GARCH models have the same number of parameters. For illustration purposes, a "BIC-landscape" is displayed in Figure 4.5.

Even though Figure 4.5 gives an idea about the evolution of BIC, other visualizations such as those in Figure 4.6 are probably more helpful. These can be interpreted as "slices" of Figure 4.5 for fixed p or q, respectively.

The pictures suggest that when fitting a CCC model, choosing $p = q = 1$ should be a reasonable and adequately parsimonious choice. Fixing q to zero would not be appropriate, and higher order models show only very little if any improvement. When looking at AIC instead of BIC, this effect is slightly less striking (since BIC penalizes large number of parameters heavier).

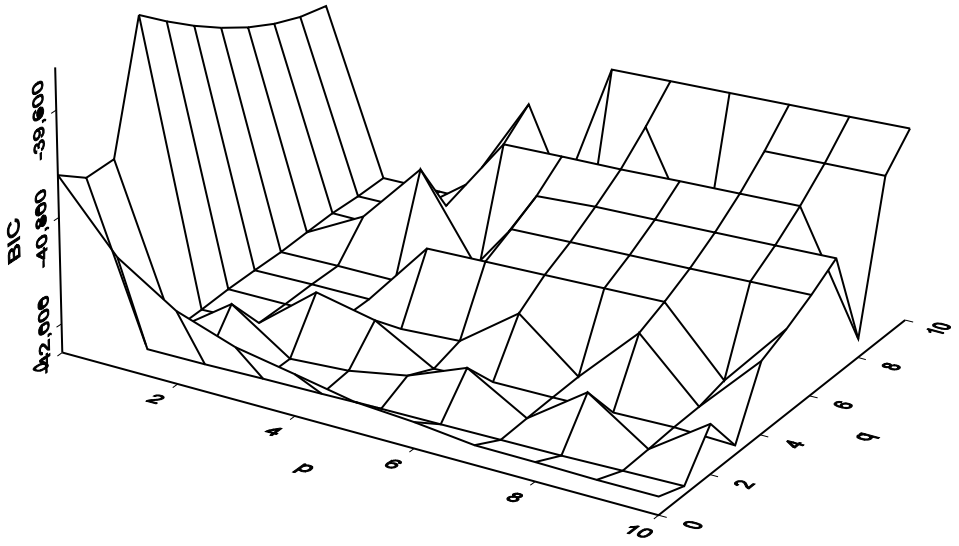

Figure 4.5: Value of BIC after fitting CCC-models to the bivariate time series consisting of ATX and DAX log-returns for different choices of $p = p_1 = p_2$ and $q = q_1 = q_2$.

4.3.2 Choice of the innovation distribution

So far, no attention has been paid to the choice of the innovation distribution. The above calculations have been performed with assumed Gaussian innovations (standard in S-PLUS). It might be reasonable to loosen this assumption and switch to a Student t distribution instead, at the cost of one extra parameter.[6] Comparing AIC, BIC and likelihood values as in Table 4.3 indicates that the Student t distribution provides a much better fit.

[6] In fact it *is* reasonable to use a Student t innovation distribution instead of a Gaussian one, as can be seen when looking at QQ-plots of residuals in Section 4.3.4. The Gaussian distribution cannot capture the heavy tails appropriately.

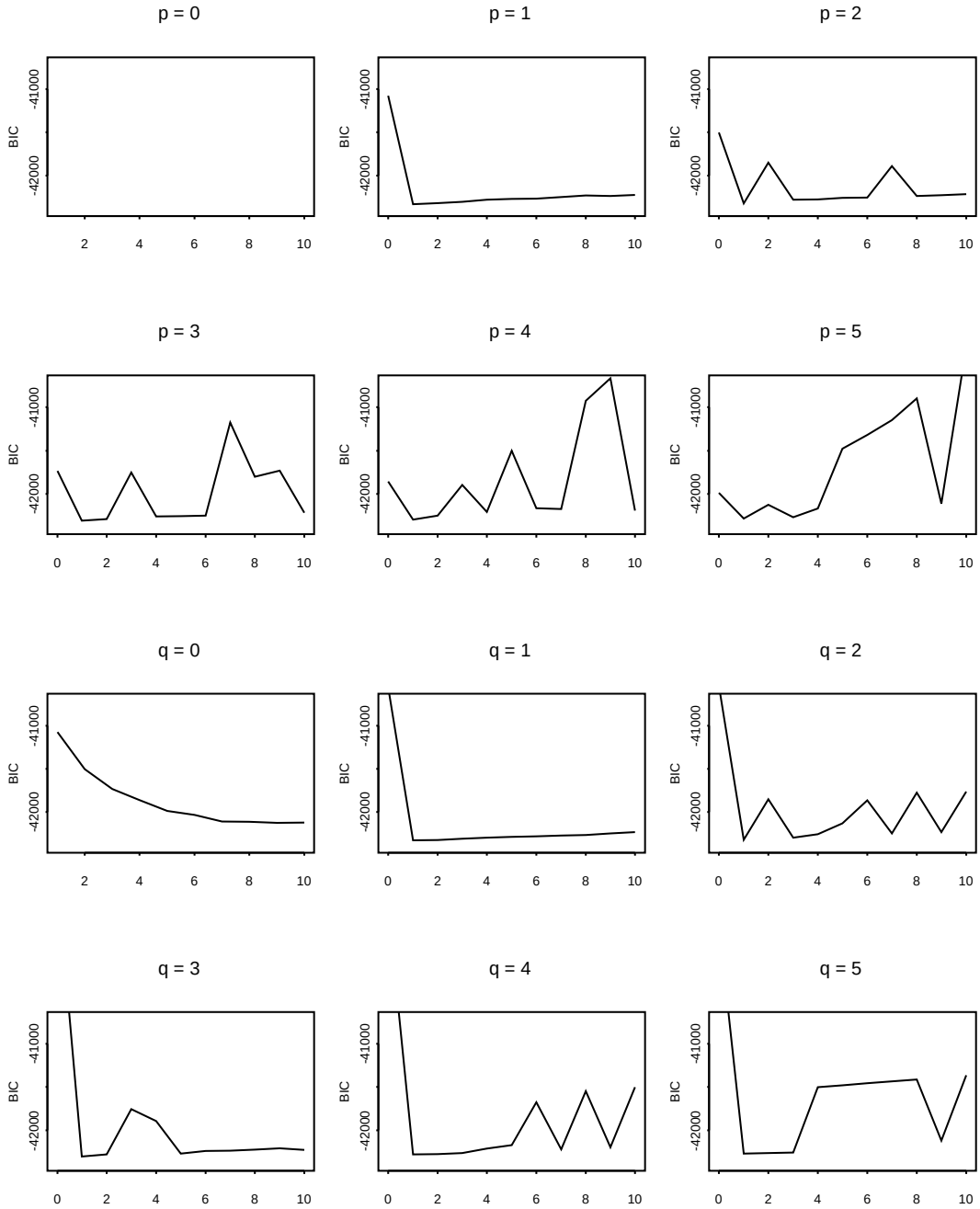

Figure 4.6: Value of BIC after fitting AR(1)-CCC-models to the bivariate time series for different choices of $p = p_1 = p_2$ and $q = q_1 = q_2$. In the top six frames, pictures show BIC-value for varying q and fixed p; in the bottom six vice versa.

	AIC		BIC		likelihood		para
(p,q)	normal	t	normal	t	normal	t	normal
(1,0)	-41125	-41681	-41076	-41626	20571	20850	8
(2,0)	-41565	-41958	-41504	-41890	20793	20990	10
(1,1)	-42389	-42586	**-42328**	**-42519**	21205	21304	**10**
(1,2)	-42390	-42582	-42317	-42502	21207	21304	12
(2,1)	**-42395**	**-42589**	-42322	-42510	21210	21308	**12**
(3,0)	-41807	-42113	-41734	-42033	20916	21069	12
(2,2)	-41937	-42231	-41852	-42139	20983	21130	14
(1,3)	-42387	-42580	-42301	-42489	21207	21305	14
(3,1)	-42393	-42587	-42308	-42495	21211	21308	14
(2,3)	-42376	-42243	-42278	-42139	21204	21138	16
(3,2)	-42392	-42582	-42294	-42479	21212	21308	16
(3,3)	-41864	-42196	-41754	-42080	20950	21117	18

Table 4.3: Comparison of AIC, BIC and likelihood value for different lag-lengths and innovation distributions. The fitted model is AR(1)-CCC with parameters (p,q) indicated as above. Last column shows number of parameters for normal innovations, for t innovations add one to each row. Best fits according to AIC and BIC are indicated in bold font.

Due to the above reasoning, the AR(1)-CCC(1,1) (short notation for diagonal AR(1) combined with MGARCH-CCC with parameters $p = p_1 = p_2 = q = q_1 = q_2 = 1$) model with a constant mean and t innovations will be considered for further investigation.

4.3.3 The fitted model

The parameters obtained through maximum likelihood estimation, as well as estimated standard errors and corresponding t-statistics with p-values are summarized in Table 4.4.

Remark. Please note that in S-PLUS the VARMA model with mean is formulated slightly different to Definition 1.9. Taking vector AR(1) as an instructive example, the S-PLUS formulation with a constant reads

$$\boldsymbol{X}_t = \boldsymbol{c} + \phi \boldsymbol{X}_{t-1} + \boldsymbol{\varepsilon_t}.$$

AR(1)-CCC(1,1)	value	std. error	t value	p value
c_1	0.00079153	0.00014644	5.404987	0.00000003
c_2	0.00091768	0.00017815	5.151118	0.00000014
ϕ_1	0.09707496	0.01563813	6.207583	$< 10^{-8}$
ϕ_2	-0.02583484	0.01609135	-1.605511	0.05423833
$\alpha_{1,0}$	0.00000313	0.00000067	4.643741	0.00000178
$\alpha_{2,0}$	0.00000080	0.00000030	2.648684	0.00405945
$\alpha_{1,1}$	0.06750091	0.00845642	7.982207	$< 10^{-8}$
$\alpha_{2,1}$	0.06098339	0.00742144	8.217191	$< 10^{-8}$
$\beta_{1,1}$	0.89674024	0.01307623	68.577882	$< 10^{-8}$
$\beta_{2,1}$	0.93710688	0.00723059	129.603145	$< 10^{-8}$

Table 4.4: Estimated AR(1)-CCC(1,1) coefficients as of ATX and DAX log-returns with standard errors. Innovation distribution: Student t.

This implies that the mean $\boldsymbol{\mu}$ as defined in Chapter 1 amounts to

$$\mathbb{E}[\boldsymbol{X}_t] = \boldsymbol{\mu} = (I - \phi)^{-1}\boldsymbol{c}.$$

According to the approximate t-statistics, all estimated parameters are more or less significantly different from zero. This is especially important for $\alpha_{1,0}$ and $\alpha_{2,0}$, which are actually rather close to zero but must be strictly positive by definition. The conditional constant correlation matrix was estimated to be (standard errors in squared brackets)

$$\widehat{P}_c = \begin{pmatrix} 1 & 0.5033 \\ 0.5033 & 1 \end{pmatrix} \quad \begin{bmatrix} & 0.0145 \\ 0.0145 & \end{bmatrix},$$

and the innovation distribution parameter ν (degrees of freedom) amounts to 9.3063 [0.8483].

The estimated conditional volatility is displayed in Figure 4.7. The pictures suggest that the constant conditional correlation matrix assumption might be too restrictive; higher correlation would for instance be expected around 1997 to 1999 rather than at other times. This justifies a different modeling approach allowing for dynamic evolvement of the conditional correlation.

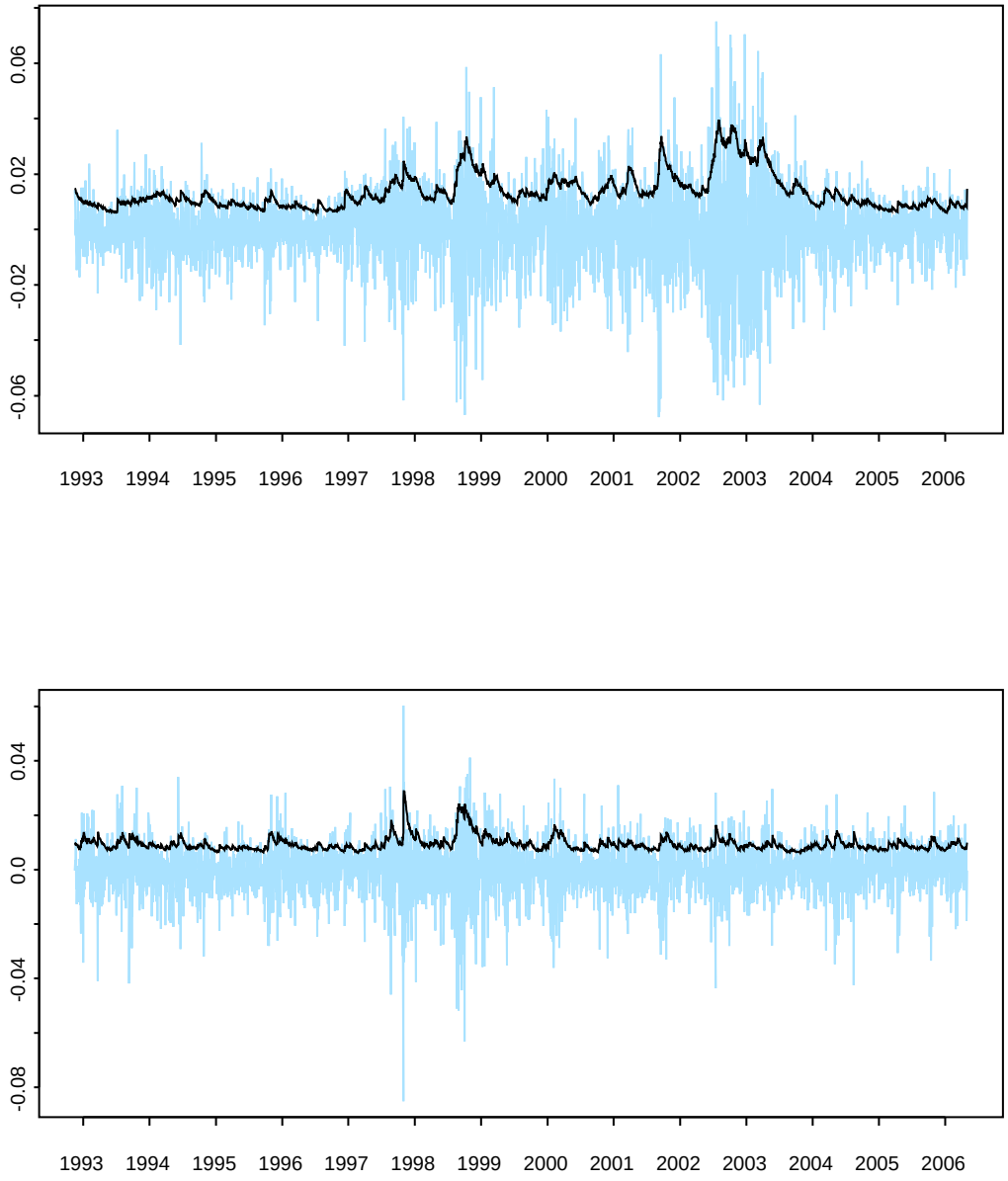

Figure 4.7: Estimated conditional volatility of ATX (bottom) and DAX (top) log-returns using AR(1)-CCC(1,1) with t-innovations. In the background the actual AR-residuals are displayed.

4.3.4 Model diagnostics

As mentioned in the previous chapters, the standardized residuals (i.e. mean-corrected residuals) should behave like a realization of a SWN($\mathbf{0}, I_d$) process. They are displayed in Figure 4.8. Clearly, with a picture like this, it is rather difficult to sense possible violations of the independence hypothesis and even harder to confirm independence.

However, it seems that (at least some of the) volatility clustering effects have successfully been removed. This can be observed when comparing the standardized residuals with the raw log-returns in Figure 4.2. Also, no striking dependence effects like simultaneous jumps or similar can be observed.

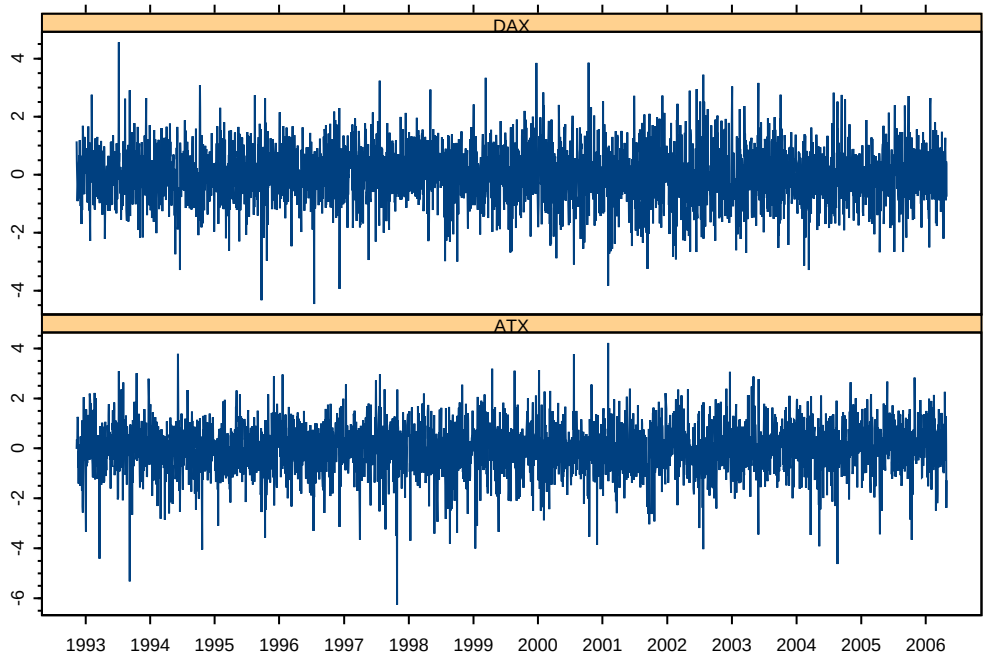

Figure 4.8: Standardized residuals after fitting a AR(1)-CCC(1,1) model with Student t innovations for log-returns of ATX and DAX.

Maybe more helpful than looking at the standardized residuals directly is the investigation of component series autocorrelation of raw and transformed values as well as the cross-correlation between them. This can comfortably be achieved using the cross correlogram, which is displayed in Figure 4.9. As before, the correlograms of raw and absolute values are shown. Using other transformations than the absolute value (such as square roots, logarithms or powers) gives similar results.

Figure 4.9 shows that both auto- and cross-correlation has been removed to a very large extent. Some lead-lag effect from DAX onto ATX remains, most likely due to the rather low-order AR-part of the model. The univariate GARCH parts seem to have done their job very well: autocorrelation of transformed values has been reduced drastically in comparison to the correlogram of the log-returns displayed in Figure 4.3.

Finally, some attention is again paid to the choice of the innovation distribution. Quantile-Quantile-plots of both the AR(1)-CCC(1,1) model with Gaussian innovations as well as of that with Student t innovations are displayed in Figure 4.10 below.

Imposing Gaussian innovation distribution underestimates the tails of the data, which can be seen by noting the *inverted S shape* of the QQ-plots in Figure 4.10. Student t innovations manage somewhat better. However, some negative outliers (i.e. downward shocks) are still not modeled correctly.

4.4 DCC Approach

The above ideas may be extended to DCC as described in Section 3.2.2, now allowing a dynamic evolution of conditional correlation while keeping the number of parameters low. In the simplest but very successful case, only two extra parameters are required.

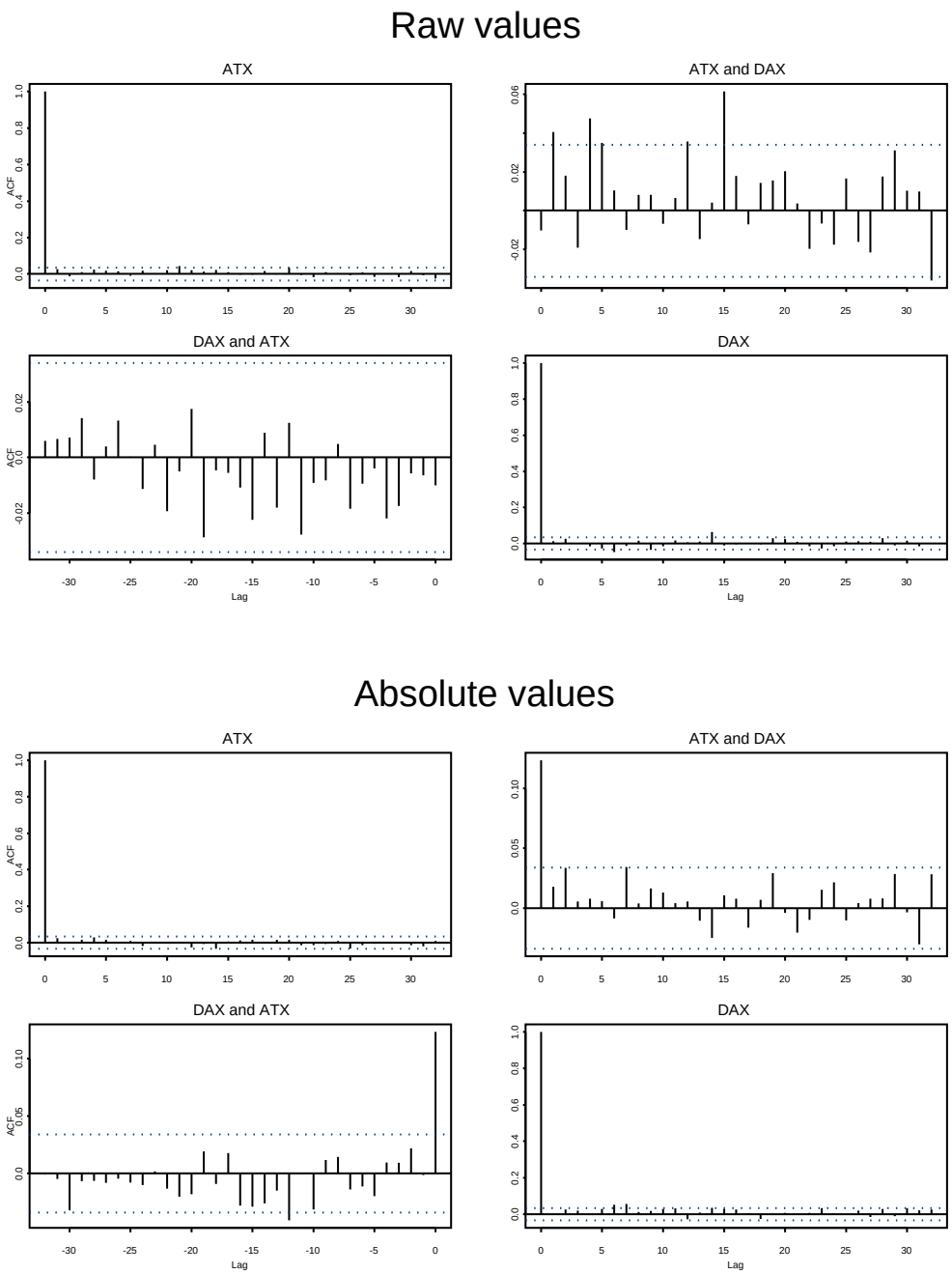

Figure 4.9: Cross Correlograms of standardized residuals after fitting an AR(1)-CCC(1,1) model with Student t innovations. Top pictures show auto- and cross-correlation of the raw values, bottom pictures of the absolute values.

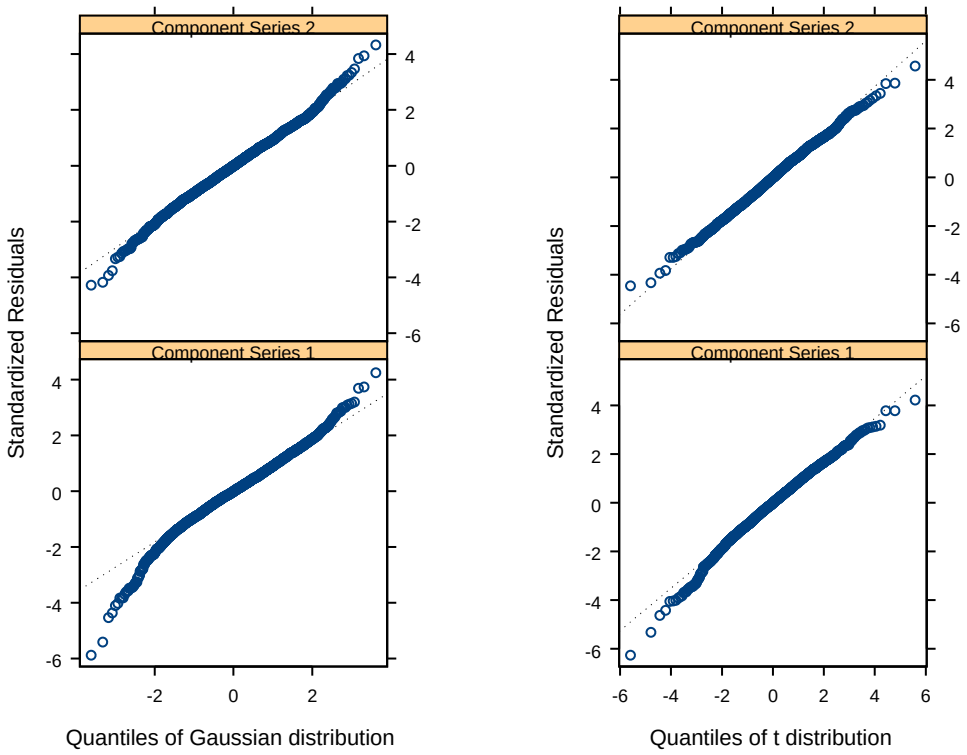

Figure 4.10: QQ-plots of innovation distribution against the empirical distribution of the standardized residuals. Left picture shows model with Gaussian innovations, right picture shows model with t innovations.

4.4.1 Stepwise estimation

Since no predefined DCC model is available in finmetrics' `MGARCH()` function for S-PLUS, the stepwise estimation approach suggested on page 55 was implemented. Univariate AR(1)-GARCH(1,1) models were fitted to the component series and the de-volatized process $\widehat{\boldsymbol{Y}}_t = \widehat{\Delta}_t^{-1}(\boldsymbol{X}_t - \widehat{\phi}\boldsymbol{X}_{t-1} - \widehat{c})$ was formed. $\overline{Q}$ was estimated by the sample correlation matrix of $\widehat{\boldsymbol{Y}}_t$. Finally, a model of structure $\boldsymbol{Y}_t = P_t^{1/2}\boldsymbol{Z}_t$ with P_t given by (3.6) and (3.7) was fitted to the de-volatized data for $p = q = 1$. The last step was performed manually[7] in S-PLUS without the use of pre-compiled code for evaluating the likelihood function, which resulted in very slow

[7]Please consult comments and source code in A.6 for remarks about finding appropriate starting values for the optimization.

but perfectly working code. Results are given in Table 4.5.

AR(1)-DCC(1,1)	value	std. error	t value	p value
c_1	0.00080285	0.00015473	5.188568	0.0000001
c_2	0.00099349	0.00018729	5.304542	0.0000001
ϕ_1	0.09003532	0.01759758	5.116346	0.0000002
ϕ_2	-0.01550337	0.01834954	-0.844891	0.1991163
$\alpha_{1,0}$	0.00000316	0.00000081	3.901673	0.0000487
$\alpha_{2,0}$	0.00000091	0.00000040	2.252767	0.0121695
$\alpha_{1,1}$	0.07057884	0.01067204	6.613435	$< 10^{-8}$
$\alpha_{2,1}$	0.06824717	0.00968003	7.050306	$< 10^{-8}$
$\beta_{1,1}$	0.89424746	0.01601047	55.853927	$< 10^{-8}$
$\beta_{2,1}$	0.93071103	0.00930861	99.983899	$< 10^{-8}$
α_1	0.01527278	0.00243383	6.275214	$< 10^{-8}$
β_1	0.96201186	0.00005660	16996.045676	$< 10^{-8}$

Table 4.5: Estimated AR(1)-DCC(1,1) coefficients as in (3.6) and (3.7) with $\overline{Q}$ equaling the sample correlation matrix of the de-volatized data. Multivariate Student t innovation distribution (estimated degrees of freedom 6.7302 [0.5228]) is chosen for modeling component series volatilities, Gaussian innovation distribution for the conditional correlation.

4.4.2 Alternative modeling

Differences in the volatility parameters in comparison with Table 4.4 result from varying internal estimation methods in S-PLUS for univariate GARCH and CCC models. It is of course also admissible to use the volatility parameter estimates from Table 4.4 and/or the constant correlation matrix estimate from above for $\overline{Q}$ instead of the sample covariance matrix. Doing so yields the very similar estimates presented in Table 4.6.

4.4.3 Time-varying conditional correlation

The in contrast to CCC now time varying conditional correlation is shown in Figure 4.11 with the constant conditional correlation from above indicated by a horizontal line. Although the conditional cross correlation between ATX and DAX

AR(1)-DCC(1,1)	value	std. error	t value	p value
α_1	0.01381020	0.00226884	6.086892	$< 10^{-8}$
β_1	0.96461046	0.00004605	20946.275590	$< 10^{-8}$

Table 4.6: Estimated MGARCH-DCC coefficients with $\overline{Q}$ equaling the CCC matrix given in (4.3.3). Student t innovation distribution (estimated degrees of freedom 9.3063 [0.8483]) is chosen for component series volatilities, Gaussian innovation distribution for modeling the conditional correlation. Values for component series volatilities can be found in Table 4.4.

usually fluctuates around 0.5, times of higher correlation around 1998 as well as times of lower correlation in the first years of the new millennium can be spotted. For visualization purposes a smoothed version of the correlation is also drawn.[8]

The DCC cross correlogram of the standardized residuals shows no noticeable difference to CCC in Figure 4.9.

4.5 DVEC Approach

Goal of this section will be the application of diagonal VEC models to the bivariate ATX-DAX log-returns time series.

4.5.1 Specifying the model

Formulations (3.13), (3.14) and (3.15) as well as combinations thereof allow many different possible fitting approaches. An excerpt of some AIC, BIC and likelihood values of selected models can be found in Table 4.7. In this table,

- "mat" stands for a parametrization of type $M_i = M_i^{1/2}(M_i^{1/2})'$,
- "vec" means that $M_i = \boldsymbol{m}_i\boldsymbol{m}_i'$, and
- "scalar" connotes $M_i = m_i I_d$.

[8]Smoothed time series obtained by lowess scatter plot smoothing as implemented in S-PLUS. Lowess uses robust locally linear fits. A window, sized 1/8 of the time series length, is placed about each time series value; points that are inside the window are weighted so that nearby points get the most weight (from S-PLUS [18] help).

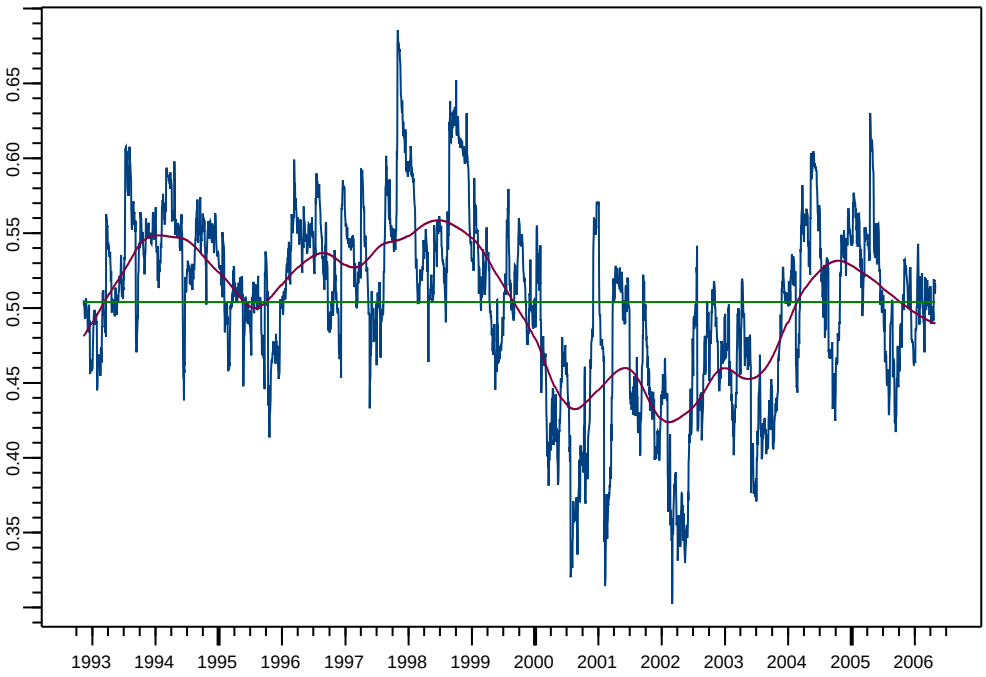

Figure 4.11: Constant and dynamic conditional correlation between ATX and DAX log returns as estimated by AR(1)-DCC(1,1). DCC is additionally visualized with a smoothed time series.

The model "mat.vec" for instance stands for a multivariate GARCH model with the following covariance matrix formulation:

$$\Sigma_t = A_0^{1/2}(A_0^{1/2})' + \sum_{i=1}^{p} A_i^{1/2}(A_i^{1/2})' \odot (\boldsymbol{X}_{t-i}\boldsymbol{X}'_{t-i}) + \sum_{j=1}^{q} \boldsymbol{b}_j\boldsymbol{b}'_j \odot \Sigma_{t-j}.$$

Please note that in all models A_0 is decomposed by $A_0^{1/2}(A_0^{1/2})'$.

When looking at Table 4.7, it appears that allowing full matrices to model B_j is unnecessary; vector or even scalar terms suffice here. For modeling A_i, however, the dynamics of "matrices" or at least "vectors" is necessary. As with CCC/DCC, already low order models capture volatility effects well, lag-lengths of one or two are

model	order	AIC	BIC	likelihood	parameters
scalar.scalar	**(1,1)**	**-42570.87**	**-42509.79**	21295.44	**10**
	(1,2)	-42568.91	-42501.72	21295.45	11
	(2,1)	-42568.68	-42501.50	21295.34	11
	(2,2)	-42270.64	-42197.34	21147.32	12
scalar.vec	**(1,1)**	**-42594.07**	**-42526.88**	21308.03	**11**
	(1,2)	-42592.38	-42512.98	21309.19	13
	(2,1)	-42591.89	-42518.59	21307.94	12
	(2,2)	-42272.97	-42187.46	21150.49	14
scalar.mat	**(1,1)**	**-42598.60**	**-42525.30**	21311.30	**12**
	(1,2)	-42270.58	-42178.97	21150.29	15
	(2,1)	-42596.70	-42517.30	21311.35	13
	(2,2)	-42268.90	-42171.18	21150.45	16
vec.scalar	(1,1)	-42586.15	-42518.96	21304.07	11
	(1,2)	-42582.74	-42509.45	21303.37	12
	(2,1)	-42614.22	**-42534.82**	21320.11	**13**
	(2,2)	**-42614.41**	-42528.90	21321.20	**14**
vec.vec	(1,1)	-42592.80	-42519.51	21308.40	12
	(1,2)	-42593.19	-42507.68	21310.59	14
	(2,1)	**-42614.81**	**-42529.30**	21321.40	**14**
	(2,2)	-42267.54	-42169.81	21149.77	16
vec.mat	**(1,1)**	**-42559.95**	**-42480.55**	21292.97	**13**
	(1,2)	-42272.19	-42174.46	21152.09	16
	(2,1)	-42545.45	-42453.83	21287.72	15
	(2,2)	-42263.48	-42153.54	21149.74	18
mat.scalar	**(1,1)**	**-42610.93**	**-42537.64**	21317.47	**12**
	(1,2)	-42607.67	-42528.27	21316.83	13
	(2,1)	-42365.67	-42274.06	21197.84	15
	(2,2)	-42552.12	-42454.40	21292.06	16
mat.vec	**(1,1)**	**-42614.23**	**-42534.83**	21320.12	**13**
	(1,2)	-42599.97	-42508.35	21314.99	15
	(2,1)	-42335.67	-42237.95	21183.84	16
	(2,2)	-42261.83	-42151.89	21148.91	18
mat.mat	**(1,1)**	**-42552.06**	**-42466.55**	21290.03	**14**
	(1,2)	-42272.26	-42168.43	21153.13	17
	(2,1)	-42300.75	-42196.91	21167.37	17
	(2,2)	-42256.88	-42134.73	21148.44	20

Table 4.7: AIC, BIC and likelihood values of different AR(1)-DVEC(p, q) models with Student t innovations. Best models according to AIC and BIC are indicated in bold font.

appropriate.[9] According to likelihood criteria, "vec.scalar" and "vec.vec" of order (2,1) as well as "mat.vec" of order (1,1) perform reasonably well when fitted to the data. They all have a total of 13 or 14 parameters. Also, "mat.scalar" of order (1,1) with only 12 parameters admits high likelihood values. Generally speaking, it can be said that those models with more parameters also work adequately with low order approaches, whereas comparably parsimonious models tend toward larger lags.

4.5.2 The "mat.scalar" model

Estimated parameters for the fitted AR(1)-DVEC(1,1) model that restricts B_1 to be scalar with asymptotic standard errors and t statistics can be found in Table 4.8.

AR(1)-DVEC(1,1) mat.scalar	value	std. error	t value	p value
c_1	0.00072540	0.00014959	4.849339	0.00000065
c_2	0.00087019	0.00017714	4.912533	0.00000047
ϕ_1	0.09253095	0.01526954	6.059840	$< 10^{-8}$
ϕ_2	-0.02613145	0.01638065	-1.595263	0.05537438
$a_{0,11}$	0.00146003	0.00013071	11.169660	$< 10^{-8}$
$a_{0,12}$	0.00103090	0.00011824	8.718653	$< 10^{-8}$
$a_{0,22}$	0.00070481	0.00019488	3.616607	0.00015147
$a_{1,11}$	0.22133182	0.01130271	19.582199	$< 10^{-8}$
$a_{1,12}$	0.20256135	0.01626014	12.457542	$< 10^{-8}$
$a_{1,22}$	0.15787331	0.01657400	9.525361	$< 10^{-8}$
b_1	0.92655579	0.00664735	139.387271	$< 10^{-8}$
ν	9.528223	0.8972094		

Table 4.8: Parameter estimates with standard errors and t statistics for AR(1)-DVEC(1,1) of type "mat.scalar". Last row displays estimated degrees of freedom for the t innovation distribution.

[9] AIC/BIC model comparison for larger lag-lengths than shown in Table 4.7 has been conducted but showed very little if any improvement.

4.5.3 Time-varying conditional correlation

In Figure 4.12, the estimated conditional and unconditional cross correlation is visualized. In comparison to Figure 4.11, it can be observed that this estimate is rougher and fluctuates more - going down to 0.1 or even less around 2001/2002. The smoothed time series, however, shows similar structure.

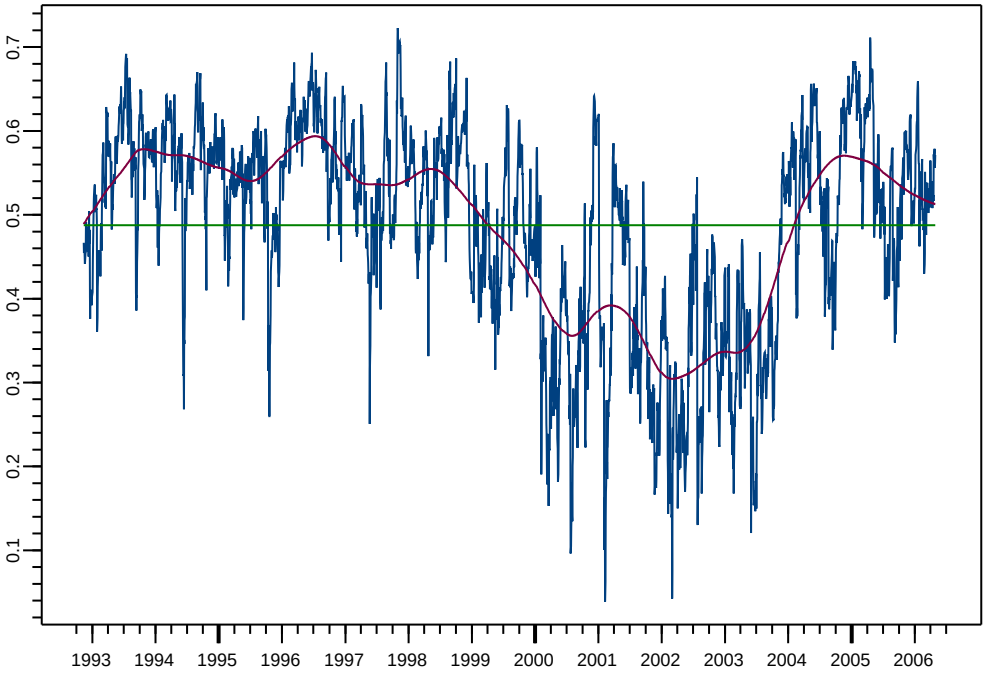

Figure 4.12: Conditional correlation between ATX and DAX log returns as estimated by AR(1)-DVEC(1,1) of type "mat.scalar" with a smoothed version. Unconditional cross correlation estimate is indicated by a horizontal line.

4.5.4 Model checking

For means of model checking, the cross correlogram of the raw and absolute standardized residuals is displayed in Figure 4.13. It is similar to that of the CCC residuals in Figure 4.9, and the above remarks apply accordingly.

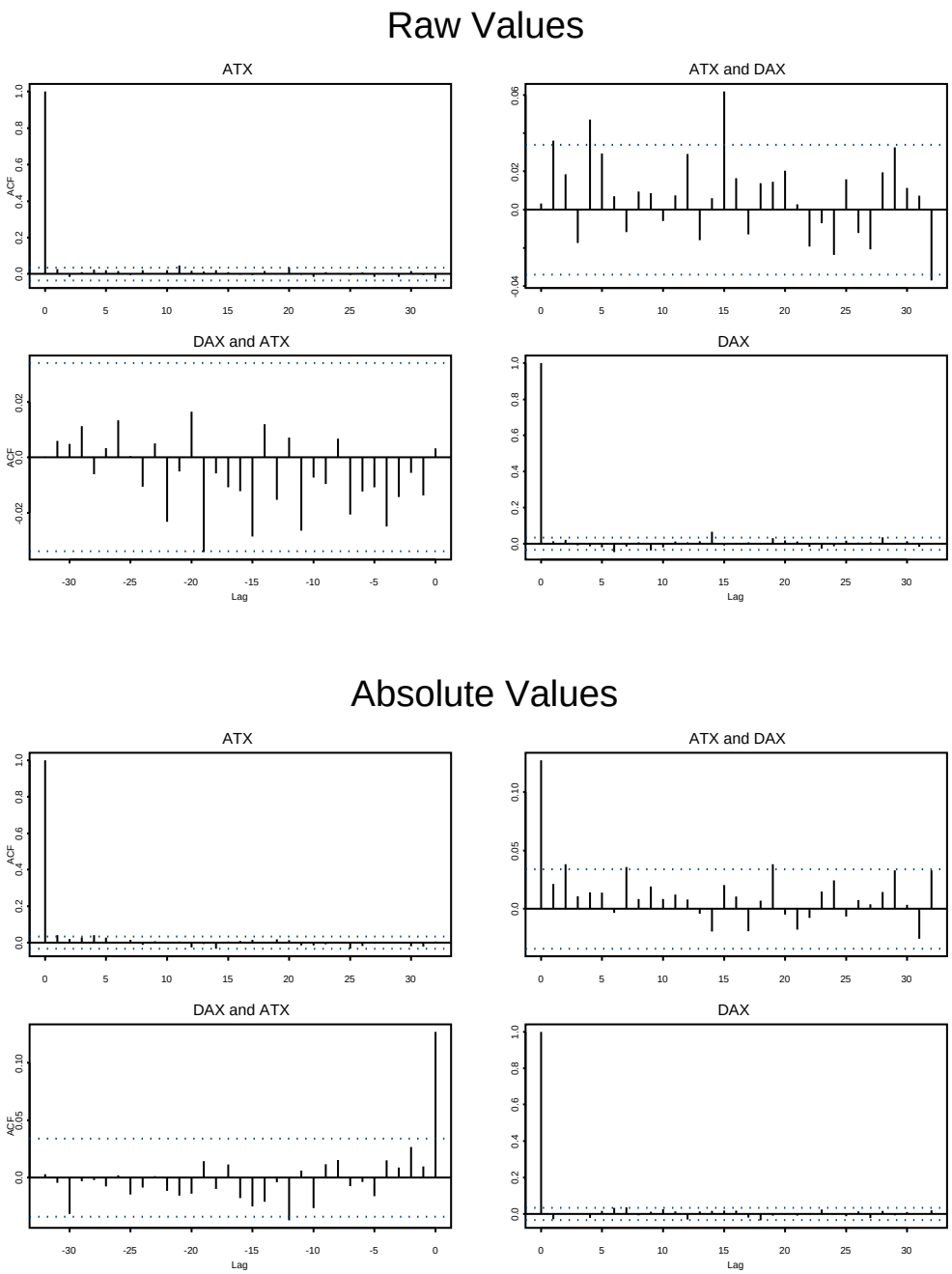

Figure 4.13: Cross Correlograms of standardized residuals after fitting an AR(1)-DVEC(1,1) model of type "mat.scalar" with Student t innovations. Top picture shows raw values, bottom picture shows absolute values.

4.6 BEKK Approach

The last model class that fitted to the data is the BEKK one with $K = 1$. Recall that the conditional covariance matrix Σ_t in this model is parameterized by

$$\Sigma_t = A_0 A_0' + \sum_{i=1}^{p} A_i \boldsymbol{X}_{t-i} \boldsymbol{X}_{t-i}' A_i' + \sum_{j=1}^{q} B_j \Sigma_{t-j} B_j'.$$

In contrast to the above models, this parameterization allows lagged values of one component series' volatility to directly effect the volatility of the second component series. However, for the bivariate BEKK(1,1) model, flexibility is achieved at the cost of two extra parameters in comparison to the full DVEC model of the same order.

4.6.1 Lag-length selection

(p, q)	AIC	BIC	likelihood	parameters
(1,0)	-41652.14	-41578.84	20838.07	12
(1,1)	-42594.28	-42496.55	21313.14	16
(2,0)	-41921.17	-41823.45	20976.59	16
(3,0)	-42052.87	-41930.71	21046.43	20
(2,1)	**-42645.14**	**-42522.99**	21342.57	**20**
(1,2)	-42611.91	-42489.75	21325.95	20
(1,3)	-42584.85	-42438.26	21316.43	24
(3,1)	-42637.57	-42490.98	21342.78	24
(2,2)	-42251.54	-42104.95	21149.77	24
(2,3)	-42243.56	-42072.54	21149.78	28
(3,2)	-42243.73	-42072.71	21149.86	28
(3,3)	-42235.75	-42040.30	21149.87	32

Table 4.9: AIC, BIC and likelihood values of different AR(1)-BEKK(p, q) models with Student t innovations. Best models according to AIC and BIC are indicated in bold font.

Model comparison may again be performed using Akaike or Bayes information criteria, the results are displayed in Table 4.9. It can be seen that even for low-order models the number of parameters is already comparably large, but due to the

large dataset certainly justifiable. A direct comparison of likelihood-related values with other models indicates that this model might provide a better fit, especially for $p = 2$ and $q = 1$. Section 4.7 will provide further insight here by comparing models according to their out-sample performance.

4.6.2 AR(1)-BEKK(2,1)

The parameters of one selected model, AR(1)-BEKK(2,1), are presented in Table 4.10. It has already 20 parameters including that of the innovation distribution, but apparently not all of them are significantly different from zero - it may be appropriate to set some parameters to zero and re-estimate the model or simply use a lower order model.

AR(1)-BEKK(2,1)	value	std. error	t value	p value
c_1	0.00071072	0.00014927	4.761202	0.0000010
c_2	0.00090894	0.00017445	5.210207	0.0000001
ϕ_1	0.09806248	0.01549255	6.329656	$< 10^{-8}$
ϕ_2	-0.02634346	0.01459442	-1.805036	0.0355799
$a_{0,11}$	0.00183625	0.00022130	8.297706	$< 10^{-8}$
$a_{0,21}$	0.00099985	0.00029887	3.345395	0.0004154
$a_{0,22}$	0.00087351	0.00016910	5.165802	0.0000001
$a_{1,11}$	0.26245554	0.02179329	12.042953	$< 10^{-8}$
$a_{1,21}$	0.12858682	0.03787498	3.395033	0.0003471
$a_{1,12}$	0.00638890	0.01751830	0.364699	0.3576798
$a_{1,22}$	0.01531479	0.05886504	0.260168	0.3973753
$a_{2,11}$	0.03741888	0.05444952	0.687221	0.2459957
$a_{2,21}$	-0.09872095	0.03564027	-2.769927	0.0028191
$a_{2,12}$	0.05923639	0.01474170	4.018288	0.0000300
$a_{2,22}$	0.30557655	0.02166363	14.105511	$< 10^{-8}$
$b_{1,11}$	0.94303484	0.00916175	102.931753	$< 10^{-8}$
$b_{1,21}$	-0.01123108	0.01386314	-0.810140	0.2089591
$b_{1,12}$	-0.00871062	0.00414487	-2.101540	0.0178345
$b_{1,22}$	0.95513583	0.00622612	153.407877	$< 10^{-8}$
ν	9.888383	0.9514615		

Table 4.10: Parameter estimates with standard errors and t statistics for an AR(1)-BEKK(2,1) model. Last row displays estimated degrees of freedom for the t innovation distribution.

4.6.3 Time-varying cross correlation

In Figure 4.14, the evolvement of estimated cross correlation over time is visualized. It again differs slightly from the cross correlation estimated before, admitting some higher peaks which go up to as high as 0.8. Nevertheless, times of lower correlation cannot be observed as strikingly as in Figure 4.12. As expected, the general structure resembles that of Figure 4.11 and Figure 4.12.

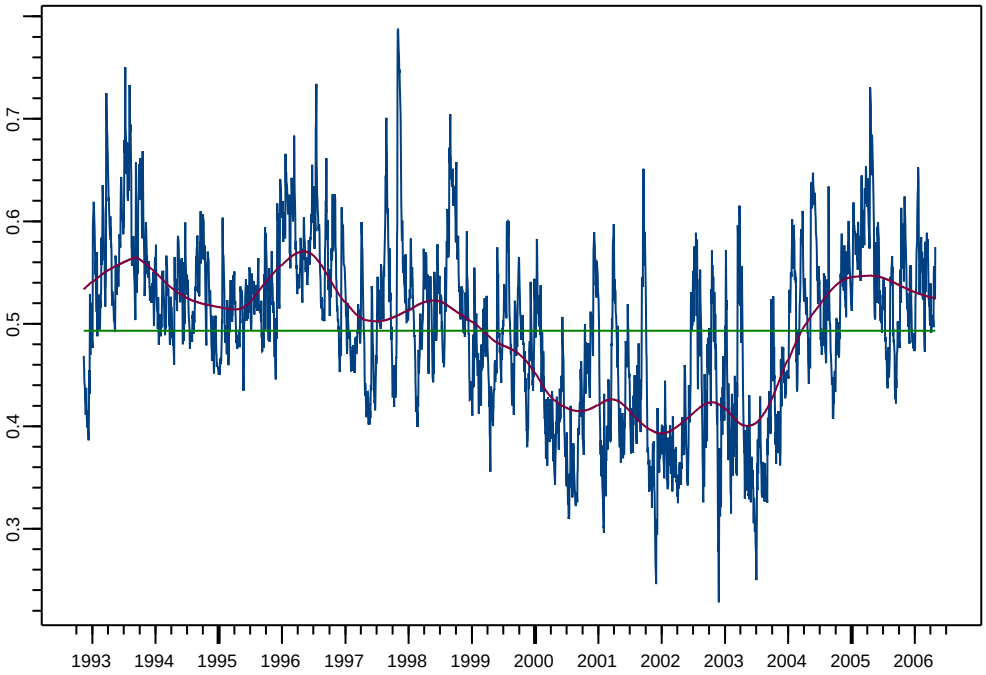

Figure 4.14: Conditional correlation between ATX and DAX log returns as estimated by AR(1)-BEKK(2,1) with a smoothed version. Unconditional cross correlation estimate is indicated by a horizontal line.

4.6.4 Model checking

The cross correlogram of absolute standardized residuals displayed in Figure 4.15 shows some slight improvement to that of the DVEC model in Figure 4.13 in modeling the lead-lag effect of DAX onto ATX partly due to an increase in lag-length, but also due to the less parsimonious model choice.

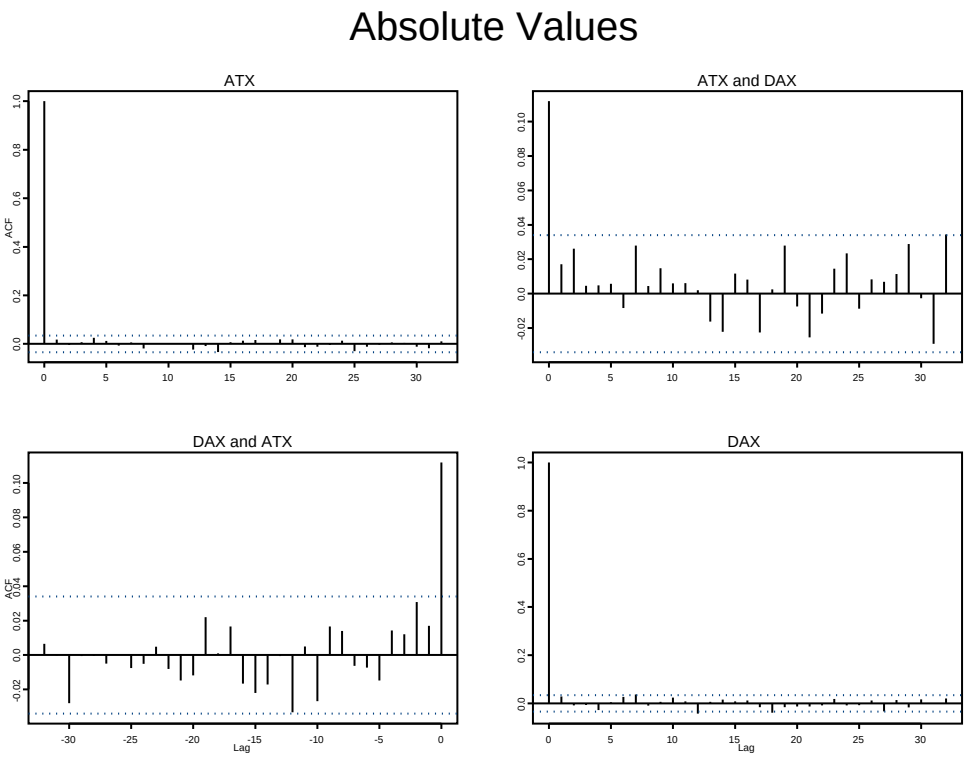

Figure 4.15: Cross correlogram of absolute standardized residuals.

4.7 Empirical Model Performance

The aim of this final section will be a direct comparison of the presented models by in- and out-sample prediction performance measures.

4.7.1 Motivation

Recalling introductory chapters, the conditional covariance matrix in the multivariate framework is given by

$$\mathrm{cov}(\boldsymbol{X}_t|\mathcal{F}_{t-1}) = \mathrm{cov}(\boldsymbol{Z}_t|\mathcal{F}_{t-1}) = \Sigma_t,$$

with Σ_t being $\mathcal{F}_{t-1}$-measurable. In other words, the estimated element $\widehat{\sigma}_{ij,t}$ of the estimated conditional covariance matrix $\widehat{\Sigma}_t$ predicts the conditional covariance between $Y_{i,t}$ and $Y_{j,t}$ for $i, j = 1 \ldots, d$. Hence, natural performance measures used to compare different estimated conditional covariance matrices may be defined for instance by

$$\begin{aligned}
\mathrm{MAE}_{ij} &= \frac{1}{n-n_0}\sum_{t=n_0}^{n} |(Y_{i,t} - \widehat{\mu}_{i,t})(Y_{j,t} - \widehat{\mu}_{j,t}) - \widehat{\Sigma}_{ij,t}|, && (4.2)\\
\mathrm{MSE}_{ij} &= \frac{1}{n-n_0}\sum_{t=n_0}^{n} ((Y_{i,t} - \widehat{\mu}_{i,t})(Y_{j,t} - \widehat{\mu}_{j,t}) - \widehat{\Sigma}_{ij,t})^2, && (4.3)\\
\mathrm{HMSE}_{ij} &= \frac{1}{n-n_0}\sum_{t=n_0}^{n} \left(\frac{(Y_{i,t} - \widehat{\mu}_{i,t})(Y_{j,t} - \widehat{\mu}_{j,t})}{\widehat{\Sigma}_{ij,t}} - 1\right)^2, && (4.4)
\end{aligned}$$

where $\widehat{\mu}_t$ denotes the vector of estimated means, n_0 is an arbitrary starting point and $i, j = 1, \ldots, d$. Total average errors are given by summing up and normalizing, i.e.

$$\begin{aligned}
\mathrm{MAE} &= \frac{1}{d^2}\sum_{i,j=1}^{d} \mathrm{MAE}_{ij},\\
\mathrm{MSE} &= \frac{1}{d^2}\sum_{i,j=1}^{d} \mathrm{MSE}_{ij},\\
\mathrm{HMSE} &= \frac{1}{d^2}\sum_{i,j=1}^{d} \mathrm{HMSE}_{ij}.
\end{aligned}$$

Note that these performance measures not only incorporate performance of conditional covariances, but also performance of conditional means. In contrast to the

mean absolute error (MAE), which penalizes deviation linearly, the *mean squared error* (MSE) weights large deviation somewhat stronger. Clearly, this makes MAE more robust to outliers than MSE. The third error measure proposed, the *heteroscedasticity-adjusted MSE* (HMSE), penalizes deviations heavier in times of low volatility (Schöftner [23]).[10]

Obviously, these statistical loss functions can be applied regardless of the distributional assumption to in- and out-sample testing. Referring to the underlying data set, dimension d equals 2 and the number of observations n equals 3321.

4.7.2 Model comparison by likelihood related values

Before turning to empirical measurement, the AIC, BIC, likelihood and parameter values for the models selected above are summarized again in Table 4.11.

	AIC	BIC	likelihood	parameters
CCC(1,1)	-42586	-42519	21304	10
mat.scalar DVEC(1,1)	-42611	-42538	21317	12
BEKK(1,2)	-42645	-42523	21343	20

Table 4.11: AIC, BIC, likelihood values and number of parameters for selected models. DCC is not directly comparable because of the stepwise estimation procedure and is therefore not included.

As expected, the more nested models achieve higher likelihood values. However, when taking the number of parameters into account, the more complicated model is not automatically the better anymore. Likelihood values of DVEC are 0.12% smaller than those of BEKK, while CCC values are about 0.18% behind.

4.7.3 In-sample measurement

One possible approach to quantifying prediction quality is *in-sample* measurement. In the underlying case, this means that:

[10] The magnitude of different HMSE_{ij}'s may differ significantly since it measures the percentage squared error. The total sum can mainly be influenced by one high value.

- Parameters are being determined once by fitting a specified model to the entire dataset.
- For measuring the prediction quality, (4.2), (4.3) and (4.4) are applied.

One striking argument for using in-sample methods is that the above procedure can be applied to (almost) the entire data, many comparisons are available and consequently lead to better estimation of the real errors. On the contrary, one might argue that the parameters are estimated knowing the "future" instead of restricting the estimation to the past. For the underlying investigation, 90% of the data have been used, i.e. $n_0 = 332$.

In the following tables, in-sample performance measures are presented. For mean modeling, a diagonal autoregressive model of order one is used. Table 4.12 shows appropriately scaled MAE_{11}, MSE_{11} and HMSE_{11} for various previously discussed models; in other words, conditional covariance prediction errors for ATX log returns.

	MAE_{11} 10^{-4}	MSE_{11} 10^{-7}	HMSE_{11} 10^{1}
CCC(1,1)	1.017404	0.587623	0.369657
DCC(1,1)	1.021291	0.588538	0.364316
alternative DCC(1,1)	1.017404	0.587623	0.369657
mat.scalar DVEC(1,1)	1.019623	0.592393	0.380945
BEKK(2,1)	**1.017310**	**0.582360**	**0.357916**

Table 4.12: In-sample performance measures for ATX conditional variance.

As can be seen in Table 4.12, AR(1)-BEKK(2,1) is superior in predicting ATX conditional variance according to all three performance measures. However, AR(1)-CCC(1,1) and equivalently AR(1)-DCC(1) are virtually equal, especially when considering the absolute error with a deviation of less than 0.01%. The slightly less parameter-parsimonious mat.scalar DVEC(1,1) model is behind about 0.23% in MAE and 1.72% in MSE. Table 4.13 presents the same numbers for DAX conditional variance.

DAX conditional variance prediction measures listed in Table 4.13 generally show significantly larger values than those of ATX. Only HMSE values are smaller, but

	MAE_{22} 10^{-4}	MSE_{22} 10^{-7}	$HMSE_{22}$ 10^{1}
CCC(1,1)	2.286205	1.986459	0.283322
DCC(1,1)	2.300142	1.979073	0.274724
alternative DCC(1,1)	2.286205	1.986459	0.283322
mat.scalar DVEC(1,1)	**2.252521**	1.976215	**0.272338**
BEKK(2,1)	2.255986	**1.960333**	0.280813

Table 4.13: In-sample performance measures for DAX conditional variance.

do not allow an equally straightforward interpretation as MAE and MSE. Depending on whether focus is laid onto absolute or squared errors, DVEC(1,1) or BEKK(2,1) exhibit smallest errors. Generally it can be said that when measuring conditional variance only, simple models such as CCC perform quite well. Considering the absolute error, CCC is about 1.5% behind DVEC, looking at squared errors the difference amounts to 0.52%.

Turning to conditional covariance prediction, this picture changes only slightly. In-sample performance measure results can be found in Table 4.14.

	MAE_{12} 10^{-4}	MSE_{12} 10^{-7}	$HMSE_{12}$ 10^{1}
CCC(1,1)	1.081838	0.530029	0.722084
DCC(1,1)	1.082260	0.523735	**0.710118**
alternative DCC(1,1)	1.079954	0.524524	0.723057
mat.scalar DVEC(1,1)	**1.063635**	0.524700	1.079326
BEKK(2,1)	1.065886	**0.522767**	0.755547

Table 4.14: In-sample performance measures for conditional covariance.

When considering absolute or squared errors, Table 4.14 gives some evidence that more nested models such as DVEC or BEKK capture conditional covariance better than simpler models. Constant correlation assumption is now outperformed somewhat clearer (1.71% in MAE and 1.02% in MSE).

For a last overview, the overall in-sample performance measures are presented in Table 4.15, which are simply created by averaging the (weighted) measures from above.

	MAE 10^{-4}	MSE 10^{-7}	HMSE 10^{1}
CCC(1,1)	1.366822	0.908535	0.524287
DCC(1,1)	1.371488	0.903770	**0.514819**
alternative DCC(1,1)	1.365879	0.905783	0.524773
mat.scalar DVEC(1,1)	**1.349854**	0.904502	0.702984
BEKK(2,1)	1.351267	**0.897057**	0.537456

Table 4.15: Overall in-sample performance measure results.

4.7.4 Out-sample measurement

Another way of measuring prediction quality is *out-sample* prediction. Here this means that

- The data is split into "past" and "future" with splitting point n_0. In order to determine the parameters, a model is fit to the "past" part of the data.
- The one-step ahead prediction is evaluated as before.
- Steps one and two are iterated until the end of the dataset is reached. The splitting point is increased by one in every step.

Obviously, this method is computationally much more expensive, since a new model has to be fitted in every iteration step.[11] Alternatively and computationally relatively inexpensive, out-sample measuring can be performed by re-using parameter estimates obtained at the first splitting point n_0. This method yields very similar results to the in-sample performance measures displayed in Table 4.12 to 4.15.

For the underlying data $n_0 = 2656$, i.e. at least 80 percent of the data are used for obtaining parameter estimates.

Similar to in-sample measurement, Table 4.16 shows that AR(1)-BEKK(2,1) is slightly ahead in ATX conditional variance prediction. According to absolute errors, it is 0.87% better than CCC and 0.56% better than DVEC. Considering

[11] Because fitting DCC is hand-made and not numerically and computationally optimized, this method is not feasible for DCC as implemented. DCC out-sample performance measure values were not obtained.

	MAE_{11} 10^{-4}	MSE_{11} 10^{-7}	$HMSE_{11}$ 10^{1}
CCC(1,1)	0.777158	0.201629	**0.298885**
mat.scalar DVEC(1,1)	0.772804	0.200974	0.306884
BEKK(2,1)	**0.770483**	**0.200417**	0.304465

Table 4.16: Out-sample performance measures for ATX conditional variance.

squared errors, this effect reduces to 0.60% and 0.20%, respectively. Paying attention to HMSE, CCC performs better by 2.68% in comparison to DVEC and 1.87% in comparison to BEKK. In Table 4.17, these numbers can be found for DAX conditional variance.

	MAE_{22} 10^{-4}	MSE_{22} 10^{-7}	$HMSE_{22}$ 10^{1}
CCC(1,1)	1.043066	0.271602	0.226310
mat.scalar DVEC(1,1)	1.049171	0.271139	**0.208190**
BEKK(2,1)	**1.031579**	**0.267115**	0.237015

Table 4.17: Out-sample performance measures for DAX conditional variance.

Again, DAX conditional variance prediction measures listed in Table 4.17 exhibit larger values than those of ATX. AR(1)-BEKK(2,1) performs best when focusing on absolute errors with 1.11% higher accuracy in comparison to CCC and 1,70% compared to DVEC. Squared errors portrait a similar picture: BEKK is 1.68% ahead of CCC and 0.17% ahead of DVEC.

	MAE_{12} 10^{-4}	MSE_{12} 10^{-7}	$HMSE_{12}$ 10^{1}
CCC(1,1)	**0.631214**	**0.110458**	0.600891
mat.scalar DVEC(1,1)	0.648044	0.110909	**0.527230**
BEKK(2,1)	0.643572	0.110856	0.584268

Table 4.18: Out-sample performance measures for conditional covariance.

Interestingly, Table 4.18 shows that when observing predicted conditional covariance, AR(1)-CCC(1,1) performs best. In comparison to DVEC, the accuracy is 2.67% higher in MAE and 0.41% higher in MSE. Compared to BEKK, it is still better by 1.96% in MAE and 0.36% in MSE. Just like before, HMSE "favours" DVEC.

Finally, in Table 4.19 the overall measure of out-sample prediction performance is displayed. No clear evidence for a "best" model can be spotted.

	MAE 10^{-4}	MSE 10^{-7}	HMSE 10^{1}
CCC(1,1)	**0.770663**	0.173536	0.431744
mat.scalar DVEC(1,1)	0.779516	0.173483	**0.392383**
BEKK(2,1)	0.772301	**0.172311**	0.427504

Table 4.19: Overall out-sample performance measure results.

Appendix A

S-PLUS Source Code

Appendix A contains used S-PLUS source code. The files used in this thesis are presented and described in order of their usage within the thesis. In S-PLUS comments are written in the "# comment" style.

A.1 firstsim.scc

Simple S-PLUS code used to create Figures 1.1, 1.2 and 1.3.

```
# Simulate and visualize time series as defined by (1.2)
# Innovation distribution: normal, t

set.seed(11)
n <- 250
t.innov <- sqrt(1/3)*rt(n+10, df=3)
g.innov <- rnorm(n+10, sd=1)
x <- y <- matrix(nrow = n, ncol = 2)

x[,1] <- g.innov[11:(n+10)]
x[,2] <- g.innov[11:(n+10)] + .75*g.innov[1:n]
y[,1] <- t.innov[11:(n+10)]
```

```
y[,2] <- t.innov[11:(n+10)] + .75*t.innov[1:n]
dimnames(x) <- list(NULL, c("Component series 1", "Component series 2"))
dimnames(y) <- list(NULL, c("Component series 1", "Component series 2"))

# Visualize the time series itself
par(mfrow=c(4,1))
tsplot(x[,1], main="Gaussian innovations: Component series 1", col=1)
tsplot(x[,2], main="Gaussian innovations: Component series 2", col=1)
tsplot(y[,1], main="t innovations: Component series 1", col=1)
tsplot(y[,2], main="t innovations: Component series 2", col=1)

# Visualizing the ACF
x.acf <- acf(x, plot=F)
y.acf <- acf(y, plot=F)
par(mfrow=c(1,1))
acf.plot(x.acf, main="")
acf.plot(y.acf, main="")
```

A.2 garchfunctions.scc

Some functions used in univariate GARCH contexts. Using these, all simulations in Chapter 2 can easily be conducted. For one example, please see A.3.

```
# Function to get the logarithmic values of given data, whereas
# also negative values are possible. Given data in modulus
# smaller than 1 is assigned zero. Used for visualization.
logit <- function(x) {
 x[-1 < x & x < 1] <- 1
 x[x<0] <- -log(-x[x<0])
 x[x>0] <- log(x[x>0])
 return(x)
}

# Similar to above. Negative values are now assigned
```

```
# the reciprocal in modulus. Used for visualization.
log.prepare <- function(x) {
 x[-1 < x & x < 1] <- 1
 x[x<0] <- -(1/x[x<0])
 return(x)
}

# Function to simulate an ARCH(1) process x with conditional
# variance sigma.
arch1sim <- function(alpha0, alpha1, lead = 10, innov = rnorm(510)) {
 tslength = length(innov)
 x <- vector(length = tslength)
 sigma <- vector(length = tslength)
 x[1] <- 0
 sigma[1] <- 0
 for (i in 2:tslength) {
  sigma[i] <- sqrt(alpha0 + alpha1 * (x[i-1])^2)
  x[i] <- sigma[i]*innov[i]
 }
 x <- x[-(1:lead)]
 sigma <- sigma[-(1:lead)]
 return(cbind(x, sigma))
}

# Function to simulate a GARCH(1,1) process x with conditional
# variance sigma.
garch11sim <- function(alpha0, alpha1, beta, lead = 10,
                       innov = rnorm(510)) {
 tslength = length(innov)
 x <- vector(length = tslength)
 sigma <- vector(length = tslength)
 x[1] <- 0
 sigma[1] <- 0
 for (i in 2:tslength) {
  sigma[i] <- sqrt(alpha0 + alpha1*(x[i-1])^2 + beta*sigma[i-1]^2)
  x[i] <- sigma[i]*innov[i]
```

```
 }
 x <- x[-(1:lead)]
 sigma <- sigma[-(1:lead)]
 return(cbind(x, sigma))
}

# Calculates kurtosis of a GARCH(1,1) process.
garch11.kurtosis <- function(alpha1, beta, kappa.z) {
 kappa.x <- (kappa.z*(1-(alpha1+beta)^2))/(1-(alpha1+beta)^2 -
            (kappa.z-1)*alpha1^2)
 return(kappa.x)
}

# Calculates real autocorrelation of squared GARCH(1,1) process, made
# possible by ARMA representation. See chapter 2.2.4.
garch11squared.acf <- function(alpha1, beta, x) {
 phi <- alpha1 + beta
 theta <- -beta
 y <- (phi^(x-1)*(phi + theta)*(1+phi*theta))/(1+theta^2+2*phi*theta)
 return(y)
}
```

A.3 arch1sim.scc

Functioning as an example, the script producing Figure 2.1 is presented here. It needs (some of) the above code in `garchfunctions.scc`.

```
# Produces Figure 2.1, uses garchfunctions.scc
set.seed(11)
n <- 500
lead <- 10
n.innov <- rnorm(n+lead)
alpha0 <- .1
alpha1 <- .5
```

```
simresult <- arch1sim(alpha0, alpha1, lead, innov = n.innov)

par(mfrow = c(2,2))
tsplot(simresult[, "x"], main = "")
tsplot(simresult[, "sigma"], type = "h", main = "")

x <- (0:200)/10
y <- alpha1^x
plot(acf(simresult[, "x"], plot=FALSE, lag.max=20), main="")
plot(acf(simresult[, "x"]^2, plot=FALSE, lag.max=20), main="")
lines(x,y, lty = 4)
```

A.4 atxdax.scc

After importing the stock return data obtained as "Comma Separated Value" (.csv) files from YAHOO!, the following code needs to be executed in order to be able to use the functionality of the specially designed `timeSeries` objects. It also illustrates the use of its high-level plotting function, `seriesPlot()`.

```
# Obtaining, converting and displaying data for ATX and DAX,
# .raw objects are imported DataFrames from .csv-files.

# converts DataFrames to timeSeries objects
toTimeSeries <- function(ts) {
 class(ts[,"Date"]) <- "character"
 td <- timeDate(ts[,"Date"], in.format="%d-%m-%y",
                format="%a %b %d, %Y")
 res <- timeSeries(pos=rev(td), data=rev(ts[,"Close"]))
}

atx <- toTimeSeries(atx.raw)      # toTimeSeries converts DataFrames to
dax <- toTimeSeries(gdaxi.raw)    # timeSeries objects
x.merged <- seriesMerge(atx,dax)  # merging the series to one multivar
```

```
colIds(x.merged) <- c("ATX", "DAX")
x.merged@title <- "ATX and DAX"
x.merged@documentation <- "Daily data for ATX and DAX (for days on which
                           all were available). Data by YAHOO!"

x <- getReturns(x.merged)         # calculating the log-return
x@title <- "Log-Returns of ATX and DAX"
x@documentation <- "Daily Log-Returns of ATX and DAX (for days on which
                    both were available). Data obtained through YAHOO!"

par(mfrow =c(1,1))
plot(x.merged, main = "")
seriesPlot(seriesMerge(x[,"DAX"],x[,"ATX"],), one.plot = F,
                                               strip.text=colIds(x))
plot(acf(x, plot = F), main = "")
plot(acf(x^2, plot = F), main = "")
```

A.5 ccc.scc

The following code illustrates exemplary how AIC/BIC tables and QQ-plots as in Figure 4.10 may easily be created. Also, conditional volatility and standardized residuals are plotted over time. In this case this is done for the CCC model.

```
i.length <- 3
j.length <- 3
ic <- data.frame(sort(rep(0:j.length, i.length+1)),
                 rep(0:j.length, i.length+1),0,0,0,0)
names(ic) <- c("p","q","AIC","BIC","likelihood","parameters")

for (i in 0:i.length) {
 for (j in 0:j.length) {
  if (i == 0 & j == 0) next
  tmp <- mgarch(x~ar(1), ~ccc(i,j), trace=F)
  ic[(i*(j.length+1))+j+1,c(3,4,6)] <- summary(tmp)$abic[c(2,3,1)]
```

```
  ic[(i*(j.length+1))+j+1,5] <- tmp$likelihood
 }
}
ic[1,] <- c(0,0,NA,NA,NA,NA)

ccc11t <- mgarch(x~ar(1), ~ccc(1,1), trace=F, cond.dist="t")

par(mfrow = c(1,1))
seriesPlot(ccc11t$sigma.t,one.plot=F,strip.text=colIds(ccc11t$sigma.t))
seriesPlot(resid(ccc11t, standardize = T), one.plot=F,
           strip.text=colIds(ccc11t$std.residuals))

print(qqPlot(resid(ccc11t, standardize = T), ccc11t$cond.dist$dist.par,
      distribution = ccc11t$cond.dist$cond.dist, strip.text =
      c("Component Series 1","Component Series 2"), id.n = 0,
      ylab = "Standardized Residuals", xlab = paste("Quantiles of t
      distribution"), main = ""), position = c(.5,0,1,1), more = T)
```

A.6 dcc.scc

For estimating the DCC model, manual likelihood evaluation needs to be performed, which is done straightforwardly and in a non-optimized way by `like()`. Even though it is possible to use this approach, the code is very slow compared to the pre-compiled Fortran-based likelihood optimization implemented in finmetrics' `MGARCH()` function. After getting an idea about possible local and global minima by placing a grid over the entire parameter space, the built-in minimizer `nlminb()` is used for final optimization. The DCC model is then embedded as a manually fitted `MGARCH` object and conditional correlation is visualized.

```
# for manually evaluating the DCC-likelihood-function
like <- function(para, qbar, x) {
 alpha <- para[1]
 beta <- para[2]
```

```
 c <- 1 - alpha - beta
 if (c <= 0) return(NA)
 q <- array(dim=c(numRows(x)+1, 2, 2))
 q[1,,] <- qbar

 for (i in 2:(numRows(x)+1)) {
  q[i,,] <- c*qbar + alpha*seriesData(x)[i-1,]%*%t(seriesData(x)[i-1,])+
            beta*q[i-1,,]
  }
 q <- q[-1,,]
 normalizer <- sqrt(q[,1,1]*q[,2,2])
 q <- q/normalizer
 q[,1,1] <- 1
 q[,2,2] <- 1
 res <- array(dim=c(numRows(x),2))
 for (i in 1:numRows(x)) {
  res[i,] <- solve(chol(q[i,,]))%*%seriesData(x)[i,]
 }
 tmp <- sum(log((1-q[,1,2]^2)^-.5*dmvnorm(res)))
 return(-tmp)
}

# fit univariate GARCH models and calculate sample correlation
garch11t <- mgarch(x~ar(1), ~garch(1,1), cond.dist = "t", trace=F)
x.devol <- garch11t$std.residuals
qbar <- cor(x.devol)

# grid for getting an idea of how like() looks like for different par
resch <- array(dim=c(50,50))
for (i in 0:49) {
 for (j in 0:49) {
  resch[i,j] <- like(i/50, j/50, qbar, x.devol)
 }
}
# results therof...
# maxi <- .01+6/1000
```

```
# maxj <- .95+11/1000

max.result <- nlminb(c(maxi,maxj), like, lower = c(0,0),
                     upper = c(1,1), x = x.devol, qbar = qbar)
alpha <- max.result$parameter[1]
beta <- max.result$parameter[2]

# manually embed the optimized model in mgarch()
dcc <- mgarch(formula.mean = x.devol~-1,  formula.var =
              ~dvec.scalar.scalar(1,1), cond.dist = "normal", trace=T)
newmodel <- dcc$model
c <- 1 - alpha - beta
newmodel$a.value <- sqrt(c)*t(chol(qbar))
newmodel$arch$value$lag.1 <- alpha
newmodel$garch$value$lag.1 <- beta
# trick for not really estimating any model, but still using mgarch()
dcc <- mgarch(series = x.devol, model = newmodel, trace=F,
              control = bhhh.control(n.iter = -1))

tmp <- timeSeries(dcc$R.t[,1,2], pos = positions(x))
smooth <- lowess(1:length(dcc$R.t[,1,2]), dcc$R.t[,1,2],
                 f = 1/8, delta = 0)$y
tmp2 <- timeSeries(smooth, pos = positions(x))
tmp3 <- timeSeries(rep(qbar[1,2], numRows(x)), pos = positions(x))
tmp <- seriesMerge(tmp, tmp2, tmp3)
plot(tmp, reference.grid=F)
```

A.7 testfunctions.scc

The following code shows how out-sample performance measuring was conducted. Both `MGARCH`'s implemented prediction methods (`outsampletest()`) as well as "manual" prediction (`outsampletest3()`) was used (faster). For different models the code may simply be adjusted accordingly.

```
# Extracts sigma from a fitted MGARCH object
getSigma <- function(model, pos) {
 sigma <- diag(as.vector(model$sigma.t[pos,]))%*%
          matrix(model$R.t[pos,,], nrow=2)%*%
          diag(as.vector(model$sigma.t[pos,]))
}
# Extracts prediction for sigma from a fitted MGARCH object
getSigmap <- function(model, pos=1) {
 sigma <- diag(as.vector(model$sigma.pred[pos,]))%*%
          matrix(model$R.pred[pos,,], nrow=2)%*%
          diag(as.vector(model$sigma.pred[pos,]))
}

outsampletest <- function(x, model = ~dvec.mat.mat(1,1),
                          meanmodel = ~ar(1), innovdist = "t",
                          testsize = numRows(x)/5, ...) {
 samplesize <- numRows(x)
 part.refcov <- part.cov <- matrix(ncol = numCols(x)^2,
                                   nrow =  testsize)

 for (i in testsize:1) {
  part.model <- mgarch(series = x[1:(samplesize-i),],
                       formula.mean = meanmodel, formula.var = model,
                       cond.dist = innovdist, trace = F, ...)
  part.pred <- predict(part.model,1)
  part.cov[i,] <- as.vector(getSigmap(part.pred, 1))
  tmp <- as.vector(x[samplesize-i+1,] - part.pred$series.pred)
  part.refcov[i,] <-  tmp %*% t(tmp)
 }

 mae <- apply(abs(part.refcov - part.cov), 2, mean)
 mse <- apply((part.refcov - part.cov)^2, 2, mean)
 hmae <- apply(abs(part.refcov/part.cov - 1), 2, mean)
 hmse <- apply((part.refcov/part.cov - 1)^2, 2, mean)
 mae.total <- mean(abs(part.refcov - part.cov))
 mse.total <- mean((part.refcov - part.cov)^2)
```

```
 hmae.total <- mean(abs(part.refcov/part.cov-1))
 hmse.total <- mean((part.refcov/part.cov-1)^2)
 res <- data.frame(t(matrix(c(mae, mse, hmae, hmse), nrow=4)),
                   c(mae.total, mse.total, hmae.total, hmse.total),
                   row.names=c("MAE", "MSE", "HMAE", "HMSE"))
}

outsampletest3 <- function(x, model = ~dvec.mat.mat(1,1),
                           meanmodel = ~ar(1), innovdist = "normal",
                           testsize = numRows(x)/5, ...) {
 samplesize <- numRows(x)
 part.refcov <- part.cov <- matrix(ncol = numCols(x)^2, nrow = testsize)

 part.model <- mgarch(series = x[1:(samplesize-testsize)],
                      formula.mean = meanmodel, formula.var = model,
                      cond.dist = innovdist, trace = F, ...)

 A0 <- part.model$model$a.value %*% t(part.model$model$a.value)
 A1 <- part.model$model$arch$value$lag.1 %*%
       t(part.model$model$arch$value$lag.1)
 B1 <- part.model$model$garch$value$lag.1 %*%
       t(part.model$model$garch$value$lag.1)
 c <- part.model$model$c.value
 ARterm <- part.model$model$AR$value$lag.1

 sigma <- getSigma(part.model, samplesize-testsize)
 eps <- as.vector(x[samplesize-testsize,]) - c -
        ARterm %*% as.vector(x[samplesize-testsize-1,])
 eps2 <- eps %*% t(eps)

 for (i in testsize:1) {
  sigma <- A0 + A1*eps2 + B1*sigma
  part.cov[i,] <- as.vector(sigma)
  eps <- as.vector(x[samplesize-i+1,]) - c -
  ARterm %*% as.vector(x[samplesize-i,])
  eps2 <- eps %*% t(eps)
```

```
  part.refcov[i,] <- as.vector(eps2)
 }

 mae <- apply(abs(part.refcov - part.cov), 2, mean)
 mse <- apply((part.refcov - part.cov)^2, 2, mean)
 hmae <- apply(abs(part.refcov/part.cov - 1), 2, mean)
 hmse <- apply((part.refcov/part.cov - 1)^2, 2, mean)
 mae.total <- mean(abs(part.refcov - part.cov))
 mse.total <- mean((part.refcov - part.cov)^2)
 hmae.total <- mean(abs(part.refcov/part.cov-1))
 hmse.total <- mean((part.refcov/part.cov-1)^2)

 res <- data.frame(t(matrix(c(mae, mse, hmae, hmse), nrow=4)),
                   c(mae.total, mse.total, hmae.total, hmse.total),
                   row.names=c("MAE", "MSE", "HMAE", "HMSE"))
}
```

Bibliography

[1] Orazio P. Attanasio. Risk, time-varying second moments and market efficiency. *Review of Economic Studies*, 58:479–494, 1991.

[2] Heinz Bauer. *Wahrscheinlichkeitstheorie*. De Gruyter Lehrbuch, 4th edition, 1991.

[3] Anil K. Bera and Matthew L. Higgins. A survey of ARCH models: Properties, estimation and testing. *Journal of Economic Surveys 7(4)*, 1993.

[4] Ernst K. Berndt, Bronwyn H. Hall, Robert E. Hall, and Jerry A. Hausman. Estimation and inference in nonlinear structural models. *Annals of Economic and Social Measurement*, 4:653–665, 1974.

[5] Tim Bollerslev. Generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics*, 31:307–327, 1986.

[6] Philippe Bougerol and Nico Picard. Stationarity of GARCH processes and of some negative time series. *Journal of Econometrics*, 52:115–127, 1992.

[7] A. Brandt. The stochastic equation $Y_{n+1} = A_n Y_n + B_n$ with stationary coefficients. *Advances in Applied Probability*, 18:211–220, 1986.

[8] Peter J. Brockwell and Richard A. Davis. *Time Series: Theory and Methods*. Springer, 2nd edition, 1998.

[9] Peter J. Brockwell and Richard A. Davis. *Introduction to Time Series and Forecasting*. Springer, 2nd edition, 2002.

[10] Richard Durrett. *Probability: Theory and Examples*. Duxbury Press, 2nd edition, 1995.

[11] Jürgen Elstrodt. *Maß- und Integrationstheorie*. Springer, 3rd edition, 2002.

[12] Robert F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation, 1982. *Econometrica*, 1982.

[13] Robert F. Engle and Kenneth F. Kroner. Multivariate simultaneous generalized ARCH. *Econometric Theory*, 11:122–150, 1995.

[14] Jürgen Franke, Wolfgang Härdle, and Christian Hafner. *Einführung in die Statistik der Finanzmärkte*. Springer-Verlag, Berlin, 2001.

[15] Wayne A. Fuller. *Introduction to Statistical Time Series*. Wiley, 1996.

[16] James D. Hamilton. *Time Series Analysis*. Princeton University Press, 1994.

[17] Insightful Corporation. *S-PLUS Reference Manual*, 2002.

[18] Insightful Corporation. *S-PLUS (version 6.1.2)*, 2002.

[19] Insightful Corporation. *S+FinMetrics (version 1.0)*, 2002.

[20] Alexander J. McNeil, Rudiger Frey, and Paul Embrechts. *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press, 2005.

[21] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2005. ISBN 3-900051-07-0.

[22] Wolfgang Scherrer and Eva Ribarits. On the parametrization of multivariate GARCH models. Unpublished article preprint.

[23] Robert Schöftner. Time-varying dependence modelling of market and credit risk. Master's thesis, Uni/ETH Zürich, 2005.

[24] Neil Shephard. Statistical aspects of ARCH and stochastic volatility. *in: David R. Cox, David V. Hinkley and Ole E. Barndorff-Nielsen (Ed.). Time Series Models in Econometrics, Finance and Other Fields. Chapman & Hall, London*, pages 1–55, 1996.

[25] Adrian Trapletti. `tseries`*: Time series analysis and computational finance*, 2005. R package version 0.10-0.

[26] Ruey S. Tsay. *Analysis of Financial Time Series.* Wiley, 2002.

[27] David Williams. *Probability with Martingales.* Cambridge University Press, 1991.

[28] Eric Zivot and Jiahui Wang. *Modeling Financial Time Series with S-PLUS.* Springer, 2002.