DISSERTATION

Multivariate Modelling of Financial Time Series

ausgeführt zum Zwecke der Erlangung des akademischen Grades eines Doktors der technischen Wissenschaften unter der Leitung von

O.Univ.Prof. Dipl.-Ing. Dr.techn. Manfred DEISTLER

E105–02
Institut für Wirtschaftsmathematik
Forschungsgruppe Ökonometrie und Systemtheorie (EOS)
A–1040 Wien, Argentinierstraße 8

eingereicht an der Technischen Universität Wien
Fakultät für Mathematik und Geoinformation
von

Eva Maria RIBARITS
Matr.Nr.: 9625411
38, Op der Heed
L–1747 Luxembourg

Wien, am 20. März 2006 Eva Ribarits
....

To my parents
Edeltraud and Walter Hamann

Contents

List of Figures v

List of Tables vii

Deutsche Kurzfassung ix

Abstract xi

Acknowledgements xiii

I Forecasting Return Series of Stock Prices 1

1 Introduction 3

2 VARX model 5

2.1 The model 5

2.2 Estimation 6

2.3 Model specification and input selection 14

2.4 Forecasting 16

3 Reduced Rank Regression 19

3.1 The model 19

3.2 Identifiability of F and G 20

3.3 Estimation 21

3.3.1 Maximum likelihood estimators for F and G: 22

3.3.2 Asymptotic properties of the ML estimators 25

3.3.3 Weighted least squares estimators for F and G: 26

3.3.4 Some alternative estimators for $B = FG'$ 26

3.3.5 Estimation under linear (or affine) restrictions on F and G: 28

3.4 Model specification and input selection 29

3.5 Forecasting 30

4 Factor Model with Idiosyncratic Noise (IN) **31**

4.1 The model 31

4.2 Estimation 33

4.3 Forecasting 34

4.4 Model specification and input selection 36

5 Empirical Analysis for Daily Share Prices **37**

5.1 Measures for out-of-sample model validation 40

5.2 Problems in developing forecasting strategies 42

5.3 Comparison of the forecasting methods 44

II Forecasting Volatilities of Stock Prices **51**

6 Introduction **53**

7 Multivariate GARCH models (MGARCH) **57**

7.1 VECH model 57

7.2 BEKK model 59

8 Parametrization of BEKK models **63**

8.1 Parametrizations with $K = n^2$ 64

8.1.1 BEKK parametrization of Scherrer and Ribarits 64

8.1.2 BEKK parametrization of Engle and Kroner 67

8.1.3 Parametrization of the positive definite orthocomplement of the $\mathcal{Q}$-sets 68

8.2 Parametrizations with $K \leq n$ 70

8.2.1 Parametrization of $\mathcal{B}_K^+$ for $K \leq n$. 71

8.2.2 The Factor GARCH model of (Engle, Ng and Rothschild, 1990) 72

8.2.3 Parametrization for a subset of the positive semidefinite orthocomplement of the $\mathcal{Q}$-sets. 73

8.2.4 The DVECH model 74

8.2.5 A restricted BEKK model 76

9 Estimation of BEKK models **77**

10 Simulation Studies and Applications on Real Data **81**

10.1 The role of starting values 82

10.2 The role of the underlying DGP 84

10.3 Application on real data 88

11 Conclusion **91**

Appendix **93**

A Random Variables and Stochastic Processes **93**

B Proofs **95**

C Basic Definitions and Frequently Used Notations **101**

C.1 Concepts of convergence of sequences of random variables 101

C.2 Matrices 101

Bibliography **105**

Curriculum Vitae **111**

List of Figures

2.1 Given 1000 samples of length 16 of the bivarite VAR(1) process defined in eq. (2.9), this figure shows the histograms of the 1000 statistics $(\hat{a}_{ij} - a_{ij})/\hat{\sigma}_{a_{ij}}$ of the respective estimated coefficients, together with the standard normal (blue) and Student's t distribution with 14 df (red). . . . 11

5.1 Return Series of the DJ EURO STOXX50 banks sector, 16.06.2000 – 13.11.2002. . . . 37

5.2 Relative differences (solid line) and the log returns expressed as a function of relative differences (dotted line). . . . 38

5.3 Coefficient of model rr5.indirect.BICF-BIC.FALSE; solid lines: $U.BBV_{t-1}$, $U.BBV_{t-2}$, $D.DBK_{t-1}$, BIX_{t-1}, IRX_{t-2}, $NDc1_{t-1}$ and $NDc1_{t-2}$; dotted lines: $E.SCH_{t-1}$, $E.SCH_{t-2}$, $F.BNP_{t-2}$, $I.UC_{t-1}$, $I.UC_{t-2}$, SKX_{t-2}, $LIEUR1YD_{t-6}$, $CH2MTRR_{t-2}$, IRX_{t-6} and $SXFE_{t-2}$. . . . 44

5.4 Boxplots of out-of-sample R^2 (first row) and hit rates (second row) of all VARX (first column), RR (second column) and IN models (third column) considered. . . . 45

5.5 Capital development of portfolios. Computation of portfolio weights is based on forecasts of VARX models; Black lines: AIC in the first step of the refined FSP; Grey lines: BIC in the first step of the refined FSP. . . . 47

5.6 Capital development of portfolios. Computation of portfolio weights is based on forecasts of IN models, respecified every 10 days; Solid lines: factor forecasts only; Dotted lines: factor plus idiosyncratic noise forecasts. . . . 48

5.7 Capital development of portfolios. Computation of portfolio weights is based on forecasts of varx5.mva.BICF-BIC.FALSE (thin solid line), rr5.indirect.BICF-BIC.FALSE (dotted line), and as benchmark the sample geometric mean (thick solid line). . . . 50

8.1 Gray: Cone containing all p.d. and p.s.d. Q matrices; Black lines: equivalence classes $\mathcal{Q}(A)$ (thick lines: contain at least one p.s.d matrix Q and therefore have a BEKK representation, thin lines: contain no p.s.d. matrix Q - have no BEKK representation). . . . 63

8.2 Given matrix A from eq. (8.1) and the corresponding $\mathcal{Q}$-set; Left: barrier function $\phi(\omega_1)$, see eq. (8.2); Right: the 4 eigenvalues of Q matrices in $\mathcal{Q}(A)$; Gray lines: the analytic center and the p.s.d. rank deficient matrices at the boundary of the conus. . . . 64

8.3 Gray: Cone containing all p.d. and p.s.d. Q matrices; Black lines: Equivalence classes $\mathcal{Q}(A)$ (thick lines: contain at least one p.s.d matrix Q and therefore have a BEKK representation, thin lines: contain no p.s.d. matrix Q - have no BEKK representation); Red: Parametrization using the analytic center. . . . 66

8.4 Gray: Cone containing all p.d. and p.s.d. Q matrices; Black: Equivalence class $\mathcal{Q}(A)$; Blue: Orthocomplement of the equivalence classes; Shaded blue plane: P.s.d. matrices within the orthocomplement. 70

10.1 The 4 eigenvalues of the Q-matrices in the $\mathcal{Q}$-sets $\mathcal{Q}(A_1)$ and $\mathcal{Q}(A_2)$; x-axis: ω_1; Gray vertical line: indicates the analytic centers. 82

10.2 Boxplot of likelihood values obtained after optimization from 100 different random starting values for each of the considered model class; Green line: likelihood value corresponding to $c = \text{vech}(\frac{1}{T}\sum_{t=1}^{T}\epsilon_t\epsilon_t')$, $A = G = 0$; Red line: likelihood value corresponding to the true parameter values; Number of parameters (from left to right): 21, 11, 17, 21, 21, 9, 15, 7, 9, 7, 9, 9. 83

10.3 Boxplots of computing times (in seconds) of the 12 estimated MGARCH models for the 12 simulated series with $n = 2$ and $T = 500$ of the respective DGP's, see also table 10.3. . 84

10.4 Boxplots of computing times (in seconds) of the 16 estimated MGARCH models for the 6 simulated series with $n = 3$ and $T = 500$ of the respective DGP's, see also table 10.4. . . . 86

10.5 Black: Absolute return series of the seven banks; Red: Two times the estimated standard deviation of model sr.1; Green: Two times the estimated standard deviation of model rbekk.1. 88

10.6 Black: Absolute return series of the portfolio returns; Red: Two times the estimated standard deviation of the portfolio returns obtained from model sr.1; Green: Two times the estimated standard deviation of the portfolio returns obtained from model rbekk.1. . . 90

List of Tables

2.1 Given 1000 samples of length 16 of the bivarite VAR(1) process defined in eq. (2.9), this table provides the sample skewness, test results of a two-sided t-test testing the Null hypothesis: "The mean of the 1000 statistics $(\hat{a}_{ij} - a_{ij})/\hat{\sigma}_{a_{ij}}$ is zero", and a two-sided Kolmogorov-Smirnov test testing the Null hypothesis: "The 1000 statistics $(\hat{a}_{ij} - a_{ij})/\hat{\sigma}_{a_{ij}}$ follow a Student's t distribution with 14 df". 12

3.1 Median of the 500 FN_B, FN_y, and log likelihood values computed for each combination of $\lambda = 0.7$ and sample size T, and each of the above presented estimation methods plus the OLS estimator in the unrestricted framework. 28

5.1 Summary statistics of the return series of the DJ EURO STOXX50 banks sector, 16.06.2000 – 13.11.2002. (q_{25} (q_{75}) is the 25% (75%) sample quantile.) 39

5.2 Input variable candidates - group two. 40

5.3 Input variable candidates - group three. 41

5.4 The out-of-sample results for varx5.mva.BICF-BIC.TRUE (exogenous variables: BIX, IRX, NDcl), rr5.direct.BICF-BIC.TRUE, in10.Bartlett.default.mva.BICF-BIC.BIC.TRUE.FALSE (considering only forecasts based on the factor part of the model) and the geometric mean as benchmark model. 45

5.5 For all VARX models considered (16 models with 3 exogenous inputs (BIX, IRK, NDcl), 16 models with whole Group 2 as exogenous explanatory variable candidates), the table shows: Fractions of positive out-of-sample R^2 and hit rates greater than 0.5 are given. . 46

5.6 For all RR models considered (8 models with new.spec = 5, 8 models with new.spec = 10), the table shows: Fractions of positive out-of-sample R^2 and hit rates greater than 0.5 are given. 46

5.7 For all IN models considered (192 models including factors only, 192 models including factors and errors), the table shows: Fractions of positive out-of-sample R^2 and hit rates greater than 0.5 are given. 46

5.8 The out-of-sample results of models varx10.mva.BICF-BIC.TRUE (exogenous variables: BIX, IRX, NDcl), rr10.indirect.BICF-BIC.TRUE and in10.regression.default.univ.BICF-BIC.AIC.TRUE.TRUE. 48

5.9 The out-of-sample results of the models rr5.direct.BICF-BIC.TRUE, rr5.direct.BICF-BIC.FALSE, rr5.indirect.BICF-BIC.TRUE, rr5.indirect.BICF-BIC.FALSE. 49

5.10 Out of sample measures for rr5.indirect.BICF-BIC.FALSE and the geometric mean as benchmark model for the first and the second half of the validation sample, i.e. 21.02.2002 to 03.07.2002 and 04.07.2002 to 13.11.2002. 49

8.1 Eigenvalues of the analytic center $\bar{Q}$ constructed from $\theta = (4.85, 2.14, 1.46, 1.75, 0.29$,-0.28,-0.72, 1.28,-1.46$)'$ as shown in section 8.1.1, the VECH-term A and the corresponding $\check{Q}$. . . 69

10.1 Coding system of the different MGARCH models. 81

10.2 Condition numbers of the analytic centers $\bar{Q}$ and the mapping $f : \theta \mapsto \text{vech}(Q)$, evaluated at the parameter vectors that correspond to the analytic centers. 83

10.3 AIC and BIC values plus the total number of parameters (Nb.P.) of the respective estimated MGARCH models (rows) for realizations of the bivariate MGARCH processes (columns); The minimum IC value for a certain DGP across the estimated models is shown in bold face. 85

10.4 AIC and BIC values plus the total number of parameters (Nb.P.) of the respective estimated MGARCH models (rows) for realizations of the 3-dimensional MGARCH processes (columns); The minimum IC value for a certain DGP across the estimated models is shown in bold face. 87

10.5 AIC and BIC values plus the total number of parameters (Nb.P.) of the respective estimated MGARCH models (rows) for observations of three bivariate processes and one 7-dimensional vector process (columns); The minimum IC value for a certain process across the estimated models is shown in bold face; For the 7-dimensional vector process only the most parsimonious model classes have been considered. 89

10.6 Portfolio weights for Strategy I (w_1) and Strategy II (w_2). 89

Deutsche Kurzfassung

Diese Dissertation beschäftigt sich mit der Modellierung und Prognose multipler Zeitreihen. Die Dissertation gliedert sich in zwei Teile:

Im ersten Teil werden verschiedene Modellklassen vorgestellt, welche den bedingten Erwartungswert einer multiplen Zeitreihe modellieren. Der Fokus dabei ist hauptsächlich auf jene Modellklassen gerichtet, welche die Information gegeben durch die Gegenwart und Vergangenheit des Vektorprozesses selbst und möglicherweise zusätzlicher exogener Inputprozesse komprimieren und dadurch die Dimension des zugrunde liegenden Parameterraumes reduzieren.
Die Anzahl der voneinander funktional unabhängigen Parameter in VARX Modellen wächst mit dem Quadrat der Dimension des Vektorprozesses, falls keine weiteren Restriktionen an die Parameter in der Koeffizientenmatrix gestellt werden. Reduced Rank Regressionsmodelle und Faktor Modelle bieten eine Möglichkeit die Anzahl der zu schätzenden Parameter zu reduzieren. Gegeben die gesamte verfügbare Information, versuchen sie jene Charakteristika in den Daten zu finden, welche in Bezug auf gewisse Verlustfunktionen den zugrunde liegenden Prozess am besten beschreiben.
Die zuvor genannten Modellklassen werden beschrieben und ihre Eigenschaften diskutiert. Es werden insbesondere Prozeduren vorgestellt, welche eine datengetriebene Modellspezifikation und Inputselektion der erwähnten Modellklassen bei relativ geringem Rechenaufwand ermöglichen. Die Prozeduren basieren auf Informationskriterien und sind im Wesentlichen Verallgemeinerungen für den multivariaten Fall der "Fast step procedure", welche von (An and Gu, 1985; An and Gu, 1989) für lineare Einzelgleichungsmodelle vorgeschlagen wurde. Ein weiteres Kapitel enthält schließlich Anwendungen. Es zeigt die Ergebnisse verschiedener Schätzungen und Prognosen aller Modellklassen und vergleicht deren Güte außerhalb der zur Modellspezifikation und -schätzung verwendeten Stichprobe.

Der zweite Teil hat die Schätzung der bedingten Varianz des Vektorprozesses zum Inhalt. Das Hauptaugenmerk liegt bei der Parametrisierung von multivariaten GARCH Modellen, insbesondere den so genannten VECH und BEKK Modellen.
VECH Modelle erlauben eine sehr flexible (affine) Modellierung der bedingten Kovarianzmatrix. Allerdings haben diese Modelle auch zwei große Nachteile: Zum einen wird nicht gewährleistet, dass die geschätzten bedingten Kovarianzmatrizen positiv definit sind und zum anderen ist die Anzahl der voneinander funktional unabhängigen zu schätzenden Parameter von der Größenordnung $O(n^4)$. Das heißt, man muss zusätzliche Bedingungen an die Parameter stellen um positiv definite Schätzer für die bedingten Kovarianzen zu erhalten und für große Vektordimensionen n ist eine Schätzung praktisch unmöglich. In weiterer Folge nennen wir VECH Modelle, welche für alle möglichen Stichprobenpfade positiv definite bedingte Kovarianzmatrizen liefern, zulässige VECH Modelle.
Die BEKK Modelle liefern per Konstruktion positiv definite Schätzer für die bedingten Kovarianzmatrizen, allerdings lastet auch auf ihnen der so genannte "Fluch der Dimensionen", denn auch hier wächst die Anzahl der unabhängigen Parameter mit $O(n^4)$. Das heißt, es ist unabdingbar die genannten Modellklassen entsprechend einzuschränken, um auch für hochdimensionale Vektorprozesse die Schätzung eines multivariaten GARCH Modells zu ermöglichen.

Im Folgenden wird gezeigt, dass im bivariaten Fall BEKK Modelle genauso allgemein sind wie VECH Modelle. Im Fall von höherdimensionalen Vektorprozessen jedoch sind die zulässigen VECH Modelle allgemeiner. Leider ist es nicht sehr einfach festzustellen, ob ein gegebenes VECH Modell zulässig ist oder nicht. Vor allem ist es schwierig eine Parametrisierung so zu definieren, dass man damit alle zulässigen VECH Modelle erhält. Es lässt sich jedoch einfach klären, ob ein gegebenes VECH Modell als BEKK Modell geschrieben werden kann. Hierfür wird eine Methode vorgeschlagen.

In weiterer Folge werden verschiedene Möglichkeiten zur Parametrisierung von BEKK Modellen in ihrer "allgemeinsten" Form diskutiert. Die Vor- und Nachteile der einzelnen Parametrisierungen werden angeführt. Mit Hilfe der zuvor erlangten Erkenntnisse über die Modellklassen werden dann bekannte restriktivere Modellklassen, wie zum Beispiel das Diagonal VECH (DVECH) Modell und das Faktor GARCH (F-GARCH) Modell analysiert. Besonders das DVECH Modell findet in der Praxis vielfach Anwendung, da sich die Parametrisierung zulässiger DVECH Modelle sehr einfach gestaltet und auch die Stationaritätsbedingungen leicht zu überprüfen sind. Es werden noch weitere alternative Modelle vorgestellt, die ebenfalls eine geringere Anzahl von Parametern zur Schätzung der bedingten Kovarianzmatrizen benötigen als das allgemeine BEKK Modell.

Zuletzt werden noch Ergebnisse von Schätzungen angeführt und die zuvor analysierten Modellklassen miteinander verglichen.

Abstract

This thesis deals with the problem of modelling and forecasting multiple time series and consists of two parts:

In the first part different model classes are presented that can be used for modelling the conditional expectation of a multiple time series. The focus is mainly on model classes that try to condense the information provided by the present and past of the vector process itself and possibly some additional exogenous input processes and thereby, shrink the dimension of the underlying parameter space.

In the framework of VARX models for instance, without imposing any restrictions on the parameter matrix, the number of functionally independent parameters is increasing with the square of the dimension of the endogenous vector process under investigation. Reduced rank regression models and factor models, both having somehow different backgrounds, provide one possibility to reduce the actual number of parameters that have to be estimated. Given all the information available they try to find the main characteristics or features in the data that subject to some quality measure explain best the underlying vector process.

The aforementioned model classes are presented and their properties are discussed. In particular, procedures will be proposed that enable data-driven model specification and input selection of the model classes at relatively low computational costs. The procedures base on information criteria and basically are generalizations to the multivariate case of the "Fast step procedure" suggested by (An and Gu, 1985; An and Gu, 1989) for linear single equation models. Finally, a section of applications shows estimation and forecasting results of all model classes and compares their out-of-sample performance.

The second part deals with estimation of the conditional variance matrix of the vector process. The main focus is on the parametrization of multivariate GARCH models, in particular, the so-called VECH and BEKK models.

VECH models allow for a quite flexible (affine) modelling of the conditional variance matrix. However, these models exhibit two disadvantages: First of all, it is not ensured that the estimated conditional variance matrices are positive definite. Second, the number of functionally independent parameters that have to be estimated is of the order $O(n^4)$. Hence, one has to impose further restrictions on the parameters in order to obtain positive definite estimates for the conditional variance matrices. Furthermore, for large vector dimensions n, estimation is infeasible in practice. In the following, VECH models that ensure positive definite conditional variance matrices for all sample paths will be called admissible.

BEKK models by construction provide positive definite estimates for the conditional variance matrices. However, they also suffer from the so-called "curse of dimensionality", since the number of functionally independent parameters is also of order $O(n^4)$. Therefore, it is indispensable to further restrict the above model classes in order to make estimation of multivariate GARCH models possible for high dimensional vector processes.

In the following, it is shown that in the bivariate case the BEKK model is as general as the VECH model. In case of higher dimensions however, the class of admissible VECH models is more general than the BEKK model class. Unfortunately, it is not easy to determine whether a given VECH model is admissible or not. It is particularly difficult to define a parametrization that parametrizes all admissible VECH models. However, it is easy to check whether the underlying VECH model may be cast as a BEKK model, thus, has a BEKK representation. For this purpose a simple method is suggested.

Next, different parametrizations for BEKK models in their most general form are presented, and their assets and drawbacks are listed. Well-known more restrictive model classes such as the diagonal VECH (DVECH) and the factor GARCH (F-GARCH) model are then analyzed. The DVECH model is often times used in practice, since it is easy to parametrize admissible DVECH models and to check whether the stationarity conditions hold or not. In addition, alternative parsimonious models are suggested.

Finally, estimation results are presented and the different model classes are compared.

Acknowledgements

There is a song, one of its rhymes just came into my mind. Translated to English the main message would be something like...

"Give me the right words, let me strike the right tone."

Writing a thesis is always a search of the right expressions. Especially within the last year I got extremely familiar with the "delete"-button on my keyboard.

I am glad that there is this page of acknowledgements, for which in contrast to the rest of the thesis words do come easy.

In the first place I would like to thank my supervisor and mentor Prof. Manfred Deistler. He was the person who encouraged me to inscribe for the PhD program, after I finished my Master program and actually decided to leave university for business life. In the past five years he made it possible that I could meet both worlds, academia and business world. I am especially grateful for his patience, his unfatiguing engagement in explaining things and his belief in me. Besides university and research, I appreciated the times where we talked about everything under the sun. In the spirit of these discussions maybe he also introduced me to interdisciplinary forums and discussion platforms, which were and are an essential enrichment for me. (I might possibly not be married now.) Thank you for your guidance, your care.

I would also like to thank all my colleagues at Vienna TU, research unit Econometrics and System Theory, for their backing and the many times of laughter. Special thanks in this respect goes to Karl Gruber with whom I shared the room over the years and during lunchtime scanned the fourth district of Vienna for a "Lokal" optimum. In addition, I would like to say thanks to Bernhard Böhm, who has been helpful throughout. Thanks also to Dietmar Bauer, who was always open for technical questions. Above all I want to thank Wolfgang Scherrer. Currently we are working together on the problem of parametrization of multivariate GARCH models. Besides the burden of having me as co-author, he also took the pain of proofreading.

Last but not least I would like to thank all my friends and my family for their love and support. Finally, I would like to thank my husband Thomas, for pushing me to keep on writing and the many fruitful discussions. Thank you for sharing your life with me.

Part I

Forecasting Return Series of Stock Prices

Chapter 1

Introduction

Forecasting of financial time series, such as asset prices, plays an important part in portfolio management. However, the question whether returns of asset prices are predictable or not, has divided the community. In the late 1960s, Fama defined the so-called *efficient market hypothesis* (EMH), one of the central propositions in finance, see (Fama, 1970; Fama, 1991). According to this hypothesis a financial market is efficient, if at any time asset prices fully reflect the available information. This implies that the market processes information rationally, in the sense that relevant information is not ignored, and systematic errors are not made. As a consequence, prices are always at levels consistent with "fundamentals". Thus, market efficiency here refers to informational efficiency of markets. Let p_t denote the price of some asset at time t, then $y_t = \frac{p_t - p_{t-1}}{p_{t-1}}$ is the return obtained within period $t-1$ to t. Let furthermore, $\mathcal{I}_t$ be some information set available at time t, then informational efficiency with respect to $\mathcal{I}_t$ means that

$$\mathbb{E}(y_{t+1}|\mathcal{I}_t) = 0.$$

Hence, any investor that at time t possesses only information that is contained in $\mathcal{I}_t$, cannot expect to gain by using this information to predict returns. According to the amount of information available Fama distinguishes between three types of efficiency:

1. *strong* form efficiency: The information set $\mathcal{I}_t$ contains all information available to any market participant. Thus, $\mathcal{I}_t$ does also include so-called insider informations.

2. *semi-strong* form efficiency: The information set $\mathcal{I}_t$ contains all publicly available information. Hence, in a semi-strong form efficient market, as soon as information becomes public, it is immediately incorporated into prices.

3. *weak* form efficiency: Here $\mathcal{I}_t = \{y_t, y_{t-1}, y_{t-2}, \ldots\}$. Thus, the information set includes only the history of the returns themselves.

For further discussions on the above hypothesis and its implications and challenges see e.g. (Kaul, 1996), (Campbell, Lo and MacKinlay, 1997), (Grinold and Kahn, 1999), (Shleifer, 2000) and (Singal, 2003).
We believe that the stock markets are weak form efficient. It is clear that the markets work and that (nontrivial) forecasting is not an easy task. In addition, it seems to be reasonable to assume that the markets have become even more efficient in the last decade by the widespread use of the Internet, the easily available data base (at least for frequently traded financial assets) and by skilled experts advising large investors. Nevertheless, we think that a necessary condition for successful forecasting is to make use of information contained in (market) variables that might be related with the financial assets of interest. Since there is no clear a priori knowledge concerning relevant explanatory variables (or inputs) available for us, data driven input selection is an important issue here. However, the problem is messed by the large number of potential input combinations (relative to sample size), by high correlation between the

potential inputs, by extremely weak correlations between future returns and present and past inputs and by time changing structures. The large number of possible input combinations or model specifications, in particular, leads to the danger of overfitting, see e.g. (White, 2000).

It is already said in the title of the thesis that the focus is on multivariate model classes for forecasting returns of asset prices and their volatilities. As stated in (Tiao, 2001), there are at least two good reasons for analyzing and modelling time series jointly. First, to understand the dynamic relationship among the series; certain structures such as comovements or common factors may be detected. Second, to improve accuracy of forecasts. If one variable contains in its historical data information on the other series, better forecasts may be obtained when multivariate model classes are applied. However, turning from single equation models to systems of equations also implies an increasing model complexity. This fact involves two problems concerning some practical limitations. First, the more parameters one has to estimate the more observations are actually needed in order to obtain statistically tenable results. This is sometimes referred to as the *curse of dimensionality.* Second, assuming that enough data is available for the investigator -invalidating the first claim- interpretability of the obtained coefficient estimates is desirable. Thus, one is interested in having sparse parameter matrices, or simplified model structures.

In the following chapters of the first part of the thesis different model classes for the conditional mean of multidimensional time series processes are presented. It will be shown how the parameter space may be restricted by imposing additional constraints. We commence with the very general class of *Vector autoregressive forecasting models including exogenous explanatory variables* (VARX) in chapter 2. We provide parameter estimates and their asymptotic properties for both cases, the unrestricted case and the case, where linear restrictions are imposed. In chapter 3 the concept of *Reduced Rank* (RR) regression is elaborated and maximum likelihood estimates are presented. The classical *Factor model* is presented in chapter 4 as a third alternative for dimension reduction of the parameter space. For each of the above model classes we propose model specification and input selection procedures that are in the style of the procedures brought up by (An and Gu, 1985; An and Gu, 1989) for the univariate case. Finally, applications are provided in chapter 5. The different forecasting methods are evaluated and compared.
The second part of the thesis is devoted to multivariate models for the conditional variance of a vector process, see chapter 6 for a short introduction. We concentrate on multivariate GARCH models, in particular, the VECH and BEKK model classes. In chapter 7 we give a description of the two model classes. Chapter 8 deals with the problem of parametrization and identifiability. A simple to check characterization of VECH models that have an equivalent BEKK representation is given. It will be shown that in the bivariate case BEKK models are as general as VECH models. In higher dimensional cases, however, VECH models allow for more flexibility. A parametrization for a *generic*, i.e. open and dense, subset of BEKK(p, q, K) models (with $K = n^2$) is presented. Furthermore, two other parametrizations (also with $K = n^2$) are analyzed. It is shown that these parametrizations both do not cover a generic set of BEKK models. In addition, several alternative parametrizations of BEKK(p, q, K) models (with $K \leq n$), thus with a small number of additive terms are presented. A short summary of estimation in the multivariate GARCH framework is given in chapter 9. Finally, chapter 10 concludes the second part of the thesis with some applications on simultated and real data.
An appendix is included for the sake of completeness and clarity. It contains an introductory chapter on random variables and stochastic processes (appendix A), appendix B provides some additional proofs and finally appendix C deals with basic definitions and frequently used notations.

Chapter 2

VARX model

Vector autoregressive models including exogenous explanatory variables (VARX) are mainly used for forecasting and structural analysis. Here, the focus is on forecasting only.

2.1 The model

VARX (or VAR(p)X) models are of the form

$$\begin{aligned} y_t &= c + A_1 y_{t-1} + A_2 y_{t-2} + \ldots + A_p y_{t-p} + Dx_{t-1} + \epsilon_t, \\ &\quad \text{or in a more compact notation} \\ A(z)y_t &= c + Dx_{t-1} + \epsilon_t, \quad t \in \mathbb{Z}, \end{aligned} \tag{2.1}$$

where (y_t) denotes the n–dimensional vector of observed endogenous variables (e.g. asset returns) and c is some n–dimensional vector of constants. $A(z) = \sum_{j=0}^{p} -A_j z^j$, with $A_0 = -I_n$, $A_j \in \mathbb{R}^{n \times n}$ and $A_p \neq 0$, is an $n \times n$–dimensional polynomial matrix of order p in a complex variable z or the *backward shift operator* z, defined by $z^j(y_t) = (y_{t-j})$. D is some $n \times k$–dimensional real parameter matrix, loading the k–dimensional exogenous inputs, x_{t-1}. The process (x_t) is stationary[1] with mean μ_x and may, for instance, contain different lags of one and the same or several candidate explanatory variables. (ϵ_t) is an n–dimensional white noise process, i.e. $\mathbb{E}\epsilon_t = 0$, $\mathbb{E}\epsilon_t\epsilon_s' = 0$ for all $s \neq t$, and $\mathbb{E}\epsilon_t\epsilon_t' = \Sigma_\epsilon$ is supposed to be positive definite unless stated otherwise. Hence, ϵ_t can be viewed as some unobservable random shock to y_t at time t.

It is assumed that the *stability condition*, $\det(A(z)) \neq 0$ for all $|z| \leq 1$, holds. Hence, the convergence of the Taylor series expansion of $A(z)^{-1}$ about 0, $A(z)^{-1} = \sum_{j=0}^{\infty} a_j z^j$, on a disk containing the unit circle is ensured and therefore, the processes $A(z)^{-1}Dx_{t-1} = \sum_{j=0}^{\infty} a_j D x_{t-1-j}$, $A(z)^{-1}\epsilon_t = \sum_{j=0}^{\infty} a_j \epsilon_{t-j}$ and thus, (y_t) are well-defined. In the case $p = 1$, the stability condition is equivalent to the condition that the eigenvalues of A_1 are smaller than one in absolute value, implying that the coefficient matrices $a_j = A_1^j$ are absolutely summable[2]. In the sequel, we consider the unique stationary solution to eq. (2.1), which is given by $y_t = A(z)^{-1}(Dx_{t-1} + \epsilon_t) = \sum_{j=0}^{\infty} a_j(Dx_{t-1-j} + \epsilon_{t-j})$. Note, that this solution is *causal*, since the series $A(z)^{-1} = \sum_{j=0}^{\infty} a_j z^j$ contains nonnegative powers of the lag operator only, implying that y_t does not depend on future shocks ϵ_{t+h}, and present and future observations x_{t-1+h}, $h > 0$.

[1] Unless stated otherwise, the term "stationary" is refering to *wide* or *weak sense stationarity* throughout the thesis, i.e. the first and second moments of the process are finite and time independent, $\mathbb{E}x_t = \mu_x$, $\mathbb{E}x_t x_s' = \Gamma_x(t-s)$ depends only on $t - s$, and $\mathbb{E}x_t x_t' = \Gamma_x(0) < \infty$.

[2] Note, that for every n–dimensional VAR(p)X process (y_t) with $p > 1$, the np–dimensional process $Y_{t-1} = (y_{t-1}, y_{t-2}, \ldots, y_{t-p})'$ is a VAR(1)X process. This representation is referred to as a *state-space representation* of the original VARX process. Thus, theoretical findings for VAR(1)X processes can easily be extended to VAR(p)X processes.

In addition, it is assumed that ϵ_t is independent of the exogenous variables x_s, for all $s < t$. Hence, the best linear prediction given $\{y_{t-1}, y_{t-2}, \ldots, x_{t-1}, x_{t-2}, \ldots\}$ for y_t in the sense of L^2–norm approximation, minimizing $\mathbb{E}(y_t - y_{t|t-1})'(y_t - y_{t|t-1})$, is given by $y_{t|t-1} = c + \sum_{j=1}^{p} A_j y_{t-j} + Dx_{t-1}$, and ϵ_t is the forecast error, see section 2.4.

The VARX model class is often applied in practice due to its simple structure, but it has also a theoretical justification: In *Wold's decomposition theorem*, see (Wold, 1938), it is shown that every n–dimensional stationary process can be decomposed into the sum of two uncorrelated processes, namely a *singular*[3] and a *regular* process[4], and the regular process has an *infinite moving average* (MA) *representation* $\sum_{j=0}^{\infty} k_j \epsilon_{t-j}$, where $k_0 = I_n$, (ϵ_t) is a white noise process and the sequence of matrices k_j is absolutely summable. Hence, the infinite sum is defined as a limit in mean square. This infinite MA process can be approximated with arbitrarily high accuracy by a VAR process, highlighting the importance of VAR and also VARX models in the framework of stationary processes.

A drawback of VARX models, however, is the fact that the number of parameters to estimate ($pn^2 + nk$, intercepts and Σ_ϵ excluded) is a quadratic function of dimension n. Hence, estimation precision might be low even for small n (and possibly known Σ_ϵ). Another problem concerns the choice of relevant explanatory variables out of a huge set of possible candidates, see section 2.3.

In the following section estimators of the real-valued parameters $c, A_1, \ldots, A_p, D$ are presented. We will, in addition, deal with the case where linear restrictions are imposed on the parameters. Finally, the asymptotic properties of the estimators will be discussed.

2.2 Estimation

In this section it is assumed throughout that the lag order p as well as the set of explanatory variables is known. So, the focus is on the estimation of the unknown coefficients $c, A_1, \ldots, A_p, D$ and Σ_ϵ, only.

Given a sample of observations $y_1, \ldots, y_T$ and $x_0, \ldots, x_T$ of the process (y_t) and the exogenous variables x_t respectively, and some presample values $y_{1-p}, \ldots, y_0$, the analyst can choose an estimation procedure according to the properties the estimators should have. The following three methods, especially the first two, are in a widespread use: the method of *multivariate least squares* (LS), the *maximum likelihood* (ML) method and the *Yule-Walker estimation* (YW) method. Apart from their different backgrounds and ways of solving the underlying estimation problem, the methods differ from each other mainly due to their different choice and treatment of starting or presample values. Especially in finite samples this may lead to different estimation results and hence, the estimators in general have different finite sample properties. Asymptotically, however, they all have the same properties. Here, the form and properties of LS estimates are discussed only. For a detailed description of all three methods mentioned refer to e.g. (Lütkepohl, 1993, chapter 3).

[3] A process (y_t) is called singular, or in other words, a process is purely deterministic, if it can be forecasted perfectly by its own past, i.e. $\hat{y}_{t+h|t} = y_{t+h}$ a.e. for all $h > 0$.

[4] A process (y_t) is called regular, when the distant past of the process has no impact on the future development of the process. In terms of forecasting this means that, if $\mathbb{E}y_t = 0$, $\underset{h\to\infty}{\text{l.i.m}}\, \hat{y}_{t+h|t} = 0$ a.e., see appendix C.1 for a definition of l.i.m, the limit in mean square.

To keep notation simple, let

$$\begin{array}{lll} \bar{k} & := 1 + pn + k & \\ y_t & := (y_{1t}, y_{2t}, \ldots, y_{nt})' & (n \times 1), \\ Y & := (y_1, y_2, \ldots, y_T) & (n \times T), \\ x_t & := (x_{1t}, x_{2t}, \ldots, x_{kt})' & (k \times 1), \\ X_t & := (1, y'_{t-1}, \ldots, y'_{t-p}, x'_{t-1})' & (\bar{k} \times 1), \\ X & := (X_1, X_2, \ldots, X_T) & (\bar{k} \times T), \\ B & := (c, A_1, \ldots, A_p, D) & (n \times \bar{k}), \\ \epsilon_t & := (\epsilon_{1t}, \epsilon_{2t}, \ldots, \epsilon_{nt})' & (n \times 1), \\ E & := (\epsilon_1, \epsilon_2, \ldots, \epsilon_T) & (n \times T). \end{array} \tag{2.2}$$

Eq. (2.1) can now be written in a compact way as

$$Y = BX + E \tag{2.3}$$

or using the vec operator it can be written as

$$\text{vec}(Y) = \text{vec}(BX) + \text{vec}(E). \tag{2.4}$$

Let $y = \text{vec}(Y)$, $\epsilon = \text{vec}(E)$ and $b = \text{vec}(B)$, then

$$y = (X' \otimes I_n)b + \epsilon, \tag{2.5}$$

where $\otimes$ denotes the Kronecker product, see appendix C.2 for a definition and basic rules. Note that the variance covariance matrix of ϵ is given by

$$\mathbb{E}(\epsilon\epsilon') = (I_T \otimes \Sigma_\epsilon).$$

Suppose for the moment that Σ_ϵ is known. The unrestricted multivariate weighted least squares estimator $\hat{b}$ then minimizes the following specially weighted quadratic loss function

$$S(b) = (y - (X' \otimes I_n)b)'(I_T \otimes \Sigma_\epsilon)^{-1}(y - (X' \otimes I_n)b). \tag{2.6}$$

Its objective is to weight the observations according to the variance of the corresponding noise term which also determines the actual precision of the underlying observation. Thus, little weight is given to terms including a noise term with high variation and much weight is given to those, whose innovations show low variation. Note that the minimizing $\hat{b}$ of eq. (2.6) due to this specific weighting is an efficient estimate of b. Eq. (2.6) can be transformed to

$$S(b) = y'(I_T \otimes \Sigma_\epsilon)y - 2b'(X \otimes \Sigma_\epsilon^{-1})y + b'(XX' \otimes \Sigma_\epsilon^{-1})b$$

using basic algebraic rules. Hence, the first and second order partial derivatives are given by

$$\frac{\partial S(b)}{\partial b} = -2(X \otimes \Sigma_\epsilon^{-1})y + 2(XX' \otimes \Sigma_\epsilon^{-1})b,$$

$$\frac{\partial^2 S(b)}{\partial b \partial b'} = 2(XX' \otimes \Sigma_\epsilon^{-1}).$$

The normal equations are obtained by setting $\frac{\partial S(b)}{\partial b} = 0$,

$$(XX' \otimes \Sigma_\epsilon^{-1})b = (X \otimes \Sigma_\epsilon^{-1})y.$$

In the theorem below it is shown that under certain assumptions on the exogenous variables and innovation process $\mathbb{E}X_tX_t'$ has full rank. Thus, for sufficiently large sample size T this should also hold for its sample counterpart $\frac{1}{T}XX' = \frac{1}{T}\sum_{t=1}^{T} X_t X_t'$. Hence, the Hessian of $S(b)$, $\frac{\partial^2 S(b)}{\partial b \partial b'}$, is positive definite and therefore, the LS estimator is a unique minimizer of $S(b)$ and given by[5]

$$\begin{aligned} \hat{b} = \text{vec}(\hat{B}) &= ((XX')^{-1}X \otimes I_n)y, \\ \hat{B} &= YX'(XX')^{-1}. \end{aligned} \tag{2.7}$$

Obviously the weighted LS estimate (2.7) is just the *ordinary least squares* (OLS) estimate that is obtained by minimizing the squared sum of residuals

$$\bar{S}(b) = (y - (X' \otimes I_n)b)'(y - (X' \otimes I_n)b).$$

The reason therefore is not only that the innovation process (ϵ_t) is assumed to be a vector white noise process and hence, is uncorrelated in time, but also the fact that the set of explanatory variables is identical in all n equations. Thus, Σ_ϵ cancels out and $\hat{b}$ is totally independent of the error variance covariance matrix. What happens if the latter of the two conditions fails to hold, will be seen in the second part of this section. First, however, it should be stressed again that if these two conditions hold, $\hat{b}$ or vec$(\hat{B})$ can be obtained by simply solving the ordinary least squares optimization problem for each single equation or endogenous variable y_{it}, $i = 1, \dots, n$.

It can be shown that the LS estimator $\hat{b}$ is consistent and asymptotically normal, if the following assumptions hold:

Assumption 2.1 (Consistency and Asymptotic Normality of $\hat{b}$)

(i) (ϵ_t) is an n-dimensional stationary process of *independently identically distributed* (iid) random variables with $\mathbb{E}\epsilon_t = 0$, $\mathbb{E}\epsilon_t\epsilon_t' = \Sigma_\epsilon > 0$, and finite 4th moments, i.e. $\mathbb{E}|\epsilon_{it}\epsilon_{jt}\epsilon_{lt}\epsilon_{mt}| < \infty$, for $i, j, l, m = 1, \dots, n$ and all $t \in \mathbb{Z}$.[6]

(ii) (x_t) is a stationary k-dimensional process that may be written as the sum of a constant and some (infinite) MA process:

$$x_t = \mu_x + \sum_{j=0}^{\infty} K_j \xi_{t-j},$$

where (ξ_t) is a k-dimensional iid white noise process with $\mathbb{E}\xi_t = 0$, $\mathbb{E}\xi_t\xi_t' = \Sigma_\xi > 0$ and finite 4th moments. Of course this implies $\mathbb{E}x_t = \mu_x$. The coefficient matrices K_j are assumed to be absolutely summable, that is $\sum_{j=0}^{\infty} \|K_j\|_F < \infty$, where $\|.\|_F$ denotes the Frobenius norm $\|K_j\|_F = \left[\text{tr}(K_j K_j')\right]^{\frac{1}{2}}$ and such that $\mathbb{E}(x_t - \mu_x)(x_t - \mu_x)' > 0$. (ξ_t) is independent of (ϵ_t), thus $\mathbb{E}\xi_t\epsilon_s' = 0$ for all $t, s \in \mathbb{Z}$.

(iii) The polynomial matrix $A(z)$ is as described in section 2.1 and fulfills the stability condition $\det(A(z)) \neq 0$ for all $|z| \leq 1$.

(iv) $\frac{1}{\sqrt{T}} \sum_{t=1}^{T} (X_t \otimes I_n)\epsilon_t \xrightarrow{d} \mathcal{N}(0, (\Gamma_X \otimes \Sigma_\epsilon))$.

Note that given assumptions (i) to (iii), the fourth assumption can be shown to hold; for instance, by application of a suitable martingale central limit theorem, see e.g. (Brown, 1971) for a list of such theorems. Standard central limit theorems are in general ruled out, since they assume that the underlying sequence of random variables is independent, which however here is not the case. (Anderson, 1971) circumvents this problem in the univariate AR case by introduction of so-called *K-dependent sequences*, see (Anderson, 1971, chapter 7) or (Amemiya, 1985, section 5.4). A proof for the VAR case is provided in (Mann and Wald, 1943).

[5] Note that the estimates $\hat{b}$ and $\hat{B}$ depend on the sample size T. Superscripts, such as $\hat{b}^T$ and $\hat{B}^T$, have been omitted for the sake of readability.

[6] In literature such processes are sometimes referred to as *standard white noise processes*.

Theorem 2.1 (Consistency and Asymptotic Normality of $\hat{b}$) *Let (y_t) be a VAR(p)X process as presented in eq. (2.1) and let the assumptions 2.1 hold, then $\Gamma_X := \mathbb{E}X_tX_t' > 0$, with X_t as in (2.2), and the multivariate LS estimate $\hat{b} = \operatorname{vec}(\hat{B})$ from eq. (2.7) is consistent and asymptotically normal, i.e.*

(i) $\operatorname*{plim}_{T\to\infty}(\hat{b} - b) = 0,$

(ii) $\sqrt{T}(\hat{b} - b) \xrightarrow{d} \mathcal{N}(0, (\Gamma_X^{-1} \otimes \Sigma_\epsilon)).$

To prove this theorem some important limit laws for sequences of random variables are needed:

Theorem 2.2 (Law of Large Numbers (LLN)) *Let (y_t) be a scalar stationary process with $\mathbb{E}y_t = \mu$, and let $\bar{y}_T = \frac{1}{T}\sum_{t=1}^{T} y_t$ and $\gamma_y(j) = \mathbb{E}y_ty_{t-j}$. If $\lim_{j\to\infty}\gamma_y(j) = 0$ holds, then*

$$\operatorname*{l.i.m}_{T\to\infty} \bar{y}_T = \mu \quad \textit{i.e.} \quad \lim_{T\to\infty} \mathbb{E}(\bar{y}_T - \mu)^2 = 0, \qquad \textit{and hence} \quad \operatorname*{plim}_{T\to\infty} \bar{y}_T = \mu.$$

Proof. See (Deistler and Scherrer, 1994, theorem 6.4). □

A similar result can of course be obtained for vector random processes. In this case the LLN has to be applied to every single component of the vector process.

Note that in the literature there exist several versions of LLN's basing on different assumptions and with possibly more general results. In our case however, the assumptions needed for the above theorem to hold and its result are sufficiently general.

Theorem 2.3 (Slutsky's theorem) *Let y_t be a sequence of n-dimensional real random variables and let $g : \mathbb{R}^n \to \mathbb{R}^m$ be a continuous function, then*

$$\operatorname*{plim}_{T\to\infty} y_t = y_0 \qquad \textit{implies} \qquad \operatorname*{plim}_{T\to\infty} g(y_t) = g(\operatorname*{plim}_{T\to\infty} y_t) = g(y_0).$$

Proof. See e.g. (Schönfeld, 1969, Satz 6.2/2) or (Davidson, 1994, theorem 18.10(ii)). □

Let us now sketch the proof of theorem 2.1:
Proof. See Lemma B.2 in the Appendix for the proof of $\Gamma_X > 0$.
(i) Consistency of $\hat{b}$:
Due to Slutsky's theorem we have

$$\begin{aligned} \operatorname{plim}(\hat{b} - b) &= \operatorname{plim}((XX')^{-1} \otimes I_n)\operatorname{plim}((X \otimes I_n)\epsilon) = \\ &= \underbrace{\operatorname{plim}\left((\frac{1}{T}\sum_{t=1}^{T} X_tX_t')^{-1} \otimes I_n\right)}_{(a)} \underbrace{\operatorname{plim}\left(\frac{1}{T}(X \otimes I_n)\epsilon\right)}_{(b)}. \end{aligned}$$

Now, consider (a) and let us show that $\operatorname{plim}((\frac{1}{T}\sum_{t=1}^{T} X_tX_t')^{-1} \otimes I_n) = (\Gamma_X \otimes I_n)$ and hence $\operatorname{plim}((\frac{1}{T}\sum_{t=1}^{T} X_tX_t')^{-1} \otimes I_n) = (\Gamma_X^{-1} \otimes I_n)$.
Let (z_t) be the process defined by $z_t := X_tX_t'$. Since the processes (ϵ_t) and (ξ_t) are independent and their 4th moments exist and since it is in addition assumed that the polynomial coefficient matrix $A(z)$ fulfills the stability condition and the coefficient matrices K_j in the MA process are absolutely summable, it can be shown that the first and second moments of z_t are finite and independent of t, hence (z_t) is stationary. In addition, due to the properties of the coefficient matrices the covariances $\gamma_z(j)$ tend to zero for j going to infinity. According to the LLN $\operatorname*{plim}_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} X_tX_t' = \Gamma_X$ and with Slutsky's theorem we have $\operatorname{plim}((\frac{1}{T}\sum_{t=1}^{T} X_tX_t')^{-1} \otimes I_n) = (\Gamma_X^{-1} \otimes I_n)$.

Next, consider (b). Note that $(X \otimes I_n)\text{vec}(E) = \text{vec}(EX') = \sum_{t=1}^{T}(X_t \otimes I_n)\epsilon_t$. Let $z_t := (X_t \otimes I_n)\epsilon_t$, then $\mathbb{E}z_t = 0$, due to the fact that ϵ_t is orthogonal to the past of y_t and x_t. Making use of the *law of iterated expectations* (LIE)[7]we obtain

$$\begin{aligned}\mathbb{E}(z_t z_t') &= \mathbb{E}\left[(X_t \otimes I_n)\mathbb{E}(\epsilon_t\epsilon_t'|\epsilon_{t-1},\dots,x_{t-1},\dots)(X_t' \otimes I_n)\right] = \\ &= \mathbb{E}((X_t \otimes I_n)\Sigma_\epsilon(X_t' \otimes I_n)) = \\ &= \mathbb{E}(X_t X_t' \otimes \Sigma_\epsilon) = \\ &= (\Gamma_X \otimes \Sigma_\epsilon) < \infty\end{aligned}$$

Hence the variance of z_t exists and is independent of t. For $j > 0$,

$$\gamma_z(j) = \gamma_z(-j)' = \mathbb{E}(z_t z_{t-j}') = \mathbb{E}((X_t \otimes I_n)\underbrace{\mathbb{E}(\epsilon_t|\epsilon_{t-1},\dots,x_{t-1},\dots)}_{=0}\epsilon_{t-j}'(X_t' \otimes I_n)) = 0.$$

Thus, the LLN applies and $\operatorname*{plim}_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}(X_t \otimes I_n)\epsilon_t = 0$, and therefore, together with the result of term (a) we obtain, $\text{plim}(\hat{b} - b) = 0$.

(ii) Asymptotic normality of $\hat{b}$:

$$\begin{aligned}&\sqrt{T}(\hat{b} - b) = \\ &= \underbrace{\left[\left((\frac{1}{T}XX')^{-1} \otimes I_n\right) - (\Gamma_X^{-1} \otimes I_n)\right]}_{(a)}\underbrace{\left(\frac{1}{\sqrt{T}}X \otimes I_n\right)\epsilon}_{(b)} + \underbrace{(\Gamma_X^{-1} \otimes I_n)\left(\frac{1}{\sqrt{T}}X \otimes I_n\right)\epsilon}_{(c)},\end{aligned}$$

where (a) converges in probability to zero, (b) converges in distribution (see assumption (iv)), and (c) converges in distribution to $\mathcal{N}(0, (\Gamma_X^{-1} \otimes \Sigma_\epsilon))$, and thus $\sqrt{T}(\hat{b} - b) \xrightarrow{d} \mathcal{N}(0, (\Gamma_X^{-1} \otimes \Sigma_\epsilon))$. □

In general, the variance matrix Σ_ϵ is unknown. It can be shown however, that the ML estimate

$$\hat{\Sigma}_\epsilon = \frac{1}{T}\sum_{t=1}^{T}(Y - \hat{B}X)(Y - \hat{B}X)' \qquad (2.8)$$

is a consistent estimate for Σ_ϵ. In small samples, however, this estimate is biased. An unbiased estimate may for instance be obtained by adjusting the degrees of freedom by the number of predetermined regressors $\frac{T}{T-\bar{k}}\hat{\Sigma}_\epsilon$. Choosing $\bar{k}$ instead of $n\bar{k}$ may be justified by the fact that the multivariate LS estimate of the coefficients corresponds to the OLS estimates obtained by estimating each of the n equations separately. For a more detailed discussion on the choice of the adjusting term in the denominator see (Lütkepohl, 1993, proposition 3.2). Anyhow, $\hat{\Sigma}_b := ((\frac{1}{T}XX')^{-1} \otimes \frac{T}{T-\bar{k}}\hat{\Sigma}_\epsilon)$ is a consistent estimate for $(\Gamma_X^{-1} \otimes \Sigma_\epsilon)$, since it has already been shown that $\frac{1}{T}XX'$ converges in propability to Γ_X.

It should also be stressed at this point that the results of theorem 2.1 are asymptotic results only. Hence, in small samples the estimate $\hat{b}$ may be biased and inference may be hard, since the real distribution of the so-called "t-statistic" $(\hat{b}_i - b_i)/\hat{\sigma}_{b_i}$, where $\hat{\sigma}_{b_i}$ is the square root of the ith diagonal element of $\hat{\Sigma}_b$, in general does not follow a Student's t distribution in small samples. To illustrate this behavior 1000 samples of length 16 of the bivariate VAR(1) process

$$y_t = \underbrace{\begin{pmatrix} 0.9 & -0.4 \\ 0.3 & 0.1 \end{pmatrix}}_{=A} y_{t-1} + \epsilon_t, \quad \epsilon_t \sim iid\mathcal{N}(0, I_2), \qquad (2.9)$$

have been generated, the coefficient matrix has been estimated by OLS and the corresponding statistics $(\hat{a}_{ij} - a_{ij})/\hat{\sigma}_{a_{ij}}$ have been computed. Figure 2.1 shows the histograms of the thus obtained 1000 statistics, the standard normal and Student's t distribution with 14 degrees of freedom (df).

[7] LIE: Given some random variable x defined on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and some information set $\mathcal{I} \subset \mathcal{A}$, the following statement holds:

$$\mathbb{E}(x) = \mathbb{E}(\mathbb{E}(x|\mathcal{I})).$$

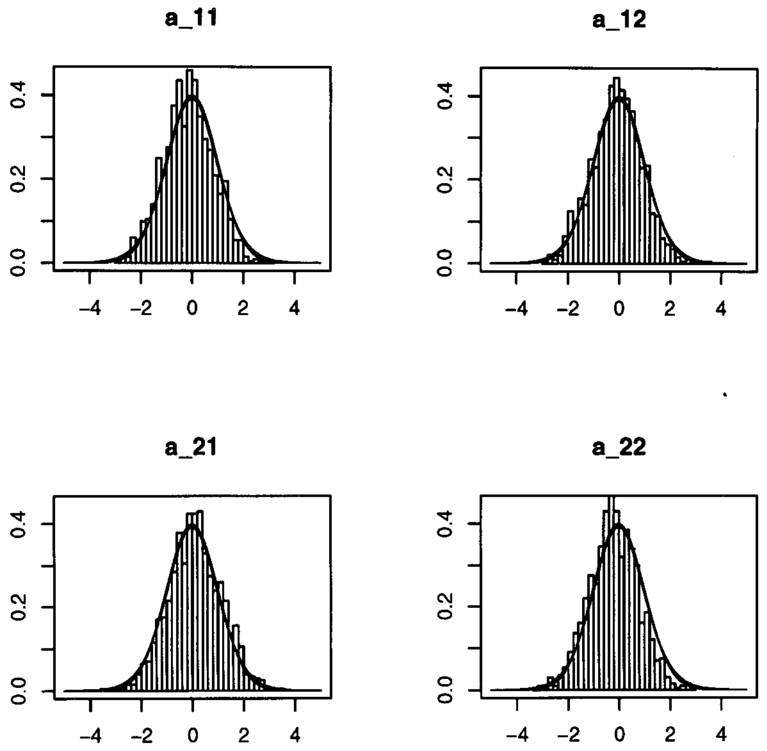

Figure 2.1: Given 1000 samples of length 16 of the bivarite VAR(1) process defined in eq. (2.9), this figure shows the histograms of the 1000 statistics $(\hat{a}_{ij} - a_{ij})/\hat{\sigma}_{a_{ij}}$ of the respective estimated coefficients, together with the standard normal (blue) and Student's t distribution with 14 df (red).

In fact, the small sample bias is conspicuous, see also table 2.1. Note however that the results of the t-tests have to be interpreted with care, since almost all of the above statistics do not follow a Student's t distribution with 14 df, as indicated by the Kolmogorov-Smirnov test. By the way, again the choice of the number of degrees of freedom has to be clarified. In the literature one can find many studies about the small sample distribution of t-statistics of AR parameters, see e.g. (Nankervis and Savin, 1988; Nicholls and Pope, 1988; Tjøstheim and Paulsen, 1983).

To sum it up, it can be said that one has to be aware of the fact that in small samples the estimates of VARX models may be biased and their corresponding test statistics may not be distributed such as in the asymptotic case. Nevertheless, the asymptotic results may be used as rugh guidelines in small sample inference, which of course is better than having nothing to go by.

Up to now, the coefficients matrix B was assumed to be totally free regardless of the number of free parameters to estimate stored therein. The more parameters one has to estimate the more observations are actually needed in order to obtain statistically tenable results[8]. Having also the aforementioned problem in mind the possibly high number of parameters is, in fact, a crucial problem in estimating VARX models. The dimension of the parameter space can be reduced by imposing constraints on the elements in B. The functional form of these constraints may be provided by some economic theory or expert knowledge. Even if there is no theory or insider knowledge available one might want to test certain hypotheses concerning the parameter space. Anyhow, estimation in a constrained setting is important and will therefore be considered next:

It should be noted however, that here, we are dealing with linear constraints, only. In our applications later on, we are, in particular, interested in zero restrictions imposed on a set of parameters in the

[8] This is sometimes referred to as the *curse of dimensionality*.

	a_{11}	a_{21}	a_{12}	a_{22}
skewness	0.049	0.010	−0.010	−0.030
t-test statistic	−4.895	3.497	−2.827	−6.835
t-test p.value	0.000	0.000	0.005	0.000
KS statistic	0.059	0.077	0.037	0.080
KS p.value	0.062	0.005	0.500	0.003

Table 2.1: Given 1000 samples of length 16 of the bivarite VAR(1) process defined in eq. (2.9), this table provides the sample skewness, test results of a two-sided t-test testing the Null hypothesis: "The mean of the 1000 statistics $(\hat{a}_{ij} - a_{ij})/\hat{\sigma}_{a_{ij}}$ is zero", and a two-sided Kolmogorov-Smirnov test testing the Null hypothesis: "The 1000 statistics $(\hat{a}_{ij} - a_{ij})/\hat{\sigma}_{a_{ij}}$ follow a Student's t distribution with 14 df".

coefficients matrix B. Thus, if there is a strong belief (accompanied at best by statistical justification) that the jth explanatory variable has no influence on the ith endogenous variable, $B_{(ij)} = b_{ij}$ is set to zero and estimation is performed on an accordingly restricted parameter space.

q linear constraints on the parameter vector $b = \text{vec}(B)$ may be formulated as

$$b = R\gamma + r, \tag{2.10}$$

where R is a known $(n\bar{k} \times n\bar{k} - q)$-dimensional real matrix with full column rank, $\text{rk}\, R = n\bar{k} - q$, γ is the $(n\bar{k} - q)$-dimensional vector of remaining free parameters and r is an $n\bar{k}$-dimensional vector of known real constants. In case of zero restrictions $r = 0$ and R is a matrix consisting of 0's and 1's only. R can then be seen as a selection matrix; it selects all variables in $(X' \otimes I_n)$ that are supposed to explain the corresponding components in y. Returning to the general case, we have,

$$\begin{aligned} y &= (X' \otimes I_n)(R\gamma + r) + \epsilon, \quad \text{or reformulated} \\ y - (X' \otimes I_n)r &= (X' \otimes I_n)R\gamma + \epsilon. \end{aligned} \tag{2.11}$$

The LS estimate of γ minimizing function $S(b(\gamma))$ from (2.6) is then given by

$$\hat{\gamma} = (R'(XX' \otimes \Sigma_\epsilon^{-1})R)^{-1} R'(X \otimes \Sigma_\epsilon^{-1})(y - (X' \otimes I_n)r). \tag{2.12}$$

In contrast to the unrestricted case the minimizer of the generalized or weighted sum of squared errors and that of the ordinary sum of squared errors may not be the same. The weighted LS estimator is preferred at this point, due to the fact that in general its asymptotic variance covariance matrix is smaller than that of the OLS estimator $\hat{\gamma}_{OLS} = (R'(XX' \otimes I_n)R)^{-1}R'(X \otimes I_n)(y - (X' \otimes I_n)r)$, where "smaller" has to be understood in the sense of the ordering of positive semidefinite matrices ($A, B \geq 0$, square matrices, then $A \geq B$, if $A - B \geq 0$).

Theorem 2.4 (Consistency and Asymptotic Normality of $\hat{\gamma}$) *Let (y_t) be a VAR(p)X process as presented in eq. (2.11) with $\text{rk}(R) = n\bar{k} - q$, and let the assumptions 2.1 hold, $\Gamma_X := \mathbb{E}X_t X_t' > 0$, with X_t as in (2.2), then the multivariate LS estimate $\hat{\gamma}$ from eq. (2.12) is consistent and asymptotically normal, i.e.*

(i) $\operatorname*{plim}_{T\to\infty} (\hat{\gamma} - \gamma) = 0$,

(ii) $\sqrt{T}(\hat{\gamma} - \gamma) \xrightarrow{d} \mathcal{N}(0, (R'(\Gamma_X \otimes \Sigma_\epsilon^{-1})R)^{-1})$.

Proof. The results can be shown analogously to those in theorem 2.1, by noting that

$$\hat{\gamma} - \gamma = \underbrace{\left(R'(\frac{1}{T}XX' \otimes \Sigma_\epsilon^{-1})R\right)^{-1}}_{(a)} \underbrace{\left(R'(\frac{1}{T}X \otimes \Sigma_\epsilon^{-1})\right)}_{(b)} \epsilon.$$

(a) converges in probability to $(R'(\Gamma_X \otimes \Sigma_\epsilon^{-1})R)^{-1}$ and (b) can be written as $R'(I_{\bar{k}} \otimes \Sigma_\epsilon^{-1})\text{vec}(\frac{1}{T}EX')$. It has already been shown that $\underset{T\to\infty}{\text{plim}}\ \text{vec}(\frac{1}{T}EX') = 0$ and $\sqrt{T}\text{vec}(\frac{1}{T}EX') \xrightarrow{d} \mathcal{N}(0, \Gamma_X \otimes \Sigma_\epsilon)$ by assumption.
□

$\hat{\gamma}$ as in eq. (2.12), however, is of little practical use, since in practice the variance covariance matrix Σ_ϵ is unknown. A two stage estimator $\tilde{\gamma}$ is used instead. In stage one a consistent estimator, $\tilde{\Sigma}_\epsilon$, for Σ_ϵ is computed, e.g. the ML estimate of the unrestricted model see eq. (2.8), which in stage two substitutes the actual error variance covariance in eq. (2.12), giving

$$\tilde{\gamma} = (R'(XX' \otimes \tilde{\Sigma}_\epsilon^{-1})R)^{-1}R'(X \otimes \tilde{\Sigma}_\epsilon^{-1})(y - (X' \otimes I_n)r). \qquad (2.13)$$

Since $\tilde{\Sigma}_\epsilon$ is a consistent estimator, the asymptotic results of theorem 2.4 translate to the two stage LS estimator $\tilde{\gamma}$ see e.g. (Lütkepohl, 1993, Proposition 5.3) and, therefore, also to $\tilde{b} = R\tilde{\gamma} + r$, for which we have

$$\begin{aligned} &\underset{T\to\infty}{\text{plim}}\,(\tilde{b} - b) = 0, \\ &\sqrt{T}(\tilde{b} - b) \xrightarrow{d} \mathcal{N}(0, R(R'(\Gamma_X \otimes \Sigma_\epsilon^{-1})R)^{-1}R'). \end{aligned}$$

An alternative consistent estimator $\tilde{\Sigma}_\epsilon$[9] is obtained by $\tilde{\Sigma}_\epsilon = \frac{1}{T}(Y - \hat{B}_{OLS}X)(Y - \hat{B}_{OLS}X)'$, where $\text{vec}(\hat{B}_{OLS}) = \hat{b}_{OLS} = R\hat{\gamma}_{OLS} + r$. Thus, the ordinary sum of least squares in the restricted framework is minimized in stage one. This might be a better choice than taking the ML estimator from the unrestricted model, since the information about the restricted parameter space is used. Thus, if there is a strong belief in the underlying restrictions one might use the latter estimator $\tilde{\Sigma}_\epsilon$.
Finally, one may also follow some iterative estimation procedure, that starts with estimating $\hat{\gamma}_{OLS}$ and iteratively performs step one and two of the two stage LS estimation procedure until convergence is obtained. Thus, $\tilde{\Sigma}_\epsilon^{(1)}$ is obtained from $\hat{\gamma}_{OLS}$ and gives the two stage LS estimate $\tilde{\gamma}^{(1)}$, which in the second iteration step is used to compute a new estimate for Σ_ϵ, $\tilde{\Sigma}_\epsilon^{(2)}$ a.s.o.

The following theorem shows that the restricted estimator $\tilde{b}$ asymptotically is as efficient or even more efficient than the unrestricted estimator $\hat{b}$.

Theorem 2.5 (Asymptotic Comparison of $\hat{b}$ and $\tilde{b}$) *Let (y_t) be as in theorem 2.4 and let $\hat{b} = ((XX')^{-1}X \otimes I_n)y$ be the unrestricted estimator and $\tilde{b} = R(R'(XX' \otimes \Sigma_\epsilon^{-1})R)^{-1}(R'(X \otimes \Sigma_\epsilon^{-1}))(y - (X' \otimes I_n)r) + r$ be the restricted estimator of the true parameter vector b. Then,*

$$(\Gamma_X^{-1} \otimes \Sigma_\epsilon) \geq R(R'(\Gamma_X \otimes \Sigma_\epsilon^{-1})R)^{-1}R'$$

Proof. Let $M := (\Gamma_X^{-1} \otimes \Sigma_\epsilon)$. Due to the assumptions 2.1, it follows that $M > 0$, see also lemma B.2. Hence,

$$\begin{pmatrix} M & R \\ R' & R'M^{-1}R \end{pmatrix} = \begin{pmatrix} I_{n\bar{k}} \\ R'M^{-1} \end{pmatrix} M \begin{pmatrix} I_{n\bar{k}} \\ R'M^{-1} \end{pmatrix}' \geq 0,$$

and therefore, application of lemma B.1 from the appendix on this matrix finalizes the proof. □

It should be stressed however, that the above result is obtained under the assumption and validity of a restricted parameter space. If the observations are generated by an unrestricted process the ordering of variances may not hold anymore. Hence, imposing restrictions may alleviate the problems arising from a high dimensional parameter space, but they should always be confirmed by theoretical findings about

[9] Consistency of $\tilde{\Sigma}_\epsilon$ here follows due the fact that under the assumptions 2.1

$$\sqrt{T}(\hat{\gamma}_{OLS} - \gamma) \xrightarrow{d} \mathcal{N}(0, (R'(\Gamma_X \otimes I_n)R)^{-1}(R'(\Gamma_X \otimes \Sigma_\epsilon)R)(R'(\Gamma_X \otimes I_n)R)^{-1}),$$

and proposition 3.2 and corrolary 3.2.1 in (Lütkepohl, 1993) apply. See also (Lütkepohl, 1993, Proposition 5.4).

the underlying system or statistical tests and information criteria, respectively. Section 2.3 will deal with the statistical procedures that detect or confirm zero restriction on elements of the parameter matrix B.

This section covers linear constraints on the coefficients vector b only. A popular model class that contains certain non-linear restrictions will be presented in chapter 3.

2.3 Model specification and input selection

Model specification in the VARX framework concerns two types of parameters: the integer valued lag order p determining the dynamic and the real valued parameters in $c, A_1, \ldots, A_p$, D, and Σ_ϵ.
The selection of the lag order p is difficult and the final choice depends highly on the analyst's, say, "selection tool". Econometric literature provides a wide range of different selection procedures, for instance, the application of *Information Criteria* (IC), see e.g. (Hannan and Deistler, 1988), or methods as the likelihood ratio test proposed by (Bartlett, 1938a) where p is determined by running a sequence of tests, or methods as described in (Tiao, 2001) basing on the sample cross-correlations, sample partial autoregression matrices and the diagonal elements of the estimated residual covariance matrix, which provide a measure of the extent to which the fit is improved as the order p is increased. G.C. Tiao himself calls these procedures "tentative" specification procedures and diagnostic checking is indispensable in order to prevent model misspecification.
Here, the focus is on totally data driven specification and input selection methods. Due to the high number of input variable candidates it is necessary to find a reasonable subset of explanatory variables and it might, in addition, be useful to impose zero restrictions on the elements in the matrices $A_1, \ldots, A_p$ and D in order to reduce the number of parameters to estimate. IC give one solution to this selection problem and for several reasons are preferable to the latter two methods mentioned above. An IC is composed as the sum of two measures, one reflecting the goodness of fit and one indicating the corresponding model complexity. Hence, if forecasting is the objective of model specification, one can choose an appropriate measure for the forecasting performance and combine it with a model complexity index. In comparison with testing sequences the actual objective may, therefore, already be part of the selection instrument, which might be meaningful. Furthermore, testing sequences have a positive probability of choosing the incorrect order p even for large samples, if their significance level does not go to zero while sample size T approaches infinity, see (Lütkepohl, 1993, section 4.3.3) for a discussion. IC, on the other hand, can be shown to be consistent, if their measure of model complexity fulfills certain limit conditions for T going to infinity, see e.g. (Lütkepohl, 1993, proposition 4.2) and references cited there for consistency of VAR order estimators. Last but not least, they are easy to compute and not as time consuming as possibly the procedure proposed by G.C. Tiao, which might give more insight into the data generating process, but might be infeasible, if model specification has to be done automatically, without any correcting interaction of the analyst.
The IC used throughout the applications of this thesis are *Akaike's Information Criterion* (AIC), see (Akaike, 1973; Akaike, 1974) and *Schwarz's Bayesian Information Criterion* (referred to as SC and BIC, respectively), see (Schwarz, 1978):

$$\begin{aligned} AIC(\tilde{k}) &= \log\det\hat{\Sigma}_\epsilon + \tfrac{2}{T}\tilde{k} \\ BIC(\tilde{k}) &= \log\det\hat{\Sigma}_\epsilon + \tfrac{\log T}{T}\tilde{k}, \end{aligned} \tag{2.14}$$

where $\tilde{k}$ is the actual number of functionally independent or free parameters, T is the underlying sample size and $\hat{\Sigma}_\epsilon$ is the sample variance covariance matrix of the estimated residuals. For $T > 7$ the "complexity" term of BIC is larger than that of AIC and, therefore, BIC tends to be more restrictive. For a single equation or $n = 1$, it can be shown however, that if the true model is element of the set of model specifications under investigation and the BIC is computed for all these model specifications, for T going to infinity the minimal BIC value is attained at the true specification, see e.g. (An and

Gu, 1985, theorem 1), while the minimal AIC value is attained at a model specification that is at least as "complex" as the true model. Anyway, in small samples AIC might be better in the sense that the true model specification might be selected more often especially in cases where the model structure is not parsimonious. Thus, it is not really clear which IC performs better or disposes of the better measure of complexity. An information criterion that is based on an adaptive measure of complexity, where "adaptive" here has to be understood in relation to the respective data set at hand, is presented in (Ye, 1998).

Since, in general, the number of all possible model specifications becomes large easily, the analyst is confronted with two problems in practice. First, it is hardly ever possible to estimate the whole set of specifications and select the one giving the minimal IC value. In other words, an *exhaustive search* is almost always impracticable. Consider for instance an n-dimensional VAR process including intercept. If the upper bound for the lag order p is assumed to be P, there are $\sum_{i=0}^{n+Pn^2} \binom{n+Pn^2}{i} = 2^{n+Pn^2}$ possible model specifications. For $n = 2$ and $P = 3$ this would already give $2^{14} = 16384$. Second, and what is even more serious than the arising computational costs, the problem of overfitting may occur, if we search for too many (relative to sample size T) model specifications using always the same dataset, see e.g. (Ye, 1998; White, 2000). This means that the goodness of fit statistics of the final model obtained from the application of a sequence of statistical and nonstatistical tools to a dataset, what is also called data mining, may be too optimistic, and hence, misleading. What can be done however, is to search for the IC-optimal specification along some especially chosen path. Several procedures of the kind have been proposed in literature: *Bottom up* and *Top Down* methods as in (Lütkepohl, 1993); fully automatic *General to Specific* selection procedures for single equation models combining diversified test batteries and IC, see (Krolzig and Hendry, 2001; Hendry and Krolzig, 2001); *Branch and Bound* or *Leaps and Bound* procedures, see e.g. (Furnival and Wilson, 1974); *Elimination of Complete Matrices* in (Penm and Terrell, 1982); versions of a *Backward algorithm*, *Forward algorithm* and *Fast Step Procedure* (FSP) for single equation models can be found in (An and Gu, 1985; An and Gu, 1989); and many more.

In this thesis we focus on the Forward algorithm and the FSP proposed by An and Gu and generalize them for the use in multivariate settings. For the sake of completeness, let us first describe in a few words the concepts of the two algorithms basing on an IC (here AIC and BIC respectively) for the univariate case, i.e. $n = 1$:

Forward algorithm: Let $\bar{k}$ be the number of candidate variables. The forward procedure looks for the IC optimal singleton, giving set $\mathcal{S}_1$ say, then for the IC optimal explanatory variable to be added to the singleton, giving set $\mathcal{S}_2$, and so on, until all $\bar{k}$ candidate variables are included and give the largest possible set $\mathcal{S}_{\bar{k}}$. Out of these thus obtained $\bar{k}$ sets, $\mathcal{S}_1, \ldots, \mathcal{S}_{\bar{k}}$, for which $\mathcal{S}_1 \subset \mathcal{S}_2 \subset \ldots \subset \mathcal{S}_{\bar{k}}$ holds, the set with the lowest criterion value is chosen to be the final set.

Thus, for the forward algorithm just $\frac{\bar{k}(\bar{k}+1)}{2}$ possible model specifications have to be considered, which is much less than for the exhaustive procedure. It is obvious that the sets $\mathcal{S}_k$, $1 < k < \bar{k}$, depend strongly on the previously selected sets $\mathcal{S}_1, \ldots, \mathcal{S}_{k-1}$ and are not necessarily the IC optimal subset of $\mathcal{S}_{\bar{k}}$ with cardinality[10] k. The following FSP procedure allows to, loosely spoken, "move" from one path of model specifications to another in order to improve the IC value and counter the shortcoming of the forward procedure.

FSP: In a first step one has to look for an IC optimal initial set, which can for instance be found by application of the above mentioned forward algorithm. Given this initial set, now in a second step, a local search is performed as follows: the initial set is enlarged by adding one single variable to the set or reduced by dropping one variable from the set and the IC optimal set out of these is compared with the initial set with respect to the IC value. The procedure is iterated until the criterion value can no longer

[10] The cardinality of a finite set is defined as the number of elements in the set.

be decreased by adding or omitting variables.

To adapt these ideas to a system of equations we investigate two different ways for determining an initial set. One way is to apply a forward algorithm to all equations at once, i.e. in each step of the forward algorithm we search for the IC optimal explanatory variable to be added to all equations. In terms of the parameter matrix B from eq. (2.3) this means, that we start with zero restrictions on the whole matrix and cancel out the restrictions in the column corresponding to the IC optimal singleton, then the restrictions are canceled out for the column corresponding to the IC optimal variable added to the singleton and so forth until there is no further column of zeros left in B or equivalently the set of explanatory variables contains all explanatory variable candidates. The set of variables out of the $\bar{k}$ sets giving the minimal IC value is then chosen as initial set. This method is referred to as *mva method.* The other method (called *univ method*) applies the forward algorithm proposed by An and Gu to each single equation. Hence, it is likely that not whole columns of B are set to zero but single elements.
Given the initial specification of B, the IC value is tried to be decreased by altering the zero restrictions on the single elements of B. So, in each iteration step one looks for that element in B for which the change in its status ("restricted to zero" or "free in $\mathbb{R}$") yields the highest decrease in IC value. The procedure stops when the IC value can no longer be improved by changing the status of one single element in B.

The above presented variable selection processes are all discrete and can hence be extremely variable, i.e. small changes in the data can result in completely different models being selected. This can decrease prediction accuracy. Continuous alternatives are for instance given by *Shrinkage estimators* such as the *Ridge* and *Lasso*[11] estimator respectively, or *Bayes estimation* methods. The first two impose (nonlinear - in case of Ridge) parameter restrictions, by bounding from above the Euclidean and L1 norm respectively of the parameter vector b. Bayesian estimators base on the prior information about the density of the parameter vector. The restriction of the parameter space is then given by the prior variance covariance matrix of the parameter vector. In any case, the analyst has to define either the upper bound for the parameter vector norm or mean and variance of the prior density of the parameter vector, which again leads to a parameter specification problem that however, is not going to be discussed here. For further details and simulation studies see (Brown, 1994; Tibshirani, 1996) and (Lütkepohl, 1993, section 5.4).

2.4 Forecasting

The problem in forecasting is to find an optimal approximation of the future values or realizations of a process by a (linear, affine or most general a measurable) function of its current and past realizations plus some exogenous variables. Optimality of course has to be understood with respect to some criterion measuring the approximation quality. A frequently used criterion is the *least squares criterion* and the resulting predictor is referred to as the *mean squared error* (MSE) predictor. Thus, the task is to solve the following minimization problem for each component of y_t,

$$\min_{y_{i,t+h|t}} MSE(y_{i,t+h|t}) = \min_{y_{i,t+h|t}} \mathbb{E}(y_{i,t+h} - y_{i,t+h|t})^2, \quad h > 0,\ i = 1, \ldots, n, \qquad (2.15)$$

where t denotes present time and the predictor $y_{t+h|t} = g_h(y_t', y_{t-1}', \ldots, x_t', x_{t-1}', \ldots, 1)$ is a function of the (possibly finite) past of (y_t) and the exogenous (x_t) and a constant. If $g_h(.)$ is taken out of the class of measurable functions, it can be shown that the conditional expectation $\mathbb{E}_t(y_{t+h}) := \mathbb{E}(y_{t+h}|y_t', y_{t-1}', \ldots, x_t', x_{t-1}', \ldots, 1)$ is the optimal predictor of y_{t+h}. This can be seen as

$$\begin{aligned} MSE(y_{t+h|t}) &= \mathbb{E}\left[(y_{t+h} - y_{t+h|t})(y_{t+h} - y_{t+h|t})'\right] = \\ &= \mathbb{E}\left[(y_{t+h} - \mathbb{E}_t(y_{t+h}) + \mathbb{E}_t(y_{t+h}) - y_{t+h|t})(y_{t+h} - \mathbb{E}_t(y_{t+h}) + \mathbb{E}_t(y_{t+h}) - y_{t+h|t})'\right] = \\ &= MSE(\mathbb{E}_t(y_{t+h})) + \mathbb{E}\left[(\mathbb{E}_t(y_{t+h}) - y_{t+h|t})(\mathbb{E}_t(y_{t+h}) - y_{t+h|t})'\right], \end{aligned}$$

[11] Standing for "Least absolute shrinkage and selection operator".

and hence, $MSE(y_{t+h|t}) \geq MSE(\mathbb{E}_t(y_{t+h}))$, (in the sense of the ordering of positive semidefinite matrices) for any predictor $y_{t+h|t}$. The term $\mathbb{E}\left[(y_{t+h} - \mathbb{E}_t(y_{t+h}))(\mathbb{E}_t(y_{t+h}) - y_{t+h|t})'\right]$ and its transpose can be shown to be zero by application of the law of iterated expectations, or by noting that the two bracket terms are uncorrelated by construction. The first bracket term is adjusted from the influence of the present and past of (y_t) and hence, depends on future ϵ_t's and ξ_t's only, and the second bracket term is dependent on the current and past ϵ_t's and ξ_t's.

Since (y_t) is assumed to be a VARX process as in eq. (2.1) and due to the assumptions 2.1, we have

$$\mathbb{E}_t(y_{t+h}) = c + A_1\mathbb{E}_t(y_{t+h-1}) + \ldots + A_p\mathbb{E}_t(y_{t+h-p}) + D\mathbb{E}_t(x_{t+h-1}) + \underbrace{\mathbb{E}_t(\epsilon_{t+h})}_{=0},$$

which for $h = 1$ is

$$\mathbb{E}_t(y_{t+1}) = c + A_1 y_t + \ldots + A_p y_{t-p+1} + Dx_t.$$

Hence, under these assumptions the conditional expectation is a linear function of the present and past observations of (y_t). It is unbiased, since its prediction error is zero in expectation, $\mathbb{E}(y_{t+h} - \mathbb{E}_t(y_{t+h})) = 0$. The $MSE(\mathbb{E}_t(y_{t+h}))$ can thus be seen as the prediction error variance.

If assumptions 2.1 (i) and (ii) are relaxed such that (ϵ_t) is no longer asked to be an independent process and to be independent of the process (ξ_t), but is just a white noise process that is uncorrelated with (ξ_t), $\mathbb{E}_t(\epsilon_{t+h}) = 0$ does not hold, in general. Hence, the conditional expectation cannot be computed without imposing further assumptions on the underlying processes. Instead of doing the latter, we might rather shrink the class of functions over which eq. (2.15) is optimized from measurable to affine functions. In case of affine functions $g_h(.)$, the problem in eq. (2.15) is equivalent to the problem of finding that $y_{t+h|t}$, element of the Hilbert space[12], $\mathbb{H}(y_t', y_{t-1}', \ldots, x_t', x_{t-1}', \ldots, 1)$, spanned by the components of the present and past y_t's, x_t's and by the constant 1, for which the distance norm, $\left\|y_{t+h} - y_{t+h|t}\right\|^2 = \langle y_{t+h} - y_{t+h|t}, y_{t+h} - y_{t+h|t}\rangle = \mathbb{E}(y_{t+h} - y_{t+h|t})(y_{t+h} - y_{t+h|t})'$, is minimal. The following theorem shows that the projection of y_{t+h} on this Hilbert space is the unique solution to this problem.

Theorem 2.6 (Projection Theorem) *Let $\mathbb{H}$, $\mathbb{M}$ be Hilbert spaces and $\mathbb{M} \subset \mathbb{H}$. Then for every $x \in \mathbb{H}$ there exists a unique decomposition*

$$x = \hat{x} + u,$$

such that $\hat{x} \in \mathbb{M}$ and $u \in \mathbb{M}^{\perp}$ (i.e. $\langle y, u\rangle = 0$ for all $y \in \mathbb{M}$). In addition, $\hat{x}$ is the unique element of $\mathbb{M}$ satisfying

$$\|x - \hat{x}\| = \min_{y \in \mathbb{M}} \|x - y\|.$$

Proof. A proof of this theorem can be found in (Brockwell and Davis, 1989, p.51 ff). □

Hence, if $P_{\mathbb{H}}$ is a projector on the above Hilbert space, $\mathbb{H}(y_t', y_{t-1}', \ldots, x_t', x_{t-1}', \ldots, 1)$, the optimal predictor for y_{t+h} is given by

$$P_{\mathbb{H}} y_{t+h} = c + P_{\mathbb{H}} A_1 y_{t+h-1} + \ldots + P_{\mathbb{H}} A_p y_{t+h-p} + P_{\mathbb{H}} D x_{t+h-1} + \underbrace{P_{\mathbb{H}} \epsilon_{t+h}}_{=0},$$

$P_{\mathbb{H}}\epsilon_{t+h} = 0$, since ϵ_{t+h}, $h > 0$, is uncorrelated with y_s, x_s for $s \leq t$ and the constant. For $h = 1$ we obtain again

$$P_{\mathbb{H}} y_{t+1} = c + A_1 y_t + \ldots + A_p y_{t-p+1} + Dx_t.$$

Note, however, that $P_{\mathbb{H}} y_{t+h}$ coincides with the conditional expectation only if (ϵ_t) is independent white noise that is, in addition, independent of (ξ_t). Otherwise, $P_{\mathbb{H}} y_{t+h}$ is "just" the optimal affine predictor of y_{t+h}. Let from now on $y_{t+1|t} := c + A_1 y_t + \ldots + A_p y_{t-p+1} + Dx_t$. If the true intercept and coefficient

[12] See appendix A for definitions and properties of Hilbert spaces of square integrable random variables.

matrices are replaced by their estimators from section 2.2, we get $\hat{y}_{t+1|t} := \hat{c}+\hat{A}_1 y_t+\ldots+\hat{A}_p y_{t-p+1}+\hat{D}x_t$. For a detailed discussion on the $MSE(\hat{y}_{t+1|t})$ in the VAR framework see (Lütkepohl, 1993, section 3.5). It should be mentioned however, that the variance of the prediction error $y_{t+1}-\hat{y}_{t+1|t}$ may be driven by three different sources of uncertainty: First of all, the uncertainty coming from the noise or innovation term; second, the uncertainty caused by estimation of the true parameters; and third, if the true model structure is unknown, the uncertainty due to model specification. Asymptotically, however, the contributions from the latter two causes to the overall forecast uncertainty should vanish, if also the specification procedure is consistent.

Chapter 3

Reduced Rank Regression

One of the major concerns in modelling multiple time series is the possibly large set of parameters involved in the underlying model. Two problems may arise in this regard: First, sample size T, in practice, may be or often is small compared to the overall number of parameters. Thus, estimation accuracy of all regression coefficients may be low. Second, irrespective of sample size, interpretation for a large number of parameters can become unwieldy. In the previous chapter we have already pointed out these problems and proposed one possibility to reduce the dimension of the parameter space in the framework of VARX models. This has been done by allowing for linear constraints to be imposed on the coefficient matrix B. These restrictions, however, were assumed to be known. In addition, as we have seen before, estimation of the full model - without imposing any restrictions - in a multivariate linear LS framework is just the same as estimating every singly equation by OLS. Hence, the fact that the multiple responses are likely to be related is not involved in estimation.

Reduced Rank (RR) regression models allows for a rank deficient regression coefficient matrix and thereby reduces the aforementioned problems. Thus on the one hand, the parameter matrix can be constructed as the product of two rectangular matrices that consist of less columns than the original coefficient matrix B, and hence, involve less parameters. On the other hand, the rank deficiency of the coefficient matrix implies that the kernel of the matrix is not empty. That is, there exists some non-zero matrix C, for which $CB' = 0$ holds. In other words, the parameters in B are again restricted by linear constraints, namely $CB' = 0$. However, in contrast to the previous chapter, here, the linear restrictions are unknown a priori. In addition, the correlation structure of the endogenous variables is taken into account during estimation. This can be seen due to its relation to *principal components* (PC) and *canonical correlation* (CC) analysis, see for instance (Reinsel and Velu, 1998).

RR was first considered by Anderson in (Anderson, 1951). More than twenty years later Izenman introduced the term "reduced rank regression" in (Izenman, 1975). RR models, in the meanwhile, are in a widespread use. The book (Reinsel and Velu, 1998) contains a detailed list of references to all sorts of disciplines working with RR models.

3.1 The model

Let in the following the n-dimensional variable of responses, y_t, and the k-dimensional exogenous variables, x_t, be mean-adjusted[1]. Thus, $\mathbb{E}y_t = 0$ and $\mathbb{E}x_t = 0$, and eq. (2.1) can be written as

[1] If y_t and x_t are not mean-adjusted and there is no extra additive constant term included (on the right hand side) in the third line of eq. (3.1), an additional restriction on the mean of the response variable y_t is introduced; namely, (given that $\mathbb{E}\epsilon_t = 0$ holds), $\mathbb{E}y_t = FG'\mathbb{E}X_t$. Thus, the mean of y_t has to be element of the space spanned by the columns of F.
A model including an additive constant term (or a whole second set of regressors, whose coefficient matrix has full rank) is examined e.g. in (Reinsel and Velu, 1998, chapter 3). Here, however, we consider the simpler case of mean-adjusted

$$\begin{aligned} y_t &= A_1 y_{t-1} + A_2 y_{t-2} + \ldots + A_p y_{t-p} + D x_{t-1} + \epsilon_t \\ &= [A_1, A_2, \ldots, A_p, D]\, X_{t-1} + \epsilon_t \\ &= B X_t + \epsilon_t, \quad t \in \mathbb{Z}, \end{aligned} \tag{3.1}$$

where (x_t), (with $\mu_x = 0$), (ϵ_t) and the $n \times n$ and $n \times k$ dimensional coefficient matrices $A_1, \ldots, A_p, D$ fulfill assumptions 2.1, $X_t = (y'_{t-1}, \ldots, y'_{t-p}, x'_{t-1})'$, $B = [A_1, \ldots, A_p, D]$ and $\text{rk}(B) = r \leq \min(n, \bar{k})$, where $\bar{k} = pn + k$, is the dimension of X_t. Hence, B can be written as the product of two rectangular matrices F and G of dimension $n \times r$ and $\bar{k} \times r$ with $\text{rk}(F) = \text{rk}(G) = r$.

The model may among others be interpreted as follows: The information provided in X_t is "compressed" or summarized in the linear combinations $G'X_t$, which is then via F transferred to y_t. RR is therefore sometimes also referred to as index (for $r = 1$), see e.g. (Sargent and Sims, 1977), or factor model, see e.g. (van der Leeden, 1990, chapter 5), where the factor f_t is given by $f_t = G'X_t$ and the matrix F stores the corresponding loadings.

3.2 Identifiability of F and G

Note that, even if the product $B = FG'$ is identified, the matrices F and G are not uniquely identified without imposing further normalizing conditions, since for any non-singular $r \times r$-dimensional matrix M it holds that

$$B = FG' = FM^{-1}MG' = \tilde{F}\tilde{G}',$$

where $\tilde{F} = FM^{-1}$ and $\tilde{G} = GM'$. Let $\mathcal{B} = \left\{(FM^{-1}, GM') \mid M \in \mathbb{R}^{r \times r}, M \text{ non-singular}\right\}$ denote the set of equivalent decompositions of B. In order to obtain unique matrices F and G, or in other words, to select a unique representative $(\bar{F}, \bar{G})$ out of $\mathcal{B}$, r^2 normalization conditions may be imposed.

One might, for instance, select matrices (F, G) that show a certain structure: Consider $\tilde{G}' = \left[\tilde{G}_1', \tilde{G}_2'\right]$, where $\tilde{G}_1$ is of dimension $r \times r$ and $\tilde{G}_2$ is of dimension $(\bar{k} - r) \times r$, and let the variables in X_t be arranged such that the square matrix $\tilde{G}_1$ is non-singular, then

$$B = \tilde{F}\left[\tilde{G}_1', \tilde{G}_2'\right] = \underbrace{\tilde{F}\tilde{G}_1'}_{F} \underbrace{\left[I_r, \tilde{G}_1'^{-1}\tilde{G}_2'\right]}_{G'}. \tag{3.2}$$

Hence, the first upper $r \times r$ sub-block matrix of G is the r-dimensional identity matrix, which also eliminates further unidentifiability concerning sign changes in the columns of F and G. (Of course, one can alternatively ask F to be of the form $[I_r, F_2']'$.)

In the following lemma we introduce an alternative set of normalization conditions that is often referred to in the literature. Note that throughout this chapter we consider symmetric square roots of symmetric matrices, see also appendix C.

Lemma 3.1 *Given $B \in \mathbb{R}^{n \times \bar{k}}$ with $\text{rk}\, B = r \leq \min(n, \bar{k})$ and two arbitrary positive definite symmetric matrices Γ and M_{XX} of order n and $\bar{k}$ respectively, there exists always a pair $(F_1, G_1) \in \mathcal{B}$ that fulfills the following r^2 normalization conditions:*

$$F'\Gamma^{-1}F = I_r, \tag{3.3}$$

$$G'M_{XX}G = \Lambda_r^2, \tag{3.4}$$

where Λ_r is a diagonal matrix.

If it holds furthermore that $\phi_{1i} > 0$, $i = 1, \ldots, r$, where ϕ_{1i} denotes the ith entry in the first row of F, and if the diagonal elements of Λ_r are distinct and ordered, i.e. $\lambda_1 > \ldots > \lambda_r > 0$, the pair (F_1, G_1) is also unique in $\mathcal{B}$.

variables and one set of regressors with a rank-deficient coefficient matrix B.

Proof. 1. Existence of a pair $(F_1, G_1) \in \mathcal{B}$ that fulfills the conditions (3.3) and (3.4): It is easy to verify that $F_1 = \Gamma^{\frac{1}{2}} U_r$, and $G_1 = B'\Gamma^{-\frac{1}{2}} U_r$, where $U_r \in \mathbb{R}^{n \times r}$ with $U_r' U_r = I_r$ contains the normalized eigenvectors corresponding to the first r eigenvalues of $\Gamma^{-\frac{1}{2}} B M_{XX} B' \Gamma^{-\frac{1}{2}}$, fulfill the above conditions (3.3) and (3.4). To see that $F_1 G_1' = B$, let $S = \Gamma^{-\frac{1}{2}} B M_{XX}^{\frac{1}{2}}$, and note that $SS' = \Gamma^{-\frac{1}{2}} B M_{XX} B' \Gamma^{-\frac{1}{2}}$. Hence, matrix Λ_r from eq. (3.4) contains the non-zero singular values of S. Since $\operatorname{rk} S = r$, the singular value decomposition of S may be written as $S = U \Lambda V' = U_r \Lambda_r V_r'$, where U_r and V_r consist of the r columns of the $n \times n$ and $\bar{k} \times \bar{k}$ dimensional orthonormal matrices U and V corresponding to the r non-zero singular values of S. In other words, U_r and V_r contain in their r columns the normalized eigenvectors of SS' and $S'S$ respectively, corresponding to the non-zero eigenvalues Λ_r^2. Hence, $F_1 G_1' = \Gamma^{\frac{1}{2}} U_r U_r' S M_{XX}^{-\frac{1}{2}} = \Gamma^{\frac{1}{2}} S M_{XX}^{-\frac{1}{2}} = B$.
2. Any $B \in \mathbb{R}^{n \times \bar{k}}$ with $\operatorname{rk} B = r \leq \min(n, \bar{k})$ can be written as the product of such matrices F_1 and G_1, i.e. given a pair $(F, G) \in \mathcal{B}$ there exists always a non-singular $M \in \mathbb{R}^{r \times r}$, such that $FM^{-1} = F_1$ and $GM' = G_1$, namely $M = U_r' \Gamma^{-\frac{1}{2}} F$.
3. Uniqueness of $(F_1, G_1) \in \mathcal{B}$: Suppose that $\phi_{1i} > 0$, $i = 1, \ldots, r$ and $\lambda_1 > \ldots > \lambda_r > 0$ holds. Due to $\lambda_1 > \ldots > \lambda_r > 0$ the eigenvalue decomposition of SS' is unique up to sign changes. This non-uniqueness is then eliminated by the condition $\phi_{1i} > 0$, $i = 1, \ldots, r$. □

Note that an alternative set of conditions to (3.3), (3.4) again with respect to the two positive definite symmetric matrices Γ and M_{XX} is given by

$$F'\Gamma^{-1}F = \Lambda_r^2, \tag{3.5}$$

$$G' M_{XX} G = I_r. \tag{3.6}$$

Note that the diagonal matrices Λ_r in (3.4) and (3.5) are identical. Analogously to above, it can be shown that $F_2 = BM_{XX}^{\frac{1}{2}} V_r$, $G_2 = M_{XX}^{-\frac{1}{2}} V_r$, with V_r defined as in the proof above, fulfill these conditions and one can always find such a pair (F_2, G_2) in $\mathcal{B}$. If $\phi_{1i} > 0$, $i = 1, \ldots, r$ and $\lambda_1 > \ldots > \lambda_r > 0$ holds, (F_2, G_2) is also unique in $\mathcal{B}$. The relation between the two corresponding to the respective set of normalization conditions unique representatives of $\mathcal{B}$ is given by

$$F_2 = F_1 \Lambda_r \quad \text{and} \quad G_2 = G_1 \Lambda_r^{-1}.$$

Given the nonlinear restriction $\operatorname{rk} B = r$ and a set of normalization conditions (3.3), (3.4) and given that $\phi_{1i} > 0$, $i = 1, \ldots, r$ and $\lambda_1 > \ldots > \lambda_r > 0$ holds, the total number of functionally independent parameters in B is $nr + r\bar{k} - r^2$. This is a considerable decrease compaired to $n\bar{k}$ in the full model.

3.3 Estimation

Throughout this section assume that r, the rank of B, is known and consider at first only estimation of the real valued parameters in F, G and B, respectively. Estimation or specification of r will be treated in the subsequent section, 3.4.

Given the notation (2.2) (excluding the constant) the reduced rank regression model (3.1) can be written as

$$\begin{aligned} y_t &= BX_t + \epsilon_t, \\ Y &= BX + E, \\ Y &= FG'X + E, \end{aligned} \tag{3.7}$$

with $\operatorname{rk}(B) = \operatorname{rk}(F) = \operatorname{rk}(G) = r \leq \min(n, \bar{k})$.

3.3.1 Maximum likelihood estimators for F and G:

Assume for the time being that (ϵ_t) is normally distributed, consider the multivariate normal likelihood function

$$\begin{aligned} L(F,G,\Sigma_\epsilon|Y,X) &= L(F,G,\Sigma_\epsilon|y_T,\dots,y_1,X_T,\dots,X_1) = \\ &= \underbrace{f(y_T|y_{T-1},\dots,y_1,X_T,\dots,X_1;\ F,G,\Sigma_\epsilon)}_{\sim\mathcal{N}(FG'X_T,\Sigma_\epsilon)}\cdot\ldots\cdot\underbrace{f(y_1|X_1;\ F,G,\Sigma_\epsilon)}_{\sim\mathcal{N}(FG'X_1,\Sigma_\epsilon)} = \\ &= \frac{1}{\sqrt{2\pi}^{Tn}}\frac{1}{\sqrt{\det(\Sigma_\epsilon)}^{T}}\prod_{t=1}^{T}\exp\left[-\tfrac{1}{2}(y_t-FG'X_t)'\Sigma_\epsilon^{-1}(y_t-FG'X_t)\right] \end{aligned} \tag{3.8}$$

and let

$$\begin{aligned} l(F,G,\Sigma_\epsilon|Y,X) &= -\tfrac{2}{T}\log(L(F,G,\Sigma_\epsilon|Y,X)) - n\log(2\pi) \\ &= \log\det(\Sigma_\epsilon) + \tfrac{1}{T}\sum_{t=1}^{T}(y_t-FG'X_t)'\Sigma_\epsilon^{-1}(y_t-FG'X_t) \\ &= \log\det(\Sigma_\epsilon) + \operatorname{tr}\left[\Sigma_\epsilon^{-1}\tfrac{1}{T}(Y-FG'X)(Y-FG'X)'\right] \end{aligned}$$

Thus, maximization of $L(F,G,\Sigma_\epsilon|Y,X)$ is equivalent to the minimization of $l(F,G,\Sigma_\epsilon|Y,X)$.

The following two lemmata are useful to interprete and prove the claims stated in the subsequent theorem 3.1:

Lemma 3.2 (Minimization of the Frobenius norm $\|S-P\|_F$) *Let S be a matrix of order $m\times n$ and of rank $\min(m,n)$ and let P be a matrix of the same size as S but of rank $r(\leq\min(m,n))$. The matrix P that minimizes the squared Frobenius norm $\|S-P\|_F^2=\operatorname{tr}\left[(S-P)(S-P)'\right]$ and that therefore is the best rank r approximation of S in the $\|.\|_F$-sense, is given by $P=U_rU_r'S$, where U_r is $m\times r$ dimensional and the columns of U_r are the r normalized eigenvectors of SS' that correspond to the r largest eigenvalues of SS'.*

Let $S=U\Lambda V'$ be the singular value decomposition of S, hence $P=U_r\Lambda_rV_r'$, where U_r is as above and V_r consists of the r columns of the orthonormal matrix V that correspond to the r largest singular values of S contained in the diagonal matrix Λ_r.

Proof. See (Reinsel and Velu, 1998, theorem 2.1). □

Lemma 3.3 (Simultaneous minimization of singular values of a rectangular matrix $(S-P)$) *Let S be as in lemma 3.2 and let $S=U\Lambda V'$ be its singular value decomposition. For any $m\times n$ dimensional matrix P of rank $r(\leq\min(m,n))$ the following inequalities hold*

$$\lambda_i(S-P)\geq\lambda_{r+i}(S)\quad i=1,\dots,\min(m,n),$$

where $\lambda_i(S)$ denotes the ith largest singular value of matrix S, and $\lambda_{r+i}(S)$ is defined to be zero for $r+i>\min(m,n)$.

The equality holds for all i if and only if $P=U_r\Lambda_rV_r'$. Thus, the singular values of the matrix $S-P$ attain their minimal value simultaneously if and only if $P=U_r\Lambda_rV_r'$.

Proof. See (Rao, 1979, theorem 2.3). □

Theorem 3.1 (ML estimators for F,G and Σ_ϵ) *Let $\mathrm{M}_{XX}:=\frac{1}{T}\sum_{t=1}^{T}X_tX_t'$, $\mathrm{M}_{Xy}=\mathrm{M}_{yX}':=\frac{1}{T}\sum_{t=1}^{T}X_ty_t'$ and $\mathrm{M}_{yy}:=\frac{1}{T}\sum_{t=1}^{T}y_ty_t'$ denote the (centered) sample second moment and cross moment matrices of y_t and X_t.*

Under the assumptions 2.1 with $\mu_x=0$ and $\epsilon_t\sim iid\mathcal{N}(0,\Sigma_\epsilon)$, two pairs of equivalent maximum likelihood estimators for the parameter matrices F and G of model 3.7 are given by

(i) $\hat{F}_1 = \Gamma^{\frac{1}{2}} U_r$ *and* $\hat{G}_1 = \mathrm{M}_{XX}^{-1} \mathrm{M}_{Xy} \Gamma^{-\frac{1}{2}} U_r$,
where U_r *with* $U_r' U_r = I_r$ *is the* $n \times r$ *dimensional matrix of normalized eigenvectors that correspond to the* r *largest eigenvalues of* $\Gamma^{-\frac{1}{2}} \mathrm{M}_{yX} \mathrm{M}_{XX}^{-1} \mathrm{M}_{Xy} \Gamma^{-\frac{1}{2}}$,

(ii) $\hat{F}_2 = \mathrm{M}_{yX} \mathrm{M}_{XX}^{-1} \mathrm{M}_{XX}^{\frac{1}{2}} V_r$ *and* $\hat{G}_2 = \mathrm{M}_{XX}^{-\frac{1}{2}} V_r$,
where V_r *with* $V_r' V_r = I_r$ *is the* $\bar{k} \times r$ *dimensional matrix of normalized eigenvectors that correspond to the* r *largest eigenvalues of* $\mathrm{M}_{XX}^{-\frac{1}{2}} \mathrm{M}_{Xy} \Gamma^{-1} \mathrm{M}_{yX} \mathrm{M}_{XX}^{-\frac{1}{2}}$,

and where $\Gamma = \hat{\Sigma}_\epsilon = \frac{1}{T} \sum_{t=1}^{T} (y_t - \mathrm{M}_{yX} \mathrm{M}_{XX}^{-1} X_t)(y_t - \mathrm{M}_{yX} \mathrm{M}_{XX}^{-1} X_t)'$ *is the ML estimator for the error variance covariance matrix in the full unrestricted linear model.*
The ML estimate for Σ_ϵ *in the RR case is given by*

$$\hat{\hat{\Sigma}}_\epsilon = \frac{1}{T} \sum_{t=1}^{T} (y_t - \hat{F}\hat{G}' X_t)(y_t - \hat{F}\hat{G}' X_t)',$$

where $\hat{F}, \hat{G}$ *is a pair of ML estimates.*

Proof. Consider function $l(F, G, \Sigma_\epsilon | Y, X)$ and let $W = W(F,G) := \frac{1}{T}(Y - FG'X)(Y - FG'X)'$. Hence, $l(W, \Sigma_\epsilon | Y, X) = \log\det(\Sigma_\epsilon) + \mathrm{tr}\left[\Sigma_\epsilon^{-1} W\right]$ and $\frac{\partial l}{\partial \Sigma_\epsilon} = \Sigma_\epsilon^{-T} - \Sigma_\epsilon^{-1} W \Sigma_\epsilon^{-1} = 0$ implies that $\hat{\hat{\Sigma}}_\epsilon(F,G) = W(F,G)$.
Substitution of Σ_ϵ by $\hat{\hat{\Sigma}}_\epsilon$ gives the concentrated objective function $l(W, \hat{\hat{\Sigma}}_\epsilon | Y, X) = \log\det(W) + n$. This function is minimal, if $\det(W)$ is minimal or equivalently if $\det(\Gamma^{-1} W)$ is minimal, for some positive definite matrix Γ. Let in the following $\Gamma = \hat{\Sigma}_\epsilon$, the ML estimate of Σ_ϵ in the full unrestriced linear model. Due to the properties of the determinant of a matrix $\det(\hat{\Sigma}_\epsilon^{-1} W) = \det(\hat{\Sigma}_\epsilon^{-\frac{1}{2}} W \hat{\Sigma}_\epsilon^{-\frac{1}{2}})$.

$$\begin{aligned} W &= \tfrac{1}{T}(Y - FG'X)(Y - FG'X)' = \\ &= \mathrm{M}_{yy} - \mathrm{M}_{yX} G F' - F G' \mathrm{M}_{Xy} + F G' \mathrm{M}_{XX} G F' = \\ &= \hat{\Sigma}_\epsilon + (\mathrm{M}_{yX} \mathrm{M}_{XX}^{-\frac{1}{2}} - F G' \mathrm{M}_{XX}^{\frac{1}{2}})(\mathrm{M}_{yX} \mathrm{M}_{XX}^{-\frac{1}{2}} - F G' \mathrm{M}_{XX}^{\frac{1}{2}})' \end{aligned}$$

The last equation follows due to $\hat{\Sigma}_\epsilon = \mathrm{M}_{yy} - \mathrm{M}_{yX} \mathrm{M}_{XX}^{-1} \mathrm{M}_{Xy}$. Hence,

$$\det(\hat{\Sigma}_\epsilon^{-\frac{1}{2}} W \hat{\Sigma}_\epsilon^{-\frac{1}{2}}) = \det(I_n + (S - P)(S - P)') = \prod_{i=1}^{n} (1 + \lambda_i (S - P)^2),$$

where $S = \hat{\Sigma}_\epsilon^{-\frac{1}{2}} \mathrm{M}_{yX} \mathrm{M}_{XX}^{-\frac{1}{2}}$, $P = \hat{\Sigma}_\epsilon^{-\frac{1}{2}} F G' \mathrm{M}_{XX}^{\frac{1}{2}}$, and $\lambda_i(S - P)$ denotes the ith singular value of $S - P$, the objective is to simultaneously minimize the singular values of $S - P$. Following lemma 3.3 the singular values of $S - P$ are simultaneously minimized, if $P = U_r \Lambda_r V_r' = U_r U_r' S = S V_r V_r'$, with Λ_r, U_r and, V_r as in lemma 3.3. Thus,

$$\begin{aligned} \hat{B} = \hat{\Sigma}_\epsilon^{\frac{1}{2}} P \mathrm{M}_{XX}^{-\frac{1}{2}} &= \hat{\Sigma}_\epsilon^{\frac{1}{2}} U_r U_r' S \mathrm{M}_{XX}^{-\frac{1}{2}} = \underbrace{\hat{\Sigma}_\epsilon^{\frac{1}{2}} U_r}_{\hat{F}_1} \underbrace{U_r' \hat{\Sigma}_\epsilon^{-\frac{1}{2}} \mathrm{M}_{yX} \mathrm{M}_{XX}^{-1}}_{\hat{G}_1'} = \\ &= \hat{\Sigma}_\epsilon^{\frac{1}{2}} S V_r V_r' \mathrm{M}_{XX}^{-\frac{1}{2}} = \underbrace{\mathrm{M}_{yX} \mathrm{M}_{XX}^{-\frac{1}{2}} V_r}_{\hat{F}_2} \underbrace{V_r' \mathrm{M}_{XX}^{-\frac{1}{2}}}_{\hat{G}_2'} \end{aligned}$$

Due to lemma 3.1 it is clear that $(\hat{F}_1, \hat{G}_1)$ and $(\hat{F}_2, \hat{G}_2)$ are pairs of matrices that correspond to the normalization conditions (3.3), (3.4), and (3.5), (3.6) respectively. Finally, $\hat{\hat{\Sigma}}_\epsilon = W(\hat{F}_i, \hat{G}_i) = \frac{1}{T} \sum_{t=1}^{T} (y_t - \hat{F}_i \hat{G}_i' X_t)(y_t - \hat{F}_i \hat{G}_i' X_t)'$, for $i = 1$ or 2. □

From the above theorem it is obvious that the ML estimators for F and G with $\Gamma = \mathrm{M}_{yy}$ are given by:

$$\tilde{F}_1 = \mathrm{M}_{yy}^{\frac{1}{2}} \tilde{U}_r \qquad \textit{and} \qquad \tilde{G}_1 = \mathrm{M}_{XX}^{-1} \mathrm{M}_{Xy} \mathrm{M}_{yy}^{-\frac{1}{2}} \tilde{U}_r,$$

where $\tilde{U}_r$ is the $n \times r$ dimensional matrix of normalized eigenvectors that correspond to the r largest eigenvalues of $\mathrm{M}_{yy}^{-\frac{1}{2}}\mathrm{M}_{yX}\mathrm{M}_{XX}^{-1}\mathrm{M}_{Xy}\mathrm{M}_{yy}^{-\frac{1}{2}}$ and

$$\tilde{F}_2 = \mathrm{M}_{yX}\mathrm{M}_{XX}^{-1}\mathrm{M}_{XX}^{\frac{1}{2}}\tilde{V}_r \qquad \textit{and} \qquad \tilde{G}_2 = \mathrm{M}_{XX}^{-\frac{1}{2}}\tilde{V}_r,$$

where $\tilde{V}_r$ is the $\bar{k} \times r$ dimensional matrix of normalized eigenvectors that correspond to the r largest eigenvalues of $\mathrm{M}_{XX}^{-\frac{1}{2}}\mathrm{M}_{Xy}\mathrm{M}_{yy}^{-1}\mathrm{M}_{yX}\mathrm{M}_{XX}^{-\frac{1}{2}}$.
Let us now study their relation to the ML estimators with $\Gamma = \hat{\Sigma}_\epsilon$ from theorem 3.1. Consider therefore the symmetric matrices $Q_\epsilon := \mathrm{M}_{XX}^{-\frac{1}{2}}\mathrm{M}_{Xy}\hat{\Sigma}_\epsilon^{-1}\mathrm{M}_{yX}\mathrm{M}_{XX}^{-\frac{1}{2}}$ and $Q_y := \mathrm{M}_{XX}^{-\frac{1}{2}}\mathrm{M}_{Xy}\mathrm{M}_{yy}^{-1}\mathrm{M}_{yX}\mathrm{M}_{XX}^{-\frac{1}{2}}$. Note that $\hat{\Sigma}_\epsilon = \mathrm{M}_{yy} - \mathrm{M}_{yX}\mathrm{M}_{XX}^{-1}\mathrm{M}_{Xy}$ and hence

$$\begin{aligned} Q_\epsilon Q_y &= \mathrm{M}_{XX}^{-\frac{1}{2}}\mathrm{M}_{Xy}(\mathrm{M}_{yy} - \mathrm{M}_{yX}\mathrm{M}_{XX}^{-1}\mathrm{M}_{Xy})^{-1}\mathrm{M}_{yX}\mathrm{M}_{XX}^{-1}\mathrm{M}_{Xy}\mathrm{M}_{yy}^{-1}\mathrm{M}_{yX}\mathrm{M}_{XX}^{-\frac{1}{2}} = \\ &= \mathrm{M}_{XX}^{-\frac{1}{2}}\mathrm{M}_{Xy}((\mathrm{M}_{yy} - \mathrm{M}_{yX}\mathrm{M}_{XX}^{-1}\mathrm{M}_{Xy})^{-1}\mathrm{M}_{yy} - I_n)\mathrm{M}_{yy}^{-1}\mathrm{M}_{yX}\mathrm{M}_{XX}^{-\frac{1}{2}} = \\ &= Q_\epsilon - Q_y. \end{aligned} \tag{3.9}$$

The second equality follows from the fact that for any two quadratic matrices A, B of the same size with $(A - B)$ non-singular, it holds that $(A-B)^{-1}B = -(A-B)^{-1}(A-B) + (A-B)^{-1}A = (A-B)^{-1}A - I$. Since two symmetric matrices can be simultaneously diagonalized by an orthogonal matrix if and only if the product of the two matrices is symmetric (for a proof see e.g. (Harville, 1997, corollary 21.13.2)) and since from the above eq. (3.9), it is easy to see that $Q_\epsilon Q_y = (Q_\epsilon Q_y)'$ holds, it follows that Q_y and Q_ϵ have the same eigenvectors, $\tilde{V} = V$, and their eigenvalues fulfill the following relation:

$$\begin{aligned} V'Q_\epsilon VV'Q_yV &= V'Q_\epsilon V - V'Q_yV \\ \Lambda^2\tilde{\Lambda}^2 &= \Lambda^2 - \tilde{\Lambda}^2, \end{aligned}$$

and hence, $\Lambda^2 = \tilde{\Lambda}^2(I_{\bar{k}} - \tilde{\Lambda}^2)^{-1}$ and $\tilde{\Lambda}^2 = \Lambda^2(I_{\bar{k}} + \Lambda^2)^{-1}$. Thus, $\tilde{V} \equiv V$ implies that $\hat{F}_2 \equiv \tilde{F}_2$ and $\hat{G}_2 \equiv \tilde{G}_2$. Finally, from lemma 3.1 it follows that

$$\hat{F}_1 = \tilde{F}_1(I_r - \tilde{\Lambda}_r^2)^{\frac{1}{2}} \text{ and } \hat{G}_1 = \tilde{G}_1(I_r - \tilde{\Lambda}_r^2)^{-\frac{1}{2}},$$

where $\tilde{\Lambda}_r^2$ is the diagonal matrix of the r largest eigenvalues of Q_y. Thus, the difference in the pairs of estimators is again just a matter of scaling of the columns of the estimators for F and G.

Let us mention just two further remarks before we state the asymptotic properties of the above ML estimators.

Remark 3.1 (*Link between RR and canonical correlation analysis.)* In fact, by introduction of Q_y one can already see the link of RR to *canonical correlation analysis* (CCA) since the square roots of the eigenvalues of $Q_y := \mathrm{M}_{XX}^{-\frac{1}{2}}\mathrm{M}_{Xy}\mathrm{M}_{yy}^{-1}\mathrm{M}_{yX}\mathrm{M}_{XX}^{-\frac{1}{2}}$, i.e. $\tilde{\Lambda}$, are the sample canonical correlations of y_t and X_t, see for instance (Brillinger, 2001) for an introduction to the canonical analysis of time series that has been introduced by (Hotelling, 1935; Hotelling, 1936). In terms of $\hat{F}_1$ and $\hat{G}_1$ the canonical variates at time t can be written as

$$\zeta_t = \Lambda_r^{-1}\hat{G}_1'X_t \text{ and } \omega_t = (I_r - \tilde{\Lambda}_r^2)^{\frac{1}{2}}\hat{F}_1^- y_t,$$

where $\hat{F}_1^- = U_r'\hat{\Sigma}_\epsilon^{-\frac{1}{2}}$ is the left inverse of $\hat{F}_1$.

Remark 3.2 (*Link to the coefficient matrix in the unrestricted full linear model.)* Note that both pairs of ML estimates contain the LS estimator of the full unrestriced linear model $\mathrm{M}_{yX}\mathrm{M}_{XX}^{-1}$. The rank reduction is achieved by pre- and postmultiplication respectively of special matrices of rank r that are constructed such that the above criterion function is optimized.

3.3.2 Asymptotic properties of the ML estimators

Consider first the case where Σ_ϵ is known, and $\Gamma = \Sigma_\epsilon$:
Let the corresponding ML estimators for F and G be denoted as $\bar{F}_1 = \Sigma_\epsilon^{\frac{1}{2}} \bar{U}_r$ and $\bar{G}_1 = \mathrm{M}_{XX}^{-1}\mathrm{M}_{Xy}\Sigma_\epsilon^{-\frac{1}{2}}\bar{U}_r$, where $\bar{U}_r$ consists of the r largest normalized eigenvectors of $\Sigma_\epsilon^{-\frac{1}{2}}\mathrm{M}_{yX}\mathrm{M}_{XX}^{-1}\mathrm{M}_{Xy}\Sigma_\epsilon^{-\frac{1}{2}}$. It can be shown that $(\bar{F}_1, \bar{G}_1)$ converge in probability to (F_1^*, G_1^*), their population counterparts, namely,

$$F_1^* = \Sigma_\epsilon^{\frac{1}{2}} U_r^* \quad \text{and} \quad G_1^* = \Gamma_X^{-1}\Gamma_{Xy}\Sigma_\epsilon^{-\frac{1}{2}} U_r^*,$$

where $\Gamma_X = \mathbb{E}X_t X_t'$, $\Gamma_{yX} = \Gamma_{Xy}' = \mathbb{E}y_t X_t'$ and U_r^* contains the eigenvectors corresponding to the r largest eigenvalues of $\Sigma_\epsilon^{-\frac{1}{2}}\Gamma_{yX}\Gamma_X^{-1}\Gamma_{Xy}\Sigma_\epsilon^{-\frac{1}{2}}$.

Theorem 3.2 (Consistency and Asymptotic Normality of $\bar{F}_1, \bar{G}_1$ and $\bar{B} = \bar{F}_1\bar{G}_1'$) *Let (y_t) be a (stable) process as presented in eq. (3.7) and let the assumptions 2.1 hold. Let in addition $\mu_x = 0$, $\epsilon_t \sim iid\mathcal{N}(0, \Sigma_\epsilon)$ with $\Sigma_\epsilon > 0$ and known, and let the r largest (and non-zero) eigenvalues of $\Sigma_\epsilon^{-\frac{1}{2}}\Gamma_{yX}\Gamma_X^{-1}\Gamma_{Xy}\Sigma_\epsilon^{-\frac{1}{2}}$ be ordered and distinct, i.e. $\lambda_1^{*2} > \lambda_2^{*2} > \ldots > \lambda_r^{*2} > 0$ and $\lambda_{r+1}^{*2} = \ldots = \lambda_n^{*2} = 0$. Then it follows that*

(i) $\operatorname*{plim}_{T\to\infty} \bar{F}_1 = F_1^*$ *and* $\operatorname*{plim}_{T\to\infty} \bar{G}_1 = G_1^*$,

(ii) $$\sqrt{T}\begin{pmatrix} \mathrm{vec}(\bar{F}_1 - F_1^*) \\ \mathrm{vec}(\bar{G}_1 - G_1^*) \end{pmatrix} \xrightarrow{d} \mathcal{N}\left[0, \begin{pmatrix} \Sigma_{F_1^*} & \Sigma_{F_1^*G_1^*} \\ \Sigma_{F_1^*G_1^*}' & \Sigma_{G_1^*} \end{pmatrix}\right],$$
where $\Sigma_{F_1^*}$ is an $nr \times nr$ dimensional matrix, whose (i,j)th $n \times n$ dimensional sub-block matrix is given by

$$\Sigma_{F_1^*,ij} = \begin{cases} \sum_{\substack{l=1\\ l\neq i}}^{r} \frac{\lambda_i^{*2}+\lambda_l^{*2}}{(\lambda_i^{*2}-\lambda_l^{*2})^2}\phi_l\phi_l' & \textit{for } i = j \\ -\frac{\lambda_i^{*2}+\lambda_j^{*2}}{(\lambda_i^{*2}-\lambda_j^{*2})^2}\phi_j\phi_i' & \textit{for } i \neq j, \end{cases}$$

where ϕ_l is the lth column of F_1^*, $\Sigma_{F_1^*G_1^*}$ is an $nr \times \bar{k}r$ dimensional matrix, whose (i,j)th $n \times \bar{k}$ dimensional sub-block matrix is given by

$$\Sigma_{F_1^*G_1^*,ij} = \begin{cases} \sum_{\substack{l=1\\ l\neq i}}^{r} \frac{2\lambda_i^{*2}}{(\lambda_i^{*2}-\lambda_l^{*2})^2}\phi_l\gamma_l' & \textit{for } i = j \\ -2\frac{\lambda_j^{*2}}{(\lambda_i^{*2}-\lambda_j^{*2})^2}\phi_j\gamma_i' & \textit{for } i \neq j, \end{cases}$$

where γ_l is the lth column of G_1^*, and finally, $\Sigma_{G_1^*}$ is an $\bar{k}r \times \bar{k}r$ dimensional matrix, whose (i,j)th $\bar{k} \times \bar{k}$ dimensional sub-block matrix is given by

$$\Sigma_{G_1^*,ij} = \begin{cases} \sum_{\substack{l=1\\ l\neq i}}^{r} \frac{3\lambda_i^{*2}-\lambda_l^{*2}}{(\lambda_i^{*2}-\lambda_l^{*2})^2}\gamma_l\gamma_l' + \Gamma_X^{-1} & \textit{for } i = j \\ -\frac{\lambda_i^{*2}+\lambda_j^{*2}}{(\lambda_i^{*2}-\lambda_j^{*2})^2}\gamma_j\gamma_i' & \textit{for } i \neq j, \end{cases}$$

and

(iii) $\operatorname*{plim}_{T\to\infty} \bar{B} = B^*$.

(iv) $\sqrt{T}\mathrm{vec}(\bar{B} - B^*) \xrightarrow{d} \mathcal{N}(0, \Sigma_{B^*})$,
where

$$\begin{aligned}\Sigma_{B^*} = \;& (G_1^* \otimes I_n)\Sigma_{F_1^*}(G_1^{*'} \otimes I_n) + (I_{\bar{k}} \otimes F_1^*)\Sigma_{G_1^*F_1^*}(G_1^{*'} \otimes I_n) + (G_1^* \otimes I_n)\Sigma_{F_1^*G_1^*}K_{\bar{k},r}'(I_{\bar{k}} \otimes F_1^{*'})+ \\ & (I_{\bar{k}} \otimes F_1^*)K_{\bar{k},r}\Sigma_{G_1^*}K_{\bar{k},r}'(I_{\bar{k}} \otimes F_1^{*'})\end{aligned}$$

and $K_{\bar{k},r}$ denotes a commutation matrix, see appendix C.2 for a definition.

Proof. For the proof of (i) and (ii) see e.g. (Reinsel and Velu, 1998, theorem 2.4). For the proof of (iii) and (iv) we refer to (Lütkepohl, 1993, corollary 5.10.1). □

Consider now the case, where Σ_ϵ is unknown:
On pages 46 ff. (Reinsel and Velu, 1998) investigate the asymptotic distribution of $\hat{F}_1, \hat{G}_1$ and $\hat{B}$, where Σ_ϵ is unknown and thus, $\Gamma = \hat{\Sigma}_\epsilon$, for the case where $\mathrm{rk}(B) = 1$. They show that the asymptotic distribution of the product $\hat{B} = \hat{F}_1\hat{G}_1'$ is unaffected by choosing $\Gamma = \hat{\Sigma}_\epsilon$ instead of $\Gamma = \Sigma_\epsilon$, whereas the asymptotic variance of the vectors $\hat{F}_1, \hat{G}_1$ changes. For further details refer to (Reinsel and Velu, 1998).

3.3.3 Weighted least squares estimators for F and G:

The objective in the weighted least squares framework is to minimize

$$\min_{F,G} \frac{1}{T} \operatorname{tr}\left[(Y - FG'X)'\Gamma^{-1}(Y - FG'X)\right],$$

where Γ is as above some positive definite symmetric matrix that here determines the weighting structure. Note that

$$\begin{aligned}
&\frac{1}{T}\operatorname{tr}\left[(Y - FG'X)'\Gamma^{-1}(Y - FG'X)\right] = \\
&= \operatorname{tr}\left[\Gamma^{-\frac{1}{2}}(\mathrm{M}_{yy} - \mathrm{M}_{yX}GF' - FG'\mathrm{M}_{Xy} + FG'\mathrm{M}_{XX}GF')\Gamma^{-\frac{1}{2}}\right] = \\
&= \operatorname{tr}\left[\Gamma^{-\frac{1}{2}}(\hat{\Sigma}_\epsilon + (\mathrm{M}_{yX}\mathrm{M}_{XX}^{-\frac{1}{2}} - FG'\mathrm{M}_{XX}^{\frac{1}{2}})(\mathrm{M}_{yX}\mathrm{M}_{XX}^{-\frac{1}{2}} - FG'\mathrm{M}_{XX}^{\frac{1}{2}})')\Gamma^{-\frac{1}{2}}\right] \\
&= \operatorname{tr}\left[\Gamma^{-\frac{1}{2}}\hat{\Sigma}_\epsilon\Gamma^{-\frac{1}{2}}\right] + \left\|\Gamma^{-\frac{1}{2}}\mathrm{M}_{yX}\mathrm{M}_{XX}^{-\frac{1}{2}} - \Gamma^{-\frac{1}{2}}FG'\mathrm{M}_{XX}^{\frac{1}{2}}\right\|_F .
\end{aligned}$$

Hence, lemma 3.2 implies that the optimal matrices (F_{wls}, G_{wls}) in the weighted least squares sense with respect to Γ and the normalization conditions (3.3) and (3.4) are $F_{wls} = \Gamma^{\frac{1}{2}}U_r$ and $G_{wls} = \mathrm{M}_{XX}^{-1}\mathrm{M}_{Xy}\Gamma^{-\frac{1}{2}}U_r$, where U_r contains the normalized eigenvectors corresponding to the r largest eigenvalues of $\Gamma^{-\frac{1}{2}}\mathrm{M}_{yX}\mathrm{M}_{XX}^{-1}\mathrm{M}_{Xy}\Gamma^{-\frac{1}{2}}$. Therefore, for $\Gamma = \hat{\Sigma}_\epsilon$ the weighted least squares estimators are equal to the ML estimators stated in theorem 3.1. Due to the above discussions analog statements can be made, if $\Gamma = \mathrm{M}_{yy}$, see section 3.3.1.

3.3.4 Some alternative estimators for $B = FG'$

Consider in the following a specially weighted form of the OLS estimator obtained from the full unrestricted linear model

$$L\mathrm{M}_{yX}\mathrm{M}_{XX}^{-1}R, \tag{3.10}$$

where L and R are some non-singular $n \times n$ and $\bar{k} \times \bar{k}$ dimensional matrices. From lemma 3.2 it follows that the best rank r approximation of $L\mathrm{M}_{yX}\mathrm{M}_{XX}^{-1}R$ in the $\|.\|_F$-sense, is given by $P = U_r\Lambda_r V_r'$, where U_r, Λ_r and V_r are the respective submatrices of U, Λ and V from the singular value decomposition of $L\mathrm{M}_{yX}\mathrm{M}_{XX}^{-1}R$ corresponding to the largest r singular values. Since

$$\min_{P \in \mathbb{R}^{n\times\bar{k}};\mathrm{rk}(P)=r} \|L\mathrm{M}_{yX}\mathrm{M}_{XX}^{-1}R - P\|_F$$

is equivalent to

$$\min_{P \in \mathbb{R}^{n\times\bar{k}};\mathrm{rk}(P)=r} \|\mathrm{M}_{yX}\mathrm{M}_{XX}^{-1} - L^{-1}PR^{-1}\|_F$$

estimators for the rank deficient matrix B that minimize this objective function may be given as $\hat{B}_{LR} = L^{-1}PR^{-1}$.

Let us now discuss briefly some different choices for the weighting matrices L and R:

1. $L = \hat{\Sigma}_\epsilon^{-\frac{1}{2}}$ or $L = \mathrm{M}_{yy}^{-\frac{1}{2}}$, $R = \mathrm{M}_{XX}^{\frac{1}{2}}$: From section 3.3.3 it is clear that with these choices for L and R we obtain the above ML and weighted least squares estimators for the rank deficient B. In (Deistler and Hamann, 2005) this method to obtain ML estimators is called *indirect* estimation method.

2. $L = I_n$, $R = I_n$: Thus, the OLS estimate of the full unrestricted method is directly approximated by a rank r matrix. In (Deistler and Hamann, 2005) this method is therefore referred to as *direct* estimation method.

3. $L = I_n$, $R = \mathrm{M}_{XX}$: This means that it is just the cross moments matrix M_{yX} that is approximated by some rank r matrix. Note however that the optimal rank deficient matrix B corresponding to this objective function minimizes in particular the following optimization problem
$$\min_{B \in \mathbb{R}^{n \times k}; \mathrm{rk}(B)=r} \|\mathrm{M}_{yX} - B\mathrm{M}_{XX}\|_F^2,$$
i.e. it minimizes the overall sum of squared residuals of the normal equations.

It has been shown before that the ML estimator of the rank deficient B is consistent and asymptotically normal. In addition, the ML estimator is asymptotically efficient. In order to investigate especially small sample properties of the ML and the alternative estimators consider the following simulation study:
Let the process be defined as

$$\begin{aligned} y_t &= Ay_{t-1} + Dx_{t-1} + \epsilon_t, \\ y_t &= B\left[y_{t-1}', \; x_{t-1}\right]' + \epsilon_t, \quad t \in \mathbb{Z}, \end{aligned}$$

with $n = 2$, $r = 1$, $k = 1$, $x_t \sim iid\mathcal{N}(0,1)$ independent of (ϵ_t), $\epsilon_t \sim iid\mathcal{N}(0, \Sigma_\epsilon)$, with $\Sigma_\epsilon = \begin{pmatrix} 6.3e-05 & -3.9e-05 \\ -3.9e-05 & 2.7e-05 \end{pmatrix}$. Due to $r = 1$, B can be written as $B = [A, D] = \left[M \begin{pmatrix} \lambda & 0 \\ 0 & 0 \end{pmatrix} M^{-1}, M \begin{pmatrix} b \\ 0 \end{pmatrix}\right]$, where $M \in \mathbb{R}^{2\times 2}$ non-singular, $\lambda \in \mathbb{R}\setminus\{0\}$ and b is free in $\mathbb{R}$. Throughout the simulations we chose $M = \begin{pmatrix} 1 & -1 \\ -2 & 1 \end{pmatrix}$, $b = 2$ and also the realizations of (x_t) were just once generated. It can easily be seen that the selection of λ, the non zero eigenvalue of the AR coefficient matrix A, determines the stability of the process (y_t). For $|\lambda| < 1$, the process will be stable and unstable otherwise.
Simulations were performed for $\lambda = 0.1,\ 0.7,\ 0.99,\ 1$ and sample size $T = 10,\ 30,\ 100,\ 1000$. In order to reduce the influence of starting values the first 50 generated data points were always dropped. The goodness of the estimates is examined by the following three different measures:

1. $FN_B = \left\|\hat{B} - B\right\|_F$, the Frobenius norm of the distance of the estimate to the true parameter matrix B,

2. $FN_y = \left\|\hat{B}X - Y\right\|_F$, where X and Y are as in 3.1 and 3.7, the Frobenius norm of the residuals, i.e. the overall sum of squared residuals, and

3. $l = \log\left\{\det\left[\frac{1}{T}\sum_{t=1}^T (y_t - \hat{B}X_t)(y_t - \hat{B}X_t)'\right]\right\} + n$, the log likelihood value (up to a constant).

For each combination of λ and T 500 data series for ϵ_t and y_t have been generated.
In short, the following conclusions may be drawn from this simulation study: The indirect estimation method performs best throughout, meaning irrespective of sample size and stability of the underlying process, while the direct estimation method lags behind its competitors. The estimator that approximates the cross moments matrix M_{yX} does a good job and due to its construction beats the ML estimator in the FN_y sense. Nevertheless, one should rather perform ML estimation. Table 3.1 shows some results of the simulation study. Due to the fact that the results indicated no extreme differences concerning the choice of λ, table 3.1 contains only summary statistics for $\lambda = 0.7$.

Measure	T	unrestr. OLS	direct	indirect	M_{yX}
FN_B	10	0.9759	0.8507	0.3285	0.8141
	30	0.9501	0.8336	0.2896	0.7794
	100	0.9516	0.8106	0.2731	0.7832
	1000	0.8743	0.7851	0.2768	0.7352
FN_y	10	0.0534	0.8768	0.0614	0.0561
	30	0.1886	1.2137	0.1945	0.1916
	100	0.6620	2.2986	0.6686	0.6666
	1000	6.7040	10.6839	6.7101	6.7054
l	10	−15.2480	−8.6116	−15.0547	−14.2906
	30	−14.6700	−9.6065	−14.6175	−14.2340
	100	−14.5064	−10.3360	−14.4888	−14.3422
	1000	−14.4543	−11.6788	−14.4531	−14.4369

Table 3.1: Median of the 500 FN_B, FN_y, and log likelihood values computed for each combination of $\lambda = 0.7$ and sample size T, and each of the above presented estimation methods plus the OLS estimator in the unrestricted framework.

3.3.5 Estimation under linear (or affine) restrictions on F and G:

Suppose we are given some information about the structure of F and G. For instance, we might have some knowledge about zero restrictions on F and G respectively, or any other linear (or affine) restriction. In some cases we might be able to find a representative in the set of equivalent multiplicative decompositions of the (aside from the rank constraint) unrestricted estimate of B that fulfills the additional constraints imposed by the new structure. However, by optimizing the "unrestricted" likelihood function an unnecessarily high number of parameters is estimated, which might deteriorate the estimation of the possibly small set of free parameters. Or in other words, optimization of a constrained setting might be advantageous concerning efficiency. Let us therefore mention a procedure that takes into account the additional restrictions during estimation:

Suppose that Σ_ϵ is known and apply the vec operator on equation

$$Y = FG'X + E.$$

According to the rules of kronecker products and the vec operator, this gives

$$\text{vec}(Y) = (X' \otimes F)\text{vec}(G') + \text{vec}(E)$$

or

$$\text{vec}(Y) = (X'G \otimes I_n)\text{vec}(F) + \text{vec}(E).$$

Now suppose that we are given the linear constraints

$$\text{vec}(G') = R_G\gamma_G + r_G \quad \text{and} \quad \text{vec}(F) = R_F\gamma_F + r_F.$$

Analogously to section 2.2 eq. (2.12) we obtain for given F

$$\hat{\gamma}_G(F) = (R_G'(XX' \otimes F'\Sigma_\epsilon^{-1}F)R_G)^{-1}R_G'(X \otimes F'\Sigma_\epsilon^{-1})(y - (X' \otimes F)r_G), \tag{3.11}$$

and reversely for given G we obtain

$$\hat{\gamma}_F(G) = (R_F'(G'XX'G \otimes \Sigma_\epsilon^{-1})R_F)^{-1}R_F'(G'X \otimes \Sigma_\epsilon^{-1})(y - (X'G \otimes I_n)r_F). \tag{3.12}$$

Thus, we might pursue the following iterative procedure: Starting with some arbitrary pair $(\hat{F}^{(0)}, \hat{G}^{(0)})$, for instance the "unrestriceted" ML estimators from theorem 3.1 (i), compute (for $i > 0$) iteratively

$$\text{vec}(\hat{G}^{(i)\prime}) = R_G\hat{\gamma}_G(\hat{F}^{(i-1)}) + r_G$$

in terms of $\hat{F}^{(i-1)}$ and

$$\mathrm{vec}(\hat{F}^{(i)}) = R_F \hat{\gamma}_F(\hat{G}^{(i)}) + r_F$$

in terms of $\hat{G}^{(i)}$ and at each step of the iteration impose suitably defined normalization conditions. Continue to iterate until convergence is reached. For further details concerning the properties of this procedure see e.g. (Reinsel and Velu, 1998, p.33).

3.4 Model specification and input selection

In case of reduced rank regression models one has to specify not only the set of explanatory variables and the dynamics but also the integer valued parameter r, namely the rank of the coefficient matrix B, and finally the real valued parameters of the rank deficient matrix B.
In literature, see e.g. (Anderson, 1951), (Reinsel and Velu, 1998, section 2.6), or (Lütkepohl, 1993, section 5.3.5), several testing procedures are proposed. Due to the relation of RR to CCA one might determine the rank of B by testing the significance of the last $(\min(n,\bar{k}) - r)$ canonical correlations between y_t and X_t. A test statistic for testing $H_0 : \mathrm{rk}(B) \leq r$ against $H_1 : \mathrm{rk}(B) > r$ is given by

$$\mathcal{M} = -\left[T - (n + \bar{k} + 1)/2\right] \sum_{j=r+1}^{\min(n,\bar{k})} \log(1 - \hat{\lambda}_j^2),$$

for $r = 1, \ldots, \min(n,\bar{k}) - 1$, where $\hat{\lambda}_j$ is the jth largest canonical correlation between y_t and X_t, i.e. the square root of the jth largest eigenvalue of Q_y. Under the null hypothesis, the test statistic converges in distribution to a χ^2 distribution with $(n-r)(\bar{k}-r)$ degrees of freedom. Thus, if $\mathcal{M}$ is greater than an upper critical value determined by the $\chi^2_{(n-r)(\bar{k}-r)}$ distribution, the null hypothesis is rejected.
Alternatively one may follow the line as proposed in (Deistler and Hamann, 2005), who simultaneously specify the rank r, the set of explanatory variables and the dynamics by a procedure involving iterative computation and comparison of IC values, similarly to the FSP for single equations proposed by (An and Gu, 1985; An and Gu, 1989). The IC values are computed as follows

$$\begin{aligned} AIC(r,\bar{k}) &= \log\det\hat{\hat{\Sigma}}_\epsilon + \tfrac{2}{T}(nr + r\bar{k} - r^2) \\ BIC(r,\bar{k}) &= \log\det\hat{\hat{\Sigma}}_\epsilon + \tfrac{\log T}{T}(nr + r\bar{k} - r^2), \end{aligned} \tag{3.13}$$

where $\hat{\hat{\Sigma}}_\epsilon = \frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{F}\hat{G}'X_t)(y_t - \hat{F}\hat{G}'X_t)'$, is the ML estimator of Σ_ϵ.

FSP in the framework of RR models: The initial rank r of B is chosen to be $\min(n,\bar{k})$ and the initial set of explanatory variables is found by application of one of the forward procedures for systems of equations described in section 2.3. A refinement of the second step of An and Gu's FSP for this model class is as follows: It is now not only allowed to add or drop variables from the set, but also to let the rank of B vary from 1 up to $\min(n,\bar{k})$ and to weight observations at time t with some weighting factor λ^{T-t}, selected from a finite grid in $(0,1]$. The latter is done in order to take into account slowly time varying parameters. Thus, in each iteration step the IC optimal variable to be added to the set, dropped from the set, the IC optimal rank r and the IC optimal weighting factor λ are determined, giving four "optimal" criterion values. These four values are compared with the criterion value corresponding to the initial setting and the IC optimal setting is chosen. The procedure is iterated and it stops, when the IC value cannot be improved anymore by any of the possibilities mentioned before.

3.5 Forecasting

Let, as in section 2.4, $P_{\mathbb{H}}$ be a projector on the Hilbert space, $\mathbb{H}(y_t', y_{t-1}', \ldots, x_t', x_{t-1}', \ldots, 1)$. The optimal predictor for y_{t+h} is then given by

$$P_{\mathbb{H}} y_{t+h} = P_{\mathbb{H}} FG' X_{t+h} + \underbrace{P_{\mathbb{H}} \epsilon_{t+h}}_{=0}.$$

The error term is mapped to zero by the projector, if assumptions 2.1 hold. Thus, for $h = 1$ we have $y_{t+1|t} = FG'X_{t+1}$, and hence, the estimate for $y_{t+1|t}$ is given by $\hat{y}_{t+1|t} = \hat{F}\hat{G}'X_{t+1}$.

Chapter 4

Factor Model with Idiosyncratic Noise (IN)

Factor analysis is an important instrument of multivariate analysis. It is motivated by the assumption that the underlying observed variables are correlated in such a way that their correlation structure can be simply reconstructed by a small set of variates, so-called factors.

Factor analysis was first introduced in the field of psychology by Burt, Spearman, Thomson and Thurstone, just to name some. In the last decades however, its use spreaded also to other disciplines. Growing data availability and computing power of ordinary personal computers made the concept of Factor models attractive to practitioners.

4.1 The model

Let for simplicity throughout this chapter all variables of interest be mean adjusted. As mentioned in the introduction the basic idea of factor models is to find a representation of the n–dimensional vector of observed variables y_t as a linear combination of a small number, say $r \ll n$, of in general unobserved factors ζ_t and an n–dimensional vector of noise, ϵ_t. Hence, y_t can be written as

$$y_t = \Lambda\zeta_t + \epsilon_t, \quad t \in \mathbb{Z}. \tag{4.1}$$

In addition, we assume that the following assumptions hold:

Assumption 4.1 (Factor Model)

(i) (ζ_t) and (ϵ_t) are linearly regular, jointly stationary and ergodic processes with mean zero.

(ii) $\Sigma_\zeta = I_r$.

(iii) $\Sigma_\epsilon > 0$ and diagonal.

(iv) $\Lambda \in \mathbb{R}^{n \times r}$ is the factor loading matrix, with $\text{rk}(\Lambda) = r$.

(v) $\mathbb{E}\zeta_t\epsilon_s' = 0$ for all $s, t \in \mathbb{Z}$.

The variance covariance matrix of y_t is therefore given by the sum

$$\Sigma_y = \Lambda\Lambda' + \Sigma_\epsilon. \tag{4.2}$$

Thus, the noise component ϵ_t has no impact on the correlation structure of y_t that is explained by a (possibly) small number of factors only, but may contain additional component individual information concerning the variances of $y_{it}, i = 1, \dots, n$. ϵ_t is therefore called *idiosyncratic*. In case of economic time series, e.g. time series of asset shares, this may have a natural interpretation, since the factors may describe the movements according to the market and the "noise" the movements according to the individual companies.

Note that the dynamics in the factor model as it is given in eq. (4.1) together with the assumptions 4.1 come from the dynamics of (ζ_t) and (ϵ_t) only, since Λ is assumed to be constant in time. The model is hence called *quasi-static* factor model. The dynamics in (ζ_t) and (ϵ_t) respectively may be modelled by some r-dimensional VARX model as presented in chapter 2 and n single equation ARX models.
For a discussion on dynamic factor models see for instance (Forni and Lippi, 1999; Forni, Hallin, Lippi and Reichlin, 2000; Forni, Hallin, Lippi and Reichlin, 2001; Forni, Hallin, Lippi and Reichlin, 2003) or (Scherrer and Deistler, 1998) and (Diebold, 2000).

Note furthermore that the above factor model involves a high order of indeterminacy in the parameters due to the fact that the factor variates are unobserved and have to be estimated themselves. Given Σ_y and the number of factors r, in estimating the factor loading matrix Λ two identifiability problems arise. The first problem is to obtain feasible $\Lambda\Lambda'$ (and feasible Σ_ϵ, i.e. $\Sigma_\epsilon \geq 0$) from Σ_y and the second is to obtain Λ from $\Lambda\Lambda'$.
Comparison of the number of underlying equations and the number of functionally independent parameters gives a first indication, whether we might expect a unique decomposition of Σ_y or not: Eq. (4.2), due to the symmetry of Σ_y, states $\frac{1}{2}n(n+1)$ single equations. The number of free parameters in Σ_ϵ and Λ is n and $nr - \frac{1}{2}r(r-1)$, respectively. The latter is because of the indeterminacy in the product $\Lambda\zeta_t = \Lambda O' O \zeta_t = \tilde{\Lambda}\tilde{\zeta}_t$, where O is an arbitrary $r \times r$ dimensional orthonormal matrix, i.e. $OO' = O'O = I_r$, $\tilde{\zeta}_t$, as a rotation of ζ_t, has the same properties as ζ_t, and $\tilde{\Lambda}$ fulfills $\Lambda\Lambda' = \tilde{\Lambda}\tilde{\Lambda}'$ and thus, eq. (4.2). Hence, Λ can be made to satisfy $\frac{1}{2}r(r-1)$ additional normalizing conditions. The solution for r of the quadratic equation

$$\frac{1}{2}n(n+1) - n - nr + \frac{1}{2}r(r-1) = 0$$

that is smaller than n is called the *Ledermann bound* and is given by $r_{upper} = \frac{2n+1}{2} - \sqrt{(\frac{2n+1}{2})^2 - (n^2 - n)}$. If the number of factors is greater than r_{upper}, then we might expect non-uniqueness of the decomposition. If $r = \lfloor r_{upper} \rfloor$[1], then we might expect uniqueness. Note however, that it is not guaranteed that this unique solution is an admissible solution (i.e. $\Sigma_\epsilon > 0$ and $\Lambda\Lambda' \geq 0$ have to hold). If however, $r < \lfloor r_{upper} \rfloor$, it has been shown that generically the decomposition of Σ_y is unique, see (Scherrer and Deistler, 1998). For further details on identifiability of the decomposition of Σ_y see (Anderson and Rubin, 1956) and (Scherrer and Deistler, 1998).
Suppose a partition of Σ_y has been found, then as mentioned before (under the assumption that Λ has rank r) Λ is uniquely determined from $\Lambda\Lambda'$ up to postmultiplication by an arbitrary orthogonal matrix O. A set of $\frac{1}{2}r(r-1)$ normalizing conditions is for instance given by

$$\Lambda' \Sigma_\epsilon^{-1} \Lambda = \Delta, \tag{4.3}$$

where $\Delta = \text{diag}(\delta_1, \delta_2, \dots, \delta_r)$ is an $(r \times r)$-dimensional diagonal matrix, with $\delta_1 > \delta_2 > \dots > \delta_r > 0$. This normalization is referred to as the *default method* in the subsequent "quasi maximum likelihood"[2] algorithm.
In many cases, however, the Λ that fulfills the default normalizing conditions may lack interpretability. This can be improved by rotating the factors in such a way that the new factor loadings become more

[1] $\lfloor . \rfloor : \mathbb{R} \to \mathbb{Z}$, denotes the *floor function*, i.e. the operator that gives the largest integer number that is smaller than the operand.

[2] The algorithm is named "quasi" here, because the variates are not explicitly asked to be normally distributed.

meaningful. Let us in this context mention two procedures, namely the *varimax* and the *promax method.* The *varimax method*, see (Kaiser, 1958; Horst, 1965), tries to rotate the factors such that the resulting loadings tend to have in each row a few relatively large entries in absolute magnitude compared with the original ones, while the others become small or close to zero. The *promax method*, see (Hendrickson and White, 1964), starts with a *varimax* rotation and continues by focussing on the columns of Λ. The aim of this second optimization step is again to increase already large entries (in absolute magnitude) in each column and to shrink the others. This time, however, the criterion function is such that the optimal transformation matrix obtained, in general, is non-orthogonal, leading to oblique, i.e. correlated factors.

4.2 Estimation

In this section estimates for the real valued parameters and the unobserved factor variates are presented under the assumption that the number of factors r is given. Consider the following function $L_T(\Lambda, \Sigma_\epsilon|\hat{\Sigma}_y^T)$, where T denotes sample size and $\hat{\Sigma}_y^T := \frac{1}{T}\sum_{t=1}^T y_t y_t'$ is the sample variance covariance matrix of $y_t, t = 1, \ldots, T$. Estimates for Λ and Σ_ϵ are obtained by iteratively maximizing,

$$\begin{aligned} L_T(\Lambda, \Sigma_\epsilon|\hat{\Sigma}_y^T) &= -\frac{T}{2}\log\det(\Lambda\Lambda' + \Sigma_\epsilon) - \frac{1}{2}\sum_{t=1}^{T}(y_t'(\Lambda\Lambda' + \Sigma_\epsilon)^{-1}y_t) = \\ &= -\frac{T}{2}\log\det(\Lambda\Lambda' + \Sigma_\epsilon) - \frac{T}{2}\text{trace}((\Lambda\Lambda' + \Sigma_\epsilon)^{-1}\hat{\Sigma}_y^T), \end{aligned} \tag{4.4}$$

subject to $\Lambda \in \mathbb{R}^{n\times r}$, $\text{rank}(\Lambda) = r$, $\Sigma_\epsilon > 0$ and the default normalization condition (4.3), see e.g. (Lawley and Maxwell, 1971, section 4.3). Note, that in case of independently identically normally distributed noise and factors, the function given in (4.4) is (up to a constant) the loglikelihood function of y_t. In case of autoregressive factors and noise as considered here, (4.4) however, is not the likelihood function. Nevertheless, the estimates $\hat{\Lambda}$ and $\hat{\Sigma}_\epsilon$ obtained from maximizing (4.4) can be shown to be consistent estimates for Λ and Σ_ϵ, if Λ and Σ_ϵ are identifiable and if $\hat{\Sigma}_y^T$ is a consistent estimate of Σ_y. The proof for this is completely analogous to the proof given in (Anderson, 1971, p. 565), since the argument only depends on $\hat{\Sigma}_y^T$ converging to Σ_y a.e. For a detailed description of the iterative maximization procedure, the choice of starting values and the case of ML estimation under constraints, on one or both matrices Σ_ϵ and Λ respectively, see (Lawley and Maxwell, 1971; Anderson, 1971).

Let us now consider estimation of the unobserved factors ζ_t. In contrast to index, RR or principal components models, here, the factors, in general, cannot be obtained directly as a function of the observed y_t and, hence, have to be approximated by some (linear) function of y_t. In the following we will present two methods, namely the *regression method*, discussed in detail by Thomson (Thomson, 1951), and *Bartlett's method*, see (Bartlett, 1937; Bartlett, 1938b):

1. *Regression method:* Here the factor process is approximated in least squares sense by some linear combination of y_t, obtained from

$$\min_{A\in\mathbb{R}^{n\times r}} \mathbb{E}(\zeta_t - A'y_t)(\zeta_t - A'y_t)'. \tag{4.5}$$

   From the assumptions 4.1 we obtain $A' = \Lambda'\Sigma_y^{-1}$ and, therefore, $\hat{\zeta}_t = \Lambda'\Sigma_y^{-1}y_t$.
   If the covariance matrix of the factors is not the identity but some symmetric positive definite matrix Φ of order r, i.e. the factors are oblique and not necessarily standardized, the factor estimates are given by $\hat{\zeta}_t = \Phi\Lambda'\Sigma_y^{-1}y_t$, since in this case $\mathbb{E}\zeta_t y_t' = \Phi\Lambda'$.

2. *Bartlett's method:* Bartlett's idea was to minimize the sum of the squared standardized residuals with respect to the r-dimensional factor process,

$$\min_{\zeta_t} \mathbb{E}(y_t - \Lambda\zeta_t)'\Sigma_\epsilon^{-1}(y_t - \Lambda\zeta_t), \tag{4.6}$$

giving $\hat{\zeta}_t = (\Lambda'\Sigma_\epsilon^{-1}\Lambda)^{-1}\Lambda'\Sigma_\epsilon^{-1}y_t$.

It is easy to see that this method is independent of the factors being orthogonal or oblique.

There is no general rule which method to apply. The decision may be based on the properties the estimates of the factor process should possess. The following remarks will point out some of them:

Remark 4.1 (*Estimation error variance.*) By application of the *matrix inversion lemma* (MIL)[3] on $\Sigma_y = \Lambda\Phi\Lambda' + \Sigma_\epsilon$, where $\Phi := \mathbb{E}\zeta_t\zeta_t'$, we obtain

$$\Sigma_y^{-1} = \Sigma_\epsilon^{-1} - \Sigma_\epsilon^{-1}\Lambda(\Phi^{-1} + \Lambda'\Sigma_\epsilon^{-1}\Lambda)^{-1}\Lambda'\Sigma_\epsilon^{-1}.$$

Hence, the second moments and the estimation error variances can be written as

	Regression method	Bartlett's method
$\mathbb{E}\hat{\zeta}_t\hat{\zeta}_t'$	$\Phi\left[(\Lambda'\Sigma_\epsilon^{-1}\Lambda)^{-1} + \Phi\right]^{-1}\Phi$	$(\Lambda'\Sigma_\epsilon^{-1}\Lambda)^{-1} + \Phi$
$\mathbb{E}\hat{\zeta}_t\zeta_t'$	$\Phi\left[(\Lambda'\Sigma_\epsilon^{-1}\Lambda)^{-1} + \Phi\right]^{-1}\Phi$	Φ
$\mathbb{E}(\hat{\zeta}_t - \zeta_t)(\hat{\zeta}_t - \zeta_t)'$	$(\Phi^{-1} + \Lambda'\Sigma_\epsilon^{-1}\Lambda)^{-1}$	$(\Lambda'\Sigma_\epsilon^{-1}\Lambda)^{-1}$.

Thus unsurprisingly due to its construction, for any $\Phi > 0$ it holds that the estimation error variance of the factor estimates obtained by the *regression method* is smaller than that of the factor estimates obtained by *Bartlett's method.*

Remark 4.2 (*Conditional unbiasdness.*) Consider the conditional expectation $\mathbb{E}(\hat{\zeta}_t|\zeta_t)$. Then, since $\mathbb{E}(\Lambda'\Sigma_y^{-1}y_t|\zeta_t) = \Lambda'\Sigma_y^{-1}\Lambda\zeta_t$ and $\mathbb{E}((\Lambda'\Sigma_\epsilon^{-1}\Lambda)^{-1}\Lambda'\Sigma_\epsilon^{-1}y_t|\zeta_t) = \zeta_t$, it follows that *Bartlett's method* is unbiased in this sense, whereas the *regression method* is biased.

Remark 4.3 (*Relation between the two estimates.*) With the MIL it follows that

$$\Lambda'\Sigma_y^{-1} = (I_r + \Lambda'\Sigma_\epsilon^{-1}\Lambda\Phi)^{-1}\Lambda'\Sigma_\epsilon^{-1}.$$

Hence, the following relation between the two estimation methods holds

$$\hat{\zeta}_t^{(regression)} = \Phi(I_r + \Lambda'\Sigma_\epsilon^{-1}\Lambda\Phi)^{-1}\Lambda'\Sigma_\epsilon^{-1}\Lambda\hat{\zeta}_t^{(Bartlett)}.$$

Note that if Φ is diagonal and if the default normalizing conditions (4.3) hold, the factor estimates differ only by scaling.

Let $\hat{\hat{\zeta}}_t = \hat{\Lambda}'\hat{\Sigma}_y^{-1}y_t$ and $\hat{\hat{\zeta}}_t = (\hat{\Lambda}'\hat{\Sigma}_\epsilon^{-1}\hat{\Lambda})^{-1}\hat{\Lambda}'\hat{\Sigma}_\epsilon^{-1}y_t$ denote the final factor estimates, where Λ,Σ_ϵ and Σ_y are substituted by $\hat{\Lambda}$, $\hat{\Sigma}_\epsilon$ and $\hat{\Sigma}_y$ $(= \hat{\Sigma}_y^T)$.

4.3 Forecasting

In order to forecast y_t we have to define forecasting models for the factor and the idiosyncratic noise component:

For forecasting the factor process (ζ_t), here, we use a VARX model of the form

$$\zeta_{t+1} = A(z)\zeta_t + D(z)x_t + u_{t+1}, \tag{4.7}$$

where $A(z)$ and $D(z)$ are polynomial matrices in the backward shift operator z of order p and q, respectively, and the stability condition

$$\det\left[I - z\ A(z)\right] \neq 0 \quad \text{for all } |z| \leq 1 \tag{4.8}$$

holds.

[3] **Matrix inversion lemma (MIL)**: Let A, D and C be non-singular quadratic matrices. A and D are $n \times n$ and C is $m \times m$ dimensional. Let B be a rectangular $n \times m$ dimensional matrix. The inverse of $D = A - BC^{-1}B'$ is then given by $D^{-1} = A^{-1} + A^{-1}B(C - B'A^{-1}B)^{-1}B'A^{-1}$.

Assumption 4.2 (Factor Forecasting Model)

(i) (u_t) is white noise.

(ii) (x_t) is a k–dimensional linearly regular, stationary and ergodic process with nonsingular spectral density and mean zero and $\mathbb{E}x_t u_s' = 0$ for all $t,\ s \in \mathbb{Z}$.

Equation (4.7) can also be written as

$$\zeta_{t+1} = F\eta_t + u_{t+1}, \tag{4.9}$$

where $F = [A_0\ A_1\ \dots\ A_p\ D_0\ D_1\ \dots\ D_q]$ and $\eta_t = (\zeta_t'\ \zeta_{t-1}'\ \cdots\ \zeta_{t-p}'\ x_t'\ x_{t-1}'\ \cdots\ x_{t-q}')'$.

If we had observed $\zeta_1, \dots, \zeta_t$, the OLS estimate for $F \in \mathbb{R}^{r\times[(p+1)r+(q+1)k]}$ would be given by $F_t = \sum_{s=\bar{t}}^{t-1}(\zeta_{s+1}\eta_s')(\sum_{s=\bar{t}}^{t-1}\eta_s\eta_s')^{-1}$, where $\bar{t} = \min(p,q)+1$. By defining $\hat{\hat{\eta}}_t$ analogously to $\hat{\hat{\zeta}}_t$ we obtain an estimate $\hat{\hat{F}}_t$ for F_t. Hence, the one-step ahead forecasts for the factors ζ_{t+1} are given by $\hat{\hat{\zeta}}_{t+1|t} = \hat{\hat{F}}_t\hat{\hat{\eta}}_t$, which finally yield the one-step ahead forecasts, $\hat{y}_{t+1|t} = \hat{\Lambda}\hat{\hat{\zeta}}_{t+1|t}$, for y_{t+1}.

Since the "noise" is assumed to be idiosyncratic and, as it has already been stated above, may be interpreted as an asset specific component, we additionally consider univariate ARX models in order to predict the idiosyncratic component, which in obvious notation are given by,

$$\epsilon_{t+1}^{(i)} = a_i(z)\epsilon_t^{(i)} + D_i(z)z_t^{(i)} + \nu_{t+1}^{(i)}, \quad i = 1, \dots, n \tag{4.10}$$

with p_i and q_i being the order of the polynomials $a_i(z)$ and $D_i(z)$ and with the assumptions

Assumption 4.3 (Forecasting the idiosyncratic component)

(i) (ν_t) is white noise.

(ii) $(z_t^{(i)})$ is an m_i–dimensional linearly regular, stationary and ergodic process with nonsingular spectral density and mean zero and $\mathbb{E}z_t^{(i)}\nu_s^{(i)'} = 0$ for all $t,\ s \in \mathbb{Z}$, and $i = 1, \dots, n$.

(iii) $|1 - za_i(z)| \neq 0$ for all $|z| \leq 1$, $i = 1, \dots, n$.

Note that the "noise" component ϵ_t is of course unobservable, and thus the coefficients of eq. (4.10) have to be computed with respect to the estimated residuals $\hat{\hat{\epsilon}}_t = y_t - \hat{\Lambda}\hat{\hat{\zeta}}_t$.

The one-step ahead forecasts for $\epsilon_{t+1}^{(i)}$ are obtained analogously to above as

$$\hat{\hat{\epsilon}}_{t+1|t}^{(i)} = \hat{\hat{F}}_i\hat{\hat{\eta}}_t^{(i)},$$

where

$$\hat{\hat{\eta}}_t^{(i)} = (\hat{\hat{\epsilon}}_t^{(i)'}, \dots, \hat{\hat{\epsilon}}_{t-p_i}^{(i)'}, z_t^{(i)'}, \dots, z_{t-q_i}^{(i)'}),$$
$$\hat{\hat{F}}_i = \textstyle\sum_{s=\bar{t}_i}^{t-1}(\hat{\hat{\epsilon}}_{s+1}^{(i)}\hat{\hat{\eta}}_s^{(i)'})(\sum_{s=\bar{t}_i}^{t-1}(\hat{\hat{\eta}}_s^{(i)}\hat{\hat{\eta}}_s^{(i)'}))^{-1}, \text{ and}$$
$$\bar{t}_i = \min(p_i, q_i) + 1.$$

Hence, this approach consists of two steps: First, estimate the factor model (4.1), and second, estimate the VARX model (4.7) from the estimated factors $\hat{\hat{\zeta}}_t$ and the univariate ARX models (4.10) from the estimated residuals $\hat{\hat{\epsilon}}_t$.

Finally, one may compute two different types of one-step ahead forecasts for y_{t+1},

$$\begin{aligned} \hat{y}_{t+1|t}^{I} &= \hat{\Lambda}\hat{\hat{\zeta}}_{t+1|t}, \\ \hat{y}_{t+1|t}^{II} &= \hat{\Lambda}\hat{\hat{\zeta}}_{t+1|t} + \hat{\hat{\epsilon}}_{t+1|t}. \end{aligned}$$

As an option in forecasting, we additionally consider to project the inputs for the models forecasting the idiosyncratic components on the orthocomplement of inputs for the VARX model forecasting the factors. This is done in order to separate the individual, idiosyncratic components from the common components.

4.4 Model specification and input selection

In (Lawley and Maxwell, 1971, section 4.4) and (Anderson and Rubin, 1956, section 8), likelihood ratio criteria testing procedures for the determination of the number of factors are proposed. These sequences of tests base only on the goodness of decomposition of Σ_y, the factor scores are totally neglected. This may be reasonable in many applications. Here, however, the aim is to forecast. Thus, it might make sense to include the forecasting ability of the forecasting models for the factors in the decision rule.

As in the previous chapters specification is done totally data-driven. In order to determine the number of factors, dynamics and the sets of explanatory variables, we use again specification procedures involving information cirteria.

The Ledermann bound, as mentioned above, provides an upper bound for the number of factors, such that the decomposition of the variance-covariance matrices Σ_y is determined generically unique from the observed data y_t. The specification procedure may hence consist of the following steps:

1. For each feasible r, $r = 1, \ldots, \lfloor r_{upper} \rfloor$, compute $\hat{\Lambda}, \hat{\Sigma}_\epsilon$, $\hat{\hat{\zeta}}_t$ and $\hat{\epsilon}_t$.
2. Specify the VARX forecasting models for the respective factor processes and the ARX forecasting models for the corresponding idiosyncratic components, as described in section 2.3.
3. Choose r such that it minimizes the underlying IC criterion. Thus, choose r such that it gives an IC-optimal trade-off between in-sample explanatory power for y_t (based on the factor (and possibly also the "noise") forecast obtained from the previous step and on the estimated loadings $\hat{\Lambda}$) and model complexity (i.e. the number of parameters to be estimated).

Note, that in this procedure input selection and dynamic specification for the VARX and ARX models respectively, is based on a goodness-of-fit measure for the factor and "noise" process, whereas determination of the number of factors is based on a goodness-of-fit measure for y_t. Of course given the number r of factors, one could alternatively specify the forecasting models for the factors and the idiosyncratic noise component jointly with respect to a goodness-of-fit measure for y_t in a first step, and in a second step just as above select the optimal r. This however, is much more time consuming due to the high number of parameters that have to be taken under consideration during the selection and specification procedure of all forecasting models, and has therefore not been considered here.

Chapter 5

Empirical Analysis for Daily Share Prices

The methods proposed in the previous chapters are applied to daily close return data for shares of the banking sector in DJ EURO STOXX50[1] from 16.06.2000 to 13.11.2002. These are in total 629 observations, weekends excluded.

The banks are: ABN AMRO (H.AAB), Banco Bilbao (U.BBV), Banco Santander (E.SCH), Hypo-Vereinsbank (D.HVM), Deutsche Bank (D.DBK), BNP Paribas (F.BNP), UniCredit (I.UC); see figure 5.1.

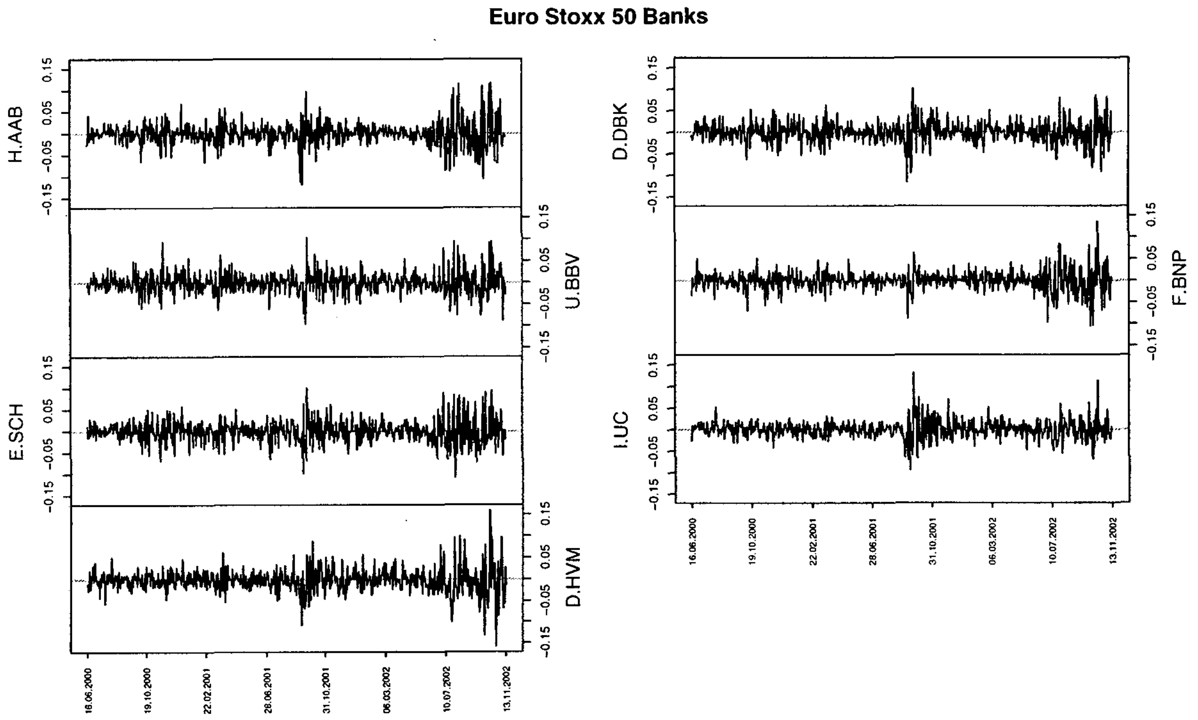

Figure 5.1: Return Series of the DJ EURO STOXX50 banks sector, 16.06.2000 – 13.11.2002.

Note that the definition of returns depends on the way of compounding. Let p_t be the price of an asset at time t, then

[1] The data were provided by Siemens-Fin4cast, Vienna.

1. *relative differences*, defined as $y_t = (p_t - p_{t-1})/p_{t-1}$, give the interest or percentage yield obtained within period $t-1$ to t using (here) daily compounding, and

2. *log returns*, defined as $r_t = \ln \frac{p_t}{p_{t-1}} = \ln(1 + y_t)$, give the interest or percentage yield obtained within period $t-1$ to t using continuous compounding.

Daily, monthly, yearly, and the like compoundings are often used by banks, financial institutions and investors in financial markets, whereas continuous compounding is used rather by researchers. Continuous compounding is also opportune in the framework of the widely used Black and Scholes option pricing model, since there one of the main assumption is that stock prices follow a geometric Brownian motion. Thus, under this assumption the differences of log prices are normally distributed. In practice however, this assumption hardly ever holds, see e.g. (Hull, 2003).
Note that for the average return $\bar{y}$ over the whole observation period, $t = 1, \ldots, T$, using daily compounding we have

$$\begin{aligned}
(1+\bar{y})^T - 1 &= \frac{p_T - p_0}{p_0}, \quad \text{and thus} \\
\bar{y} &= \left(\frac{p_T}{p_0}\right)^{1/T} - 1 = \\
&= \left(\prod_{t=1}^{T} (1 + y_t)\right)^{1/T} - 1 = \\
&= e^{\bar{r}} - 1,
\end{aligned}$$

where $\bar{r} = \frac{1}{T}\sum_{t=1}^{T} r_t$. Thus, the average return is computed as geometric mean of the y_t's, if daily compounding is used, whereas the arithmetic mean applies for log returns, hence continuous compounding. By definition (see also figure 5.2) it is clear that $y_t \geq r_t$ holds. The difference $y_t - r_t$ is small, if the y_t's are small in absolute values.

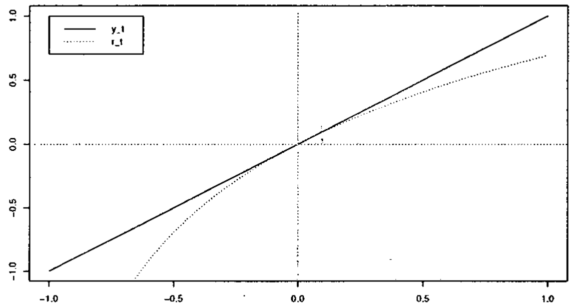

Figure 5.2: Relative differences (solid line) and the log returns expressed as a function of relative differences (dotted line).

Note that all asset price data available for us were already transformed to relative differences. Thus, the subsequent analyses are carried out on returns using daily compounding, and from now on the item "return" refers to the relative differences as defined above.

Table 5.1 contains summary statistics of the return series, and shows some stylized facts of financial return series:
First, return series tend to have a *leptokurtic* distribution. This means that compared with the normal distribution it is more peaked at the center and has fatter tails. Thus, "extremes" are more likely to occur in the framework of leptokurtic distributions than of normal distributions. For all bank return series the estimated kurtosis exceeds 3, the kurtosis of a normally distributed variable, and therefore confirms this

	H.AAB	U.BBV	E.SCH	D.HVM	D.DBK	F.BNP	I.UC
min	−0.1181	−0.0961	−0.1074	−0.1588	−0.1162	−0.1062	−0.0943
q_{25}	−0.0141	−0.0160	−0.0161	−0.0157	−0.0148	−0.0104	−0.0107
median	0.0000	0.0000	0.0000	−0.0016	0.0000	0.0000	0.0000
q_{75}	0.0120	0.0131	0.0139	0.0104	0.0134	0.0119	0.0106
max	0.1172	0.1034	0.1010	0.1595	0.1015	0.1356	0.1313
arith. mean	−0.0005	−0.0005	−0.0006	−0.0023	−0.0007	−0.0001	−0.0002
geom. mean	−0.0009	−0.0008	−0.0009	−0.0027	−0.0010	−0.0004	−0.0004
skewness	0.1745	0.2787	0.1895	0.1097	−0.1221	−0.0620	0.4197
kurtosis	6.3698	4.9782	4.4518	8.2088	4.8837	7.1862	7.7537

Table 5.1: Summary statistics of the return series of the DJ EURO STOXX50 banks sector, 16.06.2000 - 13.11.2002. (q_{25} (q_{75}) is the 25% (75%) sample quantile.)

stylized fact.
Second, volatilities tend to cluster, see figure 5.1. It can be observed that the summer of 2001 and the period from June 2002 until the end of the observed sample were periods of high volatility as compared to the rest of the time. In part II we come back to this characteristic and present model classes for the conditional variances.
It should be mentioned that the series were tested for unit roots, i.e. one type of non-stationarity. The hypothesis of the existence of a unit root in the *augmented Dickey Fuller* test, however, could be rejected for all 7 series at a significance level of 1%.

Let us now turn to the input variable candidates that are all given as relative differences and can be partitioned into three groups: The first group is given by present and lagged values of the bank returns. Throughout we consider only lags of order one and five. The second group, see table 5.2, consists of present values, first and fifth lag of 19 variables that contain general information concerning the development of the market under consideration, such as indices for the banking sector, interest rates and futures for indices. Finally, the third group pools present values, first and fifth lag of variables that are rather bank specific, like market indices of the markets or branches the banks are invested in, and stocks of companies of which the banks are important shareholders, see table 5.3.
The set of exogenous variables in the VARX models estimated is either given by whole group two or by current and lagged values of the three variables "BIX", "IRX" and "NDcl". These three variables have been selected throughout or most frequently by the two-step FSP and, if BIC was the underlying information criterion they sometimes appeared to be the only explanatory variables selected, in the framework of both, VARX and RR models. Vector X_t (see eq. (3.1)) in the RR models contains all variables provided by group one and two. Finally, for the factor models (IN) the variables in group one and two are used as "exogenous" variables in the factor forecasting models and the variables in group three are used as exogenous variables in the forecasting models for the idiosyncratic "noise" component.

Due to the fact that for each model class (i.e. the VARX method, the RR method and the IN method) considered we allow for certain design specifications, we will now introduce a coding system identifying the computed models. The code of a VARX model is as follows, 'varx[re-specification period].[initial method (mva/univ)].[criteria used in the two steps of the FSP].[logical for moving (TRUE) or expanding (FALSE) window]', e.g. varx5.univ.AICF-BIC.TRUE means that a VARX model has been respecified every 5 days using the *univ method* with AIC to determine the initial set, BIC in the second step of the refined FSP and a moving window of observations used for estimation. The code for the RR models is as follows, 'rr[re-specification period].[estimation of β (direct/indirect)]. [criteria used in the two steps of the FSP]. [logical for moving (TRUE) or expanding (FALSE) window]', e.g. rr5.direct.AICF-BIC.TRUE

Variable Name	Description
FTEUF	E300 Financials (FTSE International)
FTEUBKEC	E300 Banks EU (FTSE International)
SKX	S&P 500 BARRA Growth Special OP
GOX	Gold Index (Chicago Board Options Exchange)
TOP	EURO TOP-100 (FTSE International)
GBP1YD	GBP, 1 Year Yield (Bid)
FCUc1	Future, 3 Month Euro LIBOR traded on LIFFE, continues
BIX	S&P Bank Index (Standard & Poor's Corp)
XID	Industrial Index (American Stock Exchange)
LIEUR1YD	EURO LIBOR, 1 Year Yield
LIEUR1MD	EURO LIBOR, 1 Month Yield
CH2MTRR	Swiss, 2 Month zero-bond
DE10YTRR	BRD, 10 Years zero-bond
IRX	13-Week Treasury Bill (Chicago Board Options Exchange)
USD9MD	USD, 9 Month Yield
FTEUOFEC	E300 other Financials EU (FTSE International)
SXFE	Dow Jones Euro Stoxx Financial Services Index
NDc1	Future, NASDAQ 100, continues
FDXc1	Future, DAX INDEX, continues

Table 5.2: Input variable candidates - group two.

means that a RR model is respecified every 5 days, β is estimated by the *direct method*, AIC is used in the forward algorithm and BIC in the FSP, and a moving window is used for estimation. Finally, the code for the IN models is given by 'in[re-specification period].[factor estimates (regression/Bartlett)]. [rotation (varimax/promax/default)].[initial method (mva/univ)]. [criteria used in the two steps of the FSP]. [criterion used to specify the number of factors (AIC/BIC)]. [logical for moving (TRUE) or expanding (FALSE) window]. [orthogonal projection of the explanatory variables of the error models on the orthocomplement of the inputs of the factor models (TRUE/FALSE)]', e.g. in10.regression.default.univ.AICF-BIC.AIC.TRUE.TRUE means that a respecification period of 10 days, factor estimation by the *regression method*, normalization by the *default method*, *univ method* with AIC for the initial set, BIC for the second step of the FSP, AIC for determining the number of factors, moving window and orthogonal projection has been applied.

5.1 Measures for out-of-sample model validation

As is well known, in-sample effects can be extremely misleading in terms of notably overestimating the forecasting quality of the models. The forecast procedures presented here are "honest", in the sense of being strictly out-of-sample, i.e. for forecasting y_{t+1} only data up to time t are used, both for estimation of real valued parameters and for model specification.

Thus, throughout, for each dependent variable, $y_{i,t+1}$, the one-step (here one day) ahead predictors, $\hat{y}_{i,t+1|t}$, and the corresponding prediction errors, $\hat{e}_{i,t+1} = y_{i,t+1} - \hat{y}_{i,t+1|t}$, are calculated from a model identified from data up to time t, using both an extending and a moving window, respectively. The estimators of the real valued parameters are updated at every time instance. The specification is updated every five or every ten days, i.e. every week or every fortnight. The sample is divided into two parts, $1, \ldots, T_1$ and $T_1 + 1, \ldots, T_2$. Only the latter part is used for evaluating the out-of-sample forecasts. The evaluation sample, $T_1 + 1, \ldots, T_2$, consists of the last 30% of the whole sample.

Idiosyncratic Noise	Variable Name	Description
H.AAB	H.AEGN	Aegon, Insurance company
	H.ING	ING, Insurance company
	H.AMEV	Group Fortis AMEV, Insurance, banking, investment
	DJTNFEE	Northern European Financial Index
U.BBV	E.TEF	Telefonica
	DJLAFE	Latin America, Financial Index
	DJLABK	Latin America, Banks Index
E.SCH	VOD	Vodafone
	E.TEF	Telefonica
	DJLAFE	Latin America, Financial Index
	DJLABK	Latin America, Banks Index
D.HVM	D.MU2X	Munich Re Group
	ASTCECE	OTOB CENTRAL & EAST. EUROP.FIN.CECE
D.DBK	D.ALV	Allianz, Insurance company
	DAXBNKK	DAX 100 BANKS
	FTSE100	FTSE 100
	DAXIDXI	DAX 30
	DJI	DJI 30
F.BNP	F.MIDI	AXA, Insurance company
	F.SGE	Group Sociëté Général
	SPEURE1	S&P EURO ENERGY
	SPE35E1	S&P EUROPE 350 ENERGY
	CHEMSEX	EUROPE EX UK COMMODITY
	CHEMSFR	FRANCE-DS CHEMS. COMMODITY
	F.EADS	European Aeronautic Defence and Space Company
I.UC	I.G	Generali, Insurance Company
	ASTCECE	OTOB CENTRAL & EAST. EUROP.FIN.CECE

Table 5.3: Input variable candidates - group three.

Three measures for the quality of the forecasts are considered:

- The out-of-sample coefficient of determination $R_i^2 = 1 - \frac{\hat{\epsilon}_i'\hat{\epsilon}_i}{y_i'y_i}$, where $\hat{\epsilon}_i$ and y_i, respectively, are the vectors consisting of the components $\hat{\epsilon}_{i,t}$ and $y_{i,t}$ from the validation sample, $t = T_1 + 1, \ldots, T_2$.

- The out-of-sample hit rate given by, $h_i = \frac{1}{T_2 - T_1} \sum_{t=T_1+1}^{T_2} sign(y_{i,t}\hat{y}_{i,t|t-1})$.

- The Diebold Mariano (DM) test: Suppose a pair of h-step forecasts for an asset i have produced errors $(\hat{\epsilon}_{i,t}^{(1)}, \hat{\epsilon}_{i,t}^{(2)})$, $t = T_1 + 1, \ldots, T_2$. The quality of the forecast is to be judged on some specified function $g(\hat{\epsilon}_{i,t})$ of the forecast error, $\hat{\epsilon}_{i,t}$. Then, the Null hypothesis of equality of the expected forecast performance is

$$\mathbb{E}[g(\hat{\epsilon}_{i,t}^{(1)}) - g(\hat{\epsilon}_{i,t}^{(2)})] = 0.$$

Let $z_{i,t} = g(\hat{\epsilon}_{i,t}^{(1)}) - g(\hat{\epsilon}_{i,t}^{(2)})$. The variance of $\bar{z}_i = \frac{1}{T_2 - T_1} \sum_{t=T_1+1}^{T_2} z_{i,t}$ is asymptotically, $\mathrm{var}(\bar{z}_i) \approx \frac{1}{T_2 - T_1}(\gamma_0 + 2\sum_{k=1}^{\infty} \gamma_k)$, where γ_k is the kth autocovariance of $z_{i,t}$ that is estimated as $\hat{\gamma}_k = \frac{1}{T_2 - T_1} \sum_{t=T_1+1}^{T_2} (z_{i,t} - \bar{z}_i)(z_{i,t-k} - \bar{z}_i)$. The Diebold Mariano test statistic is then given as

$$DM = \hat{\mathrm{var}}(\bar{z}_i)^{-\frac{1}{2}} \bar{z}_i, \quad \text{with} \quad \hat{\mathrm{var}}(\bar{z}_i) = \frac{1}{T_2 - T_1} \left(\hat{\gamma}_0 + 2 \sum_{k=1}^{K} \hat{\gamma}_k \right),$$

where K is an increasing function of sample size. Under the null hypothesis, this statistic has an asymptotic standard normal distribution. Unless stated otherwise, $g(x) = x^2$, i.e. function g gives the squared residuals. See (Diebold and Mariano, 1995) and (Harvey, Leybourne and Newbold, 1997) for further details.

It should be emphasized, that these measures should be interpreted with care. An ideal measure would depend on the actual trading strategy used. Thus, a real test of the forecasting quality is context dependent and would consist in evaluating the profits made by a specific portfolio selection strategy. As an example let us consider the following criterion function:

$$C(\omega_t, \alpha) = -\omega_t' \hat{y}_{t|t-1} + \alpha \omega_t' \hat{\Sigma}_{t|t-1} \omega_t, \tag{5.1}$$

where ω_t is the n–dimensional vector of portfolio weights at time t, $\hat{\Sigma}_{t|t-1} = \frac{1}{t-1}\sum_{s=1}^{t-1} \hat{\epsilon}_s \hat{\epsilon}_s'$ is the prediction error variance covariance matrix, and α is some factor representing risk aversion (throughout we chose $\alpha = 0.5$). So, by minimizing function (5.1) with respect to ω_t, we optimize the trade-off between expected return, $\omega_t' \hat{y}_{t|t-1}$, and risk. We choose two simple strategies corresponding to the two following sets of restrictions on the portfolio weights:

Strategy I:

$$\begin{aligned} \textstyle\sum_{i=1}^{n} \omega_{i,t} &= 1 \\ \omega_{i,t} &\geq 0, \quad \text{for all } i = 1, \ldots, n \end{aligned}$$

Strategy II:

$$\sum_{i=1}^{n} \omega_{i,t} = 0$$

Note that the weights obtained from strategy II were (after optimization) always rescaled such that $\sum \omega_{i,t}^+ = -\sum \omega_{i,t}^- = 0.5$, where $\sum \omega_{i,t}^+$ ($\sum \omega_{i,t}^-$) means that we sum just the positive (negative) weights. In Strategy I the restriction $\sum_{i=1}^{n} \omega_{i,t} = 1$ gives the allocation of one monetary unit on the shares contained in the portfolio at time t. Short selling, i.e. $\omega_{i,t} < 0$ for any i, is not allowed. In the scenarios presented later on we have an investor that given a certain capital of K_t monetary units at time t invests everything in the underlying portfolio, so that at time $t+1$ he has $K_{t+1} = K_t + K_t \hat{\omega}_{t+1}' y_{t+1}$, which he reinvests in period $t+1$, and so on. Strategy II, on the other side, allows for short selling, where the money invested in certain assets equals the amount of money obtained from short selling other positions. So, at time t our investor sells positions of a total amount of $0.5K_t$ and from this money he buys other positions. At time $t+1$ he is left with $K_{t+1} = K_t + K_t \hat{\omega}_{t+1}' y_{t+1}$.
In the figures shown below we plot the time series of the capital K_t, where our investor is assumed to have $K_0 = 100$ monetary units available as starting capital.

In the following section we will summarize the results obtained concerning the main issues and problems in developing forecasting strategies for return series.

5.2 Problems in developing forecasting strategies

Input Selection: As it has been mentioned in the introduction of this part, we believe that a necessary condition for successful forecasting is to make use of information contained in inputs. In other words, we think, that there may be weak form efficiency (i.e. efficiency in relation to the history of the returns), but no semi-strong efficiency (where semi-strong efficiency means efficiency in relation to publicly available information). Since there is no clear a priori knowledge concerning relevant inputs available for us, we chose explanatory variable candidates that provide information of the financial sector or specific branches and regions in which the banks are invested. Hence, data driven input selection is a particularly important issue here.

One of the most important inputs throughout the forecasting methods appeared to be the S&P Bank Index (BIX). In any method and model it got selected nearly every specification period.

Modelling of Dynamics: Linear dynamics is modelled here by adding lagged inputs and outputs to the set of explanatory variable candidates. Thus, input selection and dynamic specification is performed in one step here. Lagged output variables, and in the IN method also lagged factor scores, seem to be essential for all three forecasting methods. This, in a certain sense, even loosens the above assumption that there may be weak form efficiency.

Specification: Throughout the methods we only computed models with a BIC-type criterion in the second step of the refined FSP (see section 2.3), as our previous investigations showed better results with BIC than with AIC.

In case of VARX models 8(1) lagged output variables have been selected on average per equation, if AIC(BIC) was the IC criterion in the first step of the refined FSP. The average number of selected (current and lagged) exogenous variables per equation ranges from 2 in case of BIC to 19 in case of AIC, when the exogenous variables were chosen from total Group two, and from 1 in case of BIC to 5 in case of AIC, when the exogenous variables were chosen from current and lagged variables of the set $\{BIX, IRX, NDc1\}$).

Due to the BIC-type criterion in the second step of the refined FSP in case of the RR method the rank r of the coefficient matrix was specified to be one throughout, whereas the number of explanatory variables $\bar{k}$ ranges from 1 up to 16 across the different models estimated.

Whenever a BIC-like penalty term was chosen to determine the IC-optimal number of factors for the IN method just one factor was selected, regardless of which estimator ("regression" or "Bartlett") for the factor is chosen. When an AIC-like penalty term was chosen, the specification procedure tended to select slightly less factors when the factors are estimated with the "regression"-method than when the factors are estimated with Bartlett's method. Especially when the "promax" method was applied to additionally transform the factors and when the factors were estimated by the "regression"-method, the procedure throughout selected only one factor.

Structural Changes: Structural changes seem to be an important issue for a number of reasons. For instance, it is quite common to distinguish between bear, bull and sideward regimes. Structural breaks and smooth transition in returns have been investigated in Krca (2002) and are not explicitly considered here. Slowly time varying parameters have been taken into account in our procedure by using adaptive identification methods: We considered a weighting factor that decreases exponentially and therefore gives less weight to past observations, and rolling and expanding estimation windows. As rolling window we chose a window of 433 observations, which corresponds to 69.5% of the whole sample, a bit less than 21 months. Finally, our models are respecified every 5 and 10 days, respectively, which also accounts for structural breaks and allows for adapting to new economic regimes or states, since the set of common components or factors and explanatory variables is allowed to alter accordingly.

Concerning design parameters like the weighting factor and rolling or expanding estimation window, we found the following: Multiplying observations at time t with a weighting factor λ^{T-t} does not seem to be crucial. Our procedure hardly ever selected a λ smaller than 1. Thus, changes in the set of explanatory variables or in the number of common components are much more striking, or the structural changes occurring in our data cannot be handled adequately by a simple weighting factor. Making use of a rolling or expanding window also seems to be of secondary importance in our context. The out-of-sample results are comparable, maybe with a slight preference for the expanding window version, see e.g. table 5.9. We should mention, however, that we only considered a fixed window size of 433 observations for the rolling windows. Different sizes could have led to different results, however we do not think that this point is important. In addition, by the introduction of an additional integer parameter that governs the rolling window size we would have made estimation and specification even more complex at the expense of computing time and the danger of overfitting. So, we did not follow this line here.

What we could observe, however, was that our procedures made great demands on the possibility of respecifying the integer valued parameters indicating that structural changes do take place.

In figure 5.3 the time path of the estimated coefficients of model rr5.indirect.BICF-BIC.FALSE is given. It shows that for every input the signs of the coefficients are the same for all bank return series except for Banco Bilbao for which the signs flip. Furthermore note that most of the inputs are not always selected as explanatory variable, as their coefficient series are given only for a certain subset of the time span considered. The numbers of selected explanatory variables, $\bar{k}$, for the underlying model ranges from 5 to 9.

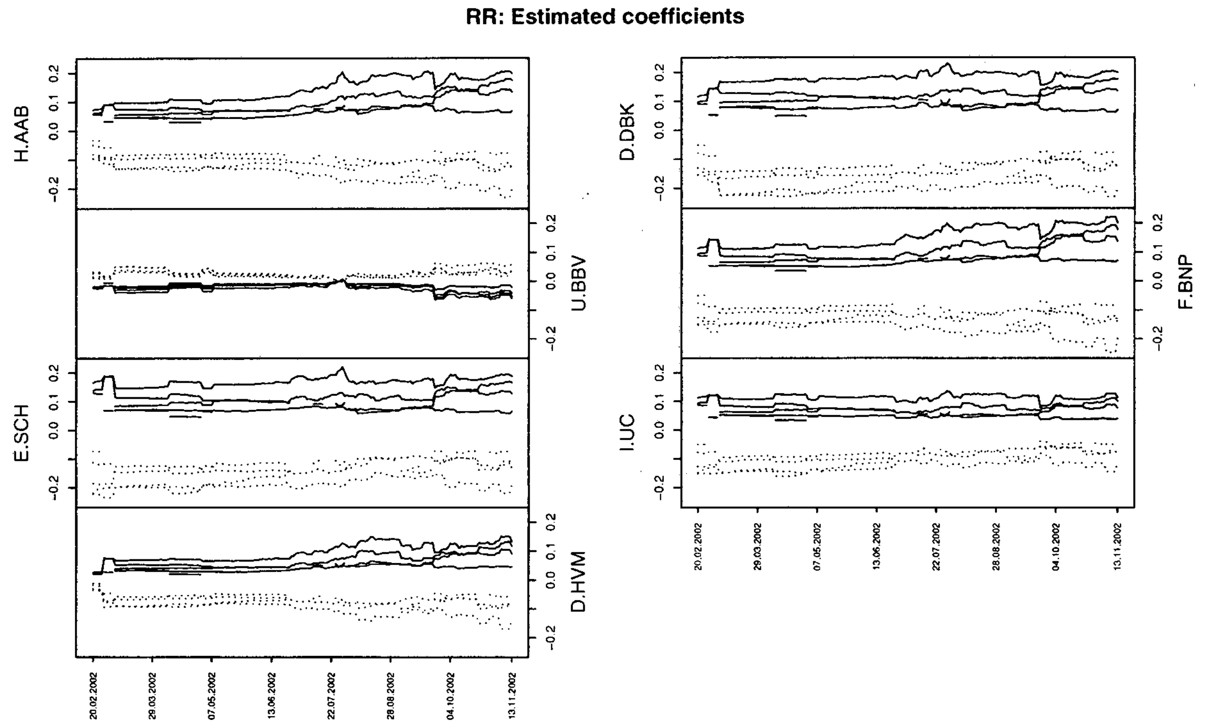

Figure 5.3: Coefficient of model rr5.indirect.BICF-BIC.FALSE; solid lines: $U.BBV_{t-1}$, $U.BBV_{t-2}$, $D.DBK_{t-1}$, BIX_{t-1}, IRX_{t-2}, $NDcl_{t-1}$ and $NDcl_{t-2}$; dotted lines: $E.SCH_{t-1}$, $E.SCH_{t-2}$, $F.BNP_{t-2}$, $I.UC_{t-1}$, $I.UC_{t-2}$, SKX_{t-2}, $LIEUR1YD_{t-6}$, $CH2MTRR_{t-2}$, IRX_{t-6} and $SXFE_{t-2}$.

Possible Nonlinearities: During the last decade modelling of financial data by neural nets has attracted great attention (see e.g. Abu-Mostafa, Atiya, Magdon-Ismail and White (2001)). Since we are interested in short term forecasting only (daily returns), we think that nonlinearity is not a big issue here, except for structural changes.

In (Deistler and Hamann, 2005) some results for neural nets are provided to the same data set. Their forecasting performance however is disappointing.

Outliers: Detection of outliers in general is an important problem in return series and needs a great amount of expertise, especially, in order to distinguish them from structural breaks. Thus, it requires specific knowledge about the data generating process and the process gathering the data of interest. The investigations in (Deistler and Hamann, 2005) concerning this topic do not lead to a clear general answer to the question whether outlier adjustment in case of the underlying data set does improve results or not.

5.3 Comparison of the forecasting methods

The largest impact on the forecasting results has the choice between VARX, RR and IN. As far as the forecasting qualities (with respect to the out-of-sample R^2 and hit rate) are concerned RR seems to be the best, whereas the performance of IN underachieves, see table 5.4 for an example. Figure 5.4 and tables 5.5, 5.6 and 5.7 give an overall impression of the methods' forecasting quality. The tables show the fractions of positive out-of-sample R^2 and hit rates greater than 0.5 for all models considered per forecasting method and confirms the ranking with respect to the out-of-sample R^2 and hit rate. Figure 5.4 gives an impression of the out-of-sample measures' distribution along the different design specifications. The relatively good results obtained using the simple -as compared to the competitors- VARX method is joyous, a bit surprising, though. Note that for VARX models using just current and lagged values of three exogenous variables on average show more positive out-of-sample R^2 and out-of-sample hit rates that are larger than 0.5 than when whole group two is provided.

	VARX		RR		IN		Benchmark	
banks	R^2	hit rate	R^2	hit rate	R^2	hit rate	R^2	hit rate
H.AAB	0.0461	0.56	0.0687	0.62	0.0144	0.54	−0.0017	0.51
U.BBV	−0.0002	0.50	0.0004	0.48	−0.0404	0.48	−0.0017	0.50
E.SCH	0.0194	0.54	0.0165	0.58	−0.0216	0.51	−0.0006	0.52
D.HVM	0.0263	0.51	0.0263	0.53	0.0292	0.48	0.0064	0.55
D.DBK	0.0456	0.52	0.0597	0.59	−0.0038	0.53	−0.0003	0.49
F.BNP	0.0799	0.53	0.1010	0.58	0.0134	0.52	−0.0017	0.46
I.UC	0.0094	0.51	0.0092	0.54	−0.0445	0.49	−0.0016	0.49

Table 5.4: The out-of-sample results for varx5.mva.BICF-BIC.TRUE (exogenous variables: BIX, IRX, NDcl), rr5.direct.BICF-BIC.TRUE, in10.Bartlett.default.mva.BICF-BIC.BIC.TRUE.FALSE (considering only forecasts based on the factor part of the model) and the geometric mean as benchmark model.

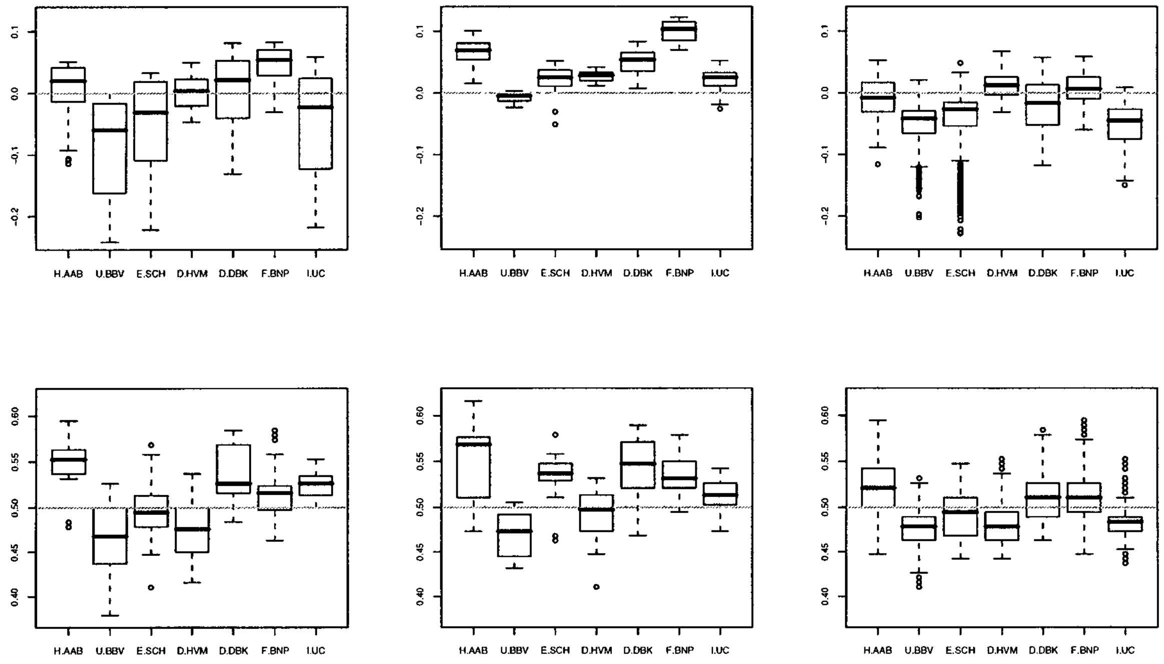

Figure 5.4: Boxplots of out-of-sample R^2 (first row) and hit rates (second row) of all VARX (first column), RR (second column) and IN models (third column) considered.

The next important design parameter is the choice of the information criterion for model selection. The best results for VARX and RR with respect to the out-of-sample measures are achieved by choosing BIC

$R^2 > 0$	H.AAB	U.BBV	E.SCH	D.HVM	D.DBK	F.BNP	I.UC
3 inputs	0.75	0.00	0.50	0.44	0.75	1.00	0.50
Group 2	0.56	0.00	0.25	0.69	0.50	0.63	0.38
all VARX models	0.66	0.00	0.38	0.56	0.63	0.81	0.44
hitrate > 0.5	H.AAB	U.BBV	E.SCH	D.HVM	D.DBK	F.BNP	I.UC
3 inputs	1.00	0.13	0.50	0.06	0.94	0.94	1.00
Group 2	0.88	0.25	0.25	0.38	0.81	0.38	0.88
all VARX models	0.94	0.19	0.38	0.22	0.88	0.66	0.94

Table 5.5: For all VARX models considered (16 models with 3 exogenous inputs (BIX, IRK, NDcl), 16 models with whole Group 2 as exogenous explanatory variable candidates), the table shows: Fractions of positive out-of-sample R^2 and hit rates greater than 0.5 are given.

$R^2 > 0$	H.AAB	U.BBV	E.SCH	D.HVM	D.DBK	F.BNP	I.UC
new.spec = 5	1.00	0.13	0.88	1.00	1.00	1.00	0.75
new.spec = 10	1.00	0.13	0.88	1.00	1.00	1.00	0.88
all RR models	1.00	0.13	0.88	1.00	1.00	1.00	0.81
hitrate > 0.5	H.AAB	U.BBV	E.SCH	D.HVM	D.DBK	F.BNP	I.UC
new.spec = 5	0.75	0.13	0.88	0.38	0.88	0.88	0.63
new.spec = 10	0.75	0.25	0.88	0.25	0.88	0.88	0.88
all RR models	0.75	0.19	0.88	0.31	0.88	0.88	0.75

Table 5.6: For all RR models considered (8 models with new.spec = 5, 8 models with new.spec = 10), the table shows: Fractions of positive out-of-sample R^2 and hit rates greater than 0.5 are given.

$R^2 > 0$	H.AAB	U.BBV	E.SCH	D.HVM	D.DBK	F.BNP	I.UC
Factors	0.52	0.00	0.02	0.64	0.38	0.75	0.11
Factors + Errors	0.32	0.04	0.18	0.68	0.43	0.59	0.00
all IN models	0.42	0.02	0.10	0.66	0.40	0.67	0.05
hitrate > 0.5	H.AAB	U.BBV	E.SCH	D.HVM	D.DBK	F.BNP	I.UC
Factors	0.78	0.12	0.26	0.11	0.60	0.53	0.06
Factors + Errors	0.71	0.07	0.54	0.09	0.62	0.78	0.23
all IN models	0.74	0.10	0.40	0.10	0.61	0.65	0.14

Table 5.7: For all IN models considered (192 models including factors only, 192 models including factors and errors), the table shows: Fractions of positive out-of-sample R^2 and hit rates greater than 0.5 are given.

in both steps of the refined FSP. The performance of portfolios of VARX models, however, seems to be somewhat better when AIC is used in the respective forward procedure, see figure 5.5. For the factor and noise component models in the IN method the combination BICF-BIC again outperforms the other IC combinations in the FSP. It seems however advantageous to choose AIC as information criterion in order to determine the number of factors.

The consequences of the choice between *mva* and *univ* give no clear picture throughout the methods of modelling and out-of-sample validation.

In table 5.8 a comparison between some selected models of all three modelling methods is given. The RR model performs slightly better than the others. In addition, it is shown for the IN models that adding the noise forecasts is helpful in some cases and may deteriorate the forecasts in other cases. Portfolio optimization, however, gives reason to prefer models that include the forecasts of idiosyncratic noise, see figure 5.6. Note that for IN the in-sample forecast errors $\hat{\epsilon}_t$, $t = 1, \ldots, T_1$, needed to estimate the forecast error variance covariance matrix, unfortunately were not saved during estimation. Anyway, the in-sample performance of the IN models is rather poor, too. The in-sample R^2 values across the seven banks range from 0.0328 to 0.3136. Hence, for IN $\frac{1}{t-1}\sum_{s=1}^{t-1} y_s y_s' \geq \hat{\Sigma}_{t|t-1} = \frac{1}{t-1}\sum_{s=1}^{t-1} \hat{\epsilon}_s \hat{\epsilon}_s'$ was used to compute the portfolio weights, see eq. 5.1.

In many cases the forecasts obtained for Banco Bilbao (U.BBV) are bad. Banco Bilbao seems to be a special case (see figure 5.3), possibly due to their engagement in South American countries. It can be seen in table 5.8 that, in particular, the results for Banco Bilbao improve by adding the noise forecasts in the IN case. The forecasts for UniCredit are reasonable for the RR model. We observe that the forecasts for UniCredit based on IN models are bad, in general. Thus, the explanatory variables for the noise component of UniCredit seem to be insufficient.

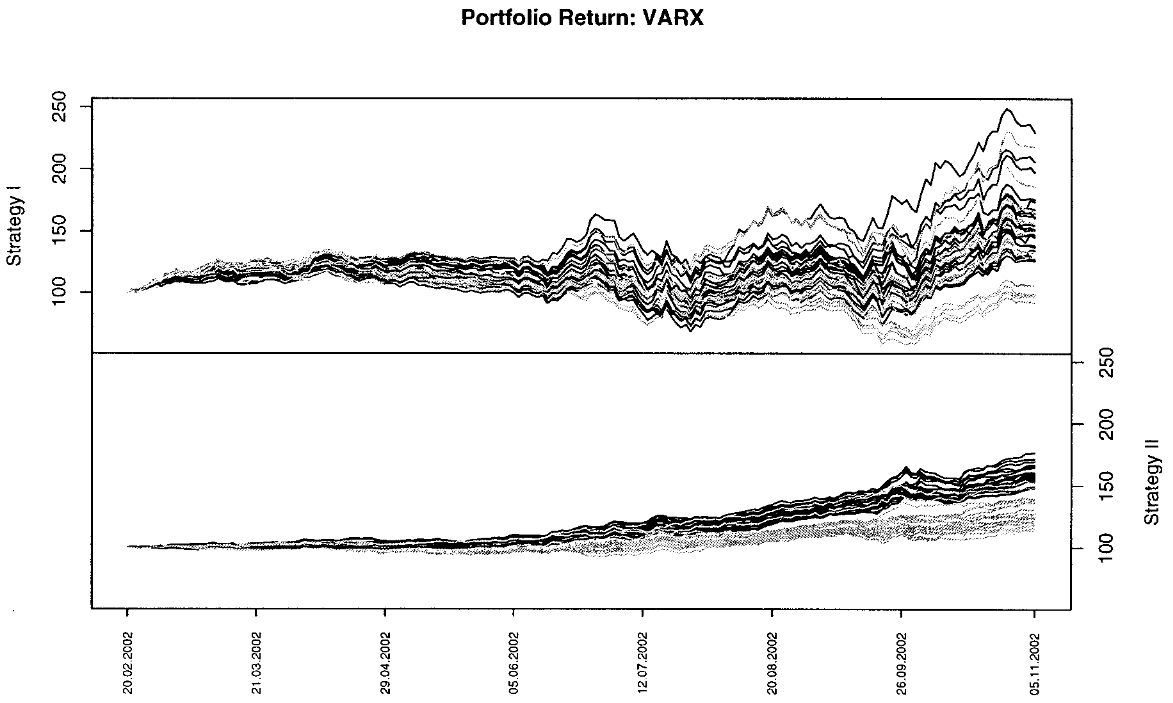

Figure 5.5: Capital development of portfolios. Computation of portfolio weights is based on forecasts of VARX models; Black lines: AIC in the first step of the refined FSP; Grey lines: BIC in the first step of the refined FSP.

Choosing between moving and expanding windows, the different re-specification periods and between the orthogonal projection of the input variables of the noise component or not, has little impact for the forecasting quality. For IN *regression* gives somewhat better results than *Bartlett* and for RR *indirect* estimation, thus MLE, gives somewhat better results than *direct* estimation. For the latter, see table 5.9.

The DM test results are disappointing, since the Null hypothesis *the zero-forecast is better than our forecast* cannot be rejected. Just in some rare cases the p-value is below 0.1. The results stay the same if we replace the zero-forecast by the sample (geometric) mean, which for all banks are slightly below zero, see table 5.1.

If we consider portfolio optimization our forecasts however appear superior to the benchmark portfolio, see figure 5.7.

In short, non-trivial forecasting of asset return series is not an easy task. However, figure 5.4 shows that RR models may yield reasonable out-of-sample R^2. RR models are easy to implement and estimate.

	VARX		RR		Forecasting with the factor part IN		Forecasting with the factor and noise part IN	
banks	R^2	hit rate	R^2	hit rate	R^2	hit rate	R^2	hit rate
H.AAB	0.0503	0.56	0.1006	0.57	0.0519	0.57	0.0433	0.56
U.BBV	−0.0076	0.51	−0.0049	0.45	−0.0170	0.47	0.0032	0.50
E.SCH	0.0183	0.52	0.0426	0.54	0.0098	0.53	0.0334	0.54
D.HVM	0.0127	0.48	0.0333	0.49	0.0607	0.52	0.0672	0.50
D.DBK	0.0514	0.51	0.0591	0.53	0.0563	0.54	0.0538	0.54
F.BNP	0.0820	0.52	0.1159	0.52	0.0591	0.54	0.0420	0.56
I.UC	0.0084	0.52	0.0251	0.51	−0.0123	0.51	−0.0282	0.50

Table 5.8: The out-of-sample results of models varx10.mva.BICF-BIC.TRUE (exogenous variables: BIX, IRX, NDcl), rr10.indirect.BICF-BIC.TRUE and in10.regression.default.univ.BICF-BIC.AIC.TRUE.TRUE.

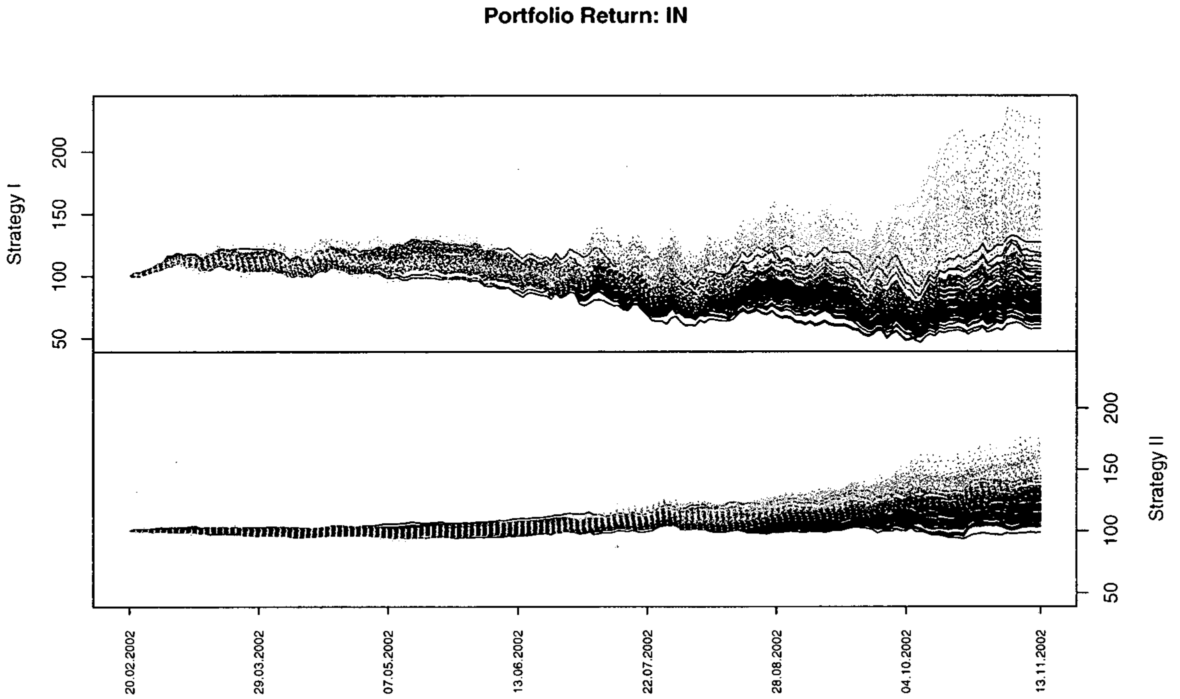

Figure 5.6: Capital development of portfolios. Computation of portfolio weights is based on forecasts of IN models, respecified every 10 days; Solid lines: factor forecasts only; Dotted lines: factor plus idiosyncratic noise forecasts.

Especially for the second half of the validation sample of the underlying data set, non-trivial forecasts seem to be worthwhile, see table 5.10 and figure 5.7.

	direct				*indirect*			
	mov. window		exp. window		mov. window		exp. window	
banks	R^2	hit rate	R^2	hit rate	R^2	hit rate	R^2	hit rate
H.AAB	0.0687	0.62	0.0690	0.57	0.0953	0.57	0.0856	0.58
U.BBV	0.0004	0.48	−0.0057	0.43	−0.0050	0.45	−0.0040	0.47
E.SCH	0.0165	0.58	0.0301	0.56	0.0351	0.54	0.0487	0.54
D.HVM	0.0263	0.53	0.0338	0.49	0.0418	0.50	0.0395	0.53
D.DBK	0.0597	0.59	0.0704	0.57	0.0734	0.54	0.0830	0.58
F.BNP	0.1010	0.58	0.0929	0.52	0.1136	0.52	0.1201	0.55
I.UC	0.0092	0.54	0.0431	0.53	0.0137	0.47	0.0524	0.54

Table 5.9: The out-of-sample results of the models rr5.direct.BICF-BIC.TRUE, rr5.direct.BICF-BIC.FALSE, rr5.indirect.BICF-BIC.TRUE, rr5.indirect.BICF-BIC.FALSE.

	RR				Benchmark			
	First half		Second half		First half		Second half	
banks	R^2	hit rate	R^2	hit rate	R^2	hit rate	R^2	hit rate
H.AAB	0.1069	0.5158	0.0836	0.6526	0.0030	0.4421	−0.0022	0.5684
U.BBV	−0.0086	0.4316	−0.0030	0.5158	−0.0022	0.4737	−0.0016	0.5263
E.SCH	0.0297	0.5579	0.0520	0.5158	0.0008	0.4737	−0.0009	0.5684
D.HVM	0.0224	0.5263	0.0422	0.5368	−0.0024	0.4947	0.0079	0.6105
D.DBK	0.0352	0.5789	0.0934	0.5789	−0.0054	0.4632	0.0008	0.5263
F.BNP	0.0720	0.5158	0.1313	0.5895	−0.0018	0.4737	−0.0017	0.4526
I.UC	−0.0700	0.4632	0.0900	0.6211	−0.0041	0.4211	−0.0009	0.5368

Table 5.10: Out of sample measures for rr5.indirect.BICF-BIC.FALSE and the geometric mean as benchmark model for the first and the second half of the validation sample, i.e. 21.02.2002 to 03.07.2002 and 04.07.2002 to 13.11.2002.

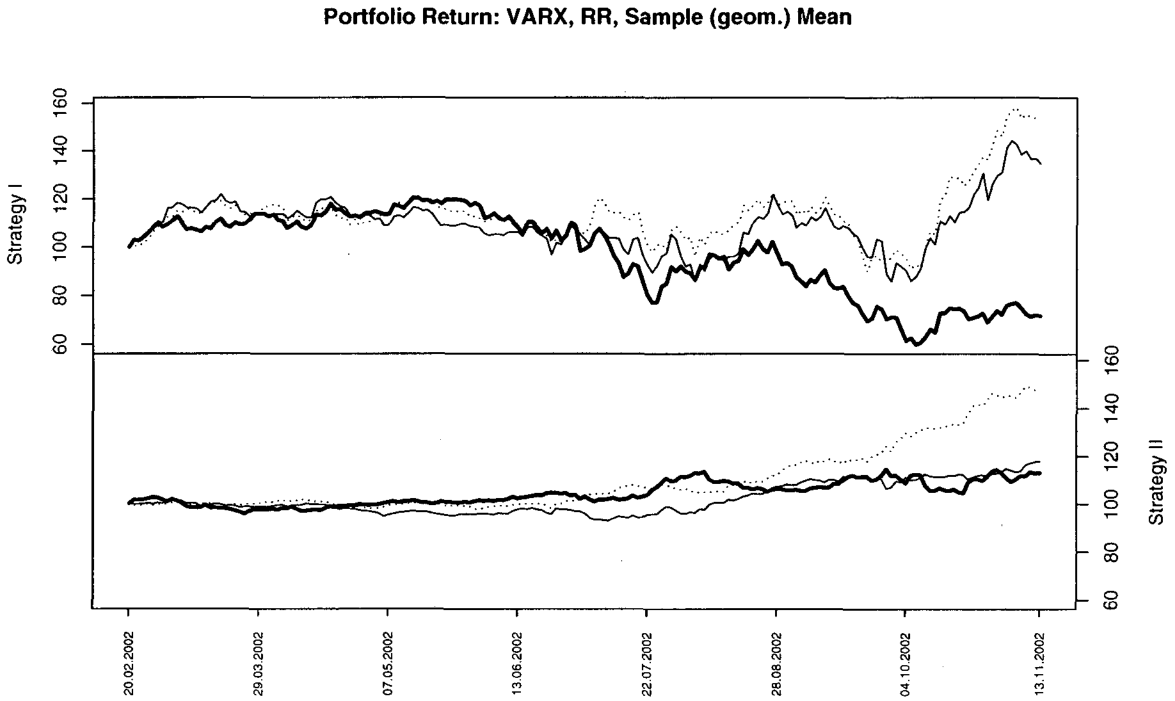

Figure 5.7: Capital development of portfolios. Computation of portfolio weights is based on forecasts of varx5.mva.BICF-BIC.FALSE (thin solid line), rr5.indirect.BICF-BIC.FALSE (dotted line), and as benchmark the sample geometric mean (thick solid line).

Part II

Forecasting Volatilities of Stock Prices

Chapter 6

Introduction

R.F. Engle describes in the introduction of his Nobel Lecture Notes, see (Engle, 2003), the central paradigm of finance as *"Optimal behavior that takes worthwhile risks"*. Knowledge of the level of risk taken is essential in portfolio optimization, pricing of assets and derivatives, computation of the *Value at Risk* (VaR), etc. Typically risk is measured by the variance of the underlying returns. Hence, reliable models for estimating variances are inevitable.

In the early eighties (Engle, 1982) proposed the so-called *autoregressive conditional heteroskedastic* (ARCH) model that obtained broad acceptance not only from researchers, but also from analysts on financial markets, and for which in 2003 he was awarded the Bank of Sweden Prize in Economic Sciences in Memory of Alfred Nobel. Let y_t denote the return of some asset and let μ_t be its expectation conditional on the sigma field, $\mathcal{I}_{t-1}$, generated by the past values of y_t, then the linear univariate ARCH(q) model can be written as

$$\begin{aligned} y_t &= \mu_t + \epsilon_t \\ \sigma_t^2 &= c + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2, \quad q > 0, \end{aligned} \tag{6.1}$$

where the conditional distibution of ϵ_t is for instance given by $\epsilon_t | \mathcal{I}_{t-1} \sim N(0, \sigma_t^2)$, if we assume normality. In this case ϵ_t may be written as the product $\epsilon_t = \sigma_t z_t$, where σ_t is the conditional standard deviation (that is also referred to as *volatility*), and z_t is some independently and identically distributed standard Gaussian random variable.

A generalization, the so-called GARCH(p, q) model, was proposed by (Bollerslev, 1986):

$$\begin{aligned} y_t &= \mu_t + \epsilon_t \\ \sigma_t^2 &= c + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2, \quad q > 0, \ p \geq 0. \end{aligned} \tag{6.2}$$

Today it is possibly the most widely used model of this type. The GARCH(1, 1) model has a particularly neat interpretation, since in this case the conditional variance is a weighted average of three different variance forecasts: First, the constant that corresponds to the long run average, second, the past estimated conditional variance, and third, a term that adjusts for the new information available.

Note that in order to guarantee that the estimate for σ_t^2 is positive for all t and that the GARCH process ϵ_t is stationary the parameters have to fulfill the following constraints:

- *Positivity constraints:*

  $$c > 0, \ \alpha_i \geq 0 \text{ and } \beta_j \geq 0 \text{ for all } i, j.$$

  It is easy to see that these constraints ensure a positive σ_t^2 for all ϵ_{t-i} and σ_{t-j}^2 in the sample space.

- *Stationarity constraints:* In (Bollerslev, 1986) it is shown that (ϵ_t) is wide sense stationary, i.e. $\mathbb{E}\epsilon_t = 0$, $\text{var}(\epsilon_t) = c(1 - \sum_{i=1}^{q} \alpha_i - \sum_{j=1}^{p} \beta_j)$ and $\text{cov}(\epsilon_t \epsilon_s) = 0$ for all $t \neq s$, if and only if
$$\sum_{i=1}^{q} \alpha_i + \sum_{j=1}^{p} \beta_j < 1.$$

In the meanwhile various extensions have been suggested due to shortcomings in the assumptions underlying the standard GARCH model, see for instance (Engle, 2003) for a list, or (Gourieroux, 1997) who provides an overview of ARCH models and their applications in financial and monetary economics. Let us here just mention one drawback of the model class that might be crucial for financial applications: In the standard GARCH framework positive and negative error terms (one can interpret these as positive and negative shocks or news) have the same effect on the volatility. The information concerning the sign of ϵ_t is lost due to the consideration of squared terms of past shocks. In practice, in particular for stock returns, one can observe however the so-called *Leverage Effect*, i.e. the volatility increases more after bad news then after good news. Thus, the assumption of symmetric effects in the standard GARCH framework is often violated in practice. The *exponential GARCH* (EGARCH) model proposed by (Nelson, 1991) or *Stochastic Volatility* models (see e.g. (Hull and White, 1987; Melino and Turnbull, 1990)) are examples for model classes, which may model this asymmetric behaviour, but are not further discussed here.
Anyway, GARCH models are successful in financial applications since financial returns do have characteristics that can be modelled by GARCH models. First of all, volatilities of returns tend to cluster, i.e. a large/small change in the asset price is very likely again followed by a large/small change. Second, extreme values appear quite often in series of asset returns. E.g., the probability for an extreme value in an ordinary return series is in general higher than for an independently and identically distributed family of Gaussian random variables. That is, the distribution underlying the process of returns has in general fatter tails than the normal distribution.

In this part of the thesis we will focus on multivariate generalizations of the above model classes. Let (y_t) be an n–dimensional vector process of returns, and let $\mathcal{I}_{t-1}$ be again the sigma field generated by the past of y_t[1]. The basic multivariate framework is then given by

$$\begin{aligned} y_t &= \mu_t + \epsilon_t \\ \epsilon_t &= H_t^{\frac{1}{2}} z_t, \end{aligned} \tag{6.3}$$

where μ_t as above denotes the conditional mean of y_t, $\mu_t = \mathbb{E}(y_t|\mathcal{I}_{t-1})$, and may for instance be modelled as shown in part I. (ϵ_t) is the stochastic error process that is not independent but uncorrelated with zero mean, (z_t) is an n-dimensional vector process independently and identically distributed with mean zero and unit variance, $z_t \sim iid(0, I_n)$, and H_t is symmetric positive definite for all t and assumed to be $\mathcal{I}_{t-1}$–measurable.
Given all these assumptions, the variance of y_t conditional on $\mathcal{I}_{t-1}$ equals the variance of ϵ_t conditional on $\mathcal{I}_{t-1}$,

$$\text{var}(y_t|\mathcal{I}_{t-1}) = \mathbb{E}\{(y_t - \mu_t)(y_t - \mu_t)'|\mathcal{I}_{t-1}\} = \text{var}(\epsilon_t|\mathcal{I}_{t-1}) = H_t^{\frac{1}{2}} \text{var}(z_t|\mathcal{I}_{t-1}) H_t^{\frac{1}{2}} = H_t. \tag{6.4}$$

The last equality follows from the fact that z_t is independent of $\mathcal{I}_{t-1}$, and hence, $\text{var}(z_t|\mathcal{I}_{t-1}) = \text{var}(z_t) = I_n$.
The model classes discussed in the subsequent chapters are the VECH model and the BEKK[2] model, see (Bollerslev, Engle and Wooldridge, 1988), (Baba, Engle, Kraft and Kroner, 1991), (Engle and Kroner, 1995). Modelling of multivariate GARCH models is challenging for several reasons:

[1] Note, that $\mathcal{I}_{t-1}$ may also contain the past of some exogenous variables x_t. Here, we suppress x_t for simplicity.

[2] Note that the name of the BEKK model is formed by the first letters of the surnames of Y. Baba, R.F. Engle, D. Kraft and K. Kroner.

1. The model structure should be parsimonious but still flexible.
2. Positivity of the conditional variances, H_t, should be ensured for all sample paths.
3. The vector GARCH process ϵ_t should be (wide sense) stationary.

The VECH model e.g. is very flexible, but in order to fulfill the positivity constraint one has to impose complicated parameter restrictions. Furthermore, the number of parameters is of the order $O(n^4)$, which makes estimation infeasible for large n. BEKK models ensure positivity by construction, but they also suffer from the curse of dimensionality. The econometric literature of the last decade comprises many attempts of dimension reduction of the parameter space of VECH and BEKK models respectively by imposing further restrictions on the structure. For an example see (Bollerslev et al., 1988) who proposed the *Diagonal VECH* model (DVECH) or (Engle, Ng and Rothschild, 1990) who suggest a *Factor GARCH* model (F-GARCH)[3]. A survey of multivariate generalizations of the basic GARCH model is provided in (Bauwens, Laurent and Rombouts, 2003).

This part proceeds as follows: In chapter 7 a detailed description of both the VECH and BEKK model class is given. Chapter 8 will then deal with the problem of parametrization and identifiability. A simple to check characterization of VECH models which have an equivalent BEKK representation will be presented. It will be shown that in the bivariate case BEKK models are as general as VECH models. In higher dimensional cases, however, VECH models allow for more flexibility. A parametrization for a *generic*, i.e. open and dense, subset of BEKK(p, q, K) models (with $K = n^2$) is presented. Furthermore, two other parametrizations (also with $K = n^2$) are analyzed. It is shown that these parametrizations both do not cover a generic set of BEKK models. In addition, several alternative parametrizations of BEKK(p, q, K) models (with $K \leq n$), thus with a small number of additive terms are presented. Estimation of the models is discussed in chapter 9. Finally, chapter 10 concludes the second part of the thesis with some applications on simulated and real data.

[3] An investigation of the restrictions that are imposed on the parameter space by the aforementioned multivariate volatility models is given in (Kroner and Ng, 1998), who point out that the choice of a multivariate volatility model can substantially affect the conclusions of the analysis.

Chapter 7

Multivariate GARCH models (MGARCH)

In the following many symbols and operators will appear that have not been used above. For a detailed list and description see Appendix C.

7.1 VECH model

The VECH(p,q) model was proposed in (Bollerslev et al., 1988) and is given by

$$\text{vech}(H_t) = c + \sum_{i=1}^{q} A_i \text{vech}(\epsilon_{t-i}\epsilon'_{t-i}) + \sum_{j=1}^{p} A_{q+j}\text{vech}(H_{t-j}), \tag{7.1}$$

where c is an $(n(n+1)/2)$–dimensional vector and the A_i's are square matrices of order $n(n+1)/2$. This is a very general formulation of the conditional second moments of (ϵ_t), since each component of H_t can be formed by a linear combination of its own past and the past of all the other elements in H_t and $\epsilon_t\epsilon'_t$.

However, at least two drawbacks of this model are apparent. First, the high number of parameters to estimate, which increases rapidly with $O(n^4)$, implies that estimation of the model is infeasible for *large* n[1]. Second, H_t has to be positive definite for all ϵ_{t-i} and all H_{t-j} in the sample space. This is ensured if we assume that

$$\text{math}(c) > 0 \tag{7.2}$$

and that the contribution of each ARCH and GARCH term, respectively, is non negative, i.e.

$$\text{math}(A_i\text{vech}(ee')) \geq 0, \quad \text{for all } e \in \mathbb{R}^n \tag{7.3}$$

holds for $i = 1, \ldots, p+q$. Here we have used the fact that any symmetric matrix $H \geq 0$ may be factorized as $H = \sum_{i=1}^{n} e_i e'_i$ and thus, the condition

$$\text{math}(A_j\text{vech}(H)) \geq 0, \quad \text{for all } H \in \mathbb{R}^{n \text{x} n}, H = H', H \geq 0$$

is equivalent to (7.3). A VECH model that satisfies the above conditions (7.2) and (7.3) will be called *admissible*[2]. However, note that admissibility is hard to check for given parameters $c, A_1, \ldots, A_{p+q}$ and thus, hard to impose during estimation.

[1] The total number of parameters is $\frac{n(n+1)}{2} + (p+q)(\frac{n(n+1)}{2})^2$.

[2] Here 'admissibility' is used only in the sense that the VECH model fulfills at least the positivity constraints. It may however not satisfy some stationarity conditions. We will treat stationarity conditions later on.

In order to further analyze the admissibility constraint let us concentrate on one of the above terms, $A\text{vech}(ee')$ say. The matrix A is *admissible* if and only if

$$f(\nu,e) = \nu'\text{math}(A\text{vech}(ee'))\nu \geq 0, \quad \text{for all } \nu, e \in \mathbb{R}^n \tag{7.4}$$

Thus A is admissible if and only if the optimization problem

$$\begin{cases} \text{minimize} & \nu'(\text{math}(A\text{vech}(ee'))\nu \\ \text{subject to} & \|\nu\| = \|e\| = 1 \end{cases} \tag{7.5}$$

has an optimal value, $\gamma_{\mathcal{V}}^*(A)$ say, which is non negative. Note that (7.5) is a non convex optimization problem, which in general is hard to solve.

If we suppose that A corresponds to the lag one ARCH term, the function $f(\nu,e)$ has the following interpretation: For a given $\epsilon_{t-1} = e$, $f(\nu,e)$ is the contribution of this lag one ARCH term to the conditional variance of the linear combination $(\nu'\epsilon_t)$. Note furthermore that for fixed ν, $f(\nu,e)$ is a quadratic form in e and that for fixed e, it is a quadratic form in ν.

By rearrangement of terms, $f(\cdot,\cdot)$ may also be written as

$$f(\nu,e) = \nu'\text{math}(A\text{vech}(ee'))\nu = (\nu \otimes e)'Q(\nu \otimes e), \tag{7.6}$$

where $Q \in \mathbb{R}^{n^2 \times n^2}$ is a symmetric matrix that fulfills the following equations:

$$\begin{array}{rcll} q_{ij,kl} & = & q_{ji,lk}, & \text{for all } i,j,k,l \text{ (symmetry of Q)} \\ q_{ii,kk} & = & a_{I(i,i),I(k,k)}, & \text{for all } i,k \\ 2q_{ii,kl} & = & a_{I(i,i),I(k,l)}, & \text{for all } i, k > l \\ q_{ij,kk} & = & a_{I(i,j),I(k,k)}, & \text{for all } i > j, k \\ q_{ij,kl} + q_{ij,lk} & = & a_{I(i,j),I(k,l)}, & \text{for all } i > j, k > l. \end{array} \tag{7.7}$$

Here we have partitioned $Q \in \mathbb{R}^{n^2 \times n^2}$ into $n \times n$ sub-blocks of size $(n \times n)$ each and used the indexing as described in eq. (C.1) and (C.2) of appendix C. Note that the first set of equations relates to the symmetry of Q, and that due to the last set of equations the matrix Q is not uniquely determined for a given matrix A. Thus, there is always a whole set of matrices Q that all correspond to one and the same VECH term A, namely

$$\mathcal{Q}(A) = \left\{ Q(\omega) = Q_0 + \sum_{k=1}^{\bar{n}} \omega_k \Delta_k \,\middle|\, \omega_k \in \mathbb{R} \right\} = \{Q_0 + \Delta \,|\, \Delta \in \mathcal{A}_{n*n}\} \tag{7.8}$$

where Q_0 is an arbitrary matrix for which, given this matrix A, the above equations (7.7) hold, $\bar{n} = (n(n-1)/2)^2$, and $\mathcal{A}_{n*n}$ is the set of symmetric matrices of dimension $n^2 \times n^2$ whose diagonal sub-blocks of dimension $n \times n$ are zero and whose off-diagonal sub-blocks are antisymmetric[3]; see appendix C for a more detailed description. Hence, the Q-set $\mathcal{Q}(A)$ corresponding to the VECH-term A is an affine subset of $\mathbb{R}^{n^2 \times n^2}$ of dimension $\bar{n}$.

One can of course always select a unique matrix Q out of $\mathcal{Q}(A)$, if further normalizing conditions are imposed. For instance, $\check{Q}$ in $\mathcal{S}_{n*n}$ is the unique representative of a $\mathcal{Q}$-set $\mathcal{Q}(A)$, where not only $\check{Q}$ itself, but also all sub-blocks of $\check{Q}$ are symmetric:

$$\check{q}_{ij,kl} = \check{q}_{ij,lk} = \frac{1}{2} a_{I(i,j),I(k,l)} \;\; \forall i > j, k > l. \tag{7.9}$$

However, as we will see later on, it might be advantageous to consider other normalizing conditions.

Note that for a given symmetric $Q \in \mathbb{R}^{n^2 \times n^2}$ the relations (7.7) uniquely define a VECH term $A \in \mathbb{R}^{n(n+1)/2 \times n(n+1)/2}$. Thus, the mapping that maps matrices Q to matrices A as displayed in eq. (7.7) is surjective but not injective.

[3] A square matrix M is said to be antisymmetric (or skew symmetric), if $M = -M'$ holds. Note that the diagonal elements of antisymmetric matrices by definition are zero.

As an example the matrices $Q(\omega_1)$, $\omega_1 \in \mathbb{R}$, that all correspond to the same matrix A are given below for the simplest case, $n = 2$:

$$Q(\omega_1) = \underbrace{\left(\begin{array}{cc|cc} a_{11} & a_{12}/2 & a_{21} & a_{22}/2 \\ a_{12}/2 & a_{13} & a_{22}/2 & a_{23} \\ \hline a_{21} & a_{22}/2 & a_{31} & a_{32}/2 \\ a_{22}/2 & a_{23} & a_{32}/2 & a_{33} \end{array}\right)}_{=\check{Q}} + \omega_1 \underbrace{\left(\begin{array}{cc|cc} 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \\ \hline 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{array}\right)}_{=\Delta_1} \tag{7.10}$$

Note that up to now we have only dealt with admissibility conditions concerning the positive definiteness of the conditional variance matrices. A necessary and sufficient condition that the aforementioned multivariate GARCH model (7.1) has a covariance stationary solution (ϵ_t) is:

$$\text{All eigenvalues of } \sum_{i=1}^{q} A_i + \sum_{j=1}^{p} A_{q+j} \text{ are less than one in modulus.} \tag{7.11}$$

A proof of the latter statement is provided in (Engle and Kroner, 1995, Proposition 2.7.)[4]. Note that (7.11) is a highly non linear restriction, which is hard to impose on a parametrization. In particular, it imposes a "cross restriction", which jointly refers to all terms A_i. In special cases however, the stationarity condition simplifies significantly, e.g. the diagonal VECH and the factor GARCH model.

7.2 BEKK model

(Baba et al., 1991) proposed the BEKK(p, q, K) model,

$$H_t = CC' + \sum_{i=1}^{q}\sum_{k=1}^{K} B'_{i,k}\epsilon_{t-i}\epsilon'_{t-i}B_{i,k} + \sum_{j=1}^{p}\sum_{k=1}^{K} B'_{q+j,k}H_{t-j}B_{q+j,k}, \tag{7.12}$$

where C is a lower triangular nonsingular $n \times n$ matrix, $B_{i,k}$ are $n \times n$ parameter matrices and K determines the generality of the process.

Note that BEKK models (with C non singular) yield by construction positive definite variance covariance matrices H_t. BEKK parameter matrices, however, are not identifiable without further normalizing restrictions. We will come back to this point in chapter 8.

Let us now examine the Q matrices corresponding to a BEKK-term[5], $\sum_{k=1}^{K} B'_k ee' B_k$:

$$\nu'\left(\sum_{k=1}^{K} B'_k ee' B_k\right)\nu = \sum_{k=1}^{K}(\nu' B'_k e)(\nu' B'_k e)' = (\nu \otimes e)' \underbrace{\sum_{k=1}^{K} \operatorname{vec}(B_k)\operatorname{vec}(B_k)'}_{=Q}(\nu \otimes e). \tag{7.13}$$

This implies that any BEKK-term has an equivalent representation as an admissible VECH-term and the corresponding $\mathcal{Q}$-set $\mathcal{Q}(A)$ contains at least one positive semidefinite element $Q = \sum_k \operatorname{vec}(B_k)\operatorname{vec}(B_k)'$. On the other hand if the $\mathcal{Q}$-set corresponding to a VECH-term, A say, contains a positive semidefinite element, $Q \geq 0$ say, then this A is admissible and has equivalent BEKK representations. To see this, just note that $Q \geq 0$ implies that Q may be factorized as $Q = \sum_{k=1}^{K} U_k U'_k$, where $U_k \in \mathbb{R}^{n^2}$ and $K \leq n^2$ is the rank of Q, and thus $B_k = \operatorname{mat}(U_k)$, $k = 1, \ldots, K$, gives a representation for the corresponding BEKK-term. Of course this representation is not unique. To obtain uniqueness one has to select a unique $Q \geq 0$ out of the corresponding $\mathcal{Q}$-set first, and then define a unique factorization of Q.

[4] Note that in fact Engle and Kroner show the above statement for the parameter matrices of the so-called VEC model. Anyway, the parameter matrices in (Engle and Kroner, 1995) may be obtained from the transformation $G_n A_i G_n^+$, where G_n denotes the so-called *duplication matrix* and $G_n^+ = (G'_n G_n)^{-1} G'_n$ is a left inverse of G_n, see also appendix (C). Due to the fact that G_n is orthogonal, these transformations have no effect on the eigenvalues. Thus, the condition stays the same.

[5] Since positive semidefinite matrices $H \geq 0$ may be factorized as $H = \sum e_i e'_i$, it follows that the "GARCH terms" $\sum_{k=1}^{K} B'_k H B_k$ may be treated completely analogous to the "ARCH terms" $\sum_{k=1}^{K} B'_k ee' B_k$.

As has been noted already it is hard to check whether a given VECH term, A say, is admissible. However, it is easy to check whether A admits a BEKK representation. To this end consider the following *semidefinite program* (s.d.p.),

$$\begin{cases} \text{maximize} & \lambda \\ \text{subject to} & (Q_0 + \sum_{k=1}^{\bar{n}} \omega_k \Delta_k) - \lambda I_{n^2} \geq 0 \end{cases} \tag{7.14}$$

where Q_0 is an arbitrary solution of (7.7). It can be shown that this s.d.p. has always an optimizer, (λ^*, ω^*) say, see Lemma B.3 in the Appendix. Furthermore, note that the optimal λ^* is the minimal eigenvalue of $Q(\omega^*) = Q_0 + \sum \omega_k^* \Delta_k$. In other words, the above s.d.p. maximizes $\lambda_{\min}(Q)$ over the set $\mathcal{Q}(A)$, where $\lambda_{\min}(Q)$ denotes the minimal eigenvalue of Q. Thus, $\mathcal{Q}(A)$ contains a positive semidefinite matrix Q if and only if the optimal λ is non negative. For a detailed description of semidefinite programs and their properties refer to (Vandenberghe and Boyd, 1996). Note that (7.14) is a convex relaxation of problem (7.5).

The theorem below summarizes some basic facts on VECH and BEKK models. For simplicity of notation only single terms will be considered, however the generalization to complete models is straightforward. Let us now define the following sets:

Let $\mathcal{V} \subseteq \mathbb{R}^{n(n+1)/2 \times n(n+1)/2}$ denote the set of all admissible VECH-terms A. Note that $A \in \mathcal{V}$ if and only if $\gamma_{\mathcal{V}}^*(A) \geq 0$ holds, see (7.5). Let $\mathcal{B} \subseteq \mathcal{V}$ denote the set of all VECH-terms A, which have a BEKK representation. Let $\lambda_{\mathcal{B}}^*(A) = \lambda^*$ denote the optimum value attained for problem (7.14), and note that $A \in \mathcal{B}$ if and only if $\mathcal{Q}(A)$ contains a positive semidefinite element $Q \geq 0$, i.e. if and only if $\lambda_{\mathcal{B}}^*(A) \geq 0$ holds. Finally, let $\mathcal{B}^+ \subset \mathcal{B} \subseteq \mathcal{V}$ denote the set of all BEKK terms A, where $\mathcal{Q}(A)$ contains a positive definite element $Q > 0$, i.e. $\lambda_{\mathcal{B}}^*(A) > 0$.

Theorem 7.1 (Facts on VECH and BEKK models)

1. *$\mathcal{V}$ is a closed, convex cone in $\mathbb{R}^{n(n+1)/2 \times n(n+1)/2}$ that contains an open subset of $\mathbb{R}^{n(n+1)/2 \times n(n+1)/2}$.*
2. *$\mathcal{B}$ is a closed, convex cone in $\mathbb{R}^{n(n+1)/2 \times n(n+1)/2}$ that contains an open subset of $\mathbb{R}^{n(n+1)/2 \times n(n+1)/2}$.*
3. *The set $\mathcal{B}^+ \subset \mathcal{B}$ is open in $\mathbb{R}^{n(n+1)/2 \times n(n+1)/2}$ and dense in $\mathcal{B}$.*
4. *$\mathcal{B} = \mathcal{V}$ holds for $n = 2$, whereas for $n > 2$ the set $(\mathcal{V} \setminus \mathcal{B}) = \{A \in \mathcal{V} \,|\, A \notin \mathcal{B}\}$ contains an open subset of $\mathbb{R}^{n(n+1)/2 \times n(n+1)/2}$.*

Proof. A set $\mathcal{X} \subseteq \mathbb{R}^{m \times m}$ is a convex cone if and only if for all $X_1, X_2 \in \mathcal{X}$ and for all $\lambda_1, \lambda_2 \geq 0$ it follows that $(\lambda_1 X_1 + \lambda_2 X_2) \in \mathcal{X}$. In particular, note that $\mathcal{S}_m^+ \subset \mathbb{R}^{m \times m}$ is a (closed) convex cone and this property is carried forward to the sets $\mathcal{V}$ and $\mathcal{B}$.

We start with the proof of statement (2): Let $A_i \in \mathcal{B}$, $Q_i \in \mathcal{Q}(A_i) \cap \mathcal{S}_{n^2}^+$ and $\lambda_i \geq 0$ for $i = 1, 2$ be given. This implies $\lambda_1 Q_1 + \lambda_2 Q_2 \geq 0$, $\lambda_1 Q_1 + \lambda_2 Q_2 \in \mathcal{Q}(\lambda_1 A_1 + \lambda_2 A_2)$ and thus $\lambda_1 A_1 + \lambda_2 A_2 \in \mathcal{B}$. Lemma B.3 shows that the function $\lambda_{\mathcal{B}}^*(A)$ is continuous and thus $\mathcal{B}$ is closed. Finally, let A_I be the VECH-term corresponding to $Q = I_{n^2}$. Then $\lambda_{\mathcal{B}}^*(A_I) = 1$ implies that an open neighborhood of A_I is contained in $\mathcal{B}$.

Statement (1) follows by a similar reasoning.

Concerning statement (3): The set $\mathcal{B}^+$ is open in $\mathbb{R}^{n(n+1)/2 \times n(n+1)/2}$ since $\lambda_{\mathcal{B}}^*(\cdot)$ is continuous and since $\mathcal{B}^+$ is non void. The set $\mathcal{B}^+$ is dense in $\mathcal{B}$ due to the fact that the set $\mathcal{S}_{n^2,n^2}^+$ is dense in $\mathcal{S}_{n^2}^+$

It remains to prove the last statement. First, consider the case $n = 2$: Let an admissible VECH-term A be given and let $\mathcal{Q}(A) = \{Q_0 + \omega_1 \Delta_1\}$, see (7.10). Let (λ^*, ω_1^*) denote the optimizer of the s.d.p. (7.14) where w.l.o.g. we assume that $\omega_1^* = 0$ holds. Furthermore let $W \in \mathbb{R}^{4 \times k}$, $1 \leq k \leq 4$ be an orthonormal basis of the eigenspace of Q_0 corresponding to its minimal eigenvalue λ^*. If $W' \Delta_1 W > 0$ then $(Q_0 - \lambda^* I_4) + \omega_1 \Delta_1 > 0$ holds for all sufficiently small $\omega_1 > 0$, which is a contradiction to the optimality of $\omega_1^* = 0$. Analogously we can rule out the case $W' \Delta_1 W < 0$. Hence, there must exist a vector $o \in \mathbb{R}^k$, $\|o\| = 1$ such that $(Wo)'\Delta_1(Wo) = \mathrm{tr}(Wo)(Wo)'\Delta_1 = 0$. By Lemma B.4 it follows that (Wo) has a representation as $(Wo) = (\nu \otimes e)$ for suitably chosen $\nu, e \in \mathbb{R}^2$. Since A is admissible

$$\lambda^* = (Wo)'Q_0(Wo) = (\nu \otimes e)'Q_0(\nu \otimes e) \geq 0$$

holds and thus Q_0 corresponds to a BEKK-term.

For the case $n = 3$ consider the matrix Q_0 given below.

$$Q_0 = \left(\begin{array}{rrr|rrr|rrr}
1.435 & -0.041 & -0.332 & -0.422 & -0.040 & -0.150 & 0.128 & 0.054 & 0.102 \\
-0.041 & 0.672 & 0.229 & -0.052 & -0.146 & -0.227 & 0.531 & -0.515 & 0.182 \\
-0.332 & 0.229 & 0.405 & -0.142 & -0.390 & -0.194 & 0.059 & 0.022 & -0.201 \\
\hline
-0.422 & -0.052 & -0.142 & 0.788 & 0.357 & -0.103 & -0.329 & -0.218 & 0.024 \\
-0.040 & -0.146 & -0.390 & 0.357 & 0.645 & 0.304 & 0.020 & -0.394 & -0.050 \\
-0.150 & -0.227 & -0.194 & -0.103 & 0.304 & 0.403 & 0.073 & -0.003 & -0.001 \\
\hline
0.128 & 0.531 & 0.059 & -0.329 & 0.020 & 0.073 & 0.688 & -0.496 & 0.210 \\
0.054 & -0.515 & 0.022 & -0.218 & -0.394 & -0.003 & -0.496 & 0.945 & 0.244 \\
0.102 & 0.182 & -0.201 & 0.024 & -0.050 & -0.001 & 0.210 & 0.244 & 1.095
\end{array}\right)$$

By means of a numerical optimization one can show that $(\nu \otimes e)'Q_0(\nu \otimes e) \geq 0.0019 > 0$ holds for all $\nu, e \in \mathbb{R}^3$ with $\|\nu\| = \|e\| = 1$. Thus, Q_0 corresponds to an admissible VECH-term. On the other hand, the s.d.p. (7.14) delivers an optimal value $\lambda^* = -0.0012 < 0$ and thus Q_0 does not correspond to a BEKK-term.

For the case $n > 3$ an example for an admissible VECH-term which does not correspond to a BEKK model may be constructed by suitably extending the above $n = 3$ example with zeroes.

Finally, let $A \in (\mathcal{V} \setminus \mathcal{B})$ and thus $\lambda^*_{\mathcal{B}}(A) = \lambda_0 < 0$ and $\gamma^*_{\mathcal{V}}(A) \geq 0$. Moreover, let A_I denote the VECH-term corresponding to $Q = I_{n^2}$. It is easy to see that $\lambda^*_{\mathcal{B}}(A - A_I\lambda_0/2) = \lambda_0/2 < 0$ and $\lambda^*_{\mathcal{V}}(A - A_I\lambda_0/2) \geq -\lambda_0/2 > 0$ holds. Thus, by continuity of $\lambda^*_{\mathcal{V}}(\cdot)$ and $\lambda^*_{\mathcal{B}}(\cdot)$ it follows that $(\mathcal{V} \setminus \mathcal{B})$ contains an open subset of $\mathbb{R}^{n(n+1)/2 \times n(n+1)/2}$. □

The above theorem shows in particular that in the bivariate case, $n = 2$, BEKK models are as general as VECH models, whereas for $n > 2$ there is a 'thick' set of admissible VECH models which have no equivalent BEKK representation.

The set $\mathcal{B}^+$ is a generic subset of BEKK models for which a parametrization will be given in the next chapter.

Concerning the stationarity condition, note that any BEKK model may be cast as a VECH model:

$$A_i = \sum_{k=1}^{K} G_n^+(B_{i,k} \otimes B_{i,k})'G_n, \quad \text{for } i = 1, \ldots, p+q,$$

where G_n denotes the respective duplication matrix, and $G_n^+ = (G_n'G_n)^{-1}G_n'$ is a left inverse of G_n, see also appendix (C). Hence, the existence of a covariance stationary process (ϵ_t) is ensured, if the coefficients matrices of the corresponding VECH model fulfill the constraint given in (7.11).

Chapter 8

Parametrization of BEKK models

Before we start with the description of possible parametrizations let us consider the geometry of the problem.

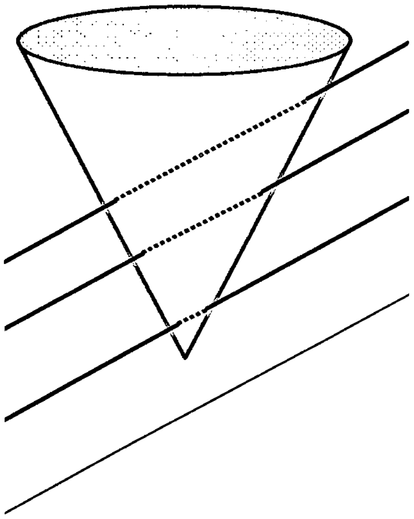

Figure 8.1: Gray: Cone containing all p.d. and p.s.d. Q matrices; Black lines: equivalence classes $\mathcal{Q}(A)$ (thick lines: contain at least one p.s.d matrix Q and therefore have a BEKK representation, thin lines: contain no p.s.d. matrix Q - have no BEKK representation).

Figure 8.1 shows a sketch of the space the $n^2 \times n^2$ dimensional symmetric matrices Q live in. Let the gray cone contain all positive definite and positive semidefinite matrices Q. Thus, the cone contains all matrices Q corresponding to VECH-terms A, which have an equivalent BEKK-term representation. Of course the positive semidefinite matrices are all on the boundary of the cone. Due to their affine structure, the equivalence classes appear as straight lines, and it is obvious that $\mathcal{Q}(\hat{A}_1)$ and $\mathcal{Q}(A_2)$ run parallel, for $A_1 \neq A_2$. Furthermore, the indefinite matrices Δ_k imply that the equivalence classes intersect the cone in a way as it is shown in the sketch. To see this, let for instance just one ω_k vary and keep the others constant, then the corresponding $Q(\omega)$ becomes indefinite, if $|\omega_k|$ is sufficiently large. This however implies that in any equivalence class $\mathcal{Q}(A)$ one can find matrices $Q \geq 0$ for which $\text{rk}(Q) = K < n^2$ holds. In figure 8.2 the eigenvalues of Q-matrices in $\mathcal{Q}(A)$ with

$$A = \begin{pmatrix} 0.662 & 0.682 & 0.333 \\ -0.074 & -0.065 & -0.045 \\ 0.561 & 0.629 & 0.324 \end{pmatrix} \tag{8.1}$$

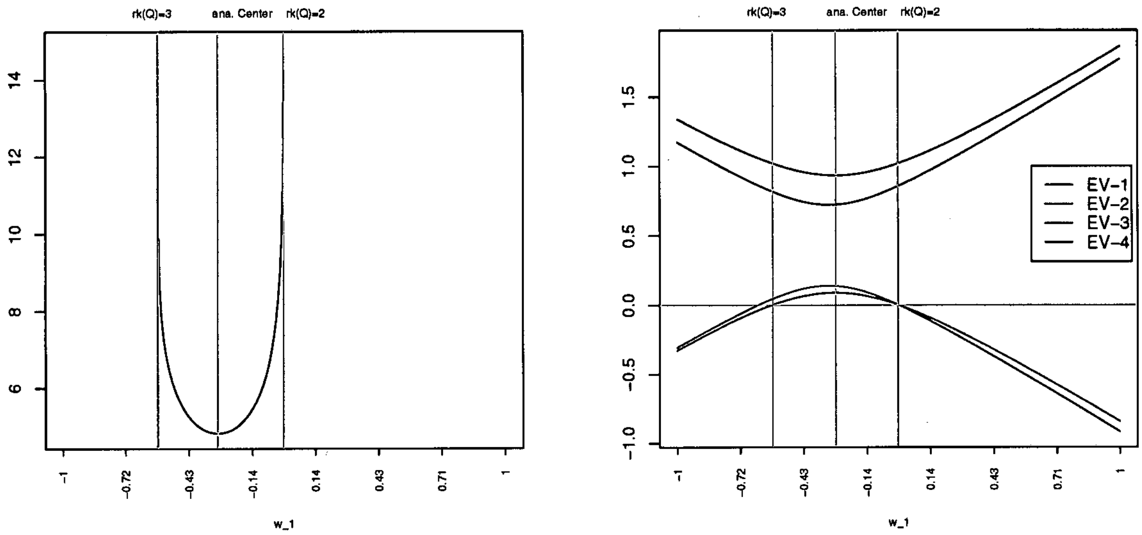

Figure 8.2: Given matrix A from eq. (8.1) and the corresponding $\mathcal{Q}$-set; Left: barrier function $\phi(\omega_1)$, see eq. (8.2); Right: the 4 eigenvalues of Q matrices in $\mathcal{Q}(A)$; Gray lines: the analytic center and the p.s.d. rank deficient matrices at the boundary of the conus.

are given as an example. Hence, for a single BEKK-term, K is not necessarily an indicator for generality. Due to the fact that any rank deficient matrix may be approximated arbitrarily well by a full rank matrix of the same size, setting $K = n^2$ will be sufficient in order to obtain a "fully general" parametrization of the BEKK model class. In fact $K = n^2 - 1$ would be sufficient as well, since one can always find rank deficient matrices in $\mathcal{Q}(A)$ of an admissible A that are positive semidefinite. However, choosing $K \leq n^2-1$ involves identifiability problems. In other words, it is not so easy to select a unique representative $Q \geq 0$ out of $\mathcal{Q}(A)$, for which $\text{rk}(Q) = K \leq n^2 - 1$ holds.

A parametrization of a BEKK-term should now be constructed such that from every equivalence class intersecting the cone a positive semidefinite matrix, $\bar{Q}$ say, is selected uniquely. The parameter matrices $B_1, \ldots, B_K$ of the corresponding BEKK-term are then given by a uniquely defined factorization of this $\bar{Q} \geq 0$.

Let us at first restrict ourselves to the generic set $\mathcal{B}^+$, i.e. to Q-sets that contain at least one positive definite element $Q > 0$. In particular, $\bar{Q}$ will be chosen to be positive definite, which implies that $K = n^2$.

8.1 Parametrizations with $K = n^2$

8.1.1 BEKK parametrization of Scherrer and Ribarits

In this parametrization, see also (Scherrer and Ribarits, 2006), the unique positive definite $\bar{Q}$ of an equivalence class $\mathcal{Q}(A)$ with $A \in \mathcal{B}^+$ is chosen such that it represents the so-called "analytic center" of all positive definite elements in $\mathcal{Q}(A)$.

Let a Q-set $\mathcal{Q}(A) = \{Q(\omega) = Q_0 + \sum_{k=1}^{\bar{n}} \omega_k \Delta_k\}$ be given, where $\mathcal{Q}^+(A) = \mathcal{Q}(A) \cap \mathcal{S}^+_{n^2,n^2} = \{Q \in \mathcal{Q}(A) \,|\, Q > 0\}$ is non void. We then define a function $\phi(\cdot)$ on $\mathbb{R}^{\bar{n}}$ by

$$\phi(\omega) = \begin{cases} \log\det Q(\omega)^{-1} & \text{if } Q(\omega) > 0 \\ +\infty & \text{otherwise} \end{cases} \tag{8.2}$$

Note that the set $\mathcal{W} = \{\omega \,|\, Q(\omega) \geq 0\}$ is a compact, convex subset of $\mathbb{R}^{\bar{n}}$, and that, by assumption, the interior of this set, i.e. the set $\mathcal{W}^+ = \{\omega \,|\, Q(\omega) > 0\}$ is non void. The function $\phi(\cdot)$, when restricted to $\mathcal{W}^+$, is convex and analytic and thus has a unique minimizer, $\bar{\omega}$ say. The corresponding $\bar{Q} = Q(\bar{\omega}) = Q_0 + \sum_k \bar{\omega}_k \Delta_k$ is called the *analytic center* of $\mathcal{Q}^+(A)$, see (Vandenberghe and Boyd, 1996) and figure 8.2 as an example. The gradient, $\nabla\phi(\omega)$, and the Hessian, $\nabla^2\phi(\omega)$, of ϕ for $\omega \in \mathcal{W}^+$ are given by

$$\begin{aligned} (\nabla\phi(\omega))_k &= -\operatorname{tr}(Q(\omega)^{-1}\Delta_k) \\ (\nabla^2\phi(\omega))_{kl} &= \operatorname{tr}(Q(\omega)^{-1}\Delta_k Q(\omega)^{-1}\Delta_l). \end{aligned} \tag{8.3}$$

Hence, in the analytic center, $\bar{Q} = Q(\bar{\omega})$, we have $\operatorname{tr}(\bar{Q}^{-1}\Delta_k) = 0$ for all $k = 1, \ldots, \bar{n}$. This is guaranteed if and only if each sub-block matrix of $\bar{Q}^{-1}$ is symmetric, see Appendix C.

Next, let $\bar{Q}^{-1} = VV'$ be the cholesky decomposition of $\bar{Q}^{-1}$. Thus, $V \in \mathbb{R}^{n^2 \times n^2}$ is a lower triangular matrix, where without loss of generality all diagonal elements are assumed to be positive. The parameter vector θ is defined as

$$\begin{aligned} \theta' = {} & (\operatorname{diag}(V_{11})', \ldots, \operatorname{diag}(V_{nn})', \underline{\operatorname{vech}}(V_{11})', \ldots, \underline{\operatorname{vech}}(V_{nn})', \\ & \operatorname{vech}(V_{21})', \ldots, \operatorname{vech}(V_{n1}), \operatorname{vech}(V_{32}), \ldots, \operatorname{vech}(V_{n(n-1)})'), \end{aligned} \tag{8.4}$$

i.e. we stack the lower triangular entries of all lower triangular blocks of V. Note that we put all the diagonal elements of V to the first n^2 positions. Thus,

$$\theta \in \Theta^+ := \mathbb{R}_+^{n^2} \times \mathbb{R}^{(n(n+1)/2)^2 - n^2} \subseteq \mathbb{R}^{(n(n+1)/2)^2}. \tag{8.5}$$

Due to the block symmetric structure of $\bar{Q}^{-1}$ these entries are sufficient to reconstruct $\bar{Q}^{-1}$.

Let us summarize the above procedure:

- *Given a VECH-term, $A \in \mathcal{B}^+$ say, compute θ and a unique set of corresponding BEKK-term parameters $B_1, \ldots, B_{n^2}$.*
  First, compute the analytic center $\bar{Q}$ of $\mathcal{Q}^+(A)$ as described above. Next, compute a cholesky factor V of $\bar{Q}^{-1}$ with strictly positive diagonal elements. The parameter vector θ is given by (8.4) and the BEKK parameter matrices are given by $B_k = \operatorname{mat}(U_k)$, where U_k is the k-th column of $U = V^{-T}$.

- *Given a parameter vector $\theta \in \Theta^+$, compute the unique parameter matrices A and $B_1, \ldots, B_{n^2}$ of the VECH-term and the corresponding BEKK-term.*
  First, note that the matrix V is a lower triangular matrix, i.e. $v_{ij,kl} = 0$ for $i < j$ and all k, l and for $i = j$ and $k < l$. In addition, the lower triangular entries $v_{ij,kl}$, $i \geq j$ and $k \geq l$ are stored in the parameter vector θ, see (8.4). Therefore, one has to construct the upper triangular elements $v_{ij,kl}$, $k < l$ for $i > j$ from the condition that the product $P = \bar{Q}^{-1} = VV'$ has symmetric sub-blocks $P_{ij} = P_{ij}'$. For $i \geq j$ such a sub-block is given by $P_{ij} = V_{i1}V_{j1}' + \cdots + V_{ij}V_{jj}'$ and since V_{jj} is a lower triangular matrix the (k, l)-th entry of such a sub-block is of the form

  $$p_{ij,kl} = \underbrace{\sum_{r=1}^{j-1}\sum_{s=1}^{n} v_{ir,ks}v_{jr,ls} + \sum_{s=1}^{l-1} v_{ij,ks}v_{jj,ls}}_{:=\bar{p}_{ij,kl}} + v_{ij,kl}v_{jj,ll}.$$

  From the symmetry condition $p_{ij,kl} = p_{ij,lk}$ and by the constraint $v_{jj,ll} > 0$ we get

  $$v_{ij,kl} = \frac{p_{ij,lk} - \bar{p}_{ij,kl}}{v_{jj,ll}} \tag{8.6}$$

  For $i > j$ and $k > l$ this expression depends only on V_{ir}, V_{jr} for $r < j$ and on $v_{ij,kr}$, $v_{ij,lr}$ for $r < l$. Thus, V may be recursively reconstructed from θ. The lower diagonal blocks of V are reconstructed in the sequence $V_{21}, \ldots, V_{n1}, V_{32}, \ldots, V_{n,n-1}$ and in each block V_{ij} the upper diagonal elements are computed in the sequence $v_{ij,12}, \ldots, v_{ij,1n}, v_{ij,23}, \ldots, v_{ij,(n-1),n}$. Finally, we get $\bar{Q} = UU'$ from $U = V^{-T}$, the BEKK parameter matrices from $B_k = \operatorname{mat}(U_k)$ and the VECH matrix A from the relations (7.7).

As an illustration consider the structure of V for the simplest case $n = 2$:

$$V = \left(\begin{array}{cc|cc} \theta_1 & 0 & 0 & 0 \\ \theta_5 & \theta_2 & 0 & 0 \\ \hline \theta_7 & \frac{\theta_1\theta_8-\theta_5\theta_7}{\theta_2} & \theta_3 & 0 \\ \theta_8 & \theta_9 & \theta_6 & \theta_4 \end{array}\right) \tag{8.7}$$

In addition, figure 8.3 shows the parametrization in the cone. Note that the parametrization by construction gets arbitrarily close to the intersection of the cone and the respective equivalence classes on the boundary. Hence the parametrization more or less cuts the cone in half as it gets close to the boundary. Inside the cone however it may have some fancy shape, since the set of analytic centers is not convex.

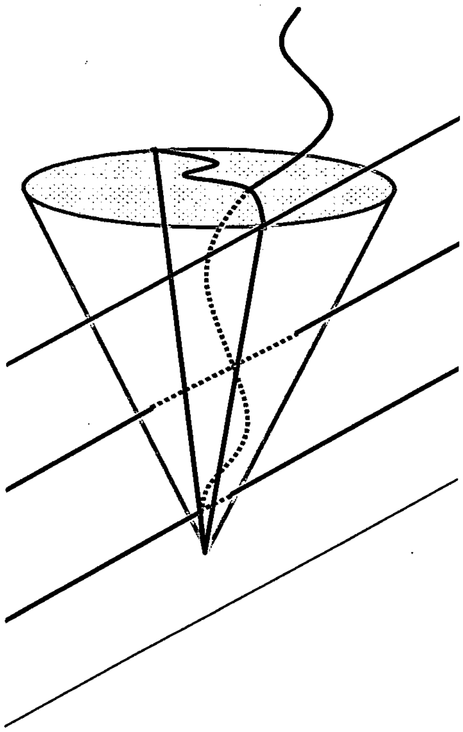

Figure 8.3: Gray: Cone containing all p.d. and p.s.d. Q matrices; Black lines: Equivalence classes $\mathcal{Q}(A)$ (thick lines: contain at least one p.s.d matrix Q and therefore have a BEKK representation, thin lines: contain no p.s.d. matrix Q - have no BEKK representation); Red: Parametrization using the analytic center.

Theorem 8.1 (Parametrization of $\mathcal{B}^+$)

1. *The above defined mapping*

$$\begin{array}{rccc} \pi^+ : & \Theta^+ & \longrightarrow & \mathcal{B}^+ \\ & \theta & \longmapsto & A \end{array}$$

is a diffeomorphism, i.e. π^+ is bijective and π^+ as well as its inverse are smooth, thus infinitely differentiable.

2. *The above defined mapping*

$$\begin{array}{rccc} \pi_B^+ : & \Theta^+ & \longrightarrow & \mathbb{R}^{n^4} \\ & \theta & \longmapsto & \mathrm{vec}(B_1, \ldots, B_{n^2}) \end{array}$$

is differentiable and injective.

Proof. By what has been said before, in particular by the restriction $\theta_i > 0$ for all indices corresponding to the diagonal elements of V, it is immediate to see that both mappings are injective and differentiable. In addition, $\pi^+(\cdot)$ is surjective by construction.

It remains to prove that the inverse mapping $(\pi^+)^{-1}(\cdot)$ is differentiable. For this end, we prove that the derivative of $\pi^+(\cdot)$ has full rank for all $\theta \in \Theta^+$. Note that $\pi^+(\cdot)$ is a concatenation of the following mappings:

$$\theta \mapsto V \mapsto P = VV' \mapsto \bar{Q} = P^{-1} \mapsto A$$

Let ∂V, ∂P, $\partial \bar{Q}$ and ∂A denote the respective derivatives of V, P, $\bar{Q}$ and A along a direction $\partial\theta \in \mathbb{R}^{(n(n+1)/2)^2}$. We have to prove that $\partial A = 0$ implies $\partial\theta = 0$. First, note that $\partial A = 0$ holds if and only if $\partial\bar{Q} = \partial\bar{Q}' = \sum_{k=1}^{\bar{n}} \Delta_k\omega_k$ for some $\omega_k \in \mathbb{R}$. Since $\partial\bar{Q} = -P^{-1}\partial P P^{-1}$ we get $\partial P = -\sum_k P\Delta_k P\omega_k$. The matrix P has symmetric sub-blocks by construction and thus the same holds for the derivative ∂P, i.e. we have $\operatorname{tr}\Delta_l\partial P = 0$ for $l = 1, \ldots, \bar{n}$. This implies $0 = \sum_k \operatorname{tr}(\Delta_l P\Delta_k P)\omega_k$ for all $l = 1, \ldots, \bar{n}$ and thus $0 = \operatorname{tr}(\partial\bar{Q}P\partial\bar{Q}P) = \operatorname{tr}(P^{T/2}\partial\bar{Q}'P\partial\bar{Q}P^{1/2})$. Since $P > 0$ one obtains $\partial\bar{Q} = 0$ and $\partial P = 0$. Next, we use $v_{jj,ll} > 0$ to show that $\partial V = 0$ as $0 = \partial P = V\partial V' + \partial V V'$. Now, since θ is composed from the entries of V we have $\partial\theta = 0$ as desired. □

It should be emphasized that this parametrization covers a generic set of BEKK models, i.e. the corresponding set is an open and dense subset of all BEKK models. In this section we considered only the case of one term, i.e. the case of a BEKK$(0, 1, K)$ model, but the generalization to the case of BEKK(p, q, K) models is straightforward.

8.1.2 BEKK parametrization of Engle and Kroner

In (Engle and Kroner, 1995, Proposition 2.3) an alternative parametrization for BEKK models is presented. The authors claim that this parametrization is *fully general* and *identifiable*. Here fully general means that the parametrization covers as many VECH models as possible and identifiable means that two different parameter vectors do not represent the same BEKK model. However, both statements are not correct as we will show in this section.

In their representation $K = n^2$ and the parameter matrices B_k are chosen such that the corresponding matrix $U = [\operatorname{vec}(B_1), \ldots, \operatorname{vec}(B_{n^2})]$ is a lower block triangular matrix where all $(n \times n)$ sub-blocks are again lower triangular, i.e. U is a cholesky factor of $Q = UU'$ with the additional zero restrictions given by $u_{ij,kl} = 0$ for $i > j$ and $k < l$. By these additional zero restrictions implicitly a unique representative, $\hat{Q}$ say, in the Q-set $\mathcal{Q}(A)$ is chosen. In addition, the entries of the last row of U are assumed to be positive, $u_{nj,nl} > 0$ for all j, l in order to get rid of the non uniqueness of the cholesky factorization of $\hat{Q}$.

Using the relations between VECH models, BEKK models and Q-sets respectively it is easy to see that two problems occur with this parametrization:

1. The positivity constraints do not guarantee uniqueness in all cases.
2. There is a "thick" set of BEKK models, which is not covered by this parametrization, since the corresponding Q-sets do not contain a positive semidefinite representative $\hat{Q}$ of this structure.

To see this consider the case $n = 2$. The Engle-Kroner parametrization leads to a U-matrix of the form

$$U = \left(\begin{array}{cc|cc} \theta_1 & 0 & 0 & 0 \\ \theta_2 & \theta_3 & 0 & 0 \\ \hline \theta_4 & 0 & \theta_7 & 0 \\ \theta_5 & \theta_6 & \theta_8 & \theta_9 \end{array}\right)$$

where $\theta_5 > 0$, $\theta_6 > 0$, $\theta_8 > 0$, $\theta_9 > 0$ and thus, $\hat{Q}$ is given by

$$\hat{Q} = UU' = \left(\begin{array}{cc|cc} \theta_1^2 & \theta_1\theta_2 & \theta_1\theta_4 & \theta_1\theta_5 \\ \theta_1\theta_2 & \theta_2^2+\theta_3^2 & \theta_2\theta_4 & \theta_2\theta_5+\theta_3\theta_6 \\ \hline \theta_1\theta_4 & \theta_2\theta_4 & \theta_4^2+\theta_7^2 & \theta_4\theta_5+\theta_7\theta_8 \\ \theta_1\theta_5 & \theta_2\theta_5+\theta_3\theta_6 & \theta_4\theta_5+\theta_7\theta_8 & \theta_5^2+\theta_6^2+\theta_8^2+\theta_9^2 \end{array}\right).$$

Now, consider a BEKK-term – in Engle-Kroner form – given by

$$U = \left(\begin{array}{cc|cc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \hline 1 & 0 & 0 & 0 \\ 1 & 1 & \theta_8 & \theta_9 \end{array}\right) \text{ and } \hat{Q} = UU' = \left(\begin{array}{cc|cc} 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ \hline 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 2+\theta_8^2+\theta_9^2 \end{array}\right)$$

In this case we lose identifiability, since $\theta_7 = 0$ and thus only the sum $\theta_8^2 + \theta_9^2$ is identified from Q. (In other words, the positivity constraint, in this case, is not sufficient to select a unique cholesky factor.) However, this problem may be circumvented by replacing the restrictions $u_{nj,nl} > 0$ by the restriction that the diagonal elements of U are positive, i.e. $u_{jj,ll} > 0$ for all j, l.

Next, consider a BEKK-term of the form:

$$U = \left(\begin{array}{cc|cc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \hline 0 & x & 1 & 0 \\ 1 & 1 & 1 & 1 \end{array}\right) \text{ and } Q = UU' = \left(\begin{array}{cc|cc} 1 & 0 & 0 & 1 \\ 0 & 1 & x & 1 \\ \hline 0 & x & 1+x^2 & 1+x \\ 1 & 1 & 1+x & 4 \end{array}\right)$$

The Engle-Kroner parameters θ_k and thus $\hat{Q} \in \mathcal{Q}$ may be computed from Q as follows:

$$\begin{aligned}
\theta_1 &= \sqrt{q_{11,11}} &&= 1 \\
\theta_2 &= q_{11,21}/\theta_1 &&= 0 \\
\theta_3 &= \sqrt{q_{11,22} - \theta_2^2} &&= 1 \\
\theta_4 &= q_{21,11}/\theta_1 &&= 0 \\
\theta_5 &= (q_{21,21} + q_{21,12} - \theta_2\theta_4)/\theta_1 &&= 1 + x \\
\theta_6 &= (q_{21,22} - \theta_2\theta_5)/\theta_3 &&= 1 \\
\theta_7 &= \sqrt{q_{22,11} - \theta_4^2} &&= \sqrt{1+x^2} \\
\theta_8 &= (q_{22,21} - \theta_4\theta_5)/\theta_7 &&= (1+x)/\sqrt{1+x^2} \\
\theta_9 &= \sqrt{q_{22,22} - \theta_5^2 - \theta_6^2 - \theta_8^2} &&= \sqrt{4 - (1+x)^2 - 1 - (1+x)^2/(1+x^2)}
\end{aligned}$$

Note that θ_5 is computed from the sum $q_{21,21} + q_{21,12}$ since only the sum is uniquely defined. However, this procedure breaks down for "large" x, since $4 - (1+x)^2 - 1 - (1+x)^2/(1+x^2) < 0$ for sufficiently large x. In this case the BEKK model defined as above has no representation in the form suggested by Engle and Kroner since the set $\mathcal{Q}(A)$ does not contain a positive semidefinite element $\hat{Q} \geq 0$ conforming to the restrictions imposed by the Engle-Kroner parametrization. It is clear that this problem occurs on a "thick" set of BEKK terms, i.e. on a set that contains an open subset of all BEKK-terms. In this sense, using the Engle-Kroner parametrization we lose a non negligible set of BEKK models.

The above examples were given for the case $n = 2$, however, the reasoning may be easily extended to the case $n > 2$.

8.1.3 Parametrization of the positive definite orthocomplement of the $\mathcal{Q}$-sets

In section 7.1 we have already mentioned that additional normalizing restrictions like symmetric sub-blocks would be sufficient in order to select a unique representative out of each equivalence class $\mathcal{Q}(A)$, denoted by $\check{Q}$, see also eq. (7.9). If we choose this selection procedure, we are in fact parametrizing the set $\mathcal{S}_{n*n}$. Note that $\mathcal{S}_{n*n}$ is the orthocomplement of $\mathcal{Q}(A)$ at the point $\check{Q} \in \mathcal{Q}(A)$. To see this, note that for any $Q \in \mathcal{Q}(A)$ it holds that $(Q - \check{Q}) \in \mathcal{A}_{n*n} \Rightarrow \text{vech}(Q - \check{Q})'\text{vech}(S) = 0$ for all $S \in \mathcal{S}_{n*n}$. As an illustration, consider the following example and let again for the sake of simplicity $n = 2$. Now, the $\mathcal{Q}$-set is given as $\mathcal{Q}(A) = \{Q(\omega_1) = \check{Q} + \omega_1\Delta_1\}$. In order to construct the orthocomplement we have to find the directions that are orthogonal to the directions that span the equivalence class. That is we have to compute $\text{vech}(\Delta_1)^\perp$. Since $\text{vech}(\Delta_1)' = (0, 0, 0, 1, 0, -1, 0, 0, 0, 0)$, the orthocomplement is spanned by the columns of

$$\text{vech}(\Delta_1)^\perp = \begin{pmatrix} 1&0&0&0&0&0&0&0&0\\ 0&1&0&0&0&0&0&0&0\\ 0&0&1&0&0&0&0&0&0\\ 0&0&0&1&0&0&0&0&0\\ 0&0&0&0&1&0&0&0&0\\ 0&0&0&1&0&0&0&0&0\\ 0&0&0&0&0&1&0&0&0\\ 0&0&0&0&0&0&1&0&0\\ 0&0&0&0&0&0&0&1&0\\ 0&0&0&0&0&0&0&0&1 \end{pmatrix} \text{ and } \text{math}(\text{vech}(\Delta_1)^\perp\varphi) = \left(\begin{array}{cc|cc} \varphi_1 & \varphi_2 & \varphi_3 & \varphi_4\\ \varphi_2 & \varphi_5 & \varphi_4 & \varphi_6\\ \hline \varphi_3 & \varphi_4 & \varphi_7 & \varphi_8\\ \varphi_4 & \varphi_6 & \varphi_8 & \varphi_9 \end{array}\right) \in \mathcal{S}_{n*n},$$

where $\varphi = (\varphi_1, \ldots, \varphi_{n(n+1)/2})' \in \mathbb{R}^{n(n+1)/2}$ is the vector of scalars that determine how far we walk along a certain direction within the orthocomplement.

Furthermore note that due to the affine structure of the equivalence classes, $\mathcal{S}_{n*n}$ is in fact an orthocomplement to all equivalence classes.

Let us just in a view words motivate this procedure: In fact, an equivalence class is a set of parameters that all correspond to the same model and, therefore, to the same likelihood value. Thus, we are given a parameter space with intrinsic non-identifiability. Given a specific parameter value, the idea now is not to parametrize directions within the equivalence class, because these are exactly the directions where the described model and, therefore, the likelihood function, does not change. Only directions in the orthogonal complement to the equivalence class are parametrized. In this way we get rid of as many parameters as is the dimension of the equivalence class, here this is $\bar{n}$.

It is however obvious that not every matrix in $\mathcal{S}_{n*n}$ is positive semidefinite. Thus, it is necessary to restrict ourselves to the part of the orthocomplement that is at least positive semidefinite in order to obtain a parametrization for terms of a BEKK model. Here however we will even ask for positive definite matrices $\check{Q}$. In section 8.1.1 the cholesky factor of some $\check{Q} > 0$ has already been presented, see eq. (8.4) to (8.6), and the example in (8.7). Hence it is easy to construct positive definite matrices $\check{Q}$. Anyway, this parametrization is not "fully general", since there is a thick set of corresponding BEKK terms that cannot be parametrized via $\check{Q}$. In other words there exist equivalence classes $\mathcal{Q}(A)$ that have a non void intersection with the cone, but whose representatives $\check{Q}(A)$ lie outside the cone. Consider for instance table 8.1 and note that $\check{Q}$ stays indefinite for small changes in the parameter vector θ. Figure 8.4 shows the implication of this fact on the position of the plane parametrized by the positive semidefinite orthocomplement of the equivalence classes within the cone. (Imagine now that you look at the cone from somewhere above.) Note furthermore that the position of this plane determines the cuttinge angle of the equivalence classes and the cone.

	Eigenvalues			
$\bar{Q}$	1.603	0.392	0.062	0.036
$\check{Q}$	1.408	0.564	0.310	−0.188
A	0.968	$-0.633 \pm 0.028i$		

Table 8.1: Eigenvalues of the analytic center $\bar{Q}$ constructed from $\theta = (4.85, 2.14, 1.46, 1.75, 0.29$,-0.28,-0.72, 1.28,-1.46)' as shown in section 8.1.1, the VECH-term A and the corresponding $\check{Q}$.

We have seen that the parametrization proposed by (Engle and Kroner, 1995) and the parametrization

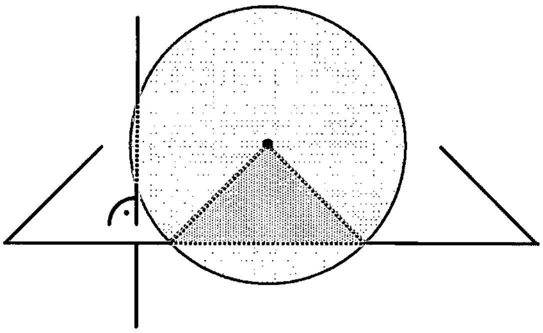

Figure 8.4: Gray: Cone containing all p.d. and p.s.d. Q matrices; Black: Equivalence class $\mathcal{Q}(A)$; Blue: Orthocomplement of the equivalence classes; Shaded blue plane: P.s.d. matrices within the orthocomplement.

of the positive definite orthocomplement of the equivalence classes need the same amount of parameters as the first parametrization but are less general. In that sense these parametrizations are sub-optimal. However, as we will see later in the applications the latter two have nicer numerical properties.

8.2 Parametrizations with $K \leq n$

The main advantages of BEKK models over VECH models is that BEKK models are admissible by construction. However, the number of free parameters to estimate for a general BEKK$(p, q, K = n^2)$ model is equal to the number of parameters for a VECH(p, q) model, which is of the order $O(n^4)$. For small vector dimensions, $n \leq 3$, estimation of a BEKK(p, q, n^2) model might still be reasonable. In case of higher dimensional vector processes however, it becomes inevitable to impose further structure and restrictions on the parameters in order to make estimation of multivariate GARCH models applicable in practice. There have been many suggestions to further restrict this model class. One natural way in doing so is to use a small number K to reduce the number of parameters.
Here, we will consider the case of a fixed $K \leq n$. In particular, we want to analyze the question, if this restriction is sufficient to guarantee identifiability of the BEKK parameter matrices B_k. As has been noted already, the non-uniqueness of the BEKK parameter matrices stems from the non-uniqueness of the Q-matrices and second from the non-uniqueness of the factorization $Q = UU'$. Here, we will mainly deal with the first problem; i.e. we want to answer the question if the additional constraint $\text{rk}(Q) = K$ is sufficient to uniquely select a positive semidefinite element $Q \in \mathcal{Q}(A)$.

Let us define the following sets: $\mathcal{B}_K \subseteq \mathcal{B}$ denotes the set of BEKK-terms for which the corresponding Q-set contains at least one positive semidefinite element Q of rank K. That means $\mathcal{B}_K$ is the set of all A such that $Q(A) \cap \mathcal{S}^+_{n^2,K}$ is non void. For the set $\mathcal{B}^+_K \subseteq \mathcal{B}_K$ we demand in addition the existence of a $Q \in Q(A) \cap \mathcal{S}^+_{n^2,K}$ such that the left upper $K \times K$ sub-block of Q has full rank K. The next proposition gives some basic topological properties of these sets:

Proposition 8.2 (Topological properties of the sets $\mathcal{B}^+_k$.)

1. *The set $\mathcal{B}^+_K$ is open and dense in $\mathcal{B}_K$.*

2. *$\overline{\mathcal{B}^+_K} = \mathcal{B}_0 \cup \mathcal{B}_1 \cup \cdots \cup \mathcal{B}_K$, where $\overline{\mathcal{B}^+_K}$ is the closure of $\mathcal{B}^+_K$ in $\mathbb{R}^{n(n+1)/2 \times n(n+1)/2}$.*

3. *For $1 \leq K_1 < K_2 \leq n$ it holds $\mathcal{B}^+_{K_1} \cap \mathcal{B}^+_{K_2} = \emptyset$.*

Proof. The proof of these properties is straightforward. □

8.2.1 Parametrization of $\mathcal{B}_K^+$ for $K \leq n$.

Let an element of $\mathcal{B}_K^+$ be given, where $Q \geq 0$ is an element of the corresponding Q-set with the desired properties; i.e. $Q \in \mathcal{Q}(A) \cap \mathcal{S}^+_{n^2,K}$ and the $K \times K$ left upper sub-block of Q has full rank. Thus, Q has a factorization as $Q = UU'$, where $U \in \mathbb{R}^{n^2 \times K}$ and where the first K rows form a lower triangular matrix with strictly positive diagonal elements. In the following U is partitioned as $U = (U_1', \ldots, U_n')'$ and U_i in turn is partitioned as $U_i = (U_{i1}', U_{i2}')'$ with $U_{i1} \in \mathbb{R}^{K \times K}$ and $U_{i2} \in \mathbb{R}^{(n-K) \times K}$. The corresponding parameter vector θ is defined as:

$$\theta = (\mathrm{diag}(U_{11})', \underline{\mathrm{vech}}(U_{11})', \mathrm{vec}(U_{12})', \mathrm{vec}(U_2)', \ldots, \mathrm{vec}(U_n)')'.$$

Note that for notational convenience the positive diagonal elements have been put to the first K positions and hence θ is an element of

$$\Theta_K^+ := \mathbb{R}_+^K \times \mathbb{R}^{(2n^2-K-1)K/2} \subseteq \mathbb{R}^{(2n^2-K+1)K/2}.$$

For the reverse direction, U is constructed from θ in an obvious way; the parameter matrices of the respective BEKK-term are computed from the columns of U and A stems from the relations (7.7), where $Q = UU'$.

Theorem 8.3 *The above defined mapping*

$$\begin{array}{cccc} \pi_K^+ : & \Theta_K^+ & \longrightarrow & \mathcal{B}_K^+ \\ & \theta & \longmapsto & A \end{array}$$

is differentiable and surjective. There is a generic subset $\Theta_K^{+g} \subset \Theta_K^+$ with the following properties:

1. *$\theta \in \Theta_K^{+g}$ is identified from A in the sense that $A = \pi_K^+(\theta) = \pi_K^+(\bar\theta)$ implies $\bar\theta = \theta$ for all $\theta \in \Theta_K^{+g}$.*
2. *The derivative of π_K^+ has full rank for all $\theta \in \Theta_K^{+g}$*

Here generic means that the complement $\Theta_K^+ \setminus \Theta_K^{+g}$ has Lebesque measure zero.

Proof. Let $\theta, \bar\theta \in \Theta_K^+$ be given and let $U, \bar U \in \mathbb{R}^{n^2 \times K}$ be constructed from θ and $\bar\theta$ as described above. Furthermore, let $H = \bar U - U$. Now, θ and $\bar\theta$ are mapped to the same VECH-term A if and only if the corresponding matrices $Q = UU'$ and $\bar Q = \bar U \bar U'$ only differ by a matrix $\Delta \in \mathcal{A}_{n*n}$. Hence, $\pi_K^+(\theta) = \pi_K^+(\bar\theta)$ holds, if and only if

$$(U + H)(U + H)' - UU' = UH' + HU' + HH' = \Delta \tag{8.8}$$

holds for some matrix $\Delta \in \mathcal{A}_{n*n}$. We will use the above described partitioning for U and H and a corresponding partitioning for Δ. The diagonal blocks of (8.8) are of the form:

$$\begin{pmatrix} U_{i1}H_{i1}' + H_{i1}U_{i1}' + H_{i1}H_{i1}' & U_{i1}H_{i2}' + H_{i1}U_{i2}' + H_{i1}H_{i2}' \\ U_{i2}H_{i1}' + H_{i2}U_{i1}' + H_{i2}H_{i1}' & U_{i2}H_{i2}' + H_{i2}U_{i2}' + H_{i2}H_{i2}' \end{pmatrix} = \Delta_{ii} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \tag{8.9}$$

For $i = 1$ the uniqueness of the Cholesky factorization implies $H_{11} = 0$. For the non diagonal blocks $i \neq j$ we get

$$\begin{pmatrix} U_{i1}H_{j1}' + H_{i1}U_{j1}' + H_{i1}H_{j1}' & U_{i1}H_{j2}' + H_{i1}U_{j2}' + H_{i1}H_{j2}' \\ U_{i2}H_{j1}' + H_{i2}U_{j1}' + H_{i2}H_{j1}' & U_{i2}H_{j2}' + H_{i2}U_{j2}' + H_{i2}H_{j2}' \end{pmatrix} = \Delta_{ij} = \begin{pmatrix} \Delta_{ij,11} & \Delta_{ij,12} \\ \Delta_{ij,21} & \Delta_{ij,22} \end{pmatrix}. \tag{8.10}$$

From the $(1,1)$ block of the above matrix equation for $i = 1$ one obtains $U_{11}H'_{j1} = \Delta_{1j,11}$ and thus

$$D_j + D'_j = 0$$

for $D_j = U_{11}^{-1}H_{j1} = U_{11}^{-1}\Delta'_{1j,11}U_{11}^{-T}$. Next, consider the $(1,1)$ block of (8.9) for $i > 1$, which delivers

$$M_i D'_i + D_i M'_i + D_i D'_i = 0$$

for $M_i = U_{11}^{-1}U_{i1}$. Therefore, Lemma B.5 implies $H_{i1} = 0$ for a generic subset of parameters $\theta \in \Theta_K^+$. Next, observe that by the $(1,2)$ blocks of (8.9) $H_{i2} = 0$ follows from $H_{i1} = 0$ if U_{i1} is a full rank matrix, which also is satisfied on a generic set of points in Θ_K^+. Of course $H = 0$ is equivalent to $\bar{\theta} = \theta$ and thus we have shown the first claim.

It is easy to see that the mapping π_K^+ has a rank deficient derivative in θ if and only if there exist (non zero) matrices H, Δ of the above described structure such that

$$UH' + HU' = \Delta \tag{8.11}$$

holds. A completely analogous reasoning shows that the above relations (8.11) imply $H = 0$ and $\Delta = 0$ for a generic set of parameters $\theta \in \Theta_K^+$. □

Consider the following remarks concerning this result:

Remark 8.1 (*Parametrization for $\mathcal{B}_K^+$.*) The Theorem shows that for $K \leq n$ the BEKK parameter matrices B_k (given some suitable zero restrictions and positivity restrictions) may be used as a parametrization for a certain subclass of BEKK-terms, namely for the set $\mathcal{B}_K^+$.

Remark 8.2 (*Exceptional points.*) This parametrization "fails" on a "thin" set of exceptional points. On this set of exceptional points one may lose identifiability; i.e. two or more parameter vectors correspond to the same model. Or one may lose "full rank derivatives", which implies that on these exceptional point the Hessian of the likelihood function may become singular. This may cause troubles for numerical optimization routines. However, as has been noted already these problems are very "unlikely".

Remark 8.3 (*The case $K = 1$.*) For the case $K = 1$, none of the above problems occurs, see also (Engle and Kroner, 1995, Proposition 2.1)

Remark 8.4 (*Rank restriction on the first K rows and columns of Q.*) The restriction on BEKK-terms where the first K rows and columns of Q have full rank may be partly relaxed. E.g. one could define an analogous parametrization for terms where the last K rows and columns of the second diagonal block have full rank. However, matrices where the full rank rows and columns are spread over several sub blocks are not easy to deal with. Also the case $K > n$ is much more involved.

8.2.2 The Factor GARCH model of (Engle, Ng and Rothschild, 1990)

A factor GARCH (F-GARCH) model is a BEKK(p,q,K) models, see (7.12), where the following restrictions are imposed:

1. $K \leq n$

2. the BEKK matrices $B_{i,k}$ are rank one matrices:

$$B_{i,k} = \mu_{i,k}(\Gamma_k \Lambda'_k),\ \Gamma_k, \Lambda_k \in \mathbb{R}^n,\ \|\Gamma_k\| = 1,\ \mu_{i,k} > 0 \tag{8.12}$$

3. Let $\Gamma = (\Gamma_1, \ldots, \Gamma_K) \in \mathbb{R}^{n \times K}$ and $\Lambda = (\Lambda_1, \ldots, \Lambda_K) \in \mathbb{R}^{n \times K}$ then
$$\Gamma'\Lambda = \Lambda'\Gamma = I_K. \tag{8.13}$$

Hence the conditional covariance matrix H_t is given by

$$H_t = CC' + \sum_{i=1}^{q}\sum_{k=1}^{K} \mu_{i,k}^2 \Lambda_k \Gamma_k' \epsilon_{t-i}\epsilon_{t-i}' \Gamma_k \Lambda_k' + \sum_{j=1}^{p}\sum_{k=1}^{K} \mu_{q+j,k}^2 \Lambda_k \Gamma_k' H_{t-j} \Gamma_k \Lambda_k'.$$

Let us first consider the case $0 < K < n$. By (8.13) there exist two matrices $\bar{\Gamma}, \bar{\Lambda} \in \mathbb{R}^{n \times (n-K)}$ such that

$$(\Gamma, \bar{\Gamma})'(\Lambda, \bar{\Lambda}) = (\Lambda, \bar{\Lambda})(\Gamma, \bar{\Gamma})' = I_n.$$

Therefore (ϵ_t) has a factor model representation of the form:

$$\epsilon_t = \sum_{k=1}^{K} \Lambda_k f_{k,t} + v_t,$$

where $(f_{k,t} = \Gamma_k' \epsilon_t)$ are the factors and where $v_t = \bar{\Lambda}\bar{\Gamma}'\epsilon_t$ is the "noise". The following relations give the conditional variances and covariances respectively:

$$\begin{aligned}
\operatorname{var}(f_{k,t}|\mathcal{I}_{t-1}) = \Gamma_k' H_t \Gamma_k &= \Gamma_k' CC' \Gamma_k + \sum_{i=1}^{q} \mu_{i,k}^2 f_{k,t}^2 + \sum_{j=1}^{p} \mu_{q+j,k}^2 \operatorname{var}(f_{k,t-j}|\mathcal{I}_{t-1-j}) \\
\operatorname{cov}(f_{k,t}, f_{l,t}|\mathcal{I}_{t-1}) = \Gamma_k' H_t \Gamma_l &= \Gamma_k' CC' \Gamma_l = \operatorname{cov}(f_{k,t}, f_{l,t}) \text{ for } k \neq l \\
\operatorname{var}(v_t|\mathcal{I}_{t-1}) = \bar{\Lambda}\bar{\Gamma}' H_t \bar{\Gamma}\bar{\Lambda}' &= \bar{\Lambda}\bar{\Gamma}' CC' \bar{\Gamma}\bar{\Lambda}' = \operatorname{var}(v_t) \\
\operatorname{cov}(v_t, f_{k,t}|\mathcal{I}_{t-1}) = \bar{\Lambda}\bar{\Gamma}' H_t \Gamma_k &= \bar{\Lambda}\bar{\Gamma}' CC' \Gamma_k = \operatorname{cov}(v_t, f_{k,t})
\end{aligned}$$

Thus, the factors $(f_{k,t})$ are univariate GARCH processes. The conditional variance of the noise (v_t) as well the conditional covariances are constant. For $K = n$ there is no noise term (v_t) and hence ϵ_t is a linear combination of scalar GARCH processes only.

Note that the above model imposes strong cross restrictions on the parameters for the respective ARCH and GARCH terms. The number of parmeters reduces to $n(n+1)/2 + (p+q)K + 2nK - K - K^2 = O(n^2)$.

The stationarity condition (7.11) simplifies to the condition that the K univariate factor processes $(f_{k,t})$ are stationary:

$$\sum_{i=1}^{p+q} \mu_{i,k}^2 < 1 \; ; \quad \text{for } k = 1, \ldots, K$$

8.2.3 Parametrization for a subset of the positive semidefinite orthocomplement of the $\mathcal{Q}$-sets.

In section 8.1.3 a parametrization for the positive definite orthocomplement of the equivalence classes given by the $\mathcal{Q}$-sets was introduced. Here we will again consider the orthocomplement, but impose the following restrictions on $\check{Q} \in \mathcal{S}_{n*n}$, the block symmetric representative in $\mathcal{Q}(A)$:

1. $\check{Q} \geq 0$,
   i.e, we restrict ourselves to VECH-terms that have an equivalent BEKK-term representation,

2. $\operatorname{rk}(\check{Q}) = K \leq n$,
   i.e. we consider only those VECH-terms A whose corresponding $\check{Q} \in \mathcal{Q}(A)$ has rank $K \leq n$ in order to reduce the number of parameters, and

3. let $S, T \in \mathbb{R}^{n\times n}$ be some matrices of full rank such that the transformed matrix

$$(S \otimes T)\check{Q}(S \otimes T)'$$

has a full rank $K \times K$ left upper block,
i.e. we want to avoid degenerated cases.

Let $\mathcal{B}^s_K \subset \mathcal{B}$ denote the set of BEKK-terms, where the block symmetric representative $\check{Q} \in \mathcal{Q}(A) \cap \mathcal{S}_{n*n}$ is a positive semidefinite rank K matrix.

Theorem 8.4 *For $K \leq n$ the set $\mathcal{B}^s_K$ is the set of all BEKK-terms which have a representation of the form*

$$B_k = \mu_k \Gamma_k \Lambda_k', \; k = 1, \ldots, K$$

where $\|\Lambda_k\| = \|\Gamma_k\| = 1$ and $\mu_k > 0$. Here Λ and Γ are two $n \times K$ dimensional matrices, where Γ has full rank K and where Λ_k and Γ_k denote the respective k-th columns. This representation is unique up to sign changes and permutations if and only if $\Lambda_k \neq \Lambda_s$ holds for all $k \neq s$.

Proof. This follows immediately from Lemma B.4 in the Appendix. □

Note that as in case of the F-GARCH model, see (8.12), the parameter matrices $B_{i,k}$ are all of rank one. However, due to the fact that here the matrices are given as

$$B_{i,k} = \mu_{i,k}(\Gamma_{i,k}\Lambda_{i,k}'), \quad \text{with} \quad \Gamma_{i,k}, \Lambda_{i,k} \in \mathbb{R}^n, \; \|\Gamma_{i,k}\| = \|\Lambda_{i,k}\| = 1, \; \mu_{i,k} > 0, \tag{8.14}$$

(plus some technical conditions to get rid of trivial non identifiabilities, like sign changes and permutations), we allow for more generality as compared to F-GARCH. The additional flexibility of course is at the cost of additional parameters. The total amount of independent parameters is now $n(n+1)/2 + (p+q)(2(n-1)K + K)$. Nevertheless, this might be a valuable alternative in cases where the F-GARCH model is too restrictive. However, one has to admit that the stationarity condition here is much more involved than for the F-GARCH model. In addition, the "factor-interpretation" gets lost, since the parameters $\Gamma_{i,k}$, $\Lambda_{i,k}$ may now depend on the lag i and since the condition (8.13) is not imposed.

8.2.4 The DVECH model

Let us now consider briefly the oftentimes used *diagonal* VECH (DVECH) model proposed by (Bollerslev et al., 1988).

A VECH model is called DVECH model, if all parameter matrices A_i in (7.1) are diagonal matrices. Thus, the elements of H_t depend only on its own past and the respective element of $\epsilon_{t-i}\epsilon_{t-i}'$. The number of parameters reduces hence to $(p+q+1)n(n+1)/2 = O(n^2)$.
Let $\mathcal{V}_{diag} \subseteq \mathcal{V} \subseteq \mathbb{R}^{n(n+1)/2\times n(n+1)/2}$ denote the set of all admissible DVECH-terms A, and let $\mathcal{B}_{diag} \subseteq \mathcal{B} \subseteq \mathcal{V}$ denote the set of all DVECH-terms A, which have a BEKK representation.

Theorem 8.5 *$\mathcal{B}_{diag} = \mathcal{V}_{diag}$ holds for all n, and for the number of additive BEKK-terms it holds that $K \leq n$.*

Proof. $\mathcal{B}_{diag} \subseteq \mathcal{V}_{diag}$ follows by definition.
Consider $\mathcal{V}_{diag} \subseteq \mathcal{B}_{diag}$: The diagonal VECH-term A is admissible if and only if $f(\nu, e) = \nu'\text{math}(A\text{vech}(ee'))\nu = (\nu \otimes e)'Q(\nu \otimes e) \geq 0$ holds for all $\nu, e \in \mathbb{R}^n$. Due to the diagonality of A it follows that $f(\nu, e) = (\nu \odot e)'\Omega(\nu \odot e)$, where Ω is the $n \times n$-dimensional symmetric matrix

$\Omega := \text{math}(\text{diag}(A))$, and the operator $\odot$ denotes the Hadamarad product[1]. Thus, $f(\nu, e) \geq 0$ holds for all $\nu, e \in \mathbb{R}^n$ if and only if $\Omega \geq 0$. Let $S \in \mathbb{R}^{n \times n^2}$ be a selection matrix designed such that $S(\nu \otimes e) = (\nu \odot e)$ holds for all $\nu, e \in \mathbb{R}^n$, i.e. the ith diagonal element of the ith $n \times n$ dimensional sub-block of S equals one, $s_{1i,ii} = 1$ for $i = 1, \dots, n$, whereas the other elements are all zero. Then $f(\nu, e)$ can be written as the quadratic form $f(\nu, e) = (\nu \otimes e)' Q^* (\nu \otimes e)$, with $Q^* := S' \Omega S \in \mathcal{Q}(A)$. By construction it follows that $Q^* \geq 0$, if $\Omega \geq 0$. Since Q^* is also the only positive semidefinite Q-matrix in the whole $\mathcal{Q}$-set $\mathcal{Q}(A)$ of an admissible DVECH-term A, it follows that $\mathcal{B}_{diag} = \mathcal{V}_{diag}$. To see this just note that the diagonal blocks of Q^* have only one non zero diagonal element, since $q^*_{ii,kk} = 0$ for all $k \neq i$. Thus, $Q = Q^* + \Delta \geq 0$, $\Delta \in \mathcal{A}_{n*n}$ implies $\Delta = 0$.

It is furthermore easy to see that the rank of Q^*, K, equals the rank of Ω and is thus bounded from above by the vector dimension n. □

As an illustration, consider the simplest case $n = 2$:

$$A = \text{diag}(a_1, a_2, a_3), \quad \Omega = \begin{pmatrix} a_1 & a_2 \\ a_2 & a_3 \end{pmatrix}, \quad S = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

$$\mathcal{Q}(A) = Q(\omega_1) = \left(\begin{array}{cc|cc} a_1 & 0 & 0 & a_2/2 + \omega_1 \\ 0 & 0 & a_2/2 - \omega_1 & 0 \\ \hline 0 & a_2/2 - \omega_1 & 0 & 0 \\ a_2/2 + \omega_1 & 0 & 0 & a_3 \end{array}\right),$$

with $\omega_1 \in \mathbb{R}$, and

$$Q^* = Q(a_2/2) = S' \Omega S = \left(\begin{array}{cc|cc} a_1 & 0 & 0 & a_2 \\ 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ a_2 & 0 & 0 & a_3 \end{array}\right).$$

Consider the following remarks:

Remark 8.5 (*Parametrization of $\mathcal{V}_{diag,K}$*) Let $\mathcal{V}_{diag,K}$ denote the set of all admissible DVECH-terms A, whose corresponding Ω-Matrix has rank $K \leq n$. Due to the fact that Q^* is the only positive semidefinite representative in the $\mathcal{Q}$-set of some admissible DVECH-term A, one can define a unique parametrization as follows: Let $K \leq n$ be the rank of $\Omega = \text{math}(\text{diag}(A))$. Then Ω can be written as the product $\Omega = LL'$, where L is the $n \times K$-dimensional cholesky factor of Ω, whose upper triangular elements are zero and whose diagonal elements are strictly positive by definition. (Strictly spoken this holds true for "generic" Ω that have a regular $K \times K$ left upper block.) Let the vector $\theta \in \mathbb{R}^K_+ \times \mathbb{R}^{nK - K(K+1)/2}$ contain the non zero elements of L - start with the diagonal and stack the remaining elements columnwise beginning with the first column. The parametrization is then a concatenation of the following mappings: $\theta \mapsto L \mapsto \Omega = LL' \mapsto A = \text{diag}(\text{vech}(\Omega))$.

Remark 8.6 (*Diagonal BEKK model*) Due to the specific structure of the positive semidefinite matrix Q^* implied by an addmissible diagonal VECH-term A, the corresponding BEKK parameter matrices B_k, $k = 1, \dots, K$, are again diagonal. This can be seen when the cholesky factor L of $\Omega = LL'$, is transformed to the $n^2 \times K$ dimensional matrix U by $U = S'L$, such that $Q^* = UU'$ holds. Let U_k and L_k denote the kth column of U and L respectively, then $B_k = \text{mat}(U_k) = \text{mat}(S'L_k)$ is again diagonal for all k.

Remark 8.7 (*Stationarity conditions*) Due to the diagonal structure the stationarity condition (7.11) is now given by

$$\sum_{i=1}^{p+q} a_{i,jj} < 1, \quad \text{for } j = 1, \dots, \leq n(n+1)/2,$$

where $a_{i,jj}$ denotes the j-th diagonal element of the DVECH term A_i.

[1] The Hadamarad product of two matrices A and B of the same size is defined as $(A \odot B)_{ij} = a_{ij} b_{ij}$.

Theorem 8.6 *$\mathcal{V}_{diag}$ is a closed, convex cone in $\mathbb{R}^{n(n+1)/2}$ that contains an open subset of $\mathbb{R}^{n(n+1)/2}$.*

Proof. In theorem 7.1 it is shown that $\mathcal{V}$ is a closed convex cone. This property is carried forward to $\mathcal{V}_{diag} \subset \mathcal{V}$.
Let $A_i \in \mathcal{V}_{diag}$, $\lambda_i \geq 0$ and $\Omega_i = \text{math}(\text{diag}(A_i))$ for $i = 1, 2$. Since $\lambda_1\Omega_1 + \lambda_2\Omega_2 \geq 0$, it follows that $\lambda_1 A_1 + \lambda_2 A_2 \in \mathcal{V}_{diag}$. Finally, let A_I be such that the corresponding Ω-matrix is the identity. Then $\Omega = I_n \geq 0$ implies that an open neighborhood of A_I is contained in $\mathcal{V}_{diag}$. □

The reduction of parameters in the DVECH as compared to the general VECH model however implies also some drawbacks. First of all, the model structure is such that a change in the volatility of one variable has no immediate or direct impact on the volatility of the other variables. Second, as it is also pointed out in (Gourieroux and Jasiak, 2001, section 6.4.1), possibly the main drawback is the fact that due to its structure the DVECH model is not invariant with respect to linear combinations. This is in particular crucial for a number of financial applications such as portfolio composition or volatility modeling of exchange rates. For instance, consider a set of exchange rates (expressed in logarithms) and assume that this set of rates has a DVECH model structure. The structure will get lost, if the currency of reference is changed.

8.2.5 A restricted BEKK model

Let us now give a final example for a parsimonious parametrization of a BEKK model, where $K = 1$ throughout. In addition, it is assumed that the BEKK matrices are all symmetric and the off diagonal elements of a BEKK parameter matrix are all equal. Thus, as an illustration consider the structure of a BEKK term B and its corresponding VECH term A for the simplest case, i.e. $n = 2$:

$$B = \begin{pmatrix} \theta_1 & \theta_2 \\ \theta_2 & \theta_3 \end{pmatrix}, U = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_2 \\ \theta_3 \end{pmatrix}, Q = \begin{pmatrix} \theta_1^2 & \theta_1\theta_2 & \theta_1\theta_2 & \theta_1\theta_3 \\ \theta_1\theta_2 & \theta_2^2 & \theta_2^2 & \theta_2\theta_3 \\ \theta_1\theta_2 & \theta_2^2 & \theta_2^2 & \theta_2\theta_3 \\ \theta_1\theta_3 & \theta_2\theta_3 & \theta_2\theta_3 & \theta_3^2 \end{pmatrix}, A = \begin{pmatrix} \theta_1^2 & 2\theta_1\theta_2 & \theta_2^2 \\ \theta_1\theta_2 & \theta_2^2 + \theta_1\theta_3 & \theta_2\theta_3 \\ \theta_2^2 & 2\theta_2\theta_3 & \theta_3^2 \end{pmatrix},$$

Note that for $\theta_2 = 0$ we are left with a DVECH model with $K = 1$ and for $\theta_2 = \theta_1\theta_3$ we may model a certain sublcass of the positive semidefinite orthocomplement of the $\mathcal{Q}$-sets. θ_2 is an additional parameter that is used directly to model the underlying correlation structure. It has such a simple structure that any VECH term is parametrized by only $n+1$ parameters. Hence, this restricted BEKK(p, q) model involves only $(p+q)(n+1)$ parameters excluding the constant. This is comparable with the number of parameters needed to parametrize the so-called *Dynamic Conditional Correlation* (DCC) model proposed by (Engle and Sheppard, 2001; Engle, 2002). The DCC model however is not a subclass of the VECH model class discussed above. But it may be compared with DVECH models, since the most suitable VECH matrices that can be imposed to meet the structure of a DCC model would show a diagonal structure.

The stationarity conditions from eq. (7.11) cannot be simplified in case of the restricted BEKK model. Nevertheless, we think that this model is a neat alternative to the models presented above and it may be applied even for moderate or large n.

Chapter 9

Estimation of BEKK models

Estimation of multivariate GARCH models is troublesome, since the number of parameters may be large also for moderate vector dimension n. Even if there is enough data available for estimation, the likelihood might be relatively "flat" as a function of many parameters. Thus, it might be hard for optimization routines to find the global maximum and therefore, constraints on the parameter space are in many cases indispensible.

Let us state again the model under consideration:

$$\begin{aligned} y_t &= \mu_t + \epsilon_t & (9.1)\\ \epsilon_t &= H_t^{\frac{1}{2}} z_t, & (9.2) \end{aligned}$$

and let us assume throughout this chapter, unless stated otherwise, that $z_t \sim iid\mathcal{N}(0, I_n)$, thus $\epsilon_t|\mathcal{I}_{t-1} \sim \mathcal{N}(0, H_t)$, and the conditional variance matrix may be modeled as

$$\text{vech}(H_t) = c + \sum_{i=1}^{q} A_i \text{vech}(\epsilon_{t-i}\epsilon'_{t-i}) + \sum_{j=1}^{p} A_{q+j} \text{vech}(H_{t-j}),$$

where the matrices A_i may have any of the structures discussed above.
Due to the fact that μ_t and ϵ_t are uncorrelated, we might apply a two step estimation procedure. Thus, estimate equation (9.1) in a first step, see part I, and in a second step estimate the GARCH structure of the error term. In order to obtain efficient estimates (Engle and Kroner, 1995) suggest to repeat these two steps until convergence, while in each step estimation is performed conditional on the estimation result of the previous step.

Let in the following for the sake of simplicity $p = q = 1$. If we assume that, as proposed in (Engle and Kroner, 1995, section 4), the presample data or initial values $\epsilon_0\epsilon'_0$ and H_0 are equal to their unconditional expectation,

$$\text{math}((I - A_1 - A_2)^{-1}c),$$

where I denotes the identity matrix of suitable dimension, i.e. $I = I_{n(n+1)/2}$, then $\epsilon_1|\mathcal{I}_0 \sim \mathcal{N}(0, H_1)$, with $H_1 = \text{math}((I - A_1 - A_2)^{-1}c)$. Let $f(.)$ denote a density function, let θ be the vector of parameters that are needed to parametrize c, A_1 and A_2, and let T be the sample size, then the likelihood function L is given by

$$\begin{aligned} L(\theta;\ & \epsilon_T, \ldots, \epsilon_0, H_0) = f(\epsilon_T, \ldots, \epsilon_0, H_0;\ \theta) = \\ &= \underbrace{f(\epsilon_T|\epsilon_{T-1}, \ldots, \epsilon_0, H_0;\ \theta)}_{\sim\mathcal{N}(0,H_T)} \cdot \ldots \cdot \underbrace{f(\epsilon_2|\epsilon_1, \epsilon_0, H_0;\ \theta)}_{\sim\mathcal{N}(0,H_2)} \cdot \underbrace{f(\epsilon_1|\epsilon_0, H_0;\ \theta)}_{\sim\mathcal{N}(0,H_1)} = \\ &= \frac{1}{\sqrt{2\pi}^{Tn}} \prod_{t=1}^{T} \frac{1}{\det(H_t)} \exp(-\tfrac{1}{2}\epsilon'_t H_t^{-1} \epsilon_t) \end{aligned}$$

If we now take the logarithm, we obtain

$$
\begin{aligned}
\log L &= -\frac{Tn}{2} log(2\pi) + \sum_{t=1}^{T} l_t, \quad \text{with} \\
l_t &= -\frac{1}{2}\log(\det(H_t)) - \frac{1}{2}\epsilon_t' H_t^{-1}\epsilon_t,
\end{aligned} \tag{9.3}
$$

The parameter vector θ is now obtained by optimization of

$$
\begin{aligned}
\min_\theta \quad & \{-\log L(\theta)\} \\
s.t.: \quad & \textit{positivity constraints } (7.2), (7.3) \\
& \textit{stationarity constraints } (7.11)
\end{aligned} \tag{9.4}
$$

and the partial derivative with respect to the ith component of θ is given by

$$
-\frac{\partial logL(\theta)}{\partial\theta_i} = \sum_{t=1}^{T} \operatorname{tr}\left\{ \left(I_n - \epsilon_t\epsilon_t' H_t^{-1}(\theta)\right) \cdot \frac{\partial H_t(\theta)}{\partial\theta_i} \cdot H_t^{-1}(\theta) \right\}. \tag{9.5}
$$

Let θ_c, θ_{A_1} and θ_{A_2} denote components of θ that are necessary to parametrize the constant, A_1 and A_2 respectively. Note that $\frac{\partial H_t(\theta)}{\partial\theta_i} = \text{math}\left(\frac{\partial h_t(\theta)}{\partial\theta_i}\right)$ and $\frac{\partial h_t(\theta)}{\partial\theta_i}$ may recursively be computed by the following equations.

$$
\begin{aligned}
\frac{\partial h_0(\theta)}{\partial\theta_c} &= (I - A_1 - A_2)^{-1} \cdot \frac{\partial c(\theta)}{\partial\theta_c} \\
\frac{\partial h_0(\theta)}{\partial\theta_{A_1}} &= (I - A_1 - A_2)^{-1} \cdot \frac{\partial A_1(\theta)}{\partial\theta_{A_1}} \cdot (I - A_1 - A_2)^{-1} c \\
\frac{\partial h_0(\theta)}{\partial\theta_{A_2}} &= (I - A_1 - A_2)^{-1} \cdot \frac{\partial A_2(\theta)}{\partial\theta_{A_2}} \cdot (I - A_1 - A_2)^{-1} c \\
\frac{\partial h_t(\theta)}{\partial\theta_c} &= \frac{\partial c(\theta)}{\partial\theta_c} + A_2 \cdot \frac{\partial h_{t-1}(\theta)}{\partial\theta_c} \\
\frac{\partial h_t(\theta)}{\partial\theta_{A_1}} &= \frac{\partial A_1(\theta)}{\partial\theta_{A_1}} \cdot \eta_{t-1} + A_2 \cdot \frac{\partial h_{t-1}(\theta)}{\partial\theta_{A_1}} \\
\frac{\partial h_t(\theta)}{\partial\theta_{A_2}} &= \frac{\partial A_2(\theta)}{\partial\theta_{A_2}} \cdot h_{t-1} + A_2 \cdot \frac{\partial h_{t-1}(\theta)}{\partial\theta_{A_2}}
\end{aligned} \tag{9.6}
$$

The partial derivatives $\frac{\partial c(\theta)}{\partial\theta_c}, \frac{\partial A_1(\theta)}{\partial\theta_{A_1}}, \frac{\partial A_2(\theta)}{\partial\theta_{A_2}}$ of course depend on the respective parametrization. For the DVECH parametrization e.g. these are easy to derive. However, for most of the other aforementioned parametrizations they are much more involved. Thus, optimization of the problem stated in eq. (9.4) may be done numerically. In the applications shown in the subsequent chapter 10 we make use of the non-linear minimization function `nlm(.)` that uses a Newton-type method and is provided in the open source R package {`stats`}, see (R Development Core Team, 2005) - `http://www.r-project.org/`.

Asymptotic properties of the quasi[1] maximum likelihood estimator: (Comte and Lieberman, 2003) provide conditions for strong consistency of the quasi maximum likelihood estimator $\hat{\theta}$ and asymptotic normality. In addition, they give a thorough survey of asymptotic results published so far for univariate as well as multivariate GARCH processes. Let us for the sake of comleteness state two of their main theorems, for the proofs and further details refer to (Comte and Lieberman, 2003):

Consider in the following a BEKK(p,q) model as defined by eq. (9.1) and (9.2) with $z_t \sim iid(0, I_n)$, in VECH notation. (Note that every BEKK model may be written as a VECH model.) And refer to the parametrizations given in section 8.1.1.

Theorem 9.1 (Consistency of quasi MLE) *For the MGARCH(p,q) process defined by eq. (6.3) with $z_t \sim iid(0, I_n)$ and (7.12), and for $\hat{\theta}_T$, the quasi maximum likelihood estimate obtained from a sample of length T, and the true parameter $\theta_0 \in \Theta$, assume that*

[1] Note that (Comte and Lieberman, 2003) call the estimate $\hat{\theta}$ quasi MLE, since they do not assume that (z_t) is Gaussion, but work with the Gaussian log-likelihood function.

1. *$\Theta \in \Theta^+$ is compact, C and $B_{i,k}$, $i = 1, \ldots, p+q$, are continuous functions of the parameters $\theta \in \Theta$, and there exists a $c > 0$ such that $\inf_{\theta \in \Theta} \det(C(\theta)C(\theta)') \geq c > 0$,*
2. *The model is identifiable,*
3. *The rescaled errors z_t admit a density absolutely continuous with respect to the Lebesgue measure and positive in a neighbourhood of the origin,*
4. *For all $\theta \in \Theta$, the largest eigenvalue in modulus of the $\sum_{i=1}^{p+q} A_i$, the sum of corresponding VECH matrices, is smaller than 1.*

Then $\hat{\theta}_T$ is strongly consistent that is, $\hat{\theta}_T \rightarrow_{T \rightarrow \infty} \theta_0$ a.s.

Remark 9.1 (*Identifiability of the model.)* Note that in their proof (Comte and Lieberman, 2003) refer to the parametrization proposed by (Engle and Kroner, 1995). The drawbacks or shortcomings of this parametrization are discussed in section 8.1.2 from above. Furthermore note that the parametrizations proposed above are at best identifiable on a generic -here i.e. open and dense- class of BEKK models.

Remark 9.2 (*Compactness of the parameter space and stationarity conditions.)* Note furthermore, that if one wants to apply theorem 9.1, one has to adjust the underlying parametrization such that assumptions 1. and 4. hold.

Theorem 9.2 (Asymptotic Normality of quasi MLE) *Under the assumptions*

1. *Ass. $(1) - (4)$ from theorem 9.1, and $C(\theta)$, $B_{i,k}(\theta)$, $i = 1, \ldots, p+q$ admit continuous derivatives up to order 3 on Θ,*
2. *The components of z_t are independent,*
3. *ϵ_t admits bounded moments of order 8,*
4. *The initial states of the process H_t are fixed*[2]*.*

The quasi MLE $\hat{\theta}_{T,init}$ given the initial state is strongly consistent and

$$\sqrt{T}(\hat{\theta}_{T,init} - \theta_0) \xrightarrow[T \rightarrow \infty]{d} \mathcal{N}(0, C_1^1 C_0 C_1^1),$$

where $C_1 = \mathbb{E}\left(\left(\frac{\partial^2 l_t(\theta_0)}{\partial\theta_i \partial\theta_j}\right)_{1 \leq i,j \leq r}\right)$, $C_0 = \mathbb{E}\left(\frac{\partial l_t(\theta_0)}{\partial\theta}\frac{\partial l_t(\theta_0)}{\partial\theta}'\right)$ and r is the length of the parameter vector θ.

[2] Note that in theorem 9.1 the initial values of H_t are assumed to be drawn randomly from the stationary law $\mathbb{P}_{\theta_0}$.

Chapter 10

Simulation Studies and Applications on Real Data

In this chapter we will present simulation studies of the above discussed model classes. Several simulated data sets are used to investigate:

- the practical feasibility to estimate the model classes under consideration in reasonable computing time,
- the sensitivity of the computational results of likelihood optimization with respect to the choice of starting values,
- the capability of each model class to approximate a *data generating process* (DGP) belonging to another model class.

We will in addition provide some estimation results of the above MGARCH model classes, where estimation is performed on the 7 series of close returns of European banks, see chapter 5 for a description of the data.
Throughout this chapter we consider only MGARCH$(1, 1)$ models. Thus, we set $q = p = 1$.

Let us first introduce some coding concerning the different models and their specification, see table 10.1. Throughout, the appendix "$.K$" indicates how many additive terms there are involved in the BEKK term specification of the underlying model.

Code	K	Parametrization	Section
sr.K	n^2	(Scherrer and Ribarits, 2006)	8.1.1
	$1, \ldots, n$	(Scherrer and Ribarits, 2006)	8.2.1
ek.K	n^2	(Engle and Kroner, 1995)	8.1.2
ortho.K	n^2	-	8.1.3
	$1, \ldots, n$	(Scherrer and Ribarits, 2006)	8.2.3
f.K	$1, \ldots, n$	(Engle et al., 1990)	8.2.2
dvech.K	$1, \ldots, n$	(Bollerslev et al., 1988)	8.2.4
rbekk.K	1	-	8.2.5

Table 10.1: Coding system of the different MGARCH models.

10.1 The role of starting values

Consider the following bivariate VECH(1, 1) process ϵ_t: $\epsilon_t = H_t^{\frac{1}{2}} z_t$ with $z_t \sim iid\mathcal{N}(0, I_2)$ and

$$\text{vech}(H_t) = \underbrace{\begin{pmatrix} 1.050 \\ 0.427 \\ 0.279 \end{pmatrix}}_{c} + \underbrace{\begin{pmatrix} 0.194 & 0.250 & 0.215 \\ 0.039 & -0.115 & -0.027 \\ 0.132 & -0.133 & 0.110 \end{pmatrix}}_{A_1} \text{vech}(\epsilon_{t-1}\epsilon_{t-1}') +$$

$$+ \underbrace{\begin{pmatrix} 0.308 & 0.229 & 0.142 \\ 0.035 & -0.028 & -0.015 \\ 0.044 & 0.004 & 0.030 \end{pmatrix}}_{A_2} \text{vech}(H_{t-1}).$$

Note that here the entries in the parameter matrices have been rounded to three digits. The absolute values of the eigenvalues of $A_1 + A_2$ are 0.654, 0.250 and 0.096. Thus, the stability condition (7.11) holds. The eigenvalues of the Q-matrices in the corresponding $\mathcal{Q}$-sets are shown in figure 10.1.

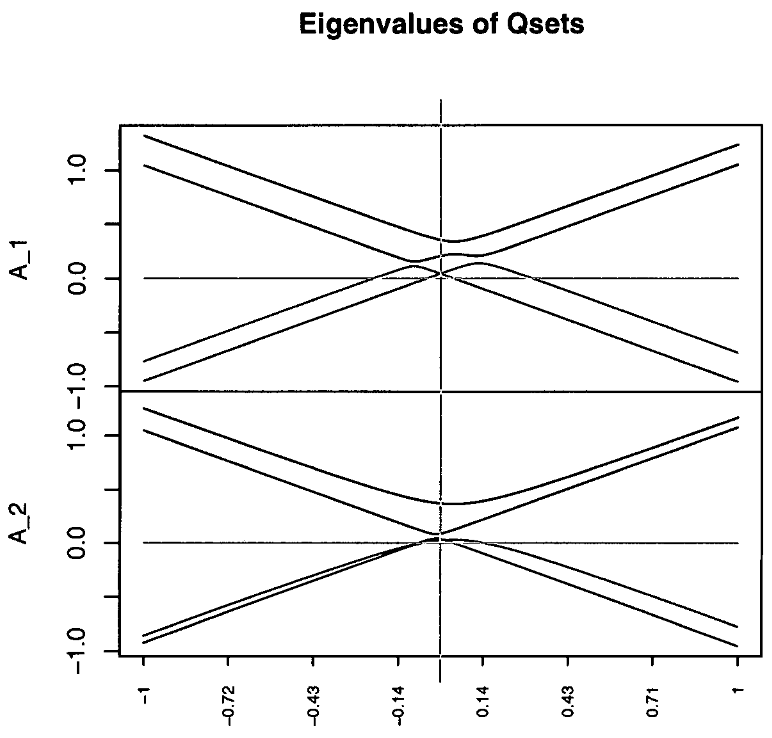

Figure 10.1: The 4 eigenvalues of the Q-matrices in the $\mathcal{Q}$-sets $\mathcal{Q}(A_1)$ and $\mathcal{Q}(A_2)$; x-axis: ω_1; Gray vertical line: indicates the analytic centers.

This process is contained in the sr.4 model class. The length of the simulated sample is $T = 300$ and the starting values for the simulation have been chosen as $H_1 = \text{math}((I_3 - A_1 - A_2)^{-1}c)$ and $\epsilon_1 = H_1^{\frac{1}{2}} z_1$, with z_1 random $\mathcal{N}(0, I_2)$.

The starting values provided for A_1 and A_2 in the Newton-like algorithm that maximizes the respective likelihood functions are all randomly chosen and of order 10^{-10} to 10^{-5}, whereas the starting value for the constant c is chosen as $\text{vech}(\frac{1}{T}\sum_{t=1}^{T} \epsilon_t\epsilon_t')$, the "vechtorized" empirical variance covariance matrix of ϵ_t. For each model class considered we provide 100 different random starting values for A_1 and A_2. Figure 10.2 shows the boxplots of the respective 100 likelihood values obtained after optimization.

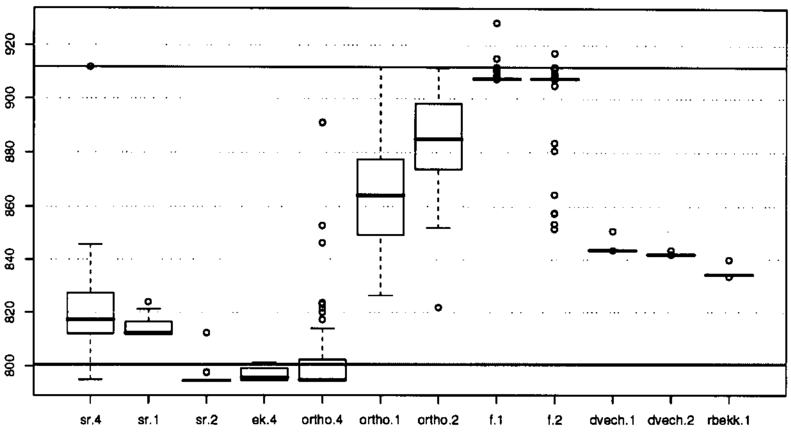

Figure 10.2: Boxplot of likelihood values obtained after optimization from 100 different random starting values for each of the considered model class; Green line: likelihood value corresponding to $c = \text{vech}(\frac{1}{T}\sum_{t=1}^{T}\epsilon_t\epsilon_t')$, $A = G = 0$; Red line: likelihood value corresponding to the true parameter values; Number of parameters (from left to right): 21, 11, 17, 21, 21, 9, 15, 7, 9, 7, 9, 9.

The performance of sr.4 is surprisingly bad as compared to its competitors that provide the same number of parameters. One or possibly the reason for sr.4 numerically lagging behind is the fact that the mapping f from the parameters θ onto the matrices Q, i.e. $f : \theta \mapsto \text{vech}(Q)$, in general is rather ill-conditioned. In the above example the condition number of the Q-matrices corresponding to the analytic centers of the $\mathcal{Q}$-sets of the true parameter matrices A_1 and A_2 and the (relative) condition number of the mapping f evaluated at the parameter vectors that correspond to the analytic centers are given in table 10.2 below. The parametrizations for ek.n^2 and ortho.n^2 are more stable concerning starting values and show considerably shorter computing times[1] , see e.g. figure 10.3. The computing times listed throughout this chapter should however be considered only as a rough guide, since sometimes several R sessions have been run at the same time.

	$\lambda_{max}(\bar{Q})/\lambda_{min}(\bar{Q})$	$\|\partial f/\partial\theta\|_F \, \|\theta\|_F \, / \, \|f\|_F$
A_1	8.47	340.58
A_2	14.59	490.28

Table 10.2: Condition numbers of the analytic centers $\bar{Q}$ and the mapping $f : \theta \mapsto \text{vech}(Q)$, evaluated at the parameter vectors that correspond to the analytic centers.

Concerning the other model classes the good performance of sr.2 is striking. If we consider the corresponding AIC values (not shown here), sr.2 still outperforms its more parsimonious competitors from the DVECH model class and rbekk.1 that are all quite stable with respect to variations in the starting values. f.1 and f.2 are also stable, but lag behind its competitors that use the same number of parameters. Finally, ortho.1 and ortho.2 show in this setting, i.e. for this particular process, the most unstable results concerning different starting values. Especially ortho.2 seems to have severe problems to find its optimum in the parameter space. Possibly its likelihood function is just too flat. ortho.1 involves less parameters and here alltogether does a better job than ortho.2.

[1] All computations have been carried out in R 2.2.0, (R Development Core Team, 2005) and were run on an Intel(R) Pentium(R) M processor 1500 MHz.

10.2 The role of the underlying DGP

Let us now consider different MGARCH processes ϵ_t in order to analyse the generality of the model classes.

For $n = 2$, data series for 12 different MGARCH processes of sample size $T = 500$ have been generated. Each of the above discussed models has been estimated on the respective data sets. In order to reduce the influence of the starting value, for each model class 10 different sets of starting values have been provided and the "best" model out of ten with respect to the likelihood value has been selected for further analysis. In table 10.3 the corresponding AIC and BIC values are listed. It can be seen that, concerning AIC, sr.2 and sr.1 show good results throughout. They are 8 and 12 times, respectively, among the three models that, given the realization of a DGP, show the lowest AIC values. What is surprising is the good performance of rbekk.1. Even though it involves a relatively small number of parameters it appears 7 times among the best (with respect to AIC) three estimated models. The results of sr.4 lag behind expectations, but as mentioned in the section before, this is due to numerical problems, see also figure 10.3. If we consider ek.4 and ortho.4, the parametrization of (Engle and Kroner, 1995) does a slightly better job than ortho.4. The F-GARCH models f.1 and f.2 perform considerably worse than their competitors even if the DGP is an F-GARCH process. DVECH models perform good, if the underlying DGP is a DVECH process, otherwise they are in the intermediate ranks. In case of a DGP of the form ortho.4, ortho.2 and ortho.1 the respective models that model the positive semidefinite orthocomplement of the $\mathcal{Q}$-sets are displaced by the more or equally parsimonious rbekk.1 model. The BIC results of course give a slight change in the ranking of the models. Anyway, sr.1 and rbekk.1 for almost all simulated series stay among the best (with respect to BIC) three estimated models.

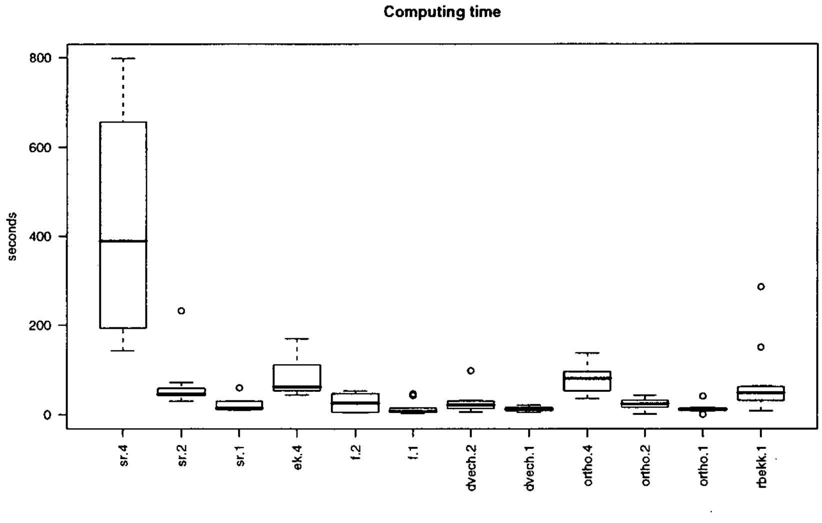

Figure 10.3: Boxplots of computing times (in seconds) of the 12 estimated MGARCH models for the 12 simulated series with $n = 2$ and $T = 500$ of the respective DGP's, see also table 10.3.

AIC			Data Generating Process											
		Nb.P.	sr.4	sr.2	sr.1	ek.4	f.2	f.1	dvech.2	dvech.1	ortho.4	ortho.2	ortho.1	rbekk.1
e	sr.4	21	6.218	5.080	7.421	2.909	5.312	2.359	1.883	5.596	3.227	5.838	2.274	4.472
s	sr.2	17	**6.186**	**5.031**	7.397	2.882	5.288	2.337	1.853	5.578	3.199	5.665	2.258	4.447
t	sr.1	11	6.192	5.044	**7.373**	**2.874**	**5.266**	**2.316**	1.834	**5.555**	3.194	5.656	2.234	4.432
i	ek.4	21	6.202	5.046	7.413	2.898	5.304	2.354	1.869	5.594	3.215	5.672	2.274	4.464
m	f.2	9	7.284	5.291	7.763	3.089	5.306	2.605	1.914	5.777	4.158	6.031	3.018	4.943
.	f.1	7	7.276	5.283	7.945	3.138	5.357	2.597	1.925	5.769	4.150	6.023	3.010	4.935
	dvech.2	9	6.567	5.230	7.582	3.020	5.321	2.481	1.837	5.573	3.304	5.828	2.438	4.815
M	dvech.1	7	6.752	5.237	7.597	3.034	5.324	2.473	**1.829**	5.565	3.304	5.857	2.447	4.815
o	ortho.4	21	6.224	5.044	7.421	2.926	5.307	2.367	1.879	5.596	3.223	5.676	2.278	4.468
d	ortho.2	15	6.320	5.121	7.695	3.131	5.342	2.635	1.861	5.801	3.197	5.874	2.254	4.581
e	ortho.1	9	6.284	5.051	7.695	3.019	5.306	2.611	1.833	5.777	3.202	5.826	**2.234**	4.493
l	rbekk.1	9	6.282	5.068	7.433	2.876	5.278	2.417	1.835	5.559	**3.193**	**5.639**	2.237	**4.431**
BIC			Data Generating Process											
		Nb.P.	sr.4	sr.2	sr.1	ek.4	f.2	f.1	dvech.2	dvech.1	ortho.4	ortho.2	ortho.1	rbekk.1
e	sr.4	21	6.656	5.518	7.859	3.347	5.750	2.797	2.321	6.034	3.665	6.276	2.712	4.910
s	sr.2	17	6.540	5.385	7.751	3.236	5.642	2.692	2.207	5.932	3.554	6.019	2.612	4.802
t	sr.1	11	**6.422**	5.273	**7.603**	3.103	5.495	**2.546**	2.064	5.784	3.424	5.886	2.464	4.661
i	ek.4	21	6.640	5.484	7.851	3.336	5.742	2.792	2.307	6.032	3.654	6.110	2.712	4.902
m	f.2	9	7.471	5.479	7.950	3.277	5.494	2.793	2.101	5.964	4.346	6.219	3.206	5.131
.	f.1	7	7.422	5.429	8.091	3.284	5.503	2.743	2.071	5.915	4.296	6.169	3.156	5.081
	dvech.2	9	6.754	5.418	7.770	3.208	5.509	2.669	2.025	5.761	3.492	6.015	2.626	5.003
M	dvech.1	7	6.898	5.383	7.743	3.180	5.470	2.619	**1.975**	**5.711**	3.450	6.003	2.593	4.962
o	ortho.4	21	6.662	5.482	7.859	3.364	5.745	2.805	2.317	6.034	3.661	6.114	2.716	4.906
d	ortho.2	15	6.633	5.434	8.008	3.444	5.655	2.948	2.174	6.114	3.510	6.187	2.567	4.894
e	ortho.1	9	6.471	**5.239**	7.882	3.207	5.494	2.799	2.021	5.964	3.390	6.013	**2.422**	4.680
l	rbekk.1	9	6.470	5.256	7.620	**3.064**	**5.466**	2.605	2.023	5.747	**3.381**	**5.827**	2.425	**4.618**

Table 10.3: AIC and BIC values plus the total number of parameters (Nb.P.) of the respective estimated MGARCH models (rows) for realizations of the bivariate MGARCH processes (columns); The minimum IC value for a certain DGP across the estimated models is shown in bold face.

For $n = 3$, data series for 6 different MGARCH processes of sample size $T = 500$ have been generated. The number of parameters needed to estimate now ranges from 12 for the dvech.1 model up to 78 for the models where $K = 9$. Figure 10.4 shows boxplots of the computing times in seconds. Due to the large differences in the computing times the models have been grouped according to the number of parameters estimated. Estimation of a rbekk.1 model in the worst case takes not even 7 minutes. This is really good, since it performs well already in the AIC sense, see table 10.4. Computation of dvech.1 takes at most 4 minutes and ortho.1, although it involves a bit more parameters than rbekk.1 and dvech.1, does not even need 3 minutes to be calculated. Note that on average these models take even less computing time, namely around 1.4 to 2.3 minutes. In fact, the IC results show that in case of $n = 3$ the parsimonious model classes can already compete with the more general model classes. sr.1, sr.2 and sr.3 for instance perform well, but at the cost of high computing times. If we compare the models that involve 78 parameters, again ek.9 outperforms sr.9 and ortho.9. The results of the F-GARCH models are again rather disappointing for the underlying data sets.

To sum it up, rbekk.1 is parsimonious, but may still be sufficiently general in order to model different kinds of MGARCH processes. The DVECH models possibly are too restrictive concerning the structure of the model and therefore tend to slightly lag behind when the DGP does not show this diagonality in the VECH parameter matrices. ortho.1 is promising, too. As it is the case for rbekk.1, it does not seem to be that dependent on the structure of the underlying DGP. ortho.2 and ortho.3 however can be found rather in the intermediate ranks.

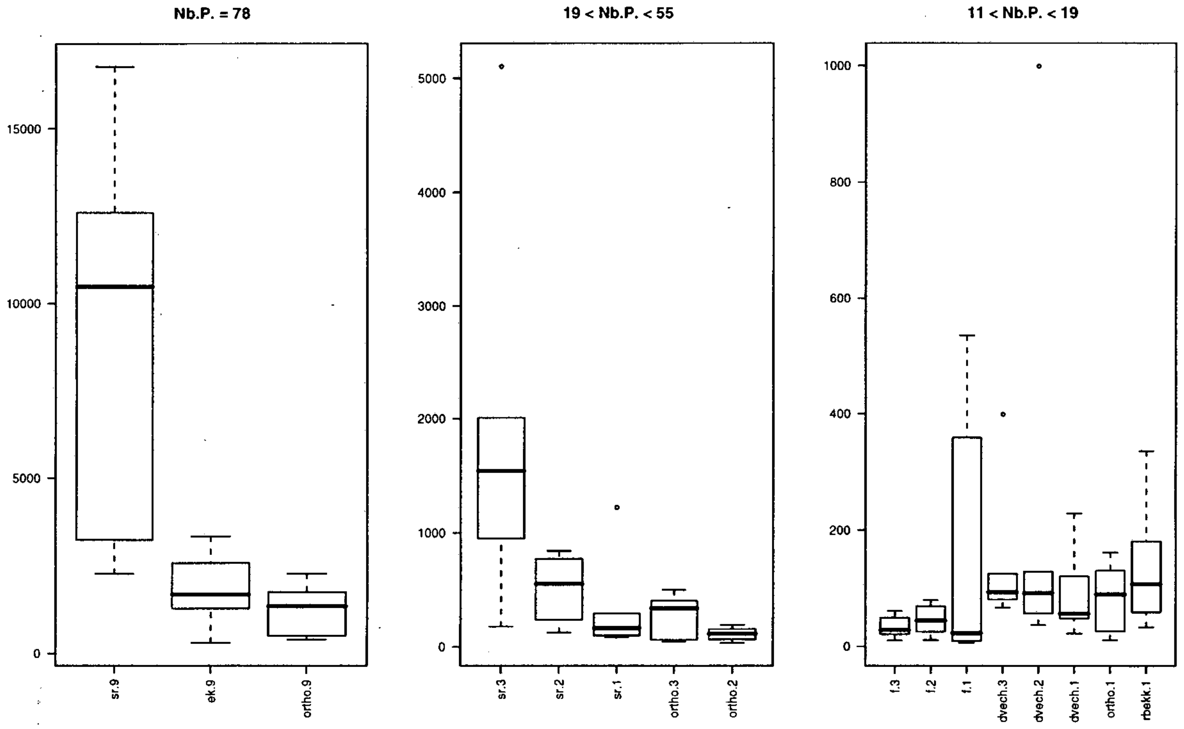

Figure 10.4: Boxplots of computing times (in seconds) of the 16 estimated MGARCH models for the 6 simulated series with $n = 3$ and $T = 500$ of the respective DGP's, see also table 10.4.

AIC			Data Generating Process					
		Nb.P.	sr.9	ek.9	ortho.9	f.1	dvech.3	rbekk.1
e	sr.9	78	6.595	7.088	7.100	4.228	6.657	6.194
s	sr.3	54	6.416	6.932	6.762	4.100	6.542	6.080
t	sr.2	40	6.367	6.885	**6.724**	4.045	6.486	6.024
i	sr.1	24	**6.359**	6.861	6.780	3.990	6.445	5.962
m	ek.9	78	6.517	7.027	6.852	4.199	6.639	6.176
.	f.3	18	6.955	6.942	7.248	4.010	7.114	6.964
	f.2	16	6.949	6.936	7.243	4.001	7.110	6.963
M	f.1	12	6.963	6.883	7.227	3.985	7.093	6.947
o	dvech.3	18	6.680	6.900	7.093	4.009	6.430	6.284
d	dvech.2	16	6.672	6.892	7.086	4.001	6.422	6.276
e	dvech.1	12	6.682	6.877	7.120	3.985	**6.421**	6.276
l	ortho.9	78	6.556	7.032	6.879	4.236	6.676	6.449
	ortho.3	36	6.837	6.961	7.152	4.080	6.496	6.316
	ortho.2	26	6.575	6.908	7.126	4.040	6.468	6.511
	ortho.1	16	6.504	6.859	6.877	4.001	6.439	6.372
	rbekk.1	14	6.576	**6.857**	7.090	**3.979**	6.428	**5.946**

BIC			Data Generating Process					
		Nb.P.	sr.9	ek.9	ortho.9	f.1	dvech.3	rbekk.1
e	sr.9	78	8.222	8.715	8.727	5.855	8.284	7.821
s	sr.3	54	7.543	8.059	7.888	5.227	7.668	7.206
t	sr.2	40	7.201	7.719	7.558	4.879	7.320	6.858
i	sr.1	24	6.860	7.361	7.280	4.490	6.946	6.463
m	ek.9	78	8.144	8.653	8.479	5.826	8.266	7.803
.	f.3	18	7.330	7.317	7.624	4.385	7.490	7.340
	f.2	16	7.282	7.270	7.576	4.335	7.443	7.297
M	f.1	12	7.213	7.133	7.477	4.235	7.343	7.198
o	dvech.3	18	7.055	7.275	7.469	4.384	6.806	6.659
d	dvech.2	16	7.005	7.225	7.420	4.335	6.756	6.609
e	dvech.1	12	6.932	**7.127**	7.371	**4.235**	**6.671**	6.527
l	ortho.9	78	8.183	8.659	8.506	5.863	8.303	8.076
	ortho.3	36	7.588	7.712	7.902	4.831	7.247	7.067
	ortho.2	26	7.118	7.451	7.669	4.583	7.010	7.053
	ortho.1	16	**6.838**	7.193	**7.211**	4.335	6.772	6.706
	rbekk.1	14	6.868	7.149	7.382	4.271	6.720	**6.238**

Table 10.4: AIC and BIC values plus the total number of parameters (Nb.P.) of the respective estimated MGARCH models (rows) for realizations of the 3-dimensional MGARCH processes (columns); The minimum IC value for a certain DGP across the estimated models is shown in bold face.

10.3 Application on real data

In this section we consider MGARCH models for observations of three bivariate processes and one 7-dimensional vector process consisting of the close returns of 7 European banks, see figure 5.1. For the latter only the model classes with the most parsimonious model structure were taken into consideration. Table 10.5 summarizes the results. In general, we can say that the models sr.K with $K \leq n$ yield good results. In case of the bivariate data sets, sr.2 is selected throughout by the AIC. If we consider all seven banks, sr.1 might be a good choice. It involves however considerably more parameters than its competitors. Following the results of the BIC values we would therefore rather choose the more parsimonious restricted BEKK model from section 8.2.5 that allows for a more general covariance structure than the DVECH models. Figures 10.5 and 10.6 illustrate the estimation results of the sr.1 and rbekk.1 models for the conditional variances and covariances of the seven return series. In fact, the changes in the conditional variances are modelled quite well. In order to see how good the conditional covariances are modelled, we consider two trivial portfolios, where the portfolio weights are held constant throughout and determined as follows: Let $\bar{y}$ denote the geometric mean and $\hat{\Sigma}_y$ the sample variance covarianc matrix of y_t, then the portfolio weights are obtained from the minimization problem

$$\min_{\omega \in \mathbb{R}^n} -\omega' \bar{y} + \frac{1}{2} \omega' \hat{\Sigma}_y \omega,$$

with respect to the two different sets of constraints or strategies presented in section 5.1 above. Table 10.6 below shows the thus obtained portfolio weights. Figure 10.6 depicts the absolute returns of the portfolios and two times the estimated conditional standard deviations obtained from the two models sr.1 and rbekk.1. It can be seen that the conditional covariances are modelled quite well, too.

To sum it up, it can be said that the above presented model classes are promising, and estimation does work even for moderate dimensions n. Especially the models sr.K with $K \leq n$, DVECH and rbekk.1 models show good results and may be general enough.

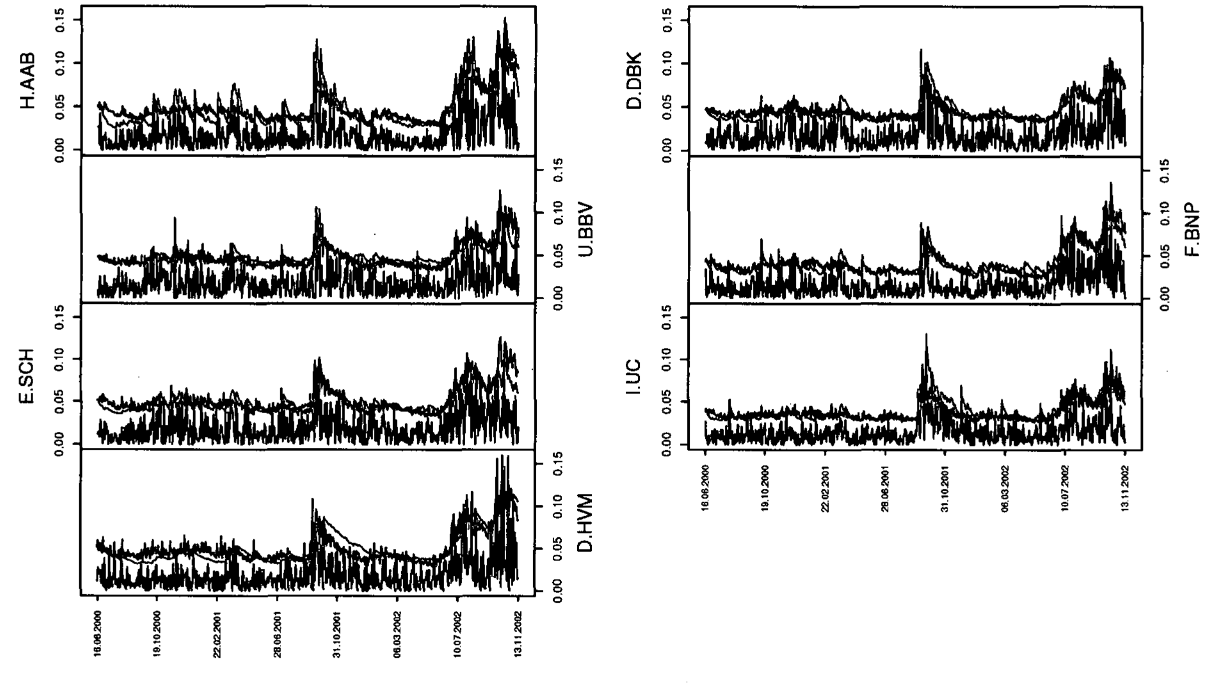

Figure 10.5: Black: Absolute return series of the seven banks; Red: Two times the estimated standard deviation of model sr.1; Green: Two times the estimated standard deviation of model rbekk.1.

AIC		Nb.P.	H.AAB-D.DBK	D.DBK-F.BNP	U.BBV-E.SCH	Nb.P.	All Banks
e	sr.4	21	4.974	4.835	5.177	-	-
s	sr.2	17	**4.720**	**4.431**	**4.978**	-	-
t	sr.1	11	4.737	4.475	5.034	126	**14.715**
i	ek.4	21	4.732	4.447	4.997	-	-
m	f.2	9	5.262	4.958	5.340	-	-
.	f.1	7	5.256	4.952	5.334	42	15.888
	dvech.2	9	4.737	4.463	5.074	54	14.814
M	dvech.1	7	4.754	4.483	5.079	42	14.993
o	ortho.4	21	4.756	4.484	4.997	-	-
d	ortho.2	15	4.911	4.460	4.984	-	-
e	ortho.1	9	4.852	4.602	5.071	54	15.259
l	rbekk.1	9	4.739	4.487	5.043	44	14.933
BIC		Nb.P.	H.AAB-D.DBK	D.DBK-F.BNP	U.BBV-E.SCH	Nb.P.	All Banks
e	sr.4	21	5.337	5.198	5.540	-	-
s	sr.2	17	5.014	4.726	5.272	-	-
t	sr.1	11	4.927	4.666	5.224	126	16.896
i	ek.4	21	5.096	4.810	5.360	-	-
m	f.2	9	5.418	5.114	5.496	-	-
.	f.1	7	5.377	5.073	5.455	42	16.615
	dvech.2	9	4.892	4.619	5.230	54	15.749
M	dvech.1	7	**4.875**	**4.604**	5.200	42	15.720
o	ortho.4	21	5.120	4.848	5.361	-	-
d	ortho.2	15	5.170	4.720	5.243	-	-
e	ortho.1	9	5.008	4.758	5.227	54	16.194
l	rbekk.1	9	4.895	4.643	**5.199**	44	**15.695**

Table 10.5: AIC and BIC values plus the total number of parameters (Nb.P.) of the respective estimated MGARCH models (rows) for observations of three bivariate processes and one 7-dimensional vector process (columns); The minimum IC value for a certain process across the estimated models is shown in bold face; For the 7-dimensional vector process only the most parsimonious model classes have been considered.

	H.AAB	U.BBV	E.SCH	D.HVM	D.DBK	F.BNP	I.UC
w_1	0.0000	0.0000	−0.0000	0.0000	0.0000	0.4591	0.5409
w_2	−0.0076	0.0472	−0.0661	−0.4263	0.0712	0.2346	0.1469

Table 10.6: Portfolio weights for Strategy I (w_1) and Strategy II (w_2).

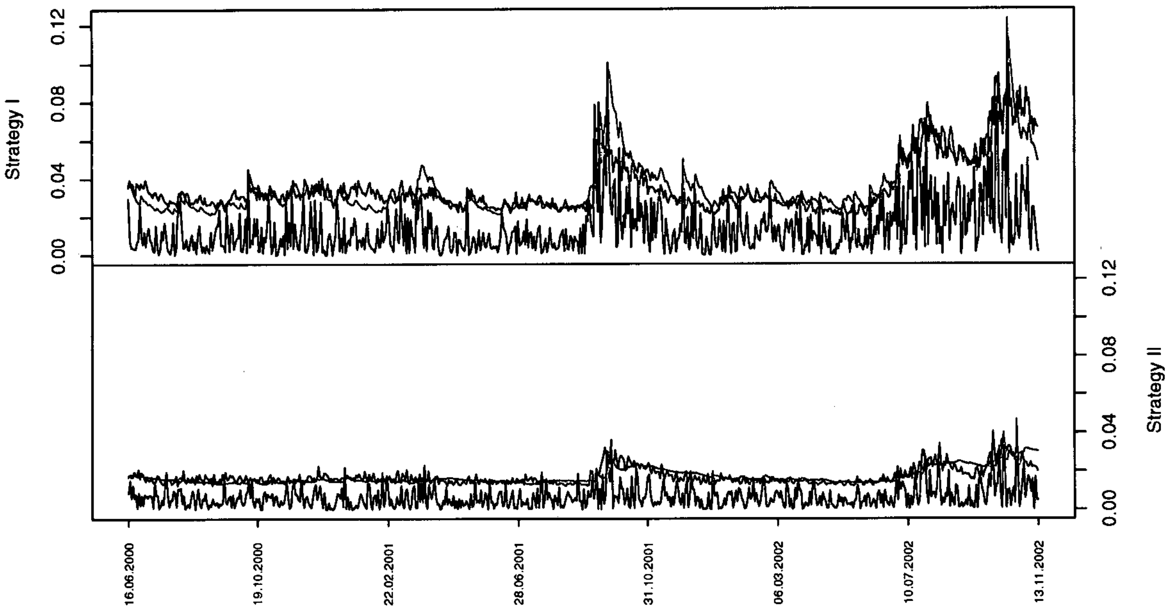

Figure 10.6: Black: Absolute return series of the portfolio returns; Red: Two times the estimated standard deviation of the portfolio returns obtained from model sr.1; Green: Two times the estimated standard deviation of the portfolio returns obtained from model rbekk.1.

Chapter 11

Conclusion

In this thesis multivariate methods for modelling the conditional first and second moments of a vector process have been analyzed.

The first part is devoted to the identification, specification and estimation of forecasting models for the conditional expectation of some vector process. In particular, the properties and main features of three model classes, namely the VARX model, the Reduced Rank model and the factor model with idiosyncratic noise, were pointed out. A special issue was to define data driven model specification and input selection procedures. In this respect methods similar to the univariate procedures suggested by (An and Gu, 1985; An and Gu, 1989), but adapted for the multivariate case and the framework of the respective model class were presented. The applications on real data showed that out-of-sample the RR models almost always outperform their competitors at least on the underlying data set of the seven European bank close returns. Here, the out-of-sample performance measures were the out-of-sample R^2 and hit rate. If we consider portfolio optimization the portfolio return obtained from portfolio weighting strategies that base on forecasts of a RR model exceeds the return of the benchmark portfolio. Thus, the model classes, in particular the RR model class, are not only useful when it comes to data analysis, but may also be of value in the area of forecasting of financial time series.

In the second part of the thesis the focus is on multivariate GARCH models. In particular, the structure and parametrization of VECH and BEKK models is investigated. Knowledge of the theory of these very general model classes is worthwhile, not only for understanding the model features, but also for the construction and parametrization of reasonable parsimonious subclasses as for instance the restricted BEKK model from section 8.2.5.
In fact, connected to each VECH parameter matrix A there is an affine subset of symmetric matrices, called $\mathcal{Q}(A)$ that may be used to analyze VECH and BEKK terms. In particular, a VECH term has an equivalent BEKK representation, if and only if this set $\mathcal{Q}(A)$ contains a positive semidefinite matrix $Q \geq 0$. The corresponding BEKK representation essentially corresponds to a factorization of such a $Q \geq 0$ and thus, the number K of additive terms in the BEKK framework is related to the rank of Q. Therefore, one may check if a BEKK representation exists via a semidefinite program. It can be shown that for the bivariate case admissible VECH models and BEKK models are equivalent. For $n > 2$ however, there is a "thick" class of admissible VECH models that have no equivalent BEKK representation. Hence, in this case VECH models are more general.
The problem of finding a unique BEKK representation may be decomposed into two steps. First, define a unique positive semidefinite element in $\mathcal{Q}(A)$ and then use a unique factorization of this element. Based on this idea a parametrization (sr.n^2) of a generic set of BEKK models is presented. In addition, it is also shown that the Engle-Kroner parametrization (ek.n^2), see (Engle and Kroner, 1995), and the parametrization of the positive semidefinite orthocomplement (ortho.n^2) of the $\mathcal{Q}$-sets do not cover a generic set.

The general BEKK model uses $O(n^4)$ free parameters which makes estimation infeasible for large n. Anyway, as can be seen from the applications above for small n the parametrization of the general BEKK model may be useful and estimation can be performed. Restricting the number of additive terms K in the BEKK framework (or equivalently restricting the rank of $Q \geq 0$) to be less than n (sr.K with $K \leq n$) dramatically reduces the number of parameters. We show that generically this rank restriction suffices to uniquely define an element in $\mathcal{Q}(A)$ and thus, the BEKK parameter matrices B_k (given some suitable zero and positivity restrictions) may be directly used as free parameters. Identifiablity problems only occur on a "thin" set of such BEKK$(p, q, K \leq n)$ models. In addition, we discuss parsimonious model classes such as the popular diagonal VECH model (dvech.K with $K \leq n$) and the F-GARCH model (f.K with $K \leq n$) using the above described methodology. We consider a model class that relaxes some of the restrictions imposed by the F-GARCH model and parametrizes the positive semidefinite orthocomplement of the $\mathcal{Q}$-sets (ortho.K with $K \leq n$). Finally, the restricted BEKK model (rbekk.1) is proposed.

The applications show that sr.K with $K \leq n$ and the parsimonious model classes dvech.1, ortho.1 and rbekk.1 are promising. The general BEKK model sr.n^2 shows severe numerical problems and in this respect lags behind its competitors involving the same number of parameters, namely ek.n^2 and ortho.n^2. The parametrization of (Engle and Kroner, 1995) performs well throughout, however it takes in general more time to estimate ek.n^2 than ortho.n^2.

Appendix A

Random Variables and Stochastic Processes

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a *probability space*, where Ω is the set of all elementary events, $\mathcal{A}$ is a sigma algebra of events or subsets of Ω, and $\mathbb{P}$ is a probability measure defined on $\mathcal{A}$. An n–dimensional *random variable* y, defined on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$, is an $\mathcal{A}$*-measurable* function, i.e. a real[1] valued function $y : \Omega \longrightarrow \mathbb{R}^n$, mapping each $\omega \in \Omega$ on $y(\omega) = (y_1(\omega), \dots, y_n(\omega))'$ such that for each $c = (c_1, \dots, c_n)' \in \mathbb{R}^n$ the set $A_c = \{\omega \mid y_1(\omega) \leq c_1, \dots, y_n(\omega) \leq c_n\}$ is an element of the sigma algebra $\mathcal{A}$. Note, that every "Borel"-measurable function of a random variable is again a random variable.

The *joint distribution function* of the components of some real valued n–dimensional random variable y is a non-negative, monotonically non-decreasing and componentwise right-continuous function F_y mapping $\mathbb{R}^n$ onto the closed intervall $[0, 1]$ and is defined by the equation $F_y(c) = \mathbb{P}(A_c) = \mathbb{P}(y_1(\omega) \leq c_1, \dots, y_n(\omega) \leq c_n)$. Considering random variables, one distinguishes between discrete and continuous random variables. A random variable is called *discrete*, if $\mathbb{P}(y(\omega) = u_k) > 0$ holds for a finite or countable infinite number of points $u_k \in \mathbb{R}^n$. In the sequel however, we will deal with continuous random variables only. A random variable is said to be *continuous*, if there exists a non-negative integrable function, $f_y : \mathbb{R}^n \longrightarrow \mathbb{R}$, $u \mapsto f_y(u_1, \dots, u_n)$, for which $F_y(c) = \int_{-\infty}^{c_1} \dots \int_{-\infty}^{c_n} f_y(u_1, \dots, u_n) \mathrm{d}u_1 \dots \mathrm{d}u_n$ holds. Function $f_y(u)$ is called the *density function* of y. Note, that every integrable non-negative function $f(u)$, whose integral over $\mathbb{R}^n$ is equal to 1, is a density function of some random variable. The density function f_y of a continuous random variable y contains the same information about the probability distribution of y as the distribution function F_y, since $\mathbb{P}(A) = \int_\Omega 1_A(\omega) \mathrm{d}\mathbb{P} = \int_A 1 \mathrm{d}\mathbb{P} = \int_B f_y(u) \mathrm{d}u$, where $1_A(\omega)$ is the indicator function, that is one for $\omega \in A$ and zero else, holds for all $A \in \mathcal{A}$, if $A = y^{-1}(B)$, $B \in \mathcal{B}^n$, where $\mathcal{B}^n$ is the sigma algebra of Borel sets in $\mathbb{R}^n$, belonging to the probability space $(\mathbb{R}^n, \mathcal{B}^n, \mathbb{P}_y)$. In many cases it is easier to work with f_y rather than with F_y.

An n–dimensional *stochastic process*, denoted by (y_t), is a family of n–dimensional random variables y_t, t in some index set $\mathbb{T}$. Thus, (y_t) is a real valued function $y : \mathbb{T} \times \Omega \longrightarrow \mathbb{R}^n$, where for each fixed $t \in \mathbb{T}$, $y(t, .)$ is an n-dimensional random variable[2]. In the following, we will only deal with time discrete processes. Hence, $\mathbb{T} \subseteq \mathbb{Z}$. The stochastic properties of a process (y_t) can be deduced from the joint distribution functions of every finite selection $\{y_{t_1}, \dots, y_{t_r}\}$ of r random variables, $t_1, \dots, t_r \in \mathbb{Z}$. The complete set of distribution functions will, however, be unknown in practice. Hence, we will be concerned with estimation of the moments characterizing the distribution functions. The distribution function of a normally distributed random variable, for instance, is totally characterized by its first and second moments. The first moment or expectation, $\mathbb{E}y_t$, and the centered second moments or variance covariance matrices, $\mathrm{cov}(y_t, y_s) = \mathbb{E}(y_t - \mathbb{E}y_t)(y_s - \mathbb{E}y_s)'$, of the underlying process are therefore in the

[1] Complex random variables are not discussed here.

[2] Note, that for time discrete (continuous) processes $\mathbb{T}$ is a subset of the integer numbers $\mathbb{Z}$ (the real numbers $\mathbb{R}$).

center of investigation. Due to the fact, that the whole information set, $\mathcal{A}$, a process (y_t) is defined on, in general, is not available or simply unknown, one is mostly restricted to conditional distributions, i.e. one investigates the distributions of the y_t's, conditioned on some available information set, $\mathcal{I}$, that is a sub sigma algebra of $\mathcal{A}$. In the following, $\mathcal{I}$ will often be the sigma field $\mathcal{I}_t = \sigma(y_t', y_{t-1}', \ldots, x_t', x_{t-1}', \ldots, 1)$, that is generated by the current and past information of (y_t) and some observed exogenous, say, k–dimensional vector process (x_t), and a constant.
What we observe, in fact, are realizations of the process (y_t), which therefore is also called a *data generating process.* If the observation intervals are equidistant, the sequence of observations ordered by time is called *time series.* Note, that with y_t we denote both, the random variable and its realization. It should be clear, however, from the context to which of the two we are refering to. In practice the sample size or the length of some time series is finite and will be denoted by T. When time series, $y_1, \ldots, y_T$, are analyzed, the ordering of observations, in general, is crucial, since it may contain additional information about the family of random variables, more precisely, the dependence structure among the random variables forming the underlying stochastic process. To understand independence of two random variables consider the following: every (n–dimensional) random variable generates a new probability space $(\mathbb{R}^n, \mathcal{B}^n, \mathbb{P}^*)$, where $\mathcal{B}^n$ is the sigma algebra of Borel sets (all half-open intervals) of the n–dimensional Euclidian space $\mathbb{R}^n$, and $\mathbb{P}^*$ is the probability measure given by $\mathbb{P}^*(B) = \mathbb{P}(y^{-1}(B))$ for all $B \in \mathcal{B}^n$, where $y^{-1}(B) = \{\omega | y(\omega) \in B\} = A_{y,B} \in \mathcal{A}$. Two n–dimensional real valued random variables y and x defined on $(\Omega, \mathcal{A}, \mathbb{P})$ are said to be *independent* if and only if the events $A_{y,B}$ and $A_{x,B}$ are independent for all $B \in \mathcal{B}^n$, i.e. $\mathbb{P}(A_{y,B} \cap A_{x,B}) = \mathbb{P}(A_{y,B})\mathbb{P}(A_{x,B})$. A measure for linear dependence is given by the covariance function, $\text{cov}(\epsilon_t, \epsilon_s)$. Two random variables are said to be *uncorrelated* if $\text{cov}(\epsilon_t, \epsilon_s) = 0$. Hence, independence implies uncorrelatedness. Let $y_t = y_{t-1} + \epsilon_t$. Then, if $\mathbb{E}(y_t|\mathcal{I}_{t-1}) = y_{t-1}$ and $\mathbb{E}(\epsilon_t|\mathcal{I}_{t-1}) = 0$ for all $t \in \mathbb{T}$, $(y_t)_{t \geq 0}$ is called *Martingale* and (ϵ_t) is a *Martingale difference sequence* (MDS). It is easy to check that the ϵ_t's of an MDS have mean zero and are uncorrelated. Furthermore note, that a process (ϵ_t), where the ϵ_t's have mean zero and are independent, is an MDS. Thus loosely spoken, one can state the following: independent process with mean zero $\Rightarrow$ MDS $\Rightarrow$ uncorrelated process.
A stochastic process (y_t) is called *(wide sense) stationary* if its first and second moments exist and are time invariant, i.e. $\mathbb{E}y_t = \mu$ (constant) for all t, $\mathbb{E}y_t'y_t < \infty$ for all t and $\mathbb{E}(y_t - \mu)(y_{t-s} - \mu)'$ does not depend on t. An important example for stationary processes are the *white noise* processes (ϵ_t), which are defined by $\mathbb{E}\epsilon_t = 0$, $\mathbb{E}\epsilon_s\epsilon_t' = \delta_{s,t}\Sigma$, where $\delta_{s,t}$ is the Kronecker delta and Σ is some $n \times n$–dimensional covariance matrix[3].

Consider again the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and let $\mathcal{L}_2(\Omega, \mathcal{A}, \mathbb{P})$ denote the set of all random variables $x : \Omega \to \mathbb{C}$ that are *square integrable*, i.e. $\mathbb{E}\,|x|^2 < \infty$. Let $x \equiv y$ be an equivalence relation defined on $\mathcal{L}_2$ such that $x \equiv y$ if and only if $x = y$ almost surely, and let $\mathbb{L}_2(\Omega, \mathcal{A}, \mathbb{P})$ denote the set of these equivalence classes. Then due to the linearity of the integral $\mathbb{L}_2$ is a linear space and $\langle x, y\rangle = \mathbb{E}x\bar{y}$ is an inner product, where $\bar{y}$ denotes the complex conjugate of y, since

1. $\langle a_1x_1 + a_2x_2, y\rangle = a_1\langle x_1, y\rangle + a_2\langle x_2, y\rangle$ for all $a_1, a_2 \in \mathbb{C}$ and $x_1, x_2 \in \mathbb{L}_2$,
2. $\langle x, y\rangle = \overline{\langle x, y\rangle}$, and
3. $\langle x, x\rangle \geq 0$ and $\langle x, x\rangle = 0 \Leftrightarrow x = 0$.

Let $\|x\|^2 = \langle x, x\rangle$ be the norm defined by the inner product, then since $\mathbb{L}_2$ is a linear space with an inner product and since any *Cauchy sequence*[4] in $\mathbb{L}_2$ converges, i.e. $\mathbb{L}_2$ is complete in the norm defined by the inner product, $\mathbb{L}_2$ is a *Hilbert space*, namely the Hilbert space of square integrable random variables.

[3] The name "white noise" is justified by the fact that the spectra of white noise processes are constant over all frequencies just like the spectrum of white light.

[4] A Cauchy sequence in $\mathbb{L}_2$ is a sequence $x_n \in \mathbb{L}_2$ for which it holds that $\lim_{n,m\to\infty} \|x_n - x_m\| = 0$. The Cauchy sequence is said to converge to some $x \in \mathbb{L}_2$, if $\lim_{n\to\infty} \|x_n - x\| = 0$.

Appendix B

Proofs

Lemma B.1 *Given matrices A,B and C of dimensions $n \times n$, $m \times n$ and $m \times m$, where A and C are symmetric. If $C > 0$,*

$$\begin{pmatrix} A & B' \\ B & C \end{pmatrix} \geq 0 \Longleftrightarrow A - B'C^{-1}B \geq 0.$$

Proof. Note that $\begin{pmatrix} A & B' \\ B & C \end{pmatrix} \geq 0$, is equivalent to $T'\begin{pmatrix} A & B' \\ B & C \end{pmatrix}T \geq 0$, for any matrix $T \in \mathbb{R}^{(n+m)\times(n+m)}$. Choose the (non-singular) T, $T = \begin{pmatrix} 0_{n\times m} & I_n \\ I_m & -C^{-1}B \end{pmatrix}$, then $T'\begin{pmatrix} A & B' \\ B & C \end{pmatrix}T = \begin{pmatrix} C & 0 \\ 0 & A - B'C^{-1}B \end{pmatrix} \geq 0$ is equivalent to $A - B'C^{-1}B \geq 0$. □

Analogously one can show that if $C > 0$, $\begin{pmatrix} A & B' \\ B & C \end{pmatrix} > 0 \Longleftrightarrow A - B'C^{-1}B > 0$.

Lemma B.2 (Non-singularity of Γ_X) *Let X_t be the $(1+pn+k)$-dimensional process as defined in eq. (2.2). Given the assumptions 2.1 from above, the second moment matrix of X_t, $\Gamma_X = \mathbb{E}X_tX_t'$ is positive definite.*

Proof. $X_t = (1, y_{t-1}', \dots, y_{t-p}', x_{t-1}')'$ and due to the fact that $A(z)$ fulfills the stability condition, y_t can be represented as

$$\begin{aligned} y_t &= \underbrace{A(z)^{-1}c + A(z)^{-1}D\mu_x}_{\mu_y} + \underbrace{A(z)^{-1}D(x_{t-1} - \mu_x)}_{y_t^x} + \underbrace{A(z)^{-1}\epsilon_t}_{y_t^\epsilon} = \\ &= \mu_y + y_t^x + y_t^\epsilon. \end{aligned}$$

Let $Y_t^x := (y_{t-1}^{x'}, \dots, y_{t-p}^{x'})'$ and $Y_t^\epsilon := (y_{t-1}^{\epsilon'}, \dots, y_{t-p}^{\epsilon'})'$. Hence,

$$X_t = \begin{pmatrix} 1 \\ \iota_p \otimes \mu_y \\ \mu_x \end{pmatrix} + \begin{pmatrix} 0 \\ Y_t^x \\ x_{t-1} - \mu_x \end{pmatrix} + \begin{pmatrix} 0 \\ Y_t^\epsilon \\ 0 \end{pmatrix},$$

where ι_p denotes the p-dimensional vector of ones. Since the ξ_t's and ϵ_t's are independent for all s, t, we have $\mathbb{E}Y_t^x Y_s^{\epsilon'} = 0$ for all t, s. Thus, Γ_X can be written as the sum of three positive semidefinite matrices:

$$\begin{aligned}\Gamma_X = \mathbb{E}X_tX_t' &= \begin{pmatrix} 1 & \iota_p'\otimes\mu_y' & \mu_x' \\ \iota_p\otimes\mu_y & 0 & 0 \\ \mu_x & 0 & 0 \end{pmatrix} + \dots \\ \dots &+ \begin{pmatrix} 0 & 0 & 0 \\ 0 & \mathbb{E}Y_t^xY_t^{x'} & \mathbb{E}Y_t^x(x_{t-1}-\mu_x)' \\ 0 & \mathbb{E}(x_{t-1}-\mu_x)Y_t^{x'} & \mathbb{E}(x_{t-1}-\mu_x)(x_{t-1}-\mu_x)' \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 0 & \mathbb{E}Y_t^\epsilon Y_t^{\epsilon'} & 0 \\ 0 & 0 & 0 \end{pmatrix}.\end{aligned}$$

Now, $\Gamma_X > 0$ holds, if $\mathbb{E}(x_{t-1}-\mu_x)(x_{t-1}-\mu_x)' > 0$ and $\Gamma := \mathbb{E}Y_t^\epsilon Y_t^{\epsilon'} > 0$ (see also lemma B.1). Since $\mathbb{E}(x_{t-1}-\mu_x)(x_{t-1}-\mu_x)' > 0$ holds by assumption, it remains to show that $\Gamma > 0$: Note that

$$Y_{t+1}^\epsilon = \begin{pmatrix} y_t^\epsilon \\ y_{t-1}^\epsilon \\ \vdots \\ y_{t-p+1}^\epsilon \end{pmatrix} = \underbrace{\begin{pmatrix} A_1 & \cdots & A_{p-1} & A_p \\ I_n & & 0 & 0 \\ & \ddots & & \vdots \\ 0 & & I_n & 0 \end{pmatrix}}_{:=A} \begin{pmatrix} y_{t-1}^\epsilon \\ y_{t-2}^\epsilon \\ \vdots \\ y_{t-p}^\epsilon \end{pmatrix} + \underbrace{\begin{pmatrix} I_n \\ 0 \\ \vdots \\ 0 \end{pmatrix}}_{:=B} \epsilon_t$$

and that the stability of $A(z)$ is equivalent to $|\lambda_i(A)| < 1$ for all i. Hence,

$$\Gamma = A\Gamma A' + B\Sigma_\epsilon B'.$$

Suppose that for some $h = (h_1', \dots, h_p')'$ with $h_i \in \mathbb{R}^n$, it holds that $h'\Gamma h = 0$. This implies that $\underbrace{h'A\Gamma A'h}_{=0} + \underbrace{h'B\Sigma_\epsilon B'h}_{=0} = 0$. Since $\Sigma_\epsilon > 0$, it follows that $B'h = h_1 = 0$. Due to $h'A\Gamma A'h = 0$, $A'h = (h_2', \dots, h_p', 0)' \in \ker\Gamma$. Thus, $h'A\Gamma A'h = h'AA\Gamma A'A'h + h'AB\Sigma_\epsilon B'A'h = 0$ together with the above reasoning implies that $h_2 = 0$. If we continue in an analogous way, we can show that $h = 0$ has to hold. □

Lemma B.3 *Consider the s.d.p. (7.14) A pair* $(\lambda \in \mathbb{R}, \Delta \in \mathcal{A}_{n*n})$ *is called* feasible *if and only if* $Q_0 + \Delta - \lambda I \geq 0$ *holds. Similarly we call* Δ *feasible for* λ *if and only if* (λ, Δ) *is feasible and finally* λ *is called feasible if and only if there exist a* $\Delta \in \mathcal{A}_{n*n}$ *s.th.* (λ, Δ) *is feasible.*

1. *The set of all* $\Delta \in \mathcal{A}_{n*n}$ *which are feasible for* λ *is bounded by* $\|\Delta\| \leq (n^2-1)\lambda_{\max}(Q_0 - \lambda I)$.

2. *The feasible* λ*'s are bounded from above by* $\lambda \leq \lambda_{\max}(Q_0)$

3. *The set of optimal points of the s.d.p. is non void, i.e. there exists a feasible pair* (λ^*, Δ^*) *such that* $\lambda^* = \sup\{\lambda \,|\, \exists \Delta \in \mathcal{A}_{n*n}$ *s.th.* $Q_0 + \Delta - \lambda I \geq 0\}$*. The optimum* λ^* *is bounded by* $\lambda_{\min}(Q_0) \leq \lambda^* \leq \lambda_{\max}(Q_0)$.

4. *The optimal value* λ^* *is a continuous function of* Q_0.

Proof. Ad (1): First, note that $\lambda_{\min}(X) + \lambda_{\min}(Y) \leq \lambda_{\min}(X+Y) \leq \lambda_{\min}(X) + \lambda_{\max}(Y)$ holds for symmetric matrices X, Y, see e.g. (Golub and VanLoan, 1996). Second, observe that $\max(|\lambda_{\min}(\Delta)|, |\lambda_{\max}(\Delta)|) = \|\Delta\| \leq -(n^2-1)\lambda_{\min}(\Delta)$ holds for all $\Delta \in \mathcal{A}_{n*n}$, since $0 = \operatorname{tr}\Delta = \sum_{k=1}^{n^2}\lambda_k(\Delta)$, where $\lambda_k(\Delta)$ denotes the kth eigenvalue of Δ. Therefore, we have $0 \leq \lambda_{\min}(Q_0 + \Delta - \lambda I) \leq \lambda_{\min}(\Delta) + \lambda_{\max}(Q_0 + \lambda I)$ and hence, $\|\Delta\| \leq -(n^2-1)\lambda_{\min}(\Delta) \leq (n^2-1)\lambda_{\max}(Q_0 - \lambda I)$.

Claim (2) follows from $0 \leq \|\Delta\|/(n^2-1) \leq \lambda_{\max}(Q_0 - \lambda I) = \lambda_{\max}(Q_0) - \lambda$.

To prove (3), let a sequence (λ_k, Δ_k) of feasible pairs be given such that $\lambda_k \leq \lambda_{k+1}$ and $\lim_k \lambda_k = \lambda^*$. Since $\|\Delta_k\| \leq (n^2-1)\lambda_{\max}(Q_0 - \lambda_k I) \leq (n^2-1)\left[\lambda_{\max}(Q_0) - \lambda_1\right]$ there exists a limiting point, Δ^* say, of the sequence (Δ_k). Since $\mathcal{S}_{n^2}^+$ is closed, it follows that $Q_0 + \Delta^* - \lambda^* I \geq 0$ and hence, (λ^*, Δ^*) is feasible. Finally, note that $\lambda^* \geq \lambda_{\min}(Q_0)$ by definition and that $\lambda^* \leq \lambda_{\max}(Q_0)$ holds by (2).

To prove (4), let a convergent sequence $Q_{0k} \to Q_{00}$ be given and let $(\lambda_k^*, \Delta_k^*)$ denote optimal points corresponding to Q_{0k} and let $(\lambda_0^*, \Delta_0^*)$ be optimal for the limit matrix Q_{00}. We have $\lambda_k^* = \lambda_{\min}(Q_{0k} + \Delta_k^*) \geq \lambda_{\min}(Q_{0k} + \Delta_0^*)$ and hence

$$\liminf_k \lambda_k^* \geq \lim \lambda_{\min}(Q_{0k} + \Delta_0^*) = \lambda_{\min}(Q_{00} + \Delta_0^*) = \lambda_0^*.$$

On the other hand, since $\lambda_{\min}(X) - \|Y\| \leq \lambda_{\min}(X+Y) \leq \lambda_{\min}(X) + \|Y\|$ holds for symmetric matrices X, Y, see again e.g. (Golub and VanLoan, 1996), $\lambda_k^* = \lambda_{\min}(Q_{0k} + \Delta_k^*) \leq \lambda_{\min}(Q_{00} + \Delta_k^*) + \|Q_{0k} - Q_{00}\| \leq \lambda_0^* + \|Q_{0k} - Q_{00}\|$ and thus

$$\limsup_k \lambda_k^* \leq \lambda_0^*.$$

□

The sets $\mathcal{S}_{n*n}$ and $\mathcal{A}_{n*n}$ are defined in appendix C. These sets are invariant under transformations of the form $X \to (S \otimes T)X(S \otimes T)'$. This means if $S, T \in \mathbb{R}^{n \times n}$ are two non singular matrices, then $Q \in \mathcal{S}_{n*n}$ holds if and only if $(S \otimes T)Q(S \otimes T)' \in \mathcal{S}_{n*n}$ and the same statement holds for matrices $\Delta \in \mathcal{A}_{n*n}$.

Suppose we have given two matrices $\Lambda, \Gamma \in \mathbb{R}^{n \times K}$ and let $U = (\Lambda_1 \otimes \Gamma_1, \ldots, \Lambda_K \otimes \Gamma_K) \in \mathbb{R}^{n^2 \times K}$ where Λ_k, Γ_k denote the k-th column of Λ, Γ respectively. It is trivial to see that $Q = UU' \in \mathcal{S}_{n*n} \cap \overline{\mathcal{S}_{n^2,K}^+}$ holds. The next lemma states that for $K \leq n$ also the reverse statement holds, given some regularity conditions:

Lemma B.4 *Let $Q \in \mathcal{S}_{n*n} \cap \mathcal{S}_{n^2,K}^+$ be given. If $K \leq n$ and if there exist non singular matrices $S, T \in \mathbb{R}^{n \times n}$ such that the first K rows and columns of $\bar{Q} = (S \otimes T)Q(S \otimes T)'$ form a positive definite $K \times K$ matrix, then there exist two matrices $\Lambda, \Gamma \in \mathbb{R}^{n \times K}$ such that Q has a factorization as $Q = UU'$, where $U = (\Lambda_1 \otimes \Gamma_1, \ldots, \Lambda_K \otimes \Gamma_K) \in \mathbb{R}^{n^2 \times K}$ and where Λ_k, Γ_k denote the k-th column of Λ, Γ respectively.*

Proof. By $Q \geq 0$ and $\operatorname{rk} Q = K$ there exists a factorization of Q as $Q = \tilde{U}\tilde{U}'$, $\tilde{U} \in \mathbb{R}^{n^2 \times K}$. The matrix $\tilde{U}$ is partitioned as

$$\tilde{U} = (\tilde{U}_1', \ldots, \tilde{U}_n')'$$

where $\tilde{U}_i = (\tilde{U}_{i1}', \tilde{U}_{i2}')'$ with $\tilde{U}_{i1} \in \mathbb{R}^{K \times K}$ and $\tilde{U}_{i2} \in \mathbb{R}^{(n-K) \times K}$.

First, note that following the above assumptions there exists a transformation $\tilde{U} \to \bar{U} = (S \otimes T)\tilde{U}$ such that $\bar{U}_{11} = I$ and $\bar{U}_{12} = 0$ holds. The block symmetry conditions imply $\bar{U}_i\bar{U}_j' = \bar{U}_j\bar{U}_i'$. For $i = 1$ these relations give $\bar{U}_{j2} = 0$ and $\bar{U}_{j1} = \bar{U}_{j1}'$ for all j. Furthermore we have $\bar{U}_{i1}\bar{U}_{j1}' = \bar{U}_{j1}\bar{U}_{i1}'$ for all i, j. From (Harville, 1997, Theorem 21.13.1.) we know that in this specific case there exists an orthogonal matrix $O \in \mathbb{R}^{K \times K}$ such that $O'\bar{U}_{i1}O = \bar{D}_i$, with $\bar{D}_i = \operatorname{diag}(\bar{d}_{i1}, \ldots, \bar{d}_{iK})$ diagonal, holds for all $i = 2, \ldots, n$. Hence,

$$\left(I_n \otimes \begin{pmatrix} O' & 0 \\ 0 & I_{n-K} \end{pmatrix}\right)(S \otimes T)\tilde{U}O = \begin{pmatrix} I_K \\ 0_{n-K \times K} \\ \hline \bar{D}_2 \\ 0_{n-K \times K} \\ \hline \vdots \\ \hline \bar{D}_n \\ 0_{n-K \times K} \end{pmatrix} = (L_1 \otimes G_1, \ldots, L_K \otimes G_K),$$

where for $j = 1, \ldots, K$, $L_j = (1, d_{2j}, \ldots, d_{nj})'$ and $G_j = e_j$, the jth unit vector of dimension n. Therefore,

$$U := \tilde{U}O = (S^{-1} \otimes T^{-1})\left(I_n \otimes \begin{pmatrix} O & 0 \\ 0 & I_{n-K} \end{pmatrix}\right)(L_1 \otimes G_1, \ldots, L_K \otimes G_K) = (\Lambda_1 \otimes \Gamma_1, \ldots, \Lambda_K \otimes \Gamma_K)$$

has the desired structure.

□

Lemma B.5 *Consider the linear matrix equation*

$$MD' + DM' = 0 \tag{B.1}$$

where for given $M \in \mathbb{R}^{n\times n}$, $n \geq 2$, we seek for antisymmetric solutions $D \in \mathcal{A}_n$. The zero matrix $D = 0$ is in any case a solution and it is the only solution for a generic set of M matrices; i.e. the set of matrices $M \in \mathbb{R}^{n\times n}$, which allow for a non zero solution $D \neq 0$ is a set of Lebesque measure zero.

The same statement holds true for the matrix equation

$$MD' + DM' + DD' = 0 \tag{B.2}$$

Proof. First, consider equation (B.1). To prove the claim we reverse the role of M and D, i.e. we assume that an antisymmetric matrix $D \in \mathcal{A}_n$ is given and we seek for the set of matrices $M \in \mathbb{R}^{n\times n}$ that satisfy (B.1). Since D is antisymmetric, it follows that D has an orthonormal eigenvector basis and that all eigenvalues of D are imaginary. The real block Schur decomposition (see e.g. (Golub and VanLoan, 1996)) of D therefore has the form

$$O'DO = \begin{pmatrix} d_1E & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & d_kE & 0 \\ 0 & \cdots & 0 & 0 \end{pmatrix}$$

where $2k$ is the rank of D, $\pm\sqrt{-1}d_i$ are the eigenvalues of D and

$$E = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

If we partition $\bar{M} = O'MO$ as

$$O'MO = \begin{pmatrix} \bar{M}_{11} & \cdots & \bar{M}_{1k} & \bar{M}_{1(k+1)} \\ \vdots & \ddots & \vdots & \vdots \\ \bar{M}_{k1} & \cdots & \bar{M}_{kk} & \bar{M}_{k(k+1)} \\ \bar{M}_{(k+1)1} & \cdots & \bar{M}_{(k+1)k} & \bar{M}_{(k+1)(k+1)} \end{pmatrix}$$

we see that M is a solution of the homogeneous equation (B.1) if and only if

$$\begin{array}{rll} \bar{M}_{ii} = m_iI_2 & \text{for} & 1 \leq i \leq k \\ \bar{M}_{ij} = -\frac{d_i}{d_j}E\bar{M}'_{ji}E & \text{for} & 1 \leq j < i \leq k \\ \bar{M}_{(k+1)j} = 0 & \text{for} & 1 \leq j \leq k \end{array}$$

where m_i, $1 \leq i \leq k$ and $\bar{M}_{ji}$, $j < i$ and $\bar{M}_{(k+1)(k+1)}$ may be chosen arbitrarily.

Now, let $\mathcal{M}_k$ denote the set of all M matrices which are solutions corresponding to an arbitrary rank $2k$ antisymmetric matrix D. By the above reasoning it follows that $\mathcal{M}_k$ may be represented as the image of a differentiable mapping, where the arguments are $O \in \mathcal{O}_n$, $d_i \neq 0$, m_i, $\bar{M}_{ij} \in \mathbb{R}^{2\times 2}$ for $1 \leq i < j \leq k$, $\bar{M}_{i(k+1)} \in \mathbb{R}^{2\times(n-2k)}$ for $1 \leq i \leq k$ and $\bar{M}_{(k+1)(k+1)} \in \mathbb{R}^{(n-2k)\times(n-2k)}$. Here $\mathcal{O}_n$ denotes the set of all orthogonal matrices $O \in \mathbb{R}^{n\times n}$, $O'O = I_n$, which is a differentiable manifold of dimension $n(n-1)/2$. Since only the $(n-2k)$–dimensional eigenspace of D corresponding to its $(n-2k)$ zero eigenvalues is uniquely defined, we may use this additional freedom to choose $\bar{M}_{(k+1)(k+1)}$ as an upper triangular matrix without losing any M matrix in $\mathcal{M}_k$. Thus, the domain of definition of this mapping is a differentiable manifold of dimension

$$n(n-1)/2 + k + k + (k(k-1)/2)4 + k(n-2k)2 + (n-2k)(n-2k+1)/2 = n^2 - k < n^2.$$

Therefore $\mathcal{M}_k$ is a set of Lebesque measure zero. Now the result follows by taking the union over all sets $\mathcal{M}_k$, $1 \leq 2k \leq n$.

For the second claim note that (B.2) is linear in M and that $M = -D/2$ is in any case a solution. Thus the solutions are of the form $M = -D/2 + M_0$, where M_0 is a solution of (B.1). Therefore we may use the same reasoning as above to prove the assertion. □

The above lemma shows that generically $D = 0$ is the only solution of the equations (B.2). However, as is shown in the proof there exist matrices M that allow for non zero solutions. Typically the number of solutions is finite which implies that there exists an open neighborhood of $D = 0$ such that there are no other solutions contained in this neighborhood. Yet, there exist also matrices M with infinitely many solutions and where in particular there exist solutions in any neighborhood of $D = 0$.

Appendix C

Basic Definitions and Frequently Used Notations

C.1 Concepts of convergence of sequences of random variables

Let in the following x_0 be a (real) scalar random variable and $(x_n)_{n\in\mathbb{N}}$ a sequence of (real) scalar random variables defined on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$.

1. Convergence in probability (plim):
   $\underset{n\to\infty}{\text{plim}}\, x_n = x_0$, if for all $\epsilon \in \mathbb{R}$, $\epsilon > 0$, $\lim_{n\to\infty} \mathbb{P}(|x_n - x_0| > \epsilon) = 0$.
2. The limit in mean square (l.i.m):
   $\underset{n\to\infty}{\text{l.i.m}}\, x_n = x_0$, if $\lim_{n\to\infty} \mathbb{E}(x_n - x_0)^2 = 0$.
3. Almost sure convergence (a.s.):
   $\lim_{n\to\infty} x_n = x_0$ a.s., if $\lim_{n\to\infty} \mathbb{P}(\omega \in \Omega \mid x_n(\omega) - x_0(\omega)) = 1$.
4. Convergence in distribution:
   $x_n \underset{n\to\infty}{\xrightarrow{d}} x_0$, if for any point $x \in \mathbb{R}$ of continuity of the distribution function $F_{x_0}(x)$ of x_0 it holds that $\lim_{n\to\infty} F_{x_n}(x) = F_{x_0}(x)$.

Consider also the following laws concerning these concepts of convergence:

- $\underset{n\to\infty}{\text{l.i.m}}\, x_n = x_0 \Rightarrow \underset{n\to\infty}{\text{plim}}\, x_n = x_0$, i.e. convergence in mean square implies convergence in probability.
- $\lim_{n\to\infty} x_n = x_0$ a.s. $\Rightarrow \underset{n\to\infty}{\text{plim}}\, x_n = x_0$, i.e. almost sure convergence implies convergence in probability.
- $\underset{n\to\infty}{\text{plim}}\, x_n = x_0 \Rightarrow x_n \underset{n\to\infty}{\xrightarrow{d}} x_0$, i.e. convergence in probability implies convergence in distribution.

For proofs of the above statements see e.g. (Davidson, 1994) or (Amemiya, 1985).

C.2 Matrices

- The symbols I_n, $0_{m\times n}$ denote the $n \times n$ identity matrix and the $m \times n$ zero matrix. If the respective size is clear from the context the indices may be omitted.

- Positive semidefinite (p.s.d.) and positive definite (p.d.) matrices: A square matrix A of dimension n, is said to be positive semidefinite, if for all $x \in \mathbb{R}^n$ $x'Ax \geq 0$ holds.
  Positive definite (p.d) matrices: A square matrix A of dimension n, is said to be positive definite, if for all $x \in \mathbb{R}^n$, $x \neq 0$, $x'Ax > 0$ holds.
  Given two matrices A and B of equal size, we say that $A \geq B$, if the difference $A - B$ is positive semidefinite, i.e. $A - B \geq 0$.

- The transposed, inverse and transposed inverse (if they exist) of a matrix A are denoted as A', A^{-1} and A^{-T}, respectively.
  For a positive semidefinite (p.s.d.) matrix, $H \geq 0$ say, a (symmetric) square root is denoted as $H^{1/2}$, i.e. $H = H^{1/2}(H^{1/2})'$.

- vec operator: The vec operator takes the columns of a matrix and stacks them column for column in a vector. Given, for instance, an $n \times n$-dimensional matrix A, vec(A) is the n^2-dimensional vector $\text{vec}(A) = (a_{11}, a_{21}, \dots, a_{n1}, a_{12}, \dots, a_{nn})'$.

- Commutation matrix $K_{m,n}$: $K_{m,n}$ is an $mn \times mn$ dimensional matrix defined such that for any $m \times n$ dimensional matrix A it holds that
$$\text{vec}(A') = K_{m,n}\text{vec}(A)$$
  or equivalentely
$$\text{vec}(A) = K_{n,m}\text{vec}(A').$$
  Note that $K_{m,n} = K_{n,m}^{-1}$.

- $\text{mat}_{m,n}$ operator: For $m \times n$-dimensional matrices, A, the $\text{mat}_{m,n}$ operator is the inverse operator to vec, i.e. $\text{mat}_{m,n}(\text{vec}(A)) = A$. In case of $m = n$, the subscripts are omitted.

- vech operator: For symmetric matrices $A \in \mathbb{R}^{n \times n}$, $A = A'$, the vech operator stacks the diagonal and lower diagonal entries, i.e. $\text{vech}(A) = (a_{11}, a_{21}, \dots, a_{n1}, a_{22}, \dots, a_{nn})' \in \mathbb{R}^{n(n+1)/2}$. The index function
$$I(k,l) = (k - l + 1) + \sum_{s=1}^{l-1}(n + 1 - s) \tag{C.1}$$
  gives the position where the (k, l)-th element of A is stored into.

- Duplication matrix G_n: G_n is the unique $n^2 \times n(n+1)/2$ matrix such that $G_n\text{vech}(M) = \text{vec}(M)$ holds for all symmetric $n \times n$ matrices M, and $G_n^+ = (G_n'G_n)^{-1}G_n'$ is a left inverse of G_n, see also (Harville, 1997, section 16.4)

- math operator: For symmetric matrices, $A = A'$, the math operator is the inverse operator to vech, i.e. math(vech(A)) = A.

- $\underline{\text{vech}}$ operator: This operator stacks the elements below the diagonal of a matrix, i.e. for a square matrix $A \in \mathbb{R}^{n \times n}$, $\underline{\text{vech}}(A) = (a_{21}, a_{31}, \dots, , a_{n1}, a_{32}, \dots, a_{n,(n-1)})'$.

- diag operator: The diagonal elements of a matrix are obtained with the diag operator, i.e. $\text{diag}(A) = (a_{11}, \dots, a_{nn})'$. With a slight abuse of notation we use diag also to construct diagonal matrices, i.e. $\text{diag}(x) = \text{diag}(x_1, \dots, x_k) \in \mathbb{R}^{k \times k}$ is a diagonal matrix with elements x_i on its diagonal.

- The Kronecker product: For two matrices A and B of dimension $m \times n$ and $p \times q$ respectively, the Kronecker product, $(A \otimes B)$, is of dimension $mp \times nq$ and defined as
$$(A \otimes B) = \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{pmatrix},$$

where a_{ij} is the (ij)-element of matrix A.

Let us just state some basic rules:

$$\begin{aligned} (A \otimes B)' &= (A' \otimes B'), \\ (A \otimes B)^{-1} &= (A^{-1} \otimes B^{-1}), \end{aligned}$$

and for matrices A, B, C, D with suitable dimensions:

$$\begin{aligned} \mathrm{vec}(ABC) &= (C' \otimes A)\mathrm{vec}(B), \\ \mathrm{tr}(ABCB') &= \mathrm{vec}(B)'(C' \otimes A)\mathrm{vec}(B). \end{aligned}$$

- The Hadamard product: For two matrices A and B of dimension $m \times n$ the Hadamard product, $(A \odot B)$, is of dimension $m \times n$ and defined as $(A \odot B)_{ij} = a_{ij} b_{ij}$, i.e. the ij-th element of $(A \odot B)$ is given by the product of the ij-th element of A and the ij-th element B.

- Eigenvalues of a symmetric matrix: Following the fundametal theorem of algebra the eigenvalues of a symmetric matrix are all real. Given a symmetric matrix A, one can always find an orthonormal matrix O, i.e. $OO' = O'O = I$, and some diagonal matrix Λ such that $A = O\Lambda O'$ holds, see e.g. (Harville, 1997, Theorem 21.5.7). Let the minimum and maximum eigenvalue of a symmetric matrix $A = A'$ be denoted as $\lambda_{\mathrm{min}}(A)$ and $\lambda_{\mathrm{max}}(A)$ respectively.

- The trace: The trace of a square matrix A of dimension n is defined to be the sum of the n diagonal elements of A, i.e. $\mathrm{tr}(A) = \sum_{i=1}^{n} a_{ii}$. Note that due to the fact that $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ for two matrices with suitable dimensions, the trace of a symmetric matrix is equal to the sum of its eigenvalues.

- The Frobenius norm: The Frobenius norm of a matrix A is defined as

  $$\|A\|_F = (\mathrm{tr}(AA'))^{\frac{1}{2}} = \left(\sum_{i,j} a_{ij}^2 \right)^{\frac{1}{2}} .$$

  Hence, the Frobenius norm of a symmetric matrix A of dimension n is given by, $\|A\|_F = \left(\sum_i \lambda_i^2\right)^{\frac{1}{2}}$, where λ_i, $i = 1, \ldots, n$, are the eigenvalues of A.

- The following sections introduce some symbols and acronyms that are in particular important for the second part of the thesis:

  Symmetric matrices: Let $\mathcal{S}_m \subset \mathbb{R}^{m \times m}$ denote the set of all symmetric $m \times m$ matrices. The set $\mathcal{S}_m^+ = \{Q \in \mathcal{S}_m, | \, Q \geq 0\}$ is the set of all positive semidefinite, symmetric matrices and $\mathcal{S}_{m,K}^+ \subseteq \mathcal{S}_m^+$ denotes the set of positive semidefinite, symmetric matrices of rank K. Note that $\overline{\mathcal{S}_{m,K}^+} = \mathcal{S}_{m,0}^+ \cup \cdots \cup \mathcal{S}_{m,K}^+$ and $\overline{\mathcal{S}_{m,m}^+} = \mathcal{S}_m^+ = \overline{\mathcal{S}_m^+}$ where $\overline{(.)}$ denotes the closure of the respective set in $\mathbb{R}^{m \times m}$.

  Antisymmetric matrices: A square matrix A of dimension n is said to be antisymmetric or skew symmetric, if $A = -A'$ holds. Note that by definition the diagonal elements of an antisymmetric matrix have to be zero.

  Let $\mathcal{A}_m \subseteq \mathbb{R}^{m \times m}$ denote the set of antisymmetric matrices.

  The Q matrices: We will often encounter square matrices Q, of size $n^2 \times n^2$, which will be partitioned into $n \times n$ square sub-blocks $Q_{ij} \in \mathbb{R}^{n \times n}$, i.e.

  $$Q = \begin{pmatrix} Q_{11} & \cdots & Q_{1n} \\ \vdots & & \vdots \\ Q_{n1} & \cdots & Q_{nn} \end{pmatrix} \tag{C.2}$$

  The elements of Q are indexed by $q_{ij,kl}$, where $q_{ij,kl}$ denotes the (k,l)-th entry of the (i,j)-th sub-block Q_{ij}.

Let $\mathcal{S}_{n*n}$ be the set of all such block-matrices Q that are symmetric and whose sub-blocks are symmetric too; i.e. $\mathcal{S}_{n*n} = \{Q \in \mathcal{S}_{n^2} \,|\, Q_{ij} = Q'_{ij} \text{ for all } 1 \leq i, j \leq n\}$.
In addition, let $\mathcal{A}_{n*n} = \{\Delta = \Delta' \,|\, \Delta_{ij} = -\Delta'_{ij} \text{ for all } 1 \leq i, j \leq n\}$; be the set of all symmetric block-matrices, whose sub-blocks are all antisymmetric.
Let $E_{ij} = u_i u'_j - u_j u'_i$, $i > j$, where $u_i = (0, \ldots, 0, 1, 0, \ldots, 0)'$ are the canonical unit vectors of $\mathbb{R}^n$. Then it is immediate that $\mathcal{A}_{n*n}$ is spanned by the matrices $(E_{ij} \otimes E_{kl})$ i.e.

$$\mathcal{A}_{n*n} = \left\{ \Delta = \sum_{1 \leq i < j \leq n} \sum_{1 \leq k < l \leq n} \omega_{ij,kl}(E_{ij} \otimes E_{kl}) \,\middle|\, \omega_{ij,kl} \in \mathbb{R} \right\}.$$

To simplify notation let $\Delta_1, \ldots, \Delta_{\bar{n}}$, $\bar{n} = (n(n-1)/2)^2$, denote the matrices $(E_{ij} \otimes E_{kl})$ for $i > j$ and $k > l$. Note that $Q = Q'$ is an element of $\mathcal{S}_{n*n}$ if and only if $\text{tr}(Q\Delta) = 0$ for all $\Delta \in \mathcal{A}_{n*n}$. Furthermore for $\Delta = \Delta'$, $\Delta \in \mathcal{A}_{n*n}$ holds if and only if $(\nu \otimes e)'\Delta(\nu \otimes e) = 0$ holds for all $\nu, e \in \mathbb{R}^n$.

Bibliography

Abu-Mostafa, Y. S., Atiya, A. F., Magdon-Ismail, M. and White, H.: 2001, Introduction to the special issue on neural networks in financial engineering, *IEEE Transactions on Neural Networks* **12**(4).

Akaike, H.: 1973, Information theory and an extension of the maximum likelihood principle, *in* B. Petrov and F. Csáki (eds), *2nd International Symposium on Information Theory*, Budapest: Acadèmiai Kiadó, pp. 267–281.

Akaike, H.: 1974, A new look at the statistical model identification, *IEEE Transactions on Automatic Control* **AC-19**, 716–723.

Amemiya, T.: 1985, *Advanced Econometrics*, Harvard University Press.

An, H. and Gu, L.: 1985, On the selection of regression variables, *Acta Mathematicae Applicatae Sinica* **2**(1), 27–36.

An, H. and Gu, L.: 1989, Fast stepwise procedures of selection of variables by using AIC and BIC criteria, *Acta Mathematicae Applicatae Sinica* **5**, 60–67.

Anderson, T. W.: 1951, Estimating linear restrictions on regression coefficients for multivariate normal distributions, *Annals of Mathematical Statistics* **22**, 327–351.

Anderson, T. W.: 1971, *An Introduction to Multivariate Statistical Analysis*, John Wiley & Sons.

Anderson, T. W. and Rubin, H.: 1956, Statistical inference in factor analysis, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* **V**, 111–150.

Baba, Y., Engle, R. F., Kraft, D. F. and Kroner, K. F.: 1991, Multivariate simultaneous generalised ARCH. University of California, San Diego: Department of Economics, Discussion Paper No. 89-57.

Bartlett, M. S.: 1937, The statistical conception of mental factors, *Br. J. Psychol.* **28**, 97–104.

Bartlett, M. S.: 1938a, Further aspects of the theory of multiple regression, *Proc. Cambridge Phil. Soc.* **34**, 33–40.

Bartlett, M. S.: 1938b, Methods of estimating mental factors, *Nature* **141**, 609–610.

Bauwens, L., Laurent, S. and Rombouts, J.: 2003, Multivariate GARCH models: A survey, *CORE Discussion Paper* **31**. revised April 2004.

Bollerslev, T.: 1986, Generalized autoregressive conditional heteroskedasticity, *Journal of Econometrics* **31**, 307–327. North-Holland.

Bollerslev, T., Engle, R. and Wooldridge, J.: 1988, A capital asset pricing model with time varying covariances, *Journal of Political Economy* **96**, 116–131.

Brillinger, D. R.: 2001, *Time series: data analysis and theory*, 2nd edn, Society for Industrial and Applied Mathematics.

Brockwell, P. and Davis, R.: 1989, *Time Series: Theory and Methods*, 2nd edn, Springer Verlag, New York.

Brown, B.: 1971, Martingale central limit theorems, *The Annals of Mathematical Statistics* **42**, 59–66.

Brown, P. J.: 1994, *Measurement, Regression, and Calibration*, Oxford University Press.

Burt, C.: 1950, The factorial analysis of categorical data, *Br. J. Psych. (Stat. Sec.)* **3**, 166–185.

Campbell, J. Y., Lo, A. W. and MacKinlay, A. C.: 1997, *The Econometrics of Financial Markets*, 2nd edn, Princeton University Press, Princeton New Jersey.

Comte, F. and Lieberman, O.: 2003, Asymptotic theory for multivariate GARCH processes, *Journal of Multivariate Analysis* **84**, 61–84.

Davidson, J.: 1994, *Stochastic Limit Theory, An Introduction for Econometricicans*, Oxford University Press.

Deistler, M. and Hamann, E.: 2005, Identification of factor models, *Journal of Financial Econometrics* **3**(2), 256–281.

Deistler, M. and Scherrer, W.: 1994, The prague lectures, ECONOMETRICS II. Lectures given at CERGE Prague 1992; mimeo Institute for Mathematical Methods in Econometrics, Research unit Econometrics and System Theory (EOS), University of Technology, Vienna.

Diebold, F.: 2000, 'big data' dynamic factor models for macroeconomic measurement and forecasting. Discussion of 'Extracting Business Cycle Indexes from Large Data Sets: Aggregation, Estimation, Identification' by Lucrezia Reichlin and 'Macroeconomic Forecasting Using Many Predictors' by Mark Watson, prepared for he World Congress of the Econometric Society, Seattle, August 2000.

Diebold, F. X. and Mariano, R. S.: 1995, Comparing predictive accuracy, *Journal of Business & Economic Statistics* **13**(3), 253–263.

Engle, R.: 1982, Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation, *Econometrica* **50**, 987–1008.

Engle, R.: 2003, Risk and volatility: Econometric models and financial practice, *Nobel Lecture* pp. 326–349.

Engle, R. F.: 2002, Dynamic conditional correlation - a simple class of multivariate GARCH models, *Journal of Business and Economic Statistics* **17**(5).

Engle, R. F. and Sheppard, K.: 2001, Theoretical and empirical properties of dynamic conditional correlation multivariate GARCH. FIN-01-027.

Engle, R. and Kroner, K.: 1995, Multivariate simultaneous generalized ARCH, *Econometric Theory* **11**, 122–150.

Engle, R., Ng, V. and Rothschild, M.: 1990, Asset pricing with a factor-ARCH covariance structure, empirical estimates for treasury bills, *Journal of Econometrics* **45**, 213–237.

Fama, E. F.: 1970, Efficient capital markets: A review of theory and empirical work., *J. Finance* **25**, 383–417.

Fama, E. F.: 1991, Efficient capital markets II, *J. Finance* **46**, 1575–1617.

Forni, M., Hallin, M., Lippi, M. and Reichlin, L.: 2000, The generalized dynamic-factor model: identification and estimation, *The Review of Economics and Statistics* **82**(4), 540–554.

Forni, M., Hallin, M., Lippi, M. and Reichlin, L.: 2001, The generalized dynamic factor model consistency and convergence rates.

Forni, M., Hallin, M., Lippi, M. and Reichlin, L.: 2003, The generalized dynamic factor model, one-sided estimation and forecasting. http://www.dynfactors.org/papers/papers.htm.

Forni, M. and Lippi, M.: 1999, The generalized dynamic factor model: representation theory. JEL Classification: C13, C33, C43.

Furnival, G. and Wilson, R.: 1974, Regressions by leaps and bounds, *Technometrics* **16**, 499–511.

Golub, G. H. and VanLoan, C. F.: 1996, *Matrix Computations*, third edn, John Hopkins University Press, Baltimore.

Gourieroux, C.: 1997, *ARCH Models and Financial Applications*, Springer Verlag.

Gourieroux, C. and Jasiak, J.: 2001, *Financial Econometrics, Problems, Models, and Methods*, Princeton Series in Finance, Princeton University Press.

Grinold, R. C. and Kahn, R. N.: 1999, *Active Protfolio Management*, 2nd edn, McGraw-Hill.

Hannan, E. J. and Deistler, M.: 1988, *The Statistical Theory of Linear Systems*, John Wiley & Sons, New York.

Harvey, D., Leybourne, S. and Newbold, P.: 1997, Testing the equality of prediction mean squared errors, *International Journal of Forecasting* **13**, 281–291.

Harville, D. A.: 1997, *Matrix Algebra from a Statistician's Perspective*, Springer.

Hendrickson, A. and White, P.: 1964, Promax: A quick method for rotation to oblique simple structure, *Br. J. statist. Psychol.* **17**, 65–70.

Hendry, D. and Krolzig, H.-M.: 2001, New developments in automatic general-to-specific modelling. Prepared for Economics and the Philosophy of Economics, edited by Bernt P. Stigum.

Horst, P.: 1965, *Factor Analysis of Data Matrices*, Holt, Rinehart and Winston, New York.

Hotelling, H.: 1935, The most predictable criterion, *Journal of Education Psychology* **26**, 139–142.

Hotelling, H.: 1936, Relations between two sets of variables, *Biometrika* **28**, 321–377.

Hull, J.: 2003, *Options, Futures and Other Derivatives*, 5 edn, Prentice Hall.

Hull, J. and White, A.: 1987, The pricing of options on assets with stochastic volatilities, *Journal of Finance* **42**, 281–300.

Izenman, A. J.: 1975, Reduced-rank regression for the multivariate linear model, *Journal of Multivariate Analysis* **5**, 248–264.

Kaiser, H.: 1958, The varimax criterion for analytic rotation in factor analysis, *Psychometrika* **23**, 187–200.

Kaul, G.: 1996, Predictable components in stock returns, *in* G. S. Maddala and C. R. Rao (eds), *Handbook of Statistics*, Vol. 14, Elsevier, Amsterdam, pp. 269–296.

Krca, M.: 2002, *Analyse und Modellierung von Regimeänderungen in Finanzzeitreihen*, Master's thesis, University of Technology of Vienna, Institute of Econometrics, Systems Theory and Operations Research.

Krolzig, H.-M. and Hendry, D.: 2001, Computer automation of general-to-specific model selection procedures, *Journal of Economic Dynamics & Control* **25**, 831–866.

Kroner, K. and Ng, V.: 1998, Modeling asymmetric comovements of asset returns, *The Review of Financial Studies* **11**(4), 817–844.

Lawley, D. N. and Maxwell, A. E.: 1971, *Factor Analysis as a Statistical Method*, 2nd edn, Butterworth & Co.

Lütkepohl, H.: 1993, *Introduction to Multiple Time Series Analysis*, 2nd edn, Springer-Verlag.

Mann, H. B. and Wald, A.: 1943, On the statistical treatment of linear stochastic difference equations, *Econometrica* **11**(3/4), 173–220.

Melino, A. and Turnbull, S.: 1990, Pricing foreign currency options with stochastic volatility, *Journal of Econometrics* **45**, 239–265.

Nankervis, J. and Savin, N.: 1988, The Student's t approximation in a stationary first order autoregressive model, *Econometrica* **56**, 119–145.

Nelson, D.: 1991, Conditional heteroskedasticity in asset returns: A new approach, *Econometrica* **59**, 347–370.

Nicholls, D. and Pope, A.: 1988, Bias in the estimation of multivariate autoregressions, *Australian Journal of Statistics* **30A**, 296–309.

Penm, J. and Terrell, R.: 1982, On the recursive fitting of subset autoregressions, *Journal of Time Series Analysis* **3**, 43–59.

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

Rao, C.: 1979, Separation theorems for singular values of matrices and their applications in multivariate analysis, *Journal of Multivariate Analysis* **9**, 326–377.

Reinsel, G. C. and Velu, R. P.: 1998, *Multivariate Reduced-Rank Regression, Theory and Applications*, Lecture Notes in Statistics 136, Springer-Verlag New York, Inc.

Sargent, T. J. and Sims, C. A.: 1977, *Business cycle modeling without pretending to have too much a priori economic theory*, Minneapolis: Federal Reserve Bank of Minneapolis, pp. 45–109.

Scherrer, W. and Deistler, M.: 1998, A strucure theory for linear dynamic errors-in-variables models, *SIAM J. Control Optim.* **36**(6), 2148–2175.

Scherrer, W. and Ribarits, E.: 2006, On the parametrization of multivariate garch models. MIMEO, Institute for Mathematical Methods in Economics Research unit Econometrics and System Theory (EOS) TU Vienna.

Schönfeld, P.: 1969, *Methoden der Ökonometrie, Band I, Lineare Regressionsmodelle*, Verlag Franz Vahlen GmbH, Berlin und Frankfurt a.M.

Schwarz, G.: 1978, Estimating the dimension of a model, *Annals of Statistics* **6**, 461–464.

Shleifer, A.: 2000, *Inefficient Markets, An introduction to behavioral finance*, Oxford University Press.

Singal, V.: 2003, *Beyond the Random Walk, A Guide to Stock Market Anomalies and Low-Risk Investing*, Oxford University Press.

Spearman, C.: 1904, General intelligence, objectively determined and measured, *Am. J. Psych.* **15**, 201–293.

Thomson, G. H.: 1951, *The Factorial Analysis of Human Ability*, London University Press.

Thurstone, L.: 1947, *Multiple Factor Analysis*, University of Chicago Press.

Tiao, G.: 2001, *A course in time series analysis*, Wiley series in probability and statistics, ed. D. Peña , G.C. Tiao and R.S. Tsay, chapter 14, Vector ARMA models.

Tibshirani, R.: 1996, Regression shrinkage and selection via the lasso, *J.R. Statist. Soc. B* **58**(1), 267–288.

Tjøstheim, D. and Paulsen, J.: 1983, Bias of some commonly-used time series estimates, *Biometrika* **70**, 389–399.

van der Leeden, R.: 1990, *Reduced Rank Regression with Structured Residuals*, Leiden: DSWO Press.

Vandenberghe, L. and Boyd, S.: 1996, Semidefinite programming, *SIAM Review* **38**(1), 49–95.

White, H.: 2000, A reality check for data snooping, *Econometrica* **68**(5), 1097–1126.

Wold, H.: 1938, *A Study in the Analysis of Stationary Time-Series*, Uppsala: Almqvist and Wiksells.

Ye, J.: 1998, On measuring and correcting the effects of data mining and model selection, *Journal of the American Statistical Association* **93**(441), 120–130.

Curriculum Vitae

Eva Maria Ribarits

Address

38, Op der Heed 38
L - 1747 - Luxembourg
Luxembourg
Email: `eva.hamann@tuwien.ac.at`

Personal Details

First names: Eva Maria
Second name: Ribarits (nee Hamann)
Gender: Female
Date of birth: 30 August 1978
Place of birth: Vöcklabruck, Austria
Citizenship: Austria
Marital Status: Married

Education

09/1988–06/1996	Comprehensive Secondary School with emphasis on languages in Vöcklabruck, Austria A-levels passed with distinction
09/1996–10/2001	Master of Science in Technical Mathematics, with emphasis on Mathematical Economics, Vienna University of Technology First and second diploma passed with distinction Master Thesis: *Time Series Models for the West German Economy* Supervisor: Ao.Univ.Prof. Dr.iur. Böhm Bernhard
Since 11/2001	Ph.D. studies in Technical Mathematics, with emphasis on Mathematical Economics, Vienna University of Technology Ph.D. Thesis: *Multivariate Modeling of Financial Time Series* Supervisor: O.Univ.Prof. Dipl.-Ing. Dr.techn. Manfred Deistler