DISSERTATION

Estimation of Constrained Factor Models

with application to

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

Institut für Wirtschaftsmathematik - E105-02
Forschungsgruppe Ökonometrie und Systemtheorie (EOS)
Technische Universität Wien

eingereicht an der Technischen Universität Wien
Fakultät für Mathematik und Geoinformation

von

Dipl.-Ing. Petra Pasching
Matrikelnummer 9625994
Kafkastr. 12/5/3/1, A-1020 Wien

Wien, am 8. Juni 2010 .....

To Nico and Lena,
who deserve my excuse
for all the time I spent
playing with econometrical models
instead of playing with them.

Contents

List of figures

List of tables

Deutsche Kurzfassung

Diese Dissertation befasst sich mit der Modellierung und der Vorhersage multivariater Zeitreihen mit einer großen Querschnittsdimension. Heutzutage betont die steigende Verfügbarkeit hochdimensionaler Daten die Notwendigkeit für die Anwendung und Entwicklung von Methodiken, um deren Informationsgehalt analysieren und erfassen zu können. Es ist hinreichend bekannt, dass Standardmethoden wie zum Beispiel Vektorautoregressive Modelle mit Exogenen Variablen (VARX Modelle) dem *Fluch der Dimensionalität* unterliegen, einem Begriff, der von Richard Bellman geprägt wurde. Das bedeutet im Fall unrestringierter VARX Modelle, dass die Anzahl der zu schätzenden Parameter mit zunehmender Zahl an endogenen Variablen quadratisch zunimmt. Eine Möglichkeit, diese Problematik zu umgehen oder abzuschwächen, besteht in der Verwendung von Modellen, die die Dimension des Parameterraums reduzieren. Im Rahmen dieser Dissertation werden zwei dieser Methoden beleuchtet, die als Faktorenanalyse zusammengefasst werden können, nämlich die *Hauptkomponentenanalyse* (*PCA*, aus dem englischen *principal component analysis*) und die *reduced rank regression analysis* (RRRA).

Im Falle der *PCA* wird eine Matrix von beobachteten Variablen durch eine Matrix niedrigerer Dimension so approximiert, dass die Höhe der erklärten Varianz maximiert wird. Die Lösung zu diesem Optimierungsproblem erhält man mithilfe einer Eigenwertzerlegung der Kovarianzmatrix der gegebenen hochdimensionalen Datenmatrix.
RRRA hingegen zerlegt die Koeffizientenmatrix eines linearen Regressionsmodells mit dem Ziel, diese durch eine Matrix von gegebener niedrigerer Dimension so anzunähern, dass möglichst viel der Variation der abhängigen Variablen erklärt wird. Die Schätzung eines solchen Modells basiert auf einer Singulärwertzerlegung der Koeffizientenmatrix.

Dennoch kann sich die Interpretation eines Faktormodells bei großer Variablenzahl trotz der deutlichen Reduktion der Parameteranzahl als schwierig herausstellen. Weiters könnten sehr kleine Einträge in einer Ladungsmatrix, deren Schätzung auf Stichprobendaten beruht, *verschmierte* Nullen des *'wahren'* Modells sein. Diese Überlegungen stellen die wesentlichen

Beweggründe für die Entwicklung von restringierten Faktormodellen mit dünn besetzten Ladungsmatrizen dar.
Oft existiert in empirischen Anwendungen zusätzliche a priori Information über die Struktur eines Faktormodells, welche voraussetzt, dass manche Faktoren nicht auf alle Variablen laden. Das bedeutet auch, dass man in so einem Fall bereits eine Vorstellung hinsichtlich der Interpretation der (latenten) Faktoren hat.
Als Beispiel für ein derartiges Vorwissen, mithilfe dessen man eine solch dünn besetzte Ladungsmatrix definiert, kann die Zugehörigkeit meherer Aktien zu zwei verschiedenen Branchen genannt werden. Hier kann man vereinfachend zwei Faktoren zugrundelegen, die jeweils eine der Branchen repräsentieren. In solch einem Fall postuliert man, dass die erste Spalte der Ladungsmatrix nur in jenen Zeilen Einträge ungleich Null enthält, die den Aktien der ersten Branche zugeordnet sind, und die zweite Spalte umgekehrt. Kann eine Ladungsmatrix jedoch vollständig in unabhängige Blöcke zerlegt werden, so besteht keine Notwendigkeit der Durchführung einer restringierten PCA, da ein unrestringiertes Modell für jede Targetgruppe getrennt berechnet werden kann. Daher beschränkt sich die eigentliche Anwendung dieser restringierter Modelle auf jene Fälle, in denen die Ladungsmatrix nicht gänzlich in einzelne Blöcke zerfällt. Aktien der Branchen *Telekommunikation* und *Technologie* eines Aktienindex können als praktisches Beispiel für die eben beschriebene Struktur einer Ladungsmatrix herangezogen werden. Es kann davon ausgegangen werden, dass einige der Aktien beiden Branchen zugeordnet werden können, wohingegen die meisten Aktien jeweils als nur zu einem der beiden Faktoren zugehörig klassifiziert werden.

Das Hauptaugenmerk dieser Dissertation liegt in der Entwicklung und Schätzung der zuvor genannten Techniken, die die Dimension des Parameterraums reduzieren, unter Berücksichtigung von zusätzlichen, a priori festgelegten Null-Restriktion an entsprechenden Positionen der Ladungsmatrizen. Es werden Optimierungsaufgaben mit Restriktionen definiert, die im unrestringierten Fall die herkömmliche Hauptkomponentenlösung oder reduced rank regression Lösung ergeben. Diese Probleme werden numerisch effizient gelöst und der Aspekt der Eindeutigkeit der erhaltenen Lösung wird analysiert. Außerdem wird sowohl für die *PCA* als auch für die *RRRA* ein Vorhersagemodell in Kombination mit einem auf Informationskriterien beruhenden Inputselektionsalgorithmus definiert.
Zum Abschluss werden anhand einer empirischen Anwendung auf Finanzzeitreihen die out-of-sample Modellanpassung und die Portfoliowertentwicklung des restringierten Hauptkomponentenmodells mit jener des unrestringierten verglichen.

Abstract

This thesis is concerned with the modeling and forecasting of multivariate time series with a large cross-sectional dimension. Nowadays, the increasing availability of high dimensional data sets underlines the necessity of applying and developing methodologies in order to analyze and administrate this huge number of variables. It is well known, that the number of parameters of standard methods such as, for example, *Vector Autoregressive Models with Exogenous Variables (VARX models)* are subject to the *curse of dimensionality*, an expression that was coined by Richard Bellman. This means in the case of unrestricted VARX models that the number of parameters, that have to be estimated, increases quadratically when additional endogenous variables are added to the model. One way to overcome this problem is given by models reducing the dimensionality of the parameter space. In this framework two of these methods, which can be summarized as *factor analysis*, are highlighted, namely *principal component analysis (PCA)* and *reduced rank regression analysis (RRRA)*.

In the case of *PCA* a matrix of observed variables is approximated by a matrix of lower dimension in such a way, that the amount of explained variance is maximized. The solution to this optimization problem is obtained with the help of the eigenvalue decomposition of the covariance matrix of the data.
RRRA is a technique that decomposes the coefficient matrix of a linear regression model with the aim of getting a coefficient matrix of a fixed lower rank than the original one and explaining as much variation of the response variables as possible. Estimation of this model class is related to a singular value decomposition.

Nevertheless, despite of a clear reduction of the number of parameters in factor models, interpretation can still be a difficult issue, if the number of response variables is relatively large. Moreover, small values in the loadings matrix of a factor model, whose estimation is based on sample data, could be blurred zeros of the *'true'* model. These aspects form the main motivation for developing restricted factor models with sparse matrices of loadings.

Abstract

In many cases of empirical applications exists additional a priori knowledge about the structure of a factor model, implying that some factors do not load on every variable. This also means that one has already a certain idea about the interpretation of the (latent) factors.
As an example for such a preknowledge defining a sparse loadings matrix, a set of assets belonging to two different branches may be considered. Then 2 factors can be expected where each factor is representing one of the branches. In such a case it could be postulated that the first column of the loadings matrix has just entries unequal to zero on those positions belonging to the assets of the first branch and zeros elsewhere, and column 2 the other way round. However, if the sparse loadings matrix of a PCA model can be decomposed entirely into separate blocks, there is no need for a restricted PCA model because an unrestricted model could be estimated for each target group separately. Thus, the main challenge consists in the estimation of models with overlapping zero blocks that cannot be decomposed entirely. As a practical example the assets of the branches telecommunication and technology of an equity index could be considered. It is natural that some assets can be assigned to both branches, whereas most of them can be classified as belonging to just one of the two branches.

The main focus of this thesis lies in developing and estimating the above mentioned dimension-reducing techniques with additional, a priori defined zero restrictions on certain entries of the corresponding loadings matrix. Optimization problems with restrictions, that lead in the unrestricted case to conventional *PCA* resp. *RRRA*, are defined and solved in a numerically efficient way. Furthermore, the aspect of uniqueness of the obtained result is analyzed and a forecasting model in combination with an input selection algorithm related to information criteria is stated both for the restricted principal component model and the restricted reduced rank regression model.
Finally, an empirical application to financial time series is presented comparing the out-of-sample fit and the performance values of a restricted versus an unrestricted principal component forecasting model.

Acknowledgements

Walk along paths not traveled by others before, in order to leave your own footprints.

(Antoine de Saint-Exupery)

Writing this thesis has not always been easy for me and that's why I want to express my deepest gratitude to some people who helped me and supported me throughout the past years.

First of all I want to thank my supervisor and mentor Prof. Manfred Deistler who assisted and encouraged me all the years with a lot of patience and in a sympathetic and understanding manner. He never lost faith in me and was abundantly helpful and offered inestimable support and guidance. Professor Deistler's comments and advice in all matters connected to this thesis are invaluable and I benefited enormously from his vast knowledge in different scientific areas. I also want to say thanks to Prof. Peter Filzmoser for his patience and help and for his valuable feedback and amendments. I'm especially grateful for the interest he took in my work.
Furthermore, I have to express my deepest gratitude to Dr. Nickolay Trendafilov, who reviewed this thesis at very short notice and who helped me to stick to my time schedule.

Moreover, I want to thank the other PHD and master students of the research unit Econometrics and System Theory, who helped me to clarify my thoughts and gave me many valuable ideas during the group seminars.

I also owe gratitude to my colleagues from *Financal Soft Computing GmbH* for all the fruitful discussions we had and for their backing I needed in order to withstand the multiple burden of being employed in a private company, writing a thesis and raising my two lovely children (without orden of importance). Many thanks to my colleagues from *C-QUADRAT* who gave me the chance to reduce my working hours to a part time employment in order to finish this thesis and who provided the data for the empirical part.

Last but not least I would like to thank all my friends and my family for their endless love and understanding. My grandfather always believed in me and I am sorry that he could not live to see the end of my studies. My parents always listened to my sorrows and doubts and supported me with their help and advice. My boyfriend Thomas and my friends Christina and Eva also deserve my gratitude for always being there for me and for pushing me to keep on writing.

Thanks to all of you for making this thesis possible!

Chapter 1

Introduction

Analysis and forecasting of high dimensional time series is a very important issue in areas such as finance, macroeconometrics or signal processing. As an example the huge quantity of assets or other financial instruments such as equities, currencies or commodities or the analysis of the behavior of consumers can be mentioned. One tool for modeling and analyzing multivariate time series is given by autoregressive models (with exogenous variables), called AR(X) models. The problem, that arises when using this model class, is called the *curse of dimensionality*. This means, that the number of parameters, which have to be estimated, depends in a quadratic way on the number of variables. Although a common way is to select appropriate subsets of the large number of variables and build smaller models, one runs the risk of overfitting, which is addressed by White [79] as the problem of *data snooping*.

So the need for dimension reduction becomes obvious. In the last century several methods have been developed for this purpose. Nowadays principal component analysis (*PCA*) and factor analysis are widely used techniques in data processing and reduction of dimensionality. The former can be interpreted as a generalized form of factor models, where the error component is correlated (no idiosyncrasy). Being aware of the fact, that different objectives are pursued when using principal component models resp. factor models, both will be addressed as *factor models* in this thesis.

Factor models were invented at the beginning of the twentieth century. Spearman [66] and Burt [11] applied these type of models in the area of psychology analyzing mental ability tests. The idea was then to find one common factor called *general intelligence* that should drive the outcome of the individual questions in such tests. Thurstone [74] generalized this framework by allowing for more than one factors.
A further generalization has been made by Geweke [31], Sargent and Sims [62], Brillinger [10]

and Engle and Watson [22] by using factor models in a time series context, which are called *dynamic factor models*.
With the so called *approximate factor model* or *generalized static factor model* Chamberlain and Rothschild [14] and Chamberlain [13] developed a new type of factor model by dropping the assumption of idiosyncrasy of the errors.
An overview on the classical factor model with idiosyncratic noise can be found in Maxwell and Lawley [50].

Nearly at the same time as the classical factor model was introduced, *PCA* was proposed by Pearson [56] and Hotelling [42] who analyzed biological relationships with this method. Pearson used *PCA* as a statistical tool for dimension reduction of multivariate data, whereas Hotelling generalized this approach to random variables instead of samples. A wider application of *PCA* has become possible in the last quarter of last century because of the increasing use of computational systems. This method is also known as *Karhunen -Loéve transform* in signal processing.

Reduced rank regression models were first developed by Anderson [4], who estimated a model by the maximum likelihood method assuming a lower rank of the matrix of coefficients in a linear model and multivariate normal distribution of the noise component. He distinguishes between the economic variables Y, which are used as dependent variables, and the noneconomic predictor variables X, that can be manipulated. Izenman [43] was the first who used the terminology *reduced rank regression* and he examined this model class besides Robinson [61] and Davies and Tso [18] in more detail. Further development of these models was proposed by Tsay and Tiao [76] and Ahn and Reinsel [1] who applied reduced rank regression in a stationary time series context. Johanson [44] estimated cointegrated reduced rank models and Stoica and Viberg [70] used this method in the area of signal processing. A quite comprehensive summary on reduced rank models was written by Reinsel and Velu [60] in 1998. Properties of the obtained estimators in the case, when the assumed rank of the coefficient matrix is misspecified, have been analyzed by Anderson [6].

Recent work by Forni et al. [25] and Forni and Lippi [28], Stock and Watson [68] and Forni et al. [27] explores the *generalized dynamic factor model* (GDFM), which is a dynamic factor model that replaces the uncorrelatedness of the noise components by a weak dependence. A generalization to state space and ARMA systems has been found by Zinner [82].

Although, in the case of a huge number of variables there are still many coefficients, which have to be estimated. For example, if a set of 50 assets is analyzed and 5 factors are specified, solely 250 parameters have to be estimated to get the so called loadings matrix, which defines the relationship between the variables and the latent factors. Moreover, it is quite a difficult

issue to interpret so many coefficients at a time in spite of the wide spread application of factor rotation, which enhances interpretability. This underlines the need for a more parsimonious model which will be the main aim of this thesis.
The idea presented here consists in imposing certain zero restrictions on predefined positions of the matrix of loadings, especially in the case of *PCA* and *reduced rank regression models*. This is one essential difference to existing literature, because up to now algorithms were developed, that find these zeros themselves, and no subjective a priori knowledge is available. As examples for existing research Jolliffe, Trendafilov and Uddin [46], Zou, Hastie and Tibshirani [85], d'Aspremont et al. [17] and [16] or Leng and Wang [51] can be mentioned. But practitioners often have an idea or the experience about the structure of such a loadings matrix. For example, in finance the fifty assets of the *Euro STOXX 50 Price Index* may depend on several factors, where one could be called the overall market and the others consist of the different branches, to which the assets belong to. So it is natural to use this additional information if it is available.
A further aspect, that distinguishes this work from other available literature, is the fact, that obtaining a structured loadings matrix is not the only focus. Apart from that, forecasting models for the response variables will be defined, estimated and evaluated. Of course the in-sample residual statistics will be worse than those of the unrestricted model, but out-of-sample an improvement of the goodness of fit of the models can be expected for certain reasons explained later on.

1.1 Summary of obtained results

In this thesis a simple and transparent but efficient algorithm (in terms of calculation time) is developed that satisfies the condition of obtaining a reasonable solution of a novel type of restricted principal component and reduced rank models. It is based on the idea of alternating least squares (ALS) and produces a sparse factor loadings matrix with a priori defined zero entries as desired, that cannot be reached by conventional methods such as factor rotation. Thus, interpretability can be enhanced in comparison with an unrestricted model provided that the definition of the structure of the loadings matrix is reasonable.
As already stated previously, further use of these models as forecasting models, in combination with an input selection algorithm related to the *Akaike* and *Schwarz information criterion*, is not common as current literature mainly limits itself to constructing sparse matrices of loadings. The proposed procedure is tested empirically with financial data, whereby the weekly returns of 14 world indices are chosen as response variables. Moreover, 17 inputs explaining the status of the economy and influencing the target variables, have been selected in order to generate future forecasts. The results of this research show via the comparison of a restricted PCA

model with an unrestricted one, that the restricted models can outperform the unrestricted ones in the sense of

- featuring better out-of-sample model statistics such as R^2 or Hitrate
- showing the tendency of producing better portfolio values if a simple long/short single asset strategy is applied.

1.2 Guide to the thesis

In the following sections of the present chapter a few comments on notation and terminology are made for better understanding. The unrestricted *PCA* model with its assumptions and properties will be explained in chapter 2. The main results of this thesis will be stated in chapter 3, where additional zero restrictions are imposed on the loadings matrix of a principal component model. An objective function for getting an optimal solution for restricted *PCA* similar to one of those defined in Okamoto [54] will be given and an algorithm for estimation of the free parameters is presented. In chapter 4 a two-step forecasting procedure for (unrestricted) *PCA* models as well as for restricted *PCA* models will be described. Moreover, an input subset selection algorithm similar to the one proposed by An and Gu [3] is introduced.
Reduced rank factor models will be pointed out in chapter 5. Analogous restrictions as in the case of *PCA* will be imposed on this model class in chapter 6. Chapter 7 contains a direct formulation of an unrestricted resp. a restricted reduced rank forecasting model for predicting the variables of interest.
Empirical results on real financial data concerning restricted principal component models are presented in chapter 8. Conclusions and further points of discussion are mentioned in chapter 9.

1.3 Notation and terminology

Let y_t be the realization of a N - dimensional random vector observed at instant in time t ($t = 1, \ldots, T$). Then a matrix of observations $Y = (y_1, ..., y_T)' \in \mathbb{R}^{T \times N}$ can be built, containing the relevant time series data, also called *targets*, *responses*, *dependent variables* or *output variables* further on. The transposition of a matrix is marked as $(.)'$. k denotes the dimension of the factor space, which leads to a factor matrix $F = (f_1, ..., f_T)'$ of dimension $T \times k$. If not stated otherwise, $X = (x_1, \ldots, x_T)' \in \mathbb{R}^{T \times s}$ refers to the matrix of *exogenous* or *explanatory variables*.
Big letters are used for matrices, small ones for vectors. I_r refers to the $r \times r$ identity matrix $\begin{pmatrix} 1 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & 1 \end{pmatrix}$. If the dimension of I_r is obvious, the subindex r can also be dropped for the convenience of the reader.

Estimators are flagged with $\hat{.}$ and $\bar{X} = \frac{1}{T}\sum_{i=1}^{T} x_i$ denotes the arithmetic mean vector of a sample matrix $X = (x_1, \ldots, x_T)'$.
$O(k)$ denotes the set of orthogonal matrices of order k, which means that for any $k \times k$ matrix $B \in O(k)$ the equality $B'B = BB' = I_k$ is valid.
With $rk(B)$ it is referred to the rank of a matrix B. $trace(B)$ stands for the trace of a square matrix B, which is calculated as the sum of its diagonal elements. The notation $card(x)$ or $card(B)$ counts the number of nonzero elements in a vector x or in a matrix B, respectively.

1.4 General framework of factor models

A model of the form

$$y_t = Lf_t + \epsilon_t, \qquad t = 1, \ldots, T \tag{1.1}$$

where the original variables y_t and the noise ϵ_t have length N, the factors f_t are of length $k < N$ and the so called loadings matrix L is of dimension $N \times k$, is called a *static factor model* applied in a time series context. The loadings matrix L as well as the factor scores f_t are unknown and therefore they are called latent variables. In a more compact way equation (1.1) can be reformulated as

$$Y = FL' + \epsilon, \tag{1.2}$$

with $Y = (y_1, \ldots, y_T)'$, $F = (f_1, \ldots, f_T)'$ and $\epsilon = (\epsilon_1, \ldots, \epsilon_T)'$. So a large number of target variables summarized in a matrix Y are approximated by a linear combination of a smaller number of factors F. The information loss obtained through this approximation is contained in ϵ. So the objective of building factor models is to approximate the original variables by lower dimensional factors in such a way, that the information loss is minimized.
Of course, such a model is not identifiable, if no additional assumptions on the parameters are made, because there are much more unknown parameters than known values. The different assumptions, that are made on the model classes, which are within the scope of this thesis, are explained in detail in the following chapters.
Before applying such a method to data one has to think about the reasonability of doing that. Naturally, the data should be homogenous, which could be expressed mathematically as having a nondiagonal covariance matrix. This can be tested with a chi-square test called *Bartlett's test of sphericity*. On the other hand there should not be too much dependency between the data, measured by the *Kaiser - Meyer - Olkin criterion*. This test is also known as *measure of sampling adequacy* and measures the relationship between correlations and partial correlations.
Having decided that a factor model is adequate for describing the data, one has to choose the number of factors k. Several methods are known that should give at least a hint about how to select the size of k. The naïvest way would be to try in an enumerative way several possible values and choose the number so, that the resulting model is the most satisfactory one.

A bit more elaborate is the so called *Kaiser* criterion, that suggest to take as many factors as there are eigenvalues of the correlation matrix of the data Y larger than one. The idea behind this criterion comes from *PCA* and can be explained by the fact, that the i^{th} eigenvalue of the correlation matrix defines the percentage of variance, explained by the i^{th} principal component. So there should be explained at least as much variance as can be explained by one of the variables itself.
With the *scree test* another well known method can be named, that determines the optimal number of factors in a graphical way with a line plot. It was first mentioned by Cattell [12] in 1966. Therefore the eigenvalues have to be ordered in terms of declining order of magnitude and then they are plotted componentwise. The number of factors is chosen by a method, which is also called *elbow criterion* and is demonstrated in figure 1.1.
Using some example data shows, that the decisions made on the different criteria are not always the same. In the case of the Kaiser criterion, 2 factors would be selected whereas the scree test suggests to select 3 factors. So these criteria give the user some hint about the size of k, but in the end the scientist has to choose with the help of his knowledge and experience the appropriate number of factors.

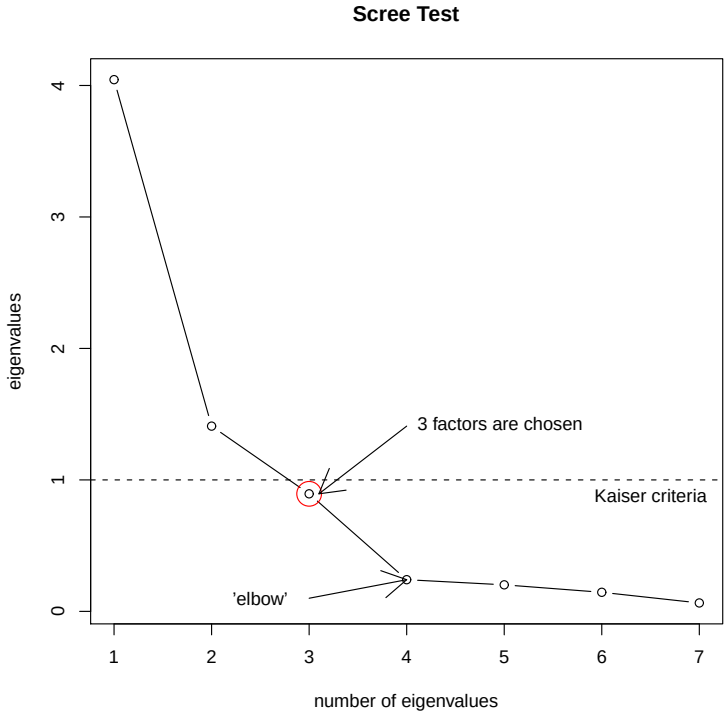

Figure 1.1: Example of a scree test in order to determine the number of factors in a factor model.

Chapter 2

Principal component analysis

Nowadays *principal component analysis (PCA)* is a widespread technique, applied in different disciplines of science, where high dimensional data sets are available and have to be analyzed. This methodology is quite famous last but not least because of its simple closed-form solution, described in the following sections.

2.1 The model

In the last century various ways of definitions and interpretations of principal components of a random vector as well as of a sample have been found.
Before pointing out the characteristics of a principal component model and its solutions, a few well known results of matrix theory will be recalled.

Given a square matrix $A \in \mathbb{C}^{N\times N}$, the (nonunique) solutions $\lambda_1, \ldots, \lambda_N \in \mathbb{C}$ resp. $\gamma_1, \ldots, \gamma_N \neq \mathbf{0} \in \mathbb{C}^N$ of the system of equations

$$A\gamma = \lambda\gamma \qquad \text{resp.} \qquad (A - \lambda I_N)\gamma = \mathbf{0}$$

are called the *eigenvalues* respectively *eigenvectors* of the matrix A.

Lemma 2.1.1. *Let A be a real, symmetric $N \times N$ matrix and let $\Lambda = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_N \end{pmatrix}$ and $\Gamma = (\gamma_1, \ldots, \gamma_N)$ be the joint matrix of eigenvalues and eigenvectors, respectively. Then*

$$A\Gamma = \Gamma\Lambda \qquad \textit{and} \qquad \Gamma'\Gamma = I_N$$

and the diagonal elements of Λ are the roots of the determinental equation

$$|A - \lambda_i I_N| = 0 \qquad i = 1, \ldots, N.$$

When restricting the elements in Λ so, that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N$, the matrix Λ is determined uniquely and Γ is determined uniquely except for postmultiplication by a matrix of orthogonal block matrices T:

$$T = \begin{pmatrix} T_1 & & 0 \\ & \ddots & \\ 0 & & T_r \end{pmatrix}, \tag{2.1}$$

where T_i, $i = 1, \ldots, r$, are orthogonal matrices of order m_i, r denotes the number of distinct eigenvalues of A and m_i their multiplicity.

If $rank(A) = k$, there exist k nonzero eigenvalues. Another property of such an eigenvalue decomposition is, that in the case of a symmetric, real matrix A all eigenvalues are real values. Moreover, eigenvectors corresponding to different eigenvalues are pairwise orthogonal. If A is additionally a positive (semi)definite matrix, then all eigenvalues are even positive (or zero), real values.

So when performing an eigenvalue decomposition of Σ, the covariance matrix of a N-dimensional random vector y, this results in a set of N nonnegative eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N \geq 0$ and a corresponding set of orthonormal eigenvectors $\gamma_1, \ldots, \gamma_N$ associated with $\lambda_1, \ldots, \lambda_N$, respectively.

For a set of eigenvalues $\{\lambda_1, \ldots, \lambda_N\}$ of a symmetric, positive definite matrix Σ with the property $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N \geq 0$, the eigenvalue λ_i is called the i^{th} *largest eigenvalue* of Σ. For any $k = 1, \ldots, N$, the set of eigenvectors $\{\gamma_1, \ldots, \gamma_k\}$ associated with the eigenvalues $\{\lambda_1, \ldots, \lambda_k\}$ is called *first k eigenvectors* and $\{\gamma_N, \gamma_{N-1}, \ldots, \gamma_{N-k+1}\}$ associated with $\{\lambda_N, \lambda_{N-1}, \ldots, \lambda_{N-k+1}\}$ *last k eigenvectors*, respectively.

With the help of these results principal components of a sample as well as of a random vector can be defined.

Definition 1
For any vector of observations $y_t \in \mathbb{R}^N$ at instant in time t ($t \in \{1, \ldots, T\}$) with mean $\hat{\mu} = \frac{1}{T}\sum_{t=1}^{T} y_t$ and for $\hat{\gamma_1}, \ldots, \hat{\gamma_N}$ being a set of N eigenvectors of its covariance matrix $\hat{\Sigma} = \frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{\mu})(y_t - \hat{\mu})'$, the scalar

$$v_j = \hat{\gamma_j}'(y_t - \hat{\mu}), \qquad j = 1, \ldots, N \tag{2.2}$$

is called the j^{th} *sample principal component of y_t*.

Definition 2
For any N-dimensional random vector y with mean $\mu = E(y)$ and for $\gamma_1, \ldots, \gamma_N$ being a set of N eigenvectors of its covariance matrix $\Sigma = E(y-\mu)(y-\mu)'$, the random variable

$$v_j = \gamma_j^{'}(y-\mu), \qquad j = 1, \ldots, N \tag{2.3}$$

is called the j^{th} *principal component of* y.

For means of simplicity, just the random version of principal components will be considered in this chapter, which can be seen as generalization of principal components of a sample. This issue can be deducted easily when considering the following:
Let $Y = (y_1, \ldots, y_T)'$ be a given $T \times N$ sample matrix, which can be regarded as simple data matrix, and define a random $N \times 1$ vector y by the probability distribution

$$Pr\{y = y_t\} = \frac{1}{T} \qquad for \quad t = 1, \ldots, T.$$

Then the sample principal components of y_t, $t = 1, \ldots, T$, are the t^{th} values taken by the principal components of the random vector y.

With the help of the definitions of the sample mean vector and the sample covariance matrix stated above this aspect can be proved easily, which is shown in more detail in [54]. Thus the following results are not only valid for random variables but also for samples.

For means of simplicity y is assumed to be centered from now on (i.e. $y_{new} = y_{old} - E(y_{old})$ and $\mu_{new} = \mathbf{0}$). This means geometrically that a non centered random variable is translated so, that its mean is a zero vector. Such a translation leaves the eigenvalues and eigenvectors of the covariance matrix unchanged. Then the following equalities hold in matrix notation:

$$\Sigma\Gamma = \Gamma\Lambda, \qquad \Gamma'\Gamma = I_N, \qquad \Lambda = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_N \end{pmatrix} \tag{2.4}$$

and

$$v = \Gamma' y, \tag{2.5}$$

where

$$v = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_N \end{bmatrix}.$$

When using all N eigenvectors of Σ, y can be reproduced exactly from the principal compo-

nents by multiplying the principal components with the transpose of the matrix of eigenvectors of Σ:

$$y = \Gamma v = \Gamma\Gamma' y. \tag{2.6}$$

The idea of reducing the possibly high dimensional random vector y to a lower dimensional space consists of neglecting those eigenvalues, which are small in order of magnitude compared to the others, and take just the first k important eigenvalues and eigenvectors.
Let $\Gamma = [\Gamma_1\Gamma_2]$ and $\Lambda = \begin{pmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{pmatrix}$, where $\Lambda_1 = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_k \end{pmatrix}$ contains the first k eigenvalues and Λ_2 the last $n-k$ eigenvalues respectively. In the same way the matrix Γ is devided into $\Gamma_1 = [\gamma_1, \ldots, \gamma_k]$ and $\Gamma_2 = [\gamma_{k+1}, \ldots, \gamma_N]$. Formally the construction of a factor model with the help of principal components can be stated as follows:

$$\begin{aligned} y = \Gamma\Gamma' y &= [\Gamma_1\Gamma_2] \begin{bmatrix} \Gamma_1{}' \\ \Gamma_2{}' \end{bmatrix} y \\ &= \underbrace{\Gamma_1}_{L} \underbrace{\Gamma_1' y}_{f} + \underbrace{\Gamma_2\Gamma_2' y}_{\epsilon} = Lf + \epsilon, \end{aligned} \tag{2.7}$$

which has the same functional form as equation (1.1). The reason why and in which context this decomposition of Σ is optimal, will be the central topic of section 2.2.

So *PCA* results in a set of uncorrelated factors and a matrix of loadings L with pairwise orthogonal columns. The amount of variance explained by each principal component can be deduced from the equation

$$\Gamma'\Sigma\Gamma = \Lambda.$$

This means that the variance of the first principal component is $var(v_1) = \gamma_1' \Sigma \gamma_1 = \lambda_1$ and thus the percentage of explained variance can be defined as

$$\frac{\lambda_1}{\lambda_1 + \ldots + \lambda_N}, \tag{2.8}$$

where $\lambda_1 + \ldots + \lambda_N$ stands for the whole variance of the multivariate variables. Taking into account the first k eigenvectors, the explained variance can be defined as $\sum_{i=1}^{k} \lambda_i$, $k = 1, \ldots, N$. Again a formula for the percentage of explained variance can be defined, according to equation (2.8):

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{N} \lambda_i}, \qquad k = 1, \ldots, N. \tag{2.9}$$

Apart from the methods explained in section 1.4, this measure of explained variance may give a hint, how to choose the number of principal components. One may select as many principal components, which are necessary to reach at least a certain level of explained variance, e.g. 90%.

Because of its property of orthogonality, *PCA* can also be interpreted as a process of finding sequentially a new orthogonal basis for the original variables so, that their variance is maximized. This means that first v_1 is calculated, which has maximal variance under all variables, that are in the space of y and that have unit length. Next v_2 is found, which has maximal variance among all variables, that are linear combinations of y with length 1, and which are orthogonal to v_1. This second principal component is also identical with the first principal component of the error component $y_1 = y - \gamma_1\gamma_1^{'}y$. Then the next principal component is obtained by requiring that it is orthogonal to v_1 and v_2 with unit length and that it maximizes the variance in $y_2 = y_1 - \gamma_2\gamma_2^{'}y_1$. This procedure can be continued until all N principal components are identified.

It was already defined before, that the random variable y will be mean adjusted so that the resulting variable has mean zero. What about the variances? If y is standardized, which means that each component of y has variance 1, the covariance matrix Σ will be replaced by the correlation matrix of y, say R. This means that each variable has the same weight in the optimization process and of course the eigenvalues of Σ and R are not identical. If the correlation matrix is used, the contribution of the j^{th} principal component to the total variation is given by

$$\frac{\lambda_j}{\sum_{i=1}^{N}\lambda_i} = \frac{\lambda_j}{N}.$$

In the same way the variance explained by the first k eigenvectors , $k = 1, \ldots, N$, can be described by the formula

$$\frac{\sum_{i=1}^{k}\lambda_i}{N}.$$

2.2 Optimality of principal components

The by now well known method of principal components may be obtained through different definitions and interpretations. Okamoto [54] classifies among the existing literature three types of objective functions, which lead as a result to the calculation of principal components and which will be described in more detail in this section. Firstly, he mentions *Variation Optimality*, which is one of the most used interpretations of principal components. This approach is also important in existing literature, when additional restrictions are imposed on the matrix of loadings. Secondly, the minimization of the so called *Information Loss* gives as a result principal components. This proposal together with a predefined structure of the loadings matrix will be the scope of research in this thesis (*see chapter 3*). Thirdly, principal components are obtained by defining the *Correlation Optimality*. This idea has not been investigated further in the context of additional restrictions on the model up to now in literature.

2.2.1 Variation optimality

At first a few definitions and lemmas are needed to formulate the principal theorem of this section.
The quotient

$$R_A(x) = \frac{x'Ax}{x'x}$$

with a square Matrix $A \in \mathbb{R}^{N\times N}$ and a $N \times 1$ vector x is called the *Rayleigh quotient* .
This quotient is strongly related to eigenvalues in the case of a Hermitian matrix A and their relationship is stated in the following two lemmas.

Lemma 2.2.1. *Let x be a real vector of dimension N and let A be a real, symmetric $N \times N$ matrix. Then*

$$\sup_x R_A(x) = \sup_x \frac{x'Ax}{x'x} = \lambda_1(A),$$

where sup denotes the supremum over all vectors $x \in \mathbb{R}^N$. This supremum is attained iff x is a first eigenvector of A.
Similarly, the following formula is valid:

$$\inf_x R_A(x) = \inf_x \frac{x'Ax}{x'x} = \lambda_N(A),$$

where inf denotes the infimum over all vectors $x \in \mathbb{R}^N$. Dually, the infimum is attained iff x is last eigenvector of A.

Lemma 2.2.2. *For any $k = 1, \ldots, N-1$, let $\{\gamma_1, \ldots, \gamma_k\}$ be a set of first k eigenvectors of a real, symmetric $N \times N$ matrix A. Then*

$$\sup_{\substack{x:\ \gamma_i' x=0 \\ i=1,\ldots,k}} \frac{x'Ax}{x'x} = \lambda_{k+1}(A).$$

The supremum is attained iff x is the eigenvector of A, which is associated with $\lambda_{k+1}(A)$.
On the other hand, if $\{\gamma_{N-k+1}, \ldots, \gamma_N\}$ is a set of last k eigenvectors of A, then

$$\inf_{\substack{x:\ \gamma_i' x=0 \\ i=N-k+1,\ldots,N}} \frac{x'Ax}{x'x} = \lambda_{N-k}(A).$$

Again, the infimum is attained iff x is an eigenvector of A, associated with the eigenvalue $\lambda_{N-k}(A)$.

With the help of these lemmas, the following theorem can be stated.

Theorem 2.2.1. *Let y be a real valued random vector of dimension N and consider the following optimization problem:*

$$\max_{\gamma \in \mathbb{R}^N} Var(\gamma' y)$$

$$s.t. \qquad \gamma'\gamma = 1.$$

Then the solution γ is given by the first eigenvector γ_1 of Σ, which is the covariance matrix of y.

Theorem 2.2.2. *Let $\{\gamma_1, \ldots, \gamma_k\}$ be a set of first k eigenvectors of Σ for fixed $k = 1, \ldots, N-1$. The solution to the problem*

$$\max_{\gamma \in \mathbb{R}^N} Var(\gamma' y)$$

$$\begin{aligned} s.t. \quad & \gamma'\gamma = 1 \\ & Cov(\gamma' y, \gamma_i' y) = 0 \qquad i = 1, \ldots, k \end{aligned}$$

is given by that eigenvector, that is associated with λ_{k+1} and that is orthogonal to $\{\gamma_1, \ldots, \gamma_k\}$.

With the help of these theorems the optimal procedure for finding sequentially $N \times 1$ vectors γ, that maximize the Rayleigh quotient, can be defined. So first the eigenvector corresponding to the first eigenvalue λ_1 will be selected. Next the eigenvector associated with λ_2 is chosen, then the one related to λ_3 and so on.

The next step consists of maximizing variation in a multivariate setup. This means, that instead of finding k vectors $\{\gamma_1, \ldots, \gamma_k\}$ subsequently, they should be optimized in one optimization process. Therefore a new matrix-valued objective function has to be defined.
Before mentioning two theorems, that give solutions to the problem of maximizing variation in a multivariate context, a lemma has to be stated in each case. Their proofs can be found in [54].

Lemma 2.2.3. *Let A be a nonnegative definite matrix of dimension $N \times N$ and let $X \in \mathbb{R}^{N \times k}$ $(k \leq N)$ be a matrix, whose columns have length 1, i.e.*

$$X'X = \begin{pmatrix} 1 & * & \ldots & * \\ * & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ * & \ldots & * & 1 \end{pmatrix}.$$

The off-diagonal elements of $X'X$, marked with an asterisk, can have any arbitrary value in $\mathbb{R}$. Then the following property is fulfilled:

$$|X'AX| \leq \prod_{i=1}^{k} \lambda_i, \tag{2.10}$$

where $|.|$ stands for the determinant of the given matrix and λ_i denotes the i^{th} eigenvalue of A, $i = 1, \ldots, k$.
If $rk(A) \geq k$, a necessary and sufficient condition for the equality sign in equation (2.10) is given by

$$X = \Gamma_k Q,$$

where $\Gamma_k \in \mathbb{R}^{N \times k}$ is a matrix of first k eigenvectors of A and $Q \in O(k)$. $O(k)$ is the set of all orthogonal matrices of order k, that is of all matrixes O with the property $O'O = OO' = I_k$.

With the help of this lemma the following theorem follows immediately with $X = B$ and $A = \Sigma = Cov(y)$:

Theorem 2.2.3. *For fixed $k \in 1, \ldots, N$ the solution of the optimization problem*

$$\max_{B \in \mathbb{R}^{N \times k}} |Var(B'y)|$$

$$s.t. \qquad B'B = \begin{pmatrix} 1 & * & \ldots & * \\ * & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ * & \ldots & * & 1 \end{pmatrix}.$$

is given by $B = \Gamma_k Q$, where the notation as well as the meaning of the parameters are explained in lemma 2.2.3.

Another possibility for defining an objective function, that results in principal components as its solution, is given by theorem 2.2.4.

Lemma 2.2.4. *Let X be an orthogonal $N \times k$ matrix with $k \leq N$, i.e. $X'X = I_k$. Then the following inequality holds:*

$$\lambda_i(X'AX) \leq \lambda_i(A) \qquad \text{for any} \quad i = 1, \ldots, k, \tag{2.11}$$

where $\lambda_i(X'AX)$ and $\lambda_i(A)$ denote the i^{th} eigenvalue of $X'AX$ and A, respectively.
A necessary and sufficient condition for obtaining equality in equation (2.11) for all i simulta-

neously is, that

$$X = \Gamma_k Q,$$

where Γ_k *and* Q *are defined as in lemma 2.2.3.*

Theorem 2.2.4. *For fixed* $k \in \{1, \ldots, N\}$*, the solution to the optimization problem*

$$\begin{aligned} \max_{B \in \mathbb{R}^{N \times k}} & \{\lambda_1, \ldots, \lambda_k\} \text{ simultaneously} \\ s.t. \quad & \{\lambda_1, \ldots, \lambda_k\} \text{ are the eigenvalues of } Var(B'y) \\ & B'B = I_k \end{aligned}$$

is given by $B = \Gamma_k Q$ *as in theorem 2.2.3.*

Note, that in theorem 2.2.4 a more restrictive side condition is needed compared to theorem 2.2.3. The aim here is not to maximize the determinant of the covariance matrix of the principal components, which is the product of its eigenvalues, but to maximize all the eigenvalues simultaneously. So for two matrices B_1 and B_2 the natural order $B_1 < B_2$ is valid, if for their eigenvalues $\{\lambda_1(B_1), \ldots, \lambda_k(B_1)\}$ resp. $\{\lambda_1(B_2), \ldots, \lambda_k(B_2)\}$ in decreasing order of magnitude the following inequalities hold:

$$\lambda_1(B_1) < \lambda_1(B_2), \ldots, \lambda_k(B_1) < \lambda_k(B_2).$$

So in all three cases objective functions are given, that result in an eigenvalue decomposition of the Covariance matrix of y. These solutions are always unique except for rotation with an orthogonal matrix Q.

2.2.2 Information loss optimality

Another category of objective functions, that gives as a result principal components, is measuring the loss of information, when reducing the dimensionality of the variables. The idea here is to approximate a given random $N \times 1$ vector y by a linear combination Ax of a $k \times 1$ random vector x with an unknown coefficient matrix $A \in \mathbb{R}^{N \times k}$. For $k < N$, the information loss can be defined as a function of the mean square error matrix $E(y - Ax)(y - Ax)'$, whereby its N eigenvalues are of special interest.
In 1964 Rao ([57]) proposes the following theorem:

Theorem 2.2.5 (Rao)**.** *Let* $k = 1, \ldots, N$ *be fixed. The solution of the problem*

$$\min_{A,B} \quad \|E(y - AB'y)(y - AB'y)'\|_F^2, \tag{2.12}$$

where A and B are real $N \times k$ matrices and $\|.\|_F$ denotes the Frobenius norm, is given by

$$\begin{aligned} AB'y &= \gamma_1 v_1 + \ldots + \gamma_k v_k = \\ &= \gamma_1 \gamma_1^{'} y + \ldots + \gamma_k \gamma_k^{'} y \\ &= \Gamma_1 \Gamma_1^{'} y, \end{aligned} \tag{2.13}$$

where Γ_1 is defined as in equation (2.7). $\{\gamma_1, \ldots, \gamma_k\}$ are the first k eigenvectors of the covariance matrix Σ and $\{v_1, \ldots, v_k\}$ denote the first k principal components.
The minimum of the objective function in equation (2.12) is given by $\lambda_{k+1}^2 + \ldots + \lambda_N^2$, where $\{\lambda_{k+1}, \ldots, \lambda_N\}$ denotes the set of the last $N - k$ eigenvalues of Σ.

Note, that here x is explicitly assumed to be a linear combination of y. Moreover, the matrix B in equation (2.12) is the same as the one in theorem 2.2.3 and in theorem 2.2.4. Thus, the equality $A = B = \Gamma_k Q$ with $Q \in O(k)$ holds.

Just one year later, Darroch [15] published a similar theorem, replacing the matrix norm in equation (2.12) by another function of the eigenvalues of a matrix: the trace.

Theorem 2.2.6 (Darroch)**.** *Let again $k \in \{1, \ldots, N\}$ be fixed. y and x are N - dimensional respectively k - dimensional random vectors.*
The minimization problem

$$\min_{Ax \in \mathbb{R}^{N \times 1}} \quad trace(E(y - Ax)(y - Ax)') \tag{2.14}$$

with $A \in \mathbb{R}^{N \times k}$ has again the solution

$$\begin{aligned} Ax &= \gamma_1 v_1 + \ldots + \gamma_k v_k = \\ &= \gamma_1 \gamma_1^{'} y + \ldots + \gamma_k \gamma_k^{'} y \\ &= \Gamma_1 \Gamma_1^{'} y, \end{aligned} \tag{2.15}$$

where Γ_1 and γ_i, $i = 1, \ldots, k$, are defined as in the previous theorem.

In Darroch's theorem x is assumed to be arbitrary, although in the optimal solution it is again a function of the original random vector y.

The most general theorem, deriving principal components as a solution of an optimization problem, that describes the loss of information of a lower dimensional approximation, dates from 1968 and was proposed by Okamoto and Kanazawa [55].

In order to prove this main theorem of the optimality of principal components in the sense of the loss of information the following two lemmas are required.

Lemma 2.2.5. *Let M be a real nonnegative definite matrix of dimension $N \times N$. A real valued function $f(M)$ is strictly increasing, i.e. $f(M_1) \geq f(M_2)$ if $M_1 \geq M_2$ and $f(M_1) > f(M_2)$ if additionally $M_1 \neq M_2$, and invariant under orthogonal transformations, i.e. $f(Q'MQ) = f(M)$ for any orthogonal matrix Q, if and only if $f(M)$ can be written as a function of the eigenvalues $\{\lambda_1(M), \ldots, \lambda_N(M)\}$ of M arranged in decreasing order of magnitude, which is strictly increasing in each argument, i.e. $f(M) = g(\lambda_1(M), \ldots, \lambda_N(M))$.*

As an example for such a function $f(M)$ the trace of a matrix $trace(M)$ or the Frobenius norm $\|M\|_F$ can be mentioned.

Lemma 2.2.6. *Let M, N and $M - N$ be real, symmetric and nonnegative definite matrices and $rk(N) \leq k$.*
Then the following properties are fulfilled:

- $$\lambda_i(M - N) \geq \lambda_{k+i}(M) \text{ for any } i \tag{2.16}$$
  and $\lambda_j(M) = 0$ for $j > N$.
- *A necessary and sufficient condition for getting equality in equation (2.16) simultaneously for all i is given by*
  $$N = \lambda_1(M)\gamma_1\gamma_1' + \ldots + \lambda_k(M)\gamma_k\gamma_k',$$

where $\lambda_i(M)$ and $\lambda_i(M-N)$ denote the i^{th} eigenvalue of M and $M-N$, respectively. $\{\gamma_1, \ldots, \gamma_k\}$ stands for the set of first k eigenvalues of M.

Theorem 2.2.7 (Okamoto and Kanazawa)**.** *Let $k \in \{1, \ldots, N\}$ be fixed. y and x are random $N \times 1$ respectively $k \times 1$ vectors. Now, consider the following problem:*

$$\begin{aligned} &\min_{Ax \in \mathbb{R}^{N\times 1}} \{\lambda_1(A,x), \ldots, \lambda_k(A,x)\} \text{ simultaneously} \\ &\quad s.t. \quad \{\lambda_1(A,x), \ldots, \lambda_k(A,x)\} \text{ are the eigenvalues of } E(y-Ax)(y-Ax)', \end{aligned}$$

where the coefficient matrix A is of dimension $N \times k$ and the eigenvalues $\lambda_i(A,x)$, $i = 1, \ldots, k$, are given as functions of the matrix A and of the random vector x.
Then the optimal approximation Ax of y is the same as in theorem 2.2.6.

Proof. The purpose of minimizing all the eigenvalues of $E(y - Ax)(y - Ax)'$ simultaneously can be reformulated as

$$\min_{A,x} \left\{ f_1(A,x) = f\left(E(y-Ax)(y-Ax)'\right)\right\} \tag{2.17}$$

with a real valued function f defined on the set of real nonnegative definite matrices as stated in lemma 2.2.5. Now it can be seen easily that the above theorem reduces to the one of Rao

for $f(.) = \|.\|_F$ and to the one of Darroch if $f(.) = trace(.)$.
If the rank of Σ is smaller than k, the solution is trivial. So let $rk(\Sigma)$ be $r > k$ from now on. Without loss of generality x can be assumed to have a covariance matrix of the form $Exx' = I_k$. If $Eyx' =: B$ the joint covariance matrix of $(y', x')'$ is given by

$$\Sigma_1 = \begin{pmatrix} \Sigma & B \\ B' & I_k \end{pmatrix} \geq 0$$

Thus the Schur complement of I_k in Σ_1, $\Sigma - BB'$, has to be nonnegative definite as well, i.e. $\Sigma - BB' \geq 0$.
Now the argument of the objective function can be modified further, namely

$$E(y - Ax)(y - Ax)' = \Sigma - BB' + (A - B)(A - B)' \geq \Sigma - BB'. \tag{2.18}$$

According to the definition of f the following inequality must hold:

$$f_1(A, x) = f(E(y - Ax)(y - Ax)') \geq f(\Sigma - BB') = f_2(B) \tag{2.19}$$

and equality is obtained if and only if $A = B$. So for $A = B$ equally $f_2(B) = f(\Sigma - BB')$ can be minimized with respect to B.
Because of the fact, that f is an increasing function of the eigenvalues of its argument, the optimum is obtained if the eigenvalues are simultaneously minimized. Applying lemma 2.2.6 gives

$$\lambda_i(\Sigma - BB') \geq \lambda_{k+i}(\Sigma) \qquad \forall i \tag{2.20}$$

and

$$\lambda_i(\Sigma - BB') = \lambda_{k+i}(\Sigma) \quad \Leftrightarrow \quad BB' = \lambda_1(\Sigma)v_1v_1' + \ldots + \lambda_k(\Sigma)\gamma_k\gamma_k', \tag{2.21}$$

where $\Gamma = (\gamma_1, \ldots, \gamma_N)$ is the matrix of eigenvectors of Σ related to the eigenvalues of Σ given in the diagonal of $\Lambda = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_N \end{pmatrix}$. Now BB' can be reformulated in a more compact way as

$$BB' = \Gamma\Lambda^{\frac{1}{2}} \begin{pmatrix} I_k & 0 \\ 0 & 0 \end{pmatrix} \Lambda^{\frac{1}{2}}V'$$

Therefore $f_2(B)$ is minimized by choosing the matrix B as

$$B = \Gamma\Lambda_1^{\frac{1}{2}} \begin{pmatrix} Q_1 \\ 0 \end{pmatrix},$$

where Λ_1 is defined equally to Λ but with ones in the diagonal on the positions where Λ

has zeros and Q_1 is a $k \times k$ orthogonal matrix. Note, that the minimum of F_2 is given by $f(\lambda_{k+1}\gamma_{k+1}\gamma'_{k+1} + \ldots + \lambda_N\gamma_N\gamma'_N)$, which is a function of the last $N-k$ eigenvalues of Σ. This matrix B is equal to the matrix A minimizing the original objective function $f_1(A,x)$. If a matrix H is defined as $H := \Gamma\Lambda_1^{-\frac{1}{2}} \begin{pmatrix} Q_1 \\ 0 \end{pmatrix}$ and a vector v as $v := H'y$, then $\Sigma H = A$ and $H'A = I_k$ are valid. The existence of a unique random vector x which satisfies the conditions $Exx' = I_k$ and $Eyx' = B$ has still to be proved. It is easy to see that the solution to x in order to minimize f_1) in equation (2.19) is given by v, because

$$Evv' = EH'yy'H = H'\Sigma H = H'A = I_k$$

and

$$Eyv' = Eyy'H = \Sigma H = A = B.$$

The uniqueness of $x = v$ follows from

$$\begin{aligned} E(v-x)(v-x)' &= Evv' - Evx' - Exv' + Exx' = I_k - EH'yx' - Exy'H + I_k = \\ &= I_k - H'A - A'H + I_k = I_k - I_k - I_k + I_k = 0. \end{aligned}$$

Thus, $f_1(A,x)$ is minimized by

$$Ax = \Gamma\Lambda_1^{\frac{1}{2}} \begin{pmatrix} Q_1 \\ 0 \end{pmatrix} \left[\Gamma\Lambda_1^{-\frac{1}{2}} \begin{pmatrix} Q_1 \\ 0 \end{pmatrix}\right]' y = \Gamma \begin{pmatrix} I_k & 0 \\ 0 & 0 \end{pmatrix} \Gamma' y = \gamma_1\gamma'_1 y + \ldots + \gamma_k\gamma'_k y.$$

□

One of the differences of this approach to the former one described in section 2.2.1 is the fact, that here a model is presented, that can be used directly in a forecasting context, and the setup as factor model becomes evident. *Variation optimality* leads to an optimal loadings matrix, called B, but there is no direct way of obtaining forecasts $\hat{y}$ for the target variables. However, within the *information loss* framework it becomes clear, that the principal components are obtained by premultiplying the original variables with a coefficient matrix B'. To get forecasts $\hat{y}$ of y, these principal components have to be multiplied by another matrix of coefficients A, which is equal to B. So, if k is chosen equal to N, no information loss would occur and the variables could be explained exactly by their principal components.

2.2.3 Correlation optimality

The third approach, that gives principal components as a result of an optimization problem, is given trough the so called correlation optimality.

Definition 3
The *multiple correlation coefficient* is a measure of the linear dependence between a one - dimensional random variable y and a certain $k \times 1$ random vector x. Let $E(y) = 0$ and $E(x) = \mathbf{0}$ and denote the common covariance matrix of y and x by

$$\tilde{\Sigma} = \begin{pmatrix} \sigma_{11}\Sigma_{12} \\ \Sigma_{21}\Sigma_{22} \end{pmatrix},$$

where $\sigma_{11} \in \mathbb{R}$, Σ_{12} and $\Sigma_{21} \in \mathbb{R}^{1\times k}$ resp. $\mathbb{R}^{k\times 1}$ and $\Sigma_{22} \in \mathbb{R}^{k\times k}$. Then the multiple coefficient of correlation $R(y, x)$ is defined as the square root of

$$R^2(y, x) = \frac{E(yx)[Var(x)]^{-1}E(xy)}{Var(y)} = \frac{\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}{\sigma_{11}}.$$

It is easy to see, that the coefficient of correlation is invariant under any nonsingular linear transformation of x.

To state the main theorem of this section, the following lemma is needed before.

Lemma 2.2.7. *Let y and x be N - dimensional respectively k - dimensional random vectors with the properties:*

$$E(y) = \mathbf{0}, \quad E(x) = \mathbf{0}, \quad E(yy') = \Sigma, \quad E(xx') = I_k, \quad \text{and} \quad E(yx') = A.$$

For $k \leq N$, $\lambda_1 \geq \ldots \geq \lambda_k > 0$ are the first k eigenvalues of Σ and $\{\gamma_1, \ldots, \gamma_k\}$ a set of first k eigenvectors of Σ associated with $\{\lambda_1, \ldots, \lambda_k\}$.
Then the existence of a matrix $Q \in O(k)$ such that

$$A = (\lambda_1^{0.5}\gamma_1, \ldots, \lambda_k^{0.5}\gamma_k)Q$$

and

$$x = Q'(\gamma_1/\lambda_1^{0.5}, \ldots, \gamma_k/\lambda_k^{0.5})'y$$

is a necessary and sufficient condition, that the equality

$$AA' = \lambda_1\gamma_1\gamma_1' + \ldots + \lambda_k\gamma_k\gamma_k' = \Gamma_1\Lambda_1\Gamma_1'$$

holds.

Now the main theorem of Okamoto [54] concerning *Correlation Optimality* can be stated.

Theorem 2.2.8. *For a fixed $k \in \{1, \dots, N\}$ assume that $E(y) = \mathbf{0}$ and $rk(\Sigma) = E(yy') \geq k$. The problem*

$$\max_{x \in \mathbb{R}^k} \quad \sum_{i=1}^{N} Var(y_i) R^2(y_i, x) \tag{2.22}$$

$$s.t. \quad E(x) = \mathbf{0}, \tag{2.23}$$

where R denotes the multiple coefficient of correlation, has the solution

$$x = T(v_1, \dots, v_k)' = Tv$$

with a regular $k \times k$ matrix T and $v = (v_1, \dots, v_k)'$ denotes the matrix of first k principal components of Σ.

Proof. Since

$$R^2(y_i, x) = \frac{E(y_i x)[Var(x)]^{-1} E(x y_i)}{Var(y_i)},$$

the objective function in equation (2.22) can be written as

$$\sum_{i=1}^{N} Var(y_i) R^2(y_i, x) = \sum_{i=1}^{N} E(y_i x)[Var(x)]^{-1} E(x y_i).$$

The coefficient of correlation is invariant under any nonsingular transformation and therefore $Var(x) = I_k$ can be assumed without loss of generality. This reduces the objective function to

$$\sum_{i=1}^{N} E(y_i x)[Var(x)]^{-1} E(x y_i) = \sum_{i=1}^{N} E(y_i x) E(x y_i) = trace(AA'),$$

if $E(yx') =: A$.
Since $Var(y - Ax) = \Sigma - AA'$ is nonnegativ definite, we deduce from lemma 2.2.6 that

$$trace(\Sigma - AA') = \sum_{i=1}^{N} \lambda_i(\Sigma - AA') \geq \sum_{i=1}^{N} \lambda_{k+i}(\Sigma) = \sum_{i=k+1}^{N} \lambda_i(\Sigma).$$

Moreover, $trace(\Sigma - AA') = trace(\Sigma) - trace(AA') = \sum_{i=1}^{N} \lambda_i(\Sigma) - trace(AA')$.
Hence,

$$trace(AA') \leq \sum_{i=1}^{k} \lambda_i(\Sigma) \tag{2.24}$$

and the equality sign holds for

$$AA' = \lambda_1\gamma_1\gamma_1^{'} + \ldots + \lambda_k\gamma_k\gamma_k^{'} = \Gamma_1\Lambda_1\Gamma_1^{'}. \tag{2.25}$$

Because of lemma 2.2.7 the optimal solution of our problem is given by

$$\begin{aligned} x &= Q'(\gamma_1/\lambda_1^{0.5}, \ldots, \gamma_k/\lambda_k^{0.5})'y = \\ &= Q'(v_1/\lambda_1^{0.5}, \ldots, v_k/\lambda_k^{0.5})' \\ &= Q'\Lambda_1^{-0.5}\Gamma_1^{'} = Q'\Lambda_1^{-0.5}v. \end{aligned}$$

Taking into account that we restricted x before so that $Var(x) = I_k$, all solutions for x are given by

$$x = T(v_1, \ldots, v_k)' = Tv$$

with a nonsingular matrix $T \in \mathbb{R}^{k\times k}$.

□

Thus the approach of Correlation Optimality is another alternative for defining an optimization problem, which leads to principal components as a result. It has to be taken into account that here the orthogonality of the principal components and of the loadings matrix is lost, if T is chosen nonorthogonal. When x is chosen as Tv with a regular matrix T, then the matrix of loadings A has to be set equal to $\Gamma_1 T^{-1}$ to get the same optimum as in the orthogonal case. Among the three approaches mentioned in this section *Correlation Optimality* is the least popular one and it has not been applied in relation with additional restrictions up to now.

2.3 Identifiability and rotation techniques

In section 2.2 it was shown, that principal components are found by performing an eigenvalue decomposition. So as already described on page 7 f. there occurs the first indeterminacy by the eigenvalue calculation itself. If γ is an eigenvector of A, then all multiples $c\gamma$ ($c \in \mathbb{R}$) are also eigenvectors of A. Therefore the eigenvectors are standardized so, that they have length 1, i.e. $\gamma'\gamma = 1$. Then there is still the possibility to change the signs of the eigenvector, which is often solved in numerical computations by making its first nonzero entry positive.
According to the finite-dimensional spectral theorem normal matrices A with the property $AA' = A'A$ with eigenvalues in $\mathbb{R}$ can be diagonalized. As a special case all symmetric matrices are normal and thus there exists for every symmetric real matrix A a real orthogonal matrix Γ such that $D = \Gamma' A \Gamma$ is a diagonal matrix. This means, the algebraic multiplicity of each eigenvalue, which is the multiplicity of a root of the characteristic polynomial, has to be

equal to the geometric multiplicity, which is the dimension of the space that is spanned by the eigenvectors, which are associated with such a multiple eigenvalue. So in spite of choosing the eigenvectors of length 1 with their first entry positive, the eigenvectors are not unique in the case of multiplicities larger that 1. As stated on page 8, all matrices $\tilde{\Gamma}$ obtained by postmultiplication of Γ with a matrix of orthogonal block matrices T_i are feasible matrices of eigenvectors.

The second source of indeterminacy was mentioned in section 2.2. In all three cases of optimality of principal components the loadings matrix A is always obtained uniquely up to an orthogonal rotation matrix Q. So using the notation of the equations (2.4) and (2.7) the following equalities hold:

$$\begin{aligned}\Sigma = \Gamma\Lambda\Gamma' &= \Gamma_1\Lambda_1\Gamma_1' + \Gamma_2\Lambda_2\Gamma_2' = \\ &= \Gamma_1\Lambda_1\Gamma_1' + \Sigma_\epsilon = \\ &= \Gamma_1 Q\Lambda_1 Q'\Gamma_1' + \Sigma_\epsilon \\ &= \tilde{\Gamma}_1\Lambda_1\tilde{\Gamma}_1' + \Sigma_\epsilon\end{aligned}$$

with $\tilde{\Gamma}_1 = \Gamma_1 Q$ and Q is an orthogonal $k \times k$ matrix. This orthogonal matrix Q should not be mistaken with the orthogonal matrix T before. In the former case T rotates the eigenvectors so, that the resulting matrix is still a solution for an eigenvalue decomposition, whereas Q rotates the eigenvectors so, that a new orthogonal basis ΓQ for Σ is found, which explains the same amount of variance as Γ, but this new basis is not necessarily a solution to the eigenvalue problem in equation (2.4). The set of feasible matrices T is a subset of the set of possible matrices Q and that's why it is sufficient to have a closer look at rotation matrices and rotation techniques of principal components or of factor models in general.

The question, that arises now, is how to choose the loadings matrix Γ_1 and thus the principal components $\Gamma_1' y$ in the infinite number of possible matrices. A possible answer lies in the interpretation of the result. The loadings matrix often lacks interpretability. Even for an advanced mathematician it is a difficult task to analyze for example a data set with 20 variables and 5 principal components, which would result in a loadings matrix with 100 entries. To overcome this problem, there exist several ways in literature to rotate the loadings matrix with a matrix in such a way that the factor loadings become more interpretable. This means that the aim is to get a more structured matrix of loadings, which makes the model easier to understand. So one may desire to obtain in each column a few large values and the others should be comparably small. So an interpretation for each factor can be found more easily by taking into account just the few variables that are correlated highly with the corresponding factors. Another way of defining structuredness may consist of finding for each variable (i.e. in each row of the loadings matrix) one factor, on which it loads high, and on the rest of the

factors it should load as low as possible.

Definition 4
A real matrix $R \in \mathbb{R}^{k \times k}$ is called a *rotation matrix*, if the following properties are fulfilled:

- the length of vectors and the angle between them remain unchanged, i.e.

$$\forall\ x, y \in \mathbb{R}^k : \qquad \langle Rx, Rx \rangle = \langle x, x \rangle$$
$$\langle Rx, Ry \rangle = \langle x, y \rangle$$

- the orientation remains unchanged, i.e. $|R| = 1$.

Thus a rotation matrix is an orthogonal matrix, whose determinant is 1.

As an example for such orthogonal rotation methods *varimax rotation*, *equimax rotation* and *quartimax rotation* can be named among others.

Sometimes such an orthogonal rotation may not be satisfactory. Consider for example the graphic in figure 2.2. It shows the coordinates of the loadings matrix corresponding to the orthogonal factors $F1$ and $F2$ for 7 variables, which are illustrated by the small black arrows. The two factors are represented by the orthogonal coordinate axes. Because of the acute angle between these arrows, an orthogonal rotation may not lead to the required result concerning interpretability. If the orthogonality of the factors is not needed, a linear transformation of the loadings matrix can be performed.

Definition 5
The premultiplication of a vector $x \in \mathbb{R}^k$ with a real nonsingular matrix $R \in \mathbb{R}^{k \times k}$ is called a *linear transformation* or *oblique rotation* of the vector x, i.e. $x' = Rx$. In general, the length of the transformed vectors and the angles between them have changed in comparison with the original ones.

Remark. In literature such a linear transformation is often called an *oblique rotation* and therefore the classical rotation will be called *orthogonal rotation* further on to stress the orthogonality of the matrix. When not specifying a certain type of rotation, just the terminology *rotation* will be used.

In figure 2.2 the red arrows indicate the new, oblique factors after rotation and it is obvious, that interpretation is easier than when applying an orthogonal rotation at the cost of obtaining correlated factors. *Promax rotation*, *oblimin rotation* or *procrustes rotation* are a few examples

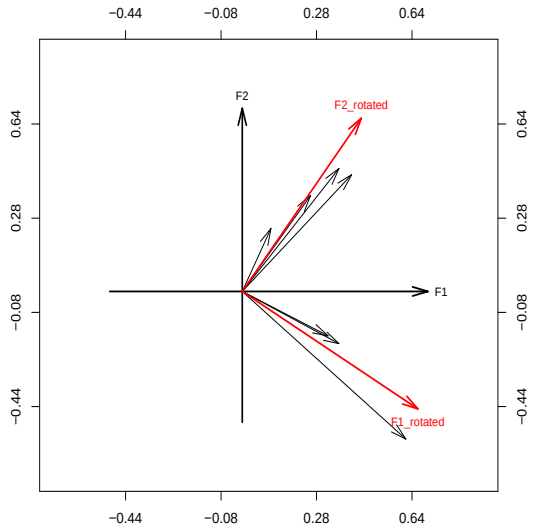

Figure 2.1: Example of an (orthogonal) varimax rotation in the case of 2 factors

of oblique rotation methods.

The following two methods are well known procedures for an orthogonal respectively oblique rotation technique of a loadings matrix and that's why they are described in more detail here.

2.3.1 Varimax rotation

This type of orthogonal rotation was developed by Kaiser [49] in 1958 and modified by Horst [41] in 1965. It seeks to rotate the factors in such a way, that the sum of the deviations of the squared entries of the loadings matrix to its corresponding column means is maximized.
Denote with $\tilde{\Gamma}_1 \in \mathbb{R}^{N\times k}$ an unrotated matrix of loadings and with γ_{ir} the element in the i^{th} row and the r^{th} column of the matrix of loadings $\Gamma_1 = \tilde{\Gamma}_1 R$, which is obtained by rotation of $\tilde{\Gamma}_1$ with an orthogonal $k \times k$ matrix R. Moreover, a scalar d_r can be defined as

$$d_r = \sum_{i=1}^{N} \gamma_{ir}^2.$$

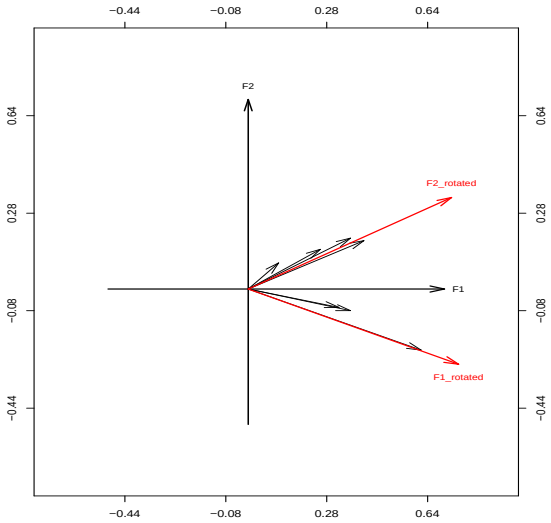

Figure 2.2: Example of an (oblique) promax rotation in the case of 2 factors

Then the maximization problem, that gives as a result the rotation matrix of varimax rotation, can be described by

$$\begin{aligned} \max_{R \in \mathbb{R}^{k \times k}} \quad & \sum_{r=1}^{k} \left[\sum_{i=1}^{N} \left(\gamma_{ir}^2 - \frac{d_r}{N} \right)^2 \right] \\ s.t. \quad & R'R = I_k. \end{aligned}$$

Due to this procedure some very high values and some very small values are obtained in each column and thus interpretation becomes easier.

2.3.2 Promax rotation

In contrast to varimax rotation, promax rotation of Hendrickson and White [39] is an oblique rotation. Here the structure of the loadings matrix is simplified further at the expense of correlated factors. Promax rotation starts with a varimax rotation resulting in a loadings matrix Γ_1. Next a Matrix S is defined whose entries are given by

$$s_{ir} = |\gamma_{ir}^{j-1}|\gamma_{ir}, \tag{2.26}$$

where j is some integer that is larger than 1 and in empirical applications normally chosen smaller or equal to 4. The elements s_{ir} have the same sign as γ_{ir} and the same absolute value as γ_{ir}^j.
Then the factors should be rotated with a regular matrix R in such a way, that for each

$r = 1, \ldots, k$ the r^{th} column of the matrix product $\Gamma_1 R$ is as similar as possible to the r^{th} column of S in a least square sense.
Thus the (oblique) rotation matrix R is given by

$$R = (\Gamma_1' \Gamma_1)^{-1} \Gamma_1' S. \tag{2.27}$$

As a consequence the covariance matrix of the factors Σ_f can be calculated by

$$\Sigma_f = (R'R)^{-1}, \tag{2.28}$$

which is different from I_k because of j being larger than 1. If the variances of the factors should still be equal to 1, it is feasible to rescale the rotation matrix R in an adequate manner.

Having a look at the definitions of the two rotation methods, it is obvious that they are not adequate if there is a notably dominating first factor. In this case it seems more reasonable to exclude this first factor from rotation and rotate just the remaining $k-1$ columns of the factor matrix.

2.4 Criticism

Up to now the structure of principal component models and its derivation have been described in detail in this chapter. To sum up, the following properties of principal components can be mentioned:

Principal component analysis (PCA)

- reduces the number of observed variables to a smaller number of principal components, which account for most of the variance of the observed variables. Components, which account for maximal variance, are retained while other components accounting for a small amount of variance are not retained. The amount of variance, that is explained by each component, is measured by the eigenvalues of the covariance matrix of the data.
- should be applied when (subsets of) variables are highly correlated.
- needs no underlying probability distribution of the data beyond the second moments; therefore it is called a *non-parametric method*.
- minimizes in the case of a given sample of observations the sum of the squared perpendicular distances to the space spanned by the principal components.
- becomes better interpretable when rotating the obtained solution in a suitable way.

What are the disadvantages or drawbacks of PCA? Which further improvements can be made? One may claim, that the data and the principal components are just connected in a linear way. To overcome that, nonlinear methods like *kernel PCA* have been developed (see [2]), which is out of the scope of this thesis. The absence of a probability distribution can both be interpreted as weakness or as strength of the method.
As already mentioned in section 2.3, one of the main difficulties of an unrotated PCA solution lies in the unability of interpreting the results in the case of high dimensional data. So rotation should ensure that afterwards there are a few large values in the matrix of loadings and that the others are small and thus unimportant. But what happens, if there is a priori knowledge available about the structure of the principal component model. An experienced scientist or economist or whatever may know, which variables load on which factor, if the meaning of the individual factors is clear.[1] For example, if an asset manager wants to analyze 20 assets, where 10 belong to the branch of technology and the others to the branch of telecommunications, it seems reasonable to define a first factor representing the market and two other factors containing the information of the two sectors mentioned before. Then one may assume, that the return[2] of an asset of the technology group may be a linear combination of the market return and some 'average return' of the technology sector, but it may not depend on the price movements of the telecommunications branch. Such a time series, measuring the average return[3] of a sector, can be interpreted as a sector index. The independence of a target variable on a factor can be forced by restricting the corresponding element of the loadings matrix to zero. Of course, the insample goodness of fit of the restricted model will be worse than the unrestricted one, but the forecasts of such a restricted model may be even better, if the true underlying model has the proposed structure with exact zeros in its matrix of loadings.

Setting such zero restrictions on the loadings matrix of a factor model will be called *sparse factor model* from now on. The model, its properties and estimation methods will be the central topics of chapter 3.

[1] The usual rotated PCA solution may, for example, give already hints on the meaning of the factors.

[2] A return is calculated as relative difference of e.g. asset prices over time. Denoting with p_t the price of an asset at time t, the return of this asset at time t, r_t, is calculated as $r_t = (p_t - p_{t-1})/p_{t-1}$.

[3] The use of 'average' does not imply, that the factor has to be calculated as arithmetic mean of different time series. More sophisticated ways of aggregating the information in the data are imaginable, when interpreting a factor.

Chapter 3

Sparse principal component analysis

Before going more into detail, an explanation about the meaning of the term *sparse principal component analysis* will be given. In literature the term 'sparse' refers to a coefficient matrix, that is used to build linear combinations either of the original variables or of the principal components, that has many zero entries and a few that are unequal zero.
Thurstone [75] suggested in 1947 five criteria to define a simple structure of a matrix of loadings. According to these criteria, a loadings matrix is simple if

- each row contains at least one element, that is zero
- in each column the number of zeros is larger or equal to the number of columns k
- for any pair of factors there are some variables with zero loadings on one factor and significant loadings on the other factor
- for any pair of factors there is a large proportion of zero loadings, if the number of factors is not too small
- for any pair of factors there are only a few variables with large loadings on both factors.

Nevertheless, the understanding of sparseness here in this thesis is slightly different from the one of Thurstone and will be explained in more detail later on.

The degree of sparsity addresses the number of elements that are not zero. Especially in small restricted PCA models the degree of sparsity can be quite large compared to bigger models taking into account the overall number of parameters. Such a sparse matrix Γ_1 may

look like

$$\begin{array}{c} \\ y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{array} \begin{array}{c} \begin{array}{ccc} f_1 & f_2 & f_3 \end{array} \\ \begin{pmatrix} * & * & 0 \\ * & * & 0 \\ * & * & 0 \\ * & 0 & * \\ * & 0 & * \\ * & 0 & * \end{pmatrix} \end{array}$$

An asterisk denotes the nonzero elements in the matrix. This would mean, that the variables 1 to 4 depend on the first and on the second factor, whereas variables 5 and 6 depend on the first and on the third factor. Note, that here the first factor can be interpreted as general factor or market factor, which explains all the variables, in contradiction to the criteria defined by Thurstone. Nevertheless, in practical applications it may make sense and the simplicity of the structure is not affected a lot if all variables load on that factor.
If the loadings matrix of a set of variables can be decomposed entirely in single blocks, a PCA for the variables of each block can be performed separately and no restricted PCA is necessary. For example, if Γ_1' is assumed to be of the form

$$\left(\begin{array}{ccc:ccc:ccc} * & * & * & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hdashline 0 & 0 & 0 & * & * & * & * & 0 & 0 & 0 \\ \hdashline 0 & 0 & 0 & 0 & 0 & 0 & 0 & * & * & * \end{array}\right),$$

one would carry out a PCA for the variables 1 to 3, a second one for the variables 4 to 7 and another one for the variables 8 to 10.

As described in section 2.3 the entries of a matrix of loadings are in general not zero, but there exists the possibility to rotate the factors and thus the loadings matrix so, that (nearly) exact zeros are obtained. With varimax or promax rotation, which are explained before, it will not be possible to get exact zeros. The following section shows an algorithm, that performs such an oblique rotation to (nearly) zeros.

3.1 Oblique rotation based on a pattern matrix

In practice it's often desirable to rotate the loadings matrix $\Gamma_1 \in \mathbb{R}^{N \times k}$ in such a way that a-priori specified elements will be or at least will come close to zero. Therefore a pattern matrix has to be constructed, that contains zeros to define restricted elements and ones otherwise. In the case of a 8×3 loadings matrix the r^{th} row of the transpose of such a pattern matrix P

could be defined by

$$p' = [1 \quad 0 \quad 1 \quad 1 \quad 0 \quad 0 \quad 1 \quad 1]$$

The aim is now to find a transformation matrix $S \in \mathbb{R}^{k \times k}$, so that the rotated loadings matrix $\Gamma_1^* = \Gamma_1 S$ has values equal or near zero on those positions, where the pattern matrix has zero entries.
To get small values in the above specified positions of the rotated loadings matrix, an optimization problem can be defined, that chooses S in such a way, that the sum of squares of the restricted elements of each column of Γ_1^* is minimized subject to the sum of squares of all elements being held constant.

Therefore matrices $\Gamma_{1,r}$ are defined that would in the above example be of the form

$$\Gamma'_{1,r} = \begin{pmatrix} * & 0 & * & * & 0 & 0 & * & * \\ * & 0 & * & * & 0 & 0 & * & * \\ * & 0 & * & * & 0 & 0 & * & * \end{pmatrix},$$

where asterisks define the original values of the loadings matrix Γ_1. It's obvious that multiplication of $\Gamma_{1,r}$ with the r^{th} column of the rotation matrix s_r produces zero values in the desired positions.
Minimizing the objective function, that models the criteria stated above, is equal to maximizing the sum of squares of the unrestricted elements subject to the sum of squares of all elements being held constant for all columns $r = 1, \ldots, k$.

This leads to the following optimization problem:

$$\begin{aligned} \max_{s_r} \quad & [s'_r(\Gamma'_{1,r}\Gamma_{1,r})s_r] \\ s.t. \quad & s'_r(\Gamma'_1\Gamma_1)s_r = \gamma'_r\gamma_r, \end{aligned}$$

where γ_r indicates the r^{th} column of the loadings matrix Γ_1. The maximization problem can be reformulated by use of a Lagrange multiplier:

$$\max_{s_r,\alpha_r} \quad \left[\left(s'_r(\Gamma'_{1,r}\Gamma_{1,r})s_r\right) - \alpha_r * \left(s'_r(\Gamma'_1\Gamma_1)s_r - \gamma'_r\gamma_r\right)\right]. \tag{3.1}$$

Thus the following derivatives have to be set equal to zero for all $r = 1, \ldots, k$:

$$\frac{\partial}{\partial s_r}[s'_r(\Gamma'_{1,r}\Gamma_{1,r})s_r - \alpha_r s'_r(\Gamma'_1\Gamma_1)s_r] = 0.$$
$$\frac{\partial}{\partial \alpha_r}[\alpha_r(s'_r(\Gamma'_1\Gamma_1)s_r - \gamma'_r\gamma_r)] = 0$$

Solving the first of the above equations we get

$$(\Gamma'_{1,r}\Gamma_{1,r})s_r = \alpha_r(\Gamma'_1\Gamma_1)s_r. \tag{3.2}$$

The second one results as expected in the side condition. Equation (3.2) can be rewritten as

$$H_r s_r = \alpha_r s_r, \tag{3.3}$$

where $H_r = (\Gamma'_1\Gamma_1)^{-1}(\Gamma'_{1,r}\Gamma_{1,r})$.

This defines an eigenvalue problem and in order to maximize the objective function in equation (3.1), α_r has to be the largest eigenvalue of H_r and s_r the eigenvector corresponding to the optimal α_r. This becomes clearer if equation (3.2) is premultiplied by s_r. Apparently α_r can be seen as the ratio of $s_r(\Gamma'_{1,r}\Gamma_{1,r})s_r$ to $s_r(\Gamma'_1\Gamma_1)s_r$, which should be as large as possible. If the r^{th} column of the pattern matrix has exactly $k-1$ zeros, the optimal α_r is 1. In the case of more than $k-1$ zeros the Lagrange multiplier reaches a value between 0 and 1.

Due to the facts, that the eigenvalues of symmetric matrices can be calculated more easily and the numeric advantage of getting real eigenvalues in the case of symmetric matrices, it seems reasonable to transform H_r into a symmetric matrix by decomposing $\Gamma'_1\Gamma_1$ as TT' with lower triangular matrices T. This can be reached by means of a QR decomposition. Now we define a matrix

$$W_r = T^{-1}(\Gamma'_{1,r}\Gamma_{1,r})T'^{-1}, \tag{3.4}$$

which is symmetric. Moreover, W_r has the same latent roots α_r as H_r and $T'^{-1}u_r = s_r$, where u_r denotes the latent vector of W_r.

The procedure described in this section is one of the well known rotation techniques in literature. But how good does it work? To analyze that aspect, the basic aim of sparse factor rotation will be recalled with the help of an example.

Example So let Γ_1 be a general 7×3 loadings matrix shown in table 3.1. Then the aim is to find an oblique rotation matrix R, whose columns are not vectors of zeros, so that $\Gamma_1 R$ is sparse. Firstly, a boundary of 0.015 is chosen to restrict all elements in this loadings matrix, that are smaller than this value, to zero. According to Thurstones criteria for simplicity, the obtained matrix is still not simple, but afterwards another example with harder restrictions will be given.

	PC1	PC2	PC3
y1	0.0194	0.6141	−0.0012
y2	−0.0015	0.0093	0.9986
y3	0.8323	−0.1106	0.0059
y4	0.3620	0.0533	−0.0152
y5	−0.0052	0.4540	0.0335
y6	0.0701	0.6320	−0.0361
y7	0.4135	0.0457	0.0116

Table 3.1: Example of a loadings matrix rotated with varimax

	PC1	PC2	PC3
y1	0.0264	0.5296	−0.0011
y2	0.0000	0.0000	0.9985
y3	0.8309	−0.4963	−0.0035
y4	0.3625	−0.1267	−0.0192
y5	0.0000	0.4006	0.0337
y6	0.0773	0.5212	−0.0366
y7	0.4141	−0.1584	+0.0070

Table 3.2: Example of a loadings matrix after rotation based on a pattern matrix

Thus the following equality should hold:

$$\Gamma_1 R = \Gamma_1 \begin{pmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{pmatrix} = \begin{pmatrix} * & * & 0 \\ 0 & 0 & * \\ * & * & 0 \\ * & * & * \\ 0 & * & * \\ * & * & * \\ * & * & 0 \end{pmatrix} \tag{3.5}$$

The result of the above described oblique rotation is shown in table 3.2. What happens? The first two columns are rotated as expected and zeros are obtained in the desired positions. But the third column has no exact zeros in the a priori defined positions.

Let equation (3.5) be written as a system of equations for each column of the loadings

matrix. Then the following system is obtained:

$$\begin{aligned} 0.0194r_{13} + 0.6141r_{23} - 0.0012r_{33} &= 0 \\ 0.8323r_{13} - 0.1106r_{23} + 0.0059r_{33} &= 0 \\ 0.4135r_{13} + 0.0457r_{23} + 0.0116r_{33} &= 0. \end{aligned} \tag{3.6}$$

In general, this system of equations will be nonsingular and thus the only vector $(r_{13}, r_{23}, r_{33})'$ that fulfills equation (3.6) exactly, would be $(0, 0, 0)'$, according to basic results of linear algebra. But this is no feasible solution and thus the algorithm described above gives some approximation as a solution. On the other hand, the first two columns of the matrix of loadings have less than $k = 3$ zeros, namely 2 and 1, respectively. In the case of the first principal component the kernel[1] of the matrix of coefficients is one - dimensional, because the corresponding system of equation consists of 2 (in general) linear independent equations. That's why the first vector of the rotation matrix is up to its sign equal to the vector of the null space, that has length one.[2] If the number of zeros in a column is $k - 2$ as in the second column of the matrix of loadings, where just 1 restriction is set, the null space of the corresponding matrix of coefficients is two - dimensional. The second column of the rotation matrix also has length 1 and is built as a linear combination of two vectors of the null space. Now it is easy to conduct, that in the case of more than k zeros, no exact zeros can be generated either.

To summarize these considerations, the oblique rotation technique presented in this section just gives exact zeros if the number of zeros in each column is less than or equal to $k-1$, which can be deduced from simple results of algebra. However, if k or more zeros are desired, which is the interesting case in practise, just small values can be achieved and there is no rule about the closeness of them to zero.

This is quite unsatisfactory and thus the aim of this thesis is to find another algorithm, that is more restrictive and that gives exact zeros, if more than $k - 1$ entries are restricted in a column of the matrix of loadings.

3.2 Historical review

First a historical overview will be given about the research that has been done in the last few decades on the topic of sparse principal component analysis. There are mainly found two types of restricting formulations. One type is founded according to the formulation of a maximization

[1] The kernel of a matrix is also called *null space*.

[2] This can be verified easily by performing a singular value decomposition $\Gamma_1 = UDV'$ and taking the k^{th} column of V.

problem, where the variance of the principal components is maximized, as described in section 2.2.1. The other one is related to the minimization of the loss of information, similar to the optimization model specified in section 2.2.2. As already mentioned earlier, there does not exist work on restricted PCA in the context of correlation optimality in literature up to now (see also section 2.2.3).

3.2.1 Variance based formulations

In 2000 Jolliffe and Uddin [47] developed the so called *simplified component technique*, which is abbreviated as *SCoT*. It can be seen as an alternative to rotated principal components. *SCoT* maximizes the variance of the principal components as in theorem 2.2.2, but with an additional penalty function, which is a multiple of one of the simplicity criteria of rotation such as e.g. varimax.

Just three years later, Jolliffe et al. [46] proposed a modified principal component technique based on the LASSO . Here *LASSO* stands for *Least Absolute Shrinkage and Selection Operator*. This method was introduced by Tibshirani [77] in 1996 in combination with regression analysis and sets a boundary to the sum of absolute values of the coefficients. This L_1 type restriction may cause that some of the coefficients of the loadings matrix are estimated as zero. The methodology of Jolliffe et al. is known as *SCoTLASS* and the name stresses the enhancement of *SCoT* by adding an additional LASSO restriction:

$$\max_{\gamma_i \in \mathbb{R}^N} Var(\gamma_i' y) \quad \text{successively for all } i = 1, \ldots, k$$

$$\begin{aligned} \text{s.t.} \quad & \gamma_i' \gamma_i = 1 \\ & \gamma_i' \gamma_j = 0 \qquad \text{for all } j < i \leq k \\ & \sum_{l=1}^{N} \gamma_{li} \leq t \end{aligned}$$

for some tuning parameter t. If $t < 1$ no solution will be obtained, whereas if $t = 1$ exactly 1 element will be unequal zero in each column. Whenever t is chosen larger or equal to $\sqrt{N}$, the optimization problem results in the unrestricted PCA solution and for values of t between 1 and $\sqrt{N}$ the number of zeros will vary between 0 and $N - 1$. This is an algorithm, that produces exact zeros. But it has the disadvantage of many local optima in optimization and high computational costs.

In 2007 D'Aspremont, Ghaoui, Jordan and Lanckriet [17] found a direct formulation for

sparse PCA using semidefinite programming. They define an optimization problem

$$\max_{\gamma_i \in \mathbb{R}^N} Var(\gamma_i' y) \quad \text{successively for all } i = 1, \ldots, k$$

$$\begin{aligned} \text{s.t.} \quad & \gamma_i' \gamma_i = 1 \\ & \gamma_i' \gamma_j = 0 \qquad \text{for all } j < i \leq k \\ & card(\gamma_i) \leq m, \end{aligned}$$

where $card(\gamma_i)$ stands for the number of elements in γ_i, that are different from zero, and m is a sparsity controlling parameter. This problem above is NP-hard[3] and that's why a semidefinite relaxation of it is derived, that contains a weaker but convex side condition:

$$\max_{\Gamma \in \mathbb{R}^{N \times N}} trace(\Sigma \Gamma)$$

$$\begin{aligned} \text{s.t.} \quad & trace(\Gamma) = 1 \\ & \mathbf{1}' |\Gamma| \mathbf{1} \leq m \\ & \Gamma' \Gamma = I_N \\ & \Gamma \succeq 0, \end{aligned}$$

where $\mathbf{1}$ stands for a N - dimensional vector of ones and $|\Gamma|$ denotes a matrix whose elements are the absolute values of Γ. Thus the cardinality or L_0 norm constraint is replaced by one using the L_1 norm.
If the optimal solution of the problem above is denoted by $\Gamma^* = (\gamma_1^*, \ldots, \gamma_N^*)$, then the first dominating sparse eigenvector γ_1^* is retained. Then the optimization algorithm is run again with $\Sigma - (\gamma_1^{*'} \Sigma \gamma_1^*) \gamma_1^* \gamma_1^{*'}$ instead of Σ. Then again the dominant sparse vector is retained as second sparse eigenvector and so on. Now the procedure is iterated until a certain stopping criterion is fulfilled. This approach of sparse PCA is called *DSPCA*.

D'Aspremont, Bach and Ghaoui [16] found in 2008 another way of defining and solving a sparse PCA problem. They start with the objective function

$$\max_{\gamma:\ ||\gamma|| \leq 1} \quad \gamma' \Sigma \gamma - \rho \ card(\gamma) \tag{3.7}$$

with the sparsity controlling parameter $\rho \in \mathbb{R}$, which should be always smaller than Σ_{11}[4]. The

[3] nondeterministic polynomial-time hard

[4] Σ_{11} denotes the element in the first row and the first column of Σ. This upper boundary ensures, that the value of the objective function will stay positive.

larger ρ, the sparser will be the vector z. Σ is the covariance matrix of y and for further computation it will be decomposed as $\Sigma = S'S$ with $S \in \mathbb{R}^{N \times N}$.
So this function can be seen as Lagrange function, where the Rayleigh quotient is maximized and constraints are set on the number of nonzero elements of the vector z. Then they reformulate the problem in equation (3.7) to a nonconvex optimization problem

$$\max_{x:\ ||x||=1} \quad \sum_{i=1}^{N} [(s_i' x)^2 - \rho]_+, \tag{3.8}$$

where s_i denotes the i^{th} column of the matrix S, $x \in \mathbb{R}^N$ and

$$[\alpha]_+ := \begin{cases} \alpha & \text{if } \alpha \geq 0 \\ 0 & \text{if } \alpha \leq 0. \end{cases}$$

Next a semidefinite, convex relaxation of equation (3.8) is proposed, that can be solved with a greedy algorithm of total complexity $O(N^3)$. Defining $X = xx'$ and $B_i = s_i s_i' - \rho I_N$, then the final convex optimization problem is given by

$$\begin{aligned} \max_{X, P_i} \quad & \sum_{i=1}^{N} trace(P_i B_i) \\ \text{s.t. } & trace(X) = 1 \\ & X \succeq 0 \\ & X \succeq P_i \succeq 0, \end{aligned}$$

where P_i is a positive semidefinite matrix and the optimal value of its objective function is an upper bound on the nonconvex problem.

Another similar, but more general approach was suggested by Journée, Nesterov, Richtárik and Sepulchre [48]. Their research is based on single factor models as well as on multifactor models. They formulate both L_0 and L_1 type penalty terms in the objective function. So when building a single unit sparse PCA model with the cardinality as penalty function, the methodology proposed by D'Aspremont, Bach and Ghaoui [16] is obtained.
The initial formulations of the optimization problems lead to nonconvex functions which are computationally intractable. Thus these functions are rewritten as convex optimization problems on a compact set, whose dimension is much smaller than the original one. So apart from making optimization easier the dimension of the search space decreases substantially. Table 3.3 opposes the original, nonconvex optimization problems and their convex reformulations for all 4 cases. Details about how they are derived, can be read in [48]. In the formulas Y is

assumed to be any rectangular data matrix of dimension $T \times N$ with sample covariance matrix $\Sigma = Y'Y$. N is a $k \times k$ diagonal matrix with positive entries $\mu_1, \ldots, \mu_k$ in the diagonal

$$N = \begin{pmatrix} \mu_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mu_k \end{pmatrix},$$

which is set to the identity matrix I_k in the empirical work of *Journée et al.* Moreover, simple first-order methods for solving the optimization problems are proposed, which give stationary points as a solution. The goal of attaining a local maximizer is in general unattainable. This methodology is called *generalized power method*.

3.2.2 Formulations based on the loss of information

In contrast to the variance based formulations there exists the second class of restricted PCA problems, which focuses on the loss of information when approximating a matrix by another of lower rank. One of the main research in that area was done by Zou, Hastie and Tibshirani [85] in 2006. Given a sample matrix $Y = (y_1, \ldots, y_T)' \in \mathbb{R}^{T \times N}$, they define the following optimization problem:

$$\begin{aligned} \min_{A,B \in \mathbb{R}^{N \times k}} \quad & \sum_{i=1}^{T} ||y_i - AB'y_i||^2 + \rho \sum_{j=1}^{k} ||b_j||^2 + \sum_{j=1}^{k} \rho_{1,j} ||b_j||_1 \\ s.t. \quad & A'A = I_k. \end{aligned}$$

This problem can be rewritten as

$$\begin{aligned} \min_{A,B \in \mathbb{R}^{N \times k}} \quad & ||Y - YBA'||_F^2 + \rho \sum_{j=1}^{k} ||b_j||^2 + \sum_{j=1}^{k} \rho_{1,j} ||b_j||_1 \\ s.t. \quad & A'A = I_k. \end{aligned} \tag{3.9}$$

which shows the similarity to the unrestricted approach proposed by Darroch (see page 16). Note, that there are two addends in the objective function. Firstly, there is the ridge penalty, which is not used to penalize the regression coefficients, but to ensure the reconstructions of the principal components. Secondly, a LASSO penalty term is added, which should control the sparseness of the $N \times k$ matrix of loadings B. As can be seen from the above objective function, different values for $\rho_{1,j}$ are allowed in each column of the loadings matrix.
Zou et al. propose an algorithm called *SPCA* (sparse PCA), which consists of iterations between the estimation of A and B. Basically, estimation reduces to generalized regression problems,

	original problem	convex reformulation	optimal γ
L_1, single	$\max_{\gamma:\gamma'\gamma\leq 1} \sqrt{\gamma'\Sigma\gamma} - \rho\|\gamma\|_1$	$\max_{x:\ x'x=1} \sum_{i=1}^{N} [\|y_i'x\| - \rho]_+^2$	$\gamma_i^* = \frac{sign(y_i'x)[\|y_i'x\|-\rho]_+}{\sqrt{\sum_{k=1}^{N}[\|y_k'x\|-\rho]_+^2}}$
L_0, single	$\max_{\gamma:\gamma'\gamma\leq 1} \gamma'\Sigma\gamma - \rho\|\gamma\|_0$	$\max_{x:\ x'x=1} \sum_{i=1}^{N} [(y_i'x)^2 - \rho]_+$	$\gamma_i^* = \frac{[sign((y_i'x)^2-\rho)]_+ y_i'x}{\sqrt{\sum_{k=1}^{N}[sign((y_k'x)^2-\rho)]_+(y_k'x)^2}}$
L_1, multi	$\max_{\substack{\Gamma:\ diag(\Gamma'\Gamma)=I_k \\ X:\ X'X=I_k}} trace(X'Y\Gamma N) - \rho \sum_{j=1}^{k}\sum_{i=1}^{N} \|\gamma_{ij}\|$	$\max_{X:\ X'X=I_k} \sum_{j=1}^{k}\sum_{i=1}^{N} [\mu_j\|y_i'x_j\| - \rho]_+^2$	$\gamma_{ij}^* = \frac{sign(y_i'x_j)[\mu_j\|y_i'x_j\|-\rho]_+}{\sqrt{\sum_{k=1}^{N}[\mu_j\|y_k'x_j\|-\rho]_+^2}}$
L_0, multi	$\max_{\substack{\Gamma:\ diag(\Gamma'\Gamma)=I_k \\ X:\ X'X=I_k}} trace(diag(X'Y\Gamma N)^2) - \rho\|\Gamma\|_0$	$\max_{X:\ X'X=I_k} \sum_{j=1}^{k}\sum_{i=1}^{N} [(\mu_j y_i'x_j)^2 - \rho]_+$	$\gamma_{ij}^* = \frac{[sign((\mu_j y_i'x_j)^2-\rho)]_+ \mu_j y_i'x_j}{\sqrt{\sum_{k=1}^{N}[sign((\mu_j y_k'x_j)^2-\rho)]_+\mu_j^2(y_k'x_j)^2}}$

Table 3.3: Sparse PCA formulations of Journée et al. [48]

which are solved by algorithms called LARS and elastic net (LARS-EN). The former was introduced in 2004 by Efron et al. [21] solving LASSO regression models, that penalize the coefficients of a regression model by adding a L_1 penalty term to the regression. In spite of wide acceptance and affirmation of the LASSO procedure, it has several drawbacks such as the inability of selecting more variables than there are observation available, which can be a problem if applied to e.g. microarray data. To overcome this limitation Zou and Hastie [84] generalized in 2005 the LASSO regression to the elastic net regression, which is a convex combination of the ridge penalty and the LASSO penalty. The estimate $\hat{\beta}_{EN}$ is given by

$$\hat{\beta}_{EN} = (1+\rho_2) arg\min_{\beta} ||y - \sum_{j=1}^{p} x_j\beta_j||^2 + \rho_2 \sum_{j=1}^{p} ||\beta_j||^2 + \rho_1||\beta_j||_1,$$

where y is a vector of dimension T, $X = (x_1, \ldots, x_p)$ is a $T \times p$ matrix of explanatory variables, $\beta = (\beta_1, \ldots, \beta_p)'$ is the vector of regression coefficients and ρ_1 and ρ_2 are nonnegative values in $\mathbb{R}$.
Note, that in the optimization problem stated in equation (3.9) A and B do not have to be equal as in the unrestricted case and that the orthogonality of the principal components BY is not required anymore.

Two years later Shen and Huang [64] introduced another sparse PCA model given by the objective function

$$\min_{u,v} \quad ||Y - uv'||_F^2 + P_\rho(v),$$

where $Y \in \mathbb{R}^{T\times N}$ is a given data matrix and u and v are T- and N-dimensional vectors, respectively. $P_\rho(v) = \sum_{j=1}^{N} p_\rho(|v_j|)$ is a penalty term with a positive tuning parameter ρ, for which three different types of penalty functions are suggested: the soft thresholding penalty or LASSO penalty, the hard tresholding penalty and the SCAD penalty[5], which can be seen as a combination of the previous two types of thresholding.
Setting $(Y'u)_j =: \tilde{y}$ and defining $(x)_+ := \max(x, 0)$, the individual penalty functions $p_\rho(|v_j|)$ and the estimates of v_j, which will be denoted by $\hat{v}_j$, are given by

- soft thresholding: $p_\rho(|v_j|) = 2\rho|v_j|$

$$\hat{v}_j = h_\rho^{soft}(\tilde{y}) = sign(\tilde{y})(|\tilde{y}| - \rho)_+$$

- hard thresholding: $p_\rho(|v_j|) = \rho^2 I(|v_j| \neq 0)$

$$\hat{v}_j = h_\rho^{hard}(\tilde{y}) = I(|\tilde{y}| > \rho)\tilde{y}$$

[5] SCAD stands for *smoothly clipped absolute deviation*

- SCAD penalty: $p_\rho(|v_j|) = 2\rho|v_j|I(|v_j| \leq \rho) - \frac{v_j^2 - 2a\rho|v_j| + \rho^2}{a-1} I(\rho < |v_j| \leq a\rho) + (a+1)\rho^2 I(|v_j| > a\rho)$

$$\hat{v}_j = h_\rho^{SCAD}(\tilde{y}) = \begin{cases} sign(\tilde{y})(|\tilde{y}| - \rho)_+ & \text{if } |\tilde{y}| \leq 2\rho \\ \frac{(a-1)\tilde{y} - sign(\tilde{y})a\rho}{a-2} & \text{if } 2\rho \leq |\tilde{y}| \leq a\rho \\ \tilde{y} & \text{if } |\tilde{y}| > a\rho \end{cases},$$

  where a is an additional tuning parameter, that takes values larger than 2. If Bayesian risk should be minimized, a value of 3.7 is recommended in the literature of *Fan and Li* [23].

By using one of the above penalty functions and an iterative algorithm called *sPCA - rSVD*, that calculates the vectors u and v in an alternating way, a sparse $\hat{v}$ is obtained, that is scaled so, that it has length 1. After obtaining this first component, the residual matrix $Y_1 = Y - \hat{u}\hat{v}'$ has to be built and the same algorithm is applied to Y_1, if a further component is desired. One may proceed in a similar way, if more than two components should be calculated.
If the parameter ρ is set to zero in the penalty function, this methodology reduces to the alternating least squares algorithm (ALS) of Gabriel and Zamir [29] in order to calculate the singular value decomposition of a sample matrix Y. Moreover, this procedure can be extended easily by adding further penalty functions.
They also introduce a measure for the cumulative percentage of explained variance ($CPEV$), which is given by

$$\frac{trace(Y_k'Y_k)}{trace(Y'Y)},$$

where Y_k denotes the projection of Y on the k-dimensional subspace spanned by the first k sparse loadings vectors $V_k = (\hat{v}_1, \ldots, \hat{v}_k)$. Thus Y_k is given by

$$Y_k = YV_k(V_k'V_k)^{-1}V_k.$$

This procedure gives k sparse loading vectors, that depend on Y only through $Y'Y$ and thus it can be applied also in the case, when just the covariance matrix is given. Nevertheless, it will not be possible to calculate sparse principal components in such a case, which is essential for the purpose of this thesis. Take also into account, that here it is also not required, that the principal components are linear combinations of the data Y. Thus, another property of unrestricted principal components is dropped besides the loss of orthogonality, which is common among all research done on sparse PCA.

Another group of researches, proposing a sparse PCA model and new estimates in 2009,

consists of Leng and Wang [51]. They reformulate and generalize the *SPCA* model of Zou et al. (see page 38) in two ways.
Firstly, a method called *simple adaptive sparse principal component analysis* (SASPCA) is proposed. It incorporates an adaptive LASSO penalty term, which has been suggested by Zou [83] in 2006, in the SPCA model:

$$\begin{aligned} \min_{A,B\in\mathbb{R}^{N\times k}} \quad & \frac{1}{T}\sum_{i=1}^{T}||y_i - AB'y_i||^2 + \sum_{i=1}^{N}\sum_{j=1}^{k}\rho_{ij}|b_{ij}| \\ s.t. \quad & A'A = I_k, \end{aligned} \tag{3.10}$$

where $B = \begin{pmatrix} b_{11} & \cdots & b_{1k} \\ \vdots & & \vdots \\ b_{N1} & \cdots & b_{Nk} \end{pmatrix}$. Thus, different shrinkage coefficients can be used for different entries of the matrix of loadings and a quite flexible way for controlling the level of sparsity is obtained. The parameter matrices A and B are calculated by applying a singular value decomposition and least angle regression (*LARS*) developed by Efron et al. [21] in 2004 iteratively. A *BIC*[6] type criterion is proposed for setting the tuning parameters. Because of the practical infeasibility of tuning so many shrinkage parameters simultaneously, the simplification

$$\rho_{ij} = \frac{\tau_j}{|\tilde{b}_{ij}|}$$

with $|\tilde{b}_{ij}|$ being the absolute value of the ij - element in the loadings matrix of the unrestricted PCA, can be made, which reduces the tuning parameter selection to choosing just k values τ_j, $j = 1, \ldots, k$. Leng and Wang [51] show, that with this method the important coefficients can be selected consistently and with high efficiency.
Secondly, within the *general adaptive sparse principal component analysis* (GASPCA) the least squares objective function of SPCA is replaced by a generalized least squares objective function, which improves the finite sample performance. If the zeros and nonzeros of the loadings matrix are not well separated, the estimates of SPCA may be poor. To overcome this problem, one of the iteration steps can be modified. According to simple linear algebra, the objective function in equation (3.10) is for fixed A equivalent to the following objective function

$$\min_{A,B\in\mathbb{R}^{N\times k}} \quad \sum_{j=1}^{k}\left\{\frac{1}{T}\sum_{i=1}^{T}(a_j'y_i - b_j'y_i)^2 + \sum_{i=1}^{N}\rho_{ij}|b_{ij}|\right\} \tag{3.11}$$

up to a constant, where $(a_1, \ldots, a_k)$ and $(b_1, \ldots, b_k)$ are the k columns of A and B, respectively.

[6] *BIC* stands for the Bayesian information criterion, which is also called *Schwarz information criterion*. It should prevent estimation from overfitting by adding a penalty term to a function of the value of the maximized likelihood L: $BIC = -2L + k\ln T$. k denotes the number of parameters, that have to be estimated, and T stands for the number of observations.

Now, this problem can again be rewritten as

$$\min_{A,B\in\mathbb{R}^{N\times k}} \quad \sum_{j=1}^{k}\left\{(a_j-b_j)'\Sigma(a_j-b_j)+\sum_{i=1}^{N}\rho_{ij}|b_{ij}|\right\}$$

with the sample covariance matrix Σ of $Y=(y_1,\ldots,y_T)'$. The idea of *GASPCA* consists of replacing the covariance matrix Σ by a positive definite matrix $\tilde{\Omega}$ with probabilistic limit Ω, which is a positive definite matrix referred to as kernel matrix, so that the following optimization problem arises:

$$\min_{A,B\in\mathbb{R}^{N\times k}} \quad \sum_{j=1}^{k}\left\{(a_j-b_j)'\tilde{\Omega}(a_j-b_j)+\sum_{i=1}^{N}\rho_{ij}|b_{ij}|\right\}. \tag{3.12}$$

The authors suggest to choose $\tilde{\Omega}$ as $cov^{-1}(\tilde{b}_j)$, which is the inverse of the covariance matrix of the unrestricted solution for the j^{th} column of B. Unfortunately, no simple formula exists for calculating this expression and so a bootstrapping method is proposed in order to calculate an estimator $\hat{cov}^{-1}(\tilde{b}_j)$.

3.3 The model

All the existing methodologies of the literature, which are summarized in the previous section, adopt a different approach to sparse principal component models. They have in common, that no information about the structure of the factor loadings matrix is available and thus the zero positions in the loadings matrix are determined in an automated way. However, in this framework a priori knowledge of the structure of the matrix of loadings exists, and this information will be considered in the estimation. The number of zeros in at least one column of the loadings matrix has to be k or larger than k in order to obtain a restricted PCA model that can not be obtained by simple rotation or transformation. The reason for formulating such sparse models is due to better interpretability of the model and enhancement of the precision of the estimation. In some practical applications a sparse PCA model can be more adequate than an unrestricted one.

Moreover, as can also be seen from existing literature, the property of orthogonality of the principal components as well as of the matrix of loadings is not assumed anymore, because this would restrict the space spanned by the principal components excessively and it does not seem reasonable from the interpretation point of view. For reasons of identifiability, the principal components will be scaled so, that they have unit length. This assumption follows from the following considerations. As a special case of linear transformations the factors can be premultiplied by any diagonal matrix $R=\begin{pmatrix} r_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & r_k \end{pmatrix}$ with $r_i\neq 0$ for all $i=1,\ldots,k$ and at least one diagonal element has to be different from 1. If the matrix of loadings is postmultiplied

by the inverse of R, the latent variables $AR^{-1}RB'y_t = AB'y_t$ stay the same and thus no real additional solution is obtained.
As already described earlier, the $N \times T$ dimensional data matrix $Y = (y_1, \ldots, y_T)'$ should be approximated by a lower dimensional matrix $\hat{Y}$ of rank $k \leq N$

$$\hat{Y} = YBA' \qquad \text{or} \qquad \hat{y}_t = AB'y_t, \quad \text{for } t = 1, \ldots, T, \tag{3.13}$$

with $rk(A) = rk(B) = k$. In the existing literature - with exception of the research done by Shen and Huang [64] - zero restrictions are just set to the matrix B, which is used to calculate the principal components as a linear combination of the original variables Y. Thus, the aim of the authors is to define principal components or factors, that are linear combination of just a few (selected) variables and not of all the variables. Shen and Huang are the only ones, who set the restriction on that matrix, that builds linear combinations of the restricted factors. However, these factors are in general not in the space spanned by the original variables Y. In this thesis zero restriction will also be set just to the matrix A, because the focus here does not lie merely in the calculation of principal components, but also in the prediction of the data. Moreover, even future values for the data should be forecasted and thus constraints on B would not be that meaningful. This will become clearer in section 4. However, the most convincing reason for setting restrictions on A and not on B is the fact, that the model should be interpretable after estimation. For example, an asset of a US company should depend on the movement of the American market, which is represented by one of the factors, and not on the Asian one, which may be another factor. So the prediction of the target variables should consist of the linear combination of just a few selected factors and not of all.
Another aspect, that changes in the case of restricted PCA, is the fact, that the coefficient matrices A and B need not be equal anymore. Equality would imply, that exact zeros would be on the same positions in the two matrices of loadings. On the other hand the equality was not forced in the case of the unrestricted PCA, but it was just the result of the optimization problems described in section 2.2. When writing down the model equations componentwise and taking into account, that the orthogonality assumption is dropped in the restricted PCA model, it becomes obvious, that there is no reason to enforce the equality of A and B.
As already mentioned before, the main interest of this thesis lies in restricted PCA models, which have k or more zeros in at least one column of their loadings matrix. In all the cases with less than k zeros in each column of the matrix of loadings, simple rotation with a regular matrix can produce the desired zeros, which has already be described in the example on page 32.

All these considerations together with the a priori information about the structure of the matrix of loadings as well as the purpose of using these restricted PCA models as forecasting

models lead to the following new definition of a sparse PCA model:

$$\begin{aligned} \min_{A,B\in\mathbb{R}^{N\times k}} \quad & \sum_{t=1}^{T} ||y_t - A\underbrace{B'y_t}_{f_t}||^2 \\ s.t. \quad & \Psi\ vec(A') = \mathbf{0} \end{aligned} \tag{3.14}$$

or in matrix notation

$$\begin{aligned} \min_{A,B\in\mathbb{R}^{N\times k}} \quad & ||Y - \underbrace{YB}_{F}A'||_F^2 \\ s.t. \quad & \Psi\ vec(A') = \mathbf{0}, \end{aligned} \tag{3.15}$$

where Ψ is a predefined sparse matrix of 0/1 entries, defining the positions of $vec(A')$, which are restricted to zero. The number of zeros in the loadings matrix is equal to the number of rows in Ψ. Let ds denote the degree of sparsity, which is defined as the number of elements in the loadings matrix that are not restricted to zero. Then Ψ is of dimension $\mathbb{R}^{(Nk-ds)\times(Nk)}$. $vec(.)$ stands for the *vec* operator, that stacks the column vectors of a matrix one below the other. Thus, a one in the $(N(j-1)+i)^{th}$ column of the matrix Ψ in any of its rows means that a_{ij}, the element in the i^{th} row and the j^{th} column of A, is restricted to be zero.

The joint covariance matrix of Y and F is given by

$$\Sigma_1 = \left(\begin{array}{c|c} \Sigma & \Sigma B \\ \hline B'\Sigma & B'\Sigma B \end{array}\right) =: \begin{pmatrix} \Sigma & \tilde{B} \\ \tilde{B}' & \tilde{C} \end{pmatrix} \geq 0.$$

This matrix Σ_1 is positive semidefinite, and thus the Schur complement of $B'\Sigma B$ in Σ_1, which is $\Sigma - \tilde{B}\tilde{C}^{-1}\tilde{B}' = \Sigma - \Sigma B(B'\Sigma B)^{-1}B'\Sigma$, also has to be positive semidefinite.
Since

$$\begin{aligned} \frac{1}{T}\epsilon'\epsilon &= \frac{1}{T}(Y - YBA')'(Y - YBA') = \\ &= \Sigma - \underbrace{\Sigma B}_{\tilde{B}}A' - A\underbrace{B'\Sigma}_{\tilde{B}'} + A\underbrace{B'\Sigma B}_{\tilde{C}}A' = \\ &= \Sigma - \tilde{B}\tilde{C}^{-1}\tilde{B}' + (A\tilde{C}^{\frac{1}{2}} - \tilde{B}\tilde{C}^{-\frac{1}{2}})(A\tilde{C}^{\frac{1}{2}} - \tilde{B}\tilde{C}^{-\frac{1}{2}})' \geq \\ &\geq \Sigma - \tilde{B}\tilde{C}^{-1}\tilde{B}' \geq 0. \end{aligned} \tag{3.16}$$

Equality in equation (3.16) is obtained for $A\tilde{C} = \tilde{B}$ or $AB'\Sigma B = \Sigma B$, which is the case if $B = A(A'A)^{-1} = (A')^+$. $(A')^+$ is the Moore-Penrose pseudoinverse of A'.

Thus instead of minimizing $trace((Y - FA')'(Y - FA'))$ as given in equation (3.15) equally $trace(\Sigma - \tilde{B}\tilde{C}^{-1}\tilde{B}')$ with $A\tilde{C} = \tilde{B}$ can be minimized.
This leads to

$$\begin{aligned}
\min_{\tilde{B}\in\mathbb{R}^{N\times k}} \quad trace(\Sigma - \tilde{B}\tilde{C}^{-1}\tilde{B}') &= \min_{A,\tilde{B}\in\mathbb{R}^{N\times k}} \quad trace(\Sigma - A\tilde{C}A') = \\
&= \min_{A,\tilde{B}\in\mathbb{R}^{N\times k}} \quad trace(\Sigma) - trace(A\tilde{C}A') \\
&= \min_{A,B\in\mathbb{R}^{N\times k}} \quad trace(\Sigma) - trace(AB'\Sigma BA') \\
&= \min_{A,B\in\mathbb{R}^{N\times k}} \quad trace(\Sigma) - trace(\Sigma BA'). \qquad (3.17)
\end{aligned}$$

The solution $\hat{A}$ of the optimization problem in equation (3.17) is equal to the one of the following maximization problem:

$$\begin{aligned}
\max_{A\in\mathbb{R}^{N\times k}} \quad & trace(\Sigma A(A'A)^{-1}A') \\
s.t. \quad & \Psi\ vec(A') = \mathbf{0},
\end{aligned}$$

which leads to the optimum $\hat{B} = \hat{A}(\hat{A}'\hat{A})^{-1} = (\hat{A}')^{+}$.

Obviously, the objective function of the optimization problem above is nonlinear and neither concave nor convex. So it cannot be expected to get a global optimum or a closed form solution. Of course, some 'black box' algorithm can compute a local optimum, but that is not the goal of this thesis. Here rather attention will be payed to develop a transparent simple algorithm for obtaining a reasonable solution of the problem of interest. Running this procedure for several sets of different starting values should ensure the quality of the solution.

3.4 Numerical solution

The sparse PCA problem in equation (3.15) based on the minimization of the loss of information can be described by the following system of equations:

$$Y = YBA' + \epsilon \quad s.t. \quad \Psi\ vec(A') = \mathbf{0}. \qquad (3.18)$$

If B would be known in equation (3.18), a usual least squares estimate with restrictions on $vec(A')$ could be performed to get an estimate $\hat{A}$ for A. For this problem a closed-form solution

exists. So rewrite equation (3.18) as univariate model

$$\underbrace{vec(Y)}_{\tilde{Y}} = \underbrace{(I_N \otimes (YB))}_{\tilde{F}} \underbrace{vec(A')}_{\tilde{a}} + \underbrace{vec(\epsilon)}_{\tilde{\epsilon}} \quad s.t. \quad \Psi\ vec(A') = \mathbf{0}, \tag{3.19}$$

which can be simplified as

$$\tilde{Y} = \tilde{F}\tilde{a} + \tilde{\epsilon} \quad s.t. \quad \Psi\ \tilde{a} = \mathbf{0}. \tag{3.20}$$

The symbol $\otimes$ is known as Kronecker product, that concatenates a rectangular matrix $G = \begin{pmatrix} g_{11} & \cdots & g_{1n} \\ \vdots & & \vdots \\ g_{m1} & \cdots & g_{mn} \end{pmatrix}$ of dimension $m \times n$ and a $r \times q$ matrix H to a matrix of dimension $mr \times nq$ in the following way:

$$G \otimes H = \begin{pmatrix} g_{11}H & \cdots & g_{1n}H \\ \vdots & & \vdots \\ g_{m1}H & \cdots & g_{mn}H \end{pmatrix}.$$

Denoting by $\hat{a}$ the unrestricted least squares estimator of the model $\tilde{Y} = \tilde{F}a + \tilde{\epsilon}$, the constrained least squares solution for the estimator of $\tilde{a}$ is given by

$$\hat{\tilde{a}} = \hat{a} - (\tilde{F}'\tilde{F})^{-1}\Psi'[\Psi(\tilde{F}'\tilde{F})^{-1}\Psi']^{-1}\Psi\hat{a}.$$

On the other hand, if A would be known, equation (3.18) can be postmultiplied by the Moore-Penrose pseudoinverse $(A')^+$. Then $Y(A')^+$ has to be regressed on Y, which gives an estimate $\hat{B} = (A')^+ = A(A'A)^{-1}$, which is equal to the solution that was obtained before, when building the derivatives of the optimization problem.

Now it seems natural to alternate these two least squares steps to get final estimates for A and B. So an initial estimate for B, say B^1, is needed which is first held fixed. One may choose the unrestricted loadings matrix as a starting value for B, but, as the empirical examples later on show, any random matrix can be taken and convergence properties are still unchanged. Afterwards a constrained estimate A^1 can be calculated as described above. In the next step the obtained A^1 is fixed and a new estimate B^2 is calculated as the Moore-Penrose pseudoinverse of A^1. Next B_2 is rescaled, so that the columns of YB have length 1 and so on.
Because of performing just linear regressions in each step, it is clear, that this algorithm converges monotonically. Defining $\sum_{t=1}^{T} ||y_t - AB'y_t||^2$ as function f(A,B), the following inequalities must hold:

$$f(A^1, B^1) \geq f(A^1, B^2) \geq f(A^2, B^2) \geq f(A^2, B^3) \geq \ldots,$$

which ensures, that the above defined alternating least squares algorithm converges, because f is bounded below by the value of the objective function of the unrestricted solution, which has no sparsity constraints. If f_k stands for the value of the objective function in the k^{th} iteration,

a possible common stopping criterion of the algorithm proposed here, is, that the value of the objective function in iteration step k changes relatively to the value obtained in iteration step $k-1$ less than a certain threshold τ:

$$\frac{f_k - f_{k-1}}{f_{k-1}} < \tau.$$

In the empirical applications of this thesis another stopping criterion is used, that is also considering the stability of the solution, which is measured by a function of the coefficients. Let A^k and A^{k+1} be two consecutive sparse loadings matrices, ds the degree of sparsity and τ a threshold for convergence as defined above. Then an alternative stopping criterion can be defined by

$$||A^k - A^{k-1}||_F^2 < ds\ \tau,$$

where ds is just a scaling parameter taking into account the number of free parameters in A.

Finally, when applying this methodology with a set of m different starting values B_i^1, $i = 1, \ldots, m$, a reasonable solution can be calculated. Obviously, the finally obtained estimate for A is a sparse loadings matrix, whereas B is just sparse if A is an orthogonal matrix, which is not the case in general. This coincides exactly with the requirements on the sparse PCA problem, that were defined previously.

Furthermore, the question about uniqueness of the obtained solution arises. Which conditions have to be met, so that a with a regular matrix transformed loadings matrix is still a solution to the restricted PCA problem? In the case of usual PCA without restrictions the equality

$$BA' = (BS^{-1})(SA') =: \tilde{B}\tilde{A}'$$

holds for all regular matrices S of full rank k.
When imposing restrictions on the PCA model, not only the equality $\hat{Y} = YBA' = Y\tilde{B}\tilde{A}'$ has to hold, but also the additional condition that

$$\Psi vec(\tilde{A}') = \Psi\ vec(SA') = 0. \tag{3.21}$$

Because of

$$vec(SA') = (A \otimes I_k) vec(S),$$

where I_k defines the $k \times k$ identity matrix, equation (3.21) can be written as

$$\Psi(A \otimes I_k) vec(S) = 0. \tag{3.22}$$

Thus in the restricted case of PCA the solution is unique up to a regular matrix S, whose vectorized form $vec(S)$ is in the kernel of the map $\Psi(A \otimes I_k)$.

Another way of interpretation is obtained if equation (3.21) is rewritten as

$$\Psi(I_N \otimes S)vec(A') = 0. \tag{3.23}$$

So when splitting $\Psi = [\Psi_1 \dots \Psi_N]$ into N blocks , whereby Ψ_i denotes the i^{th} block of Ψ that contains those coefficients of the matrix of restrictions with which the i^{th} row of A is multiplied, the equation above can be simplified to

$$[\Psi_1 S \ \Psi_2 S \ \dots \ \Psi_N S]vec(A') = 0. \tag{3.24}$$

That means that for any feasible regular matrix S the vector $vec(A')$ lies not only in the kernel of $\Psi = [\Psi_1 \dots \Psi_N]$ but also in the kernel of $[\Psi_1 S \ \dots \ \Psi_N S]$.

Moreover, it has to be mentioned that, when applying the proposed methodology without restrictions on A, a rotated solution of usual PCA is obtained and the equality $AB' = \Gamma\Gamma'$ for the unrestricted loadings matrix Γ holds, which again points out the reasonability of this algorithm.

Chapter 4

Forecasting with PCA and sparse PCA models

4.1 The forecast model

As already mentioned earlier, the focus of this thesis does not merely lie in obtaining a restricted matrix of loadings but in building a model, which is able to calculate forecasts for future values of a time series. The basic sparse PCA model, which is the solution to the optimization problem given in equation (3.14), is as follows:

$$\begin{aligned} y_t &= A_{\tilde{t}} \underbrace{B_{\tilde{t}}' y_t}_{f_t} + \epsilon_t \qquad \text{for } t = 1, \ldots, \tilde{t} \text{ and } \tilde{t} \leq T \\ s.t. \quad & \Psi \; vec(A_{\tilde{t}}') = \mathbf{0}. \end{aligned} \tag{4.1}$$

The index $\tilde{t}$ in $A_{\tilde{t}}$ and $B_{\tilde{t}}$ indicates, that data up to time point $\tilde{t}$ is used for calculating these matrices of rank k. It is up to the practitioner to decide whether to choose a moving or an extending window in the calculation. So one may select data from 1 to $\tilde{t}$ in the first step, then data from 2 to $\tilde{t}+1$, next from 3 to $\tilde{t}+2$ and so on, which is called rolling or moving window. Another possibility consists of taking data from 1 to $\tilde{t}$, then from 1 to $\tilde{t}+1$, next from 1 to $\tilde{t}+2$, which means that the number of data points increases by 1 in each step.
To calculate a single forecast based on a (restricted) PCA model for a particular instant in time $\tilde{t}+1$ based on the data up to $\tilde{t}$ the following procedure can be applied.
First a PCA model as in equation (4.1) has to be build to obtain a (sparse) loadings matrix $A_{\tilde{t}}$ and the factors $f_t = B_{\tilde{t}}' y_t$. As can be seen in the subscripts the dynamic of the model is represented by the factors f_t. Due to the fact, that strong correlation between the loadings matrices of subsequent points in time has been found in empirical applications, the forecast of $A_{\tilde{t}}$ is chosen as naive forecast $\hat{A}_{\tilde{t}+1} = A_{\tilde{t}}$ in this work. Thus the focus lies solely in forecasting

the factors $\hat{f}_{\tilde{t}+1}|\tilde{t}$ based on the information available at $\tilde{t}$. Once the forecasts of the principal components $\hat{f}_{\tilde{t}+1}|\tilde{t}$ are calculated, the forecasts of the original variates $y_{\tilde{t}+1}|\tilde{t}$ can be computed by the formula

$$\hat{y}_{\tilde{t}+1}|\tilde{t} = \hat{A}_{\tilde{t}+1}|\tilde{t}\ \hat{f}_{\tilde{t}+1}|\tilde{t} = A_{\tilde{t}}\hat{f}_{\tilde{t}+1}|\tilde{t}. \tag{4.2}$$

There are numerous ways of building forecasting models for the factors. As an example vector autoregressive models with exogenous variables (*VARX* models) are chosen in the empirical work of this thesis with a special input selection algorithm based on the one proposed by An and Gu [3], which are described in the next two sections in more detail.

4.2 VARX models

Vector autoregressive models with exogenous variables of order p are a special type of multivariate linear models, that take into account lags of the targeted variable up to a maximum lag p as well as a set of s exogenous variables $\mathbf{x}_t = (x_{1t}, x_{2t}, \ldots, x_{st})'$ in order to explain the output variable. This vector $\mathbf{x}_t$ contains values of variables at time t or prior to t and $(\mathbf{x}_t)$ is supposed to be a stationary process in the sense of weak stationarity with mean μ_x. Because of calculating a VARX model for the factors in this thesis, the dependent variable will be called $f_t = (f_{1t}, f_{2t}, \ldots, f_{kt})'$ in this context. So the following model will be considered:

$$\mathbf{f}_t = \mathbf{c} + A_1\mathbf{f}_{t-1} + A_2\mathbf{f}_{t-2} + \ldots + A_p\mathbf{f}_{t-p} + B\mathbf{x}_{t-1} + \mathbf{e}_t, \qquad t = 1, \ldots, T \tag{4.3}$$

or more compactly as

$$A(z)\mathbf{f}_t = \mathbf{c} + B\mathbf{x}_{t-1} + \mathbf{e}_t, \qquad t = 1, \ldots, T \tag{4.4}$$

where $\mathbf{c} \in \mathbb{R}^k$ denotes a constant vector, A_i are real coefficient matrices of dimension $k \times k$ $(i = 1, \ldots, p)$ and $\mathbf{e}_t$ is the k dimensional noise vector at time t, which is a white noise process, i.e. $E(\mathbf{e}_t) = 0$, $E(\mathbf{e}_s\mathbf{e}_t') = 0$ for $s \neq t$ and $E(\mathbf{e}_t\mathbf{e}_t') = \Sigma_e$ with a positive definite matrix Σ_e. The impact of the exogenous variables $\mathbf{x}_t$ is given through the coefficient matrix $B \in \mathbb{R}^{k\times s}$. Moreover, $\mathbf{e}_t$ is required to be independent of the exogenous variables $\mathbf{x}_s$ for all s smaller than t. $A(z)$ is a lag polynomial which is given by $A(z) = \sum_{i=0}^{p} -A_i z^i$ with $A_0 = -I_k$ and $A_p \neq 0$. z can be interpreted as complex variable or as backward shift operator, whereby the latter is defined as:

$$z\{\mathbf{f}_t | t \in \mathbb{Z}\} = \{\mathbf{f}_{t-1} | t \in \mathbb{Z}\},$$

where $\{\mathbf{f}_t | t \in \mathbb{Z}\}$ is the series of factor values. Because of

$$\mathbf{f}_t = A(z)^{-1}(\mathbf{c} + B\mathbf{x}_{t-1} + \mathbf{e}_t), \qquad t = 1, \ldots, T \tag{4.5}$$

the convergence of the Taylor series expansion of $A(z)^{-1}$ about the point 0 in an area that contains the unit circle has to be guaranteed, which can be reached if the stability condition $|A(z)| \neq 0$ for all $|z| \leq 1$ holds.

There exist basically three ways of estimating the unknown parameters $\mathbf{c}, A_1, \ldots, A_p, B$ and Σ_e. One would be to estimate them by ordinary least squares which minimizes the residual sum of squares of equation (4.3). The predicted value for $\mathbf{f}_{t+1}$ based on information known at time t can be easily calculated by

$$\hat{\mathbf{f}}_{t+1}|t = \hat{\mathbf{c}} + \hat{A}_1\mathbf{f}_t + \hat{A}_2\mathbf{f}_{t-1} + \ldots + \hat{A}_p\mathbf{f}_{t-p+1} + \hat{B}\mathbf{x}_t, \tag{4.6}$$

where $\hat{\cdot}$ denotes the estimated OLS parameters.
The second possibility for estimating equation (4.3) would be to estimate the autoregressive part of the equation by maximum likelihood (ML) with the help of a Kalman filter and regress then the remaining error vector on the exogenous variables $\mathbf{x}_{t-1}$.
Thirdly, the Yule Walker equations are another approach to get parameter estimates for a VARX model. This methodology is widespread and one of the most popular ones in practice. All these estimation methods have similar asymptotic properties and they differ mainly in their finite sample behavior. Details concerning VARX models and their estimation can be found for example in [52].

4.3 Inputselection

In finance as well as in other scientific applications there exists a huge universe of explanatory variables, which can be used as exogenous variables when using not only the target time series' own history. It's quite a difficult task to select a subset of those variables, that explains the targets in a satisfying way, because

- economical data are often not very informative concerning the target
- the a priori info about the choice of variables is uncertain; in practice one often has to select among a huge number of candidate inputs.

In any case a preselection has to be performed based on prior knowledge, which could be based on economic relationships in the case of financial forecasting. But even if it would be done by one of the top economists he/she will not be able to define a manageable set of input variables because of the complexity of the markets.[1]
Thus a way for further reduction of the number of possible candidates has to be applied often

[1] And if one is able to do that, he/she will not tell others and thus the problem of reducing the number of variables is still present.

in empirical work. One possibility to do that is to select a subset of the inputs according to statistical criteria to get a feasible number of input variables for the prediction of each factor. Therefore an algorithm based on information criteria similar to that introduced by An and Gu [3] is applied here. The algorithm will be explained for a univariate model first and at the end a generalization to multivariate models is given.

The model under consideration is

$$f_j = x\theta_j + u_j \tag{4.7}$$

where f_j denotes the j^{th} column of the factor matrix, $x = (x_1, \ldots, x_s)$ is the $T \times s$ matrix of explanatory variables, θ_j the least squares estimator and u_j the white noise error process. The matrix x consists of s candidates of predictor variables and it is not distinguished here between autoregressive terms and exogenous variables. To simplify notation the index j will be omitted from now on. Note, that for forecasting the factor at instant in time $\tilde{t}+1$ based on the information available at $\tilde{t}$, which will be called $\hat{f}_{\tilde{t}+1}|\tilde{t}$ just data up to time $\tilde{t}$ can be used. That's why the matrix of explanatory variables x in equation (4.7) has to contain only data up to the point in time $\tilde{t}-1$ or earlier when estimating the parameter vector θ, say $\hat{\theta}$. Then the forecasts are calculated as $\hat{f}_{\tilde{t}+1}|\tilde{t} = x_{new}\hat{\theta}$ with x_{new} containing the variables in x shifted 1 period ahead, i.e. x_{new} contains information up to $\tilde{t}$. Otherwise calculating a forecast would not be possible.

The aim is now to find those variables in x that have predictive power. Let us assume that there exists a true model

$$f = x(I_k)\theta(I_k) + u \tag{4.8}$$

with $I_k = (i_1, \ldots, i_k)$ is the index set of the k true predictor variables. Then $x(I_k) = (x_{i_1}, \ldots, x_{i_k})$ and $\theta(I_k) = (\theta_{i_1} \ldots \theta_{i_k})$. Suppose that $x'(I_k)x(I_k)$ is not singular, then the least squares estimator of $\theta(I_k)$ is given by

$$\hat{\theta}(I_k) = (x'(I_k)x(I_k))^{-1}x'(I_k)f \tag{4.9}$$

and the corresponding *mean squared error* (MSE) is equal to the residual sum of squares (RSS) divided by the number of observations T:

$$\begin{aligned} MSE(I_k) &= \frac{1}{T}(f - x(I_k)\hat{\theta}(I_k))'(f - x(I_k)\hat{\theta}(I_k)) \\ &= \frac{1}{T}(\|f\| - f'x(I_k)(x'(I_k)x(I_k))^{-1}x'(I_k)f), \end{aligned} \tag{4.10}$$

where $\|.\|$ indicates the L^2 norm.
Choosing another index set $J_l = (j_1, \ldots, j_l)$ instead of I_k leads to a different estimator $\hat{\theta}(J_l)$ and the mean squared error denoted by $MSE(J_l)$ will be calculated analogous to equation (4.10). Altogether there are $2^s - 1$ possible subsets of the s possible predictor variables $(x_1, \ldots, x_s)$. Model selection can be accomplished by means of information criteria as the *Akaike Information Criterion* (AIC) or the *Bayesian Information Criterion* (BIC) , which is also called *Schwarz Information Criterion*.
These are defined as

$$AIC(J_l) = \log MSE(J_l) + \frac{2l}{T} \tag{4.11}$$

and

$$BIC(J_l) = \log MSE(J_l) + \frac{l \log T}{T}, \tag{4.12}$$

where $l = 0, \ldots, s$, $1 \leq j_1 < \ldots < j_l \leq s$ and T denotes the number of observations over time in f and x, respectively. It is intuitively clear, that the number of possible models, that have to be compared $(2^s - 1)$, is often far too high in practical applications and represents one of the main disadvantages of this approach. The risk of overfitting is not negligible if many hypothesis are tested in comparison to a relatively small sample size.
Thus a search algorithm has to be found that evaluates just the promising subsets of the whole input space and neglects those that seem to lead to bad results or that are dispensable.

The procedure of comparing the explanatory power of all different subsets of the available inputs can also be structured in the following way:

Step1 For each l from 0 to s find out the index set J_l^* satisfying

$$MSE(J_l^*) = \min_{J_l} MSE(J_l), \qquad l = 0, \ldots, s. \tag{4.13}$$

where 'min' stands for the minimum value of $MSE(J_l)$ over all J_l having l elements belonging to the complete set $J_s = \{1, \ldots, s\}$.

Step2 Let $J_0^* = \emptyset$, $MSE(\emptyset) = \log \|f\|^2 = \sum_{t=1}^{T} f_t^2$ for $f = (f_1, \ldots, f_T)'$ and

$$BIC(l) = \log MSE(J_l^*) + \frac{l \log T}{T}, \qquad l = 0, \ldots, s. \tag{4.14}$$

This leads to a series $\langle BIC(0), \ldots, BIC(s) \rangle$. The aim is then to find that l and thus that J_l^*, that produces the minimal BIC value in equation (4.14) above.

It is obvious, that this two-step procedure is an exhaustive search, calculating the mean squared error of all the $2^s - 1$ subsets.
In order to search just a subset of the power set of possible input candidates, An and Gu [3] considered similar to the algorithm above the following two steps, which will be presented in the next two subsections. There the procedure of obtaining an optimal subset with l explanatory variables, J_l^*, in *Step1* is replaced by an approximation.

4.3.1 Forward and backward search

Definition 6 (Forward order)
A set M_l with elements $\{m_1, \ldots, m_l\} \subset \{1, \ldots, s\}$ is called forward order index set, if $M_0 = \emptyset$ and $M_l = \{m_1, \ldots, m_l\}$ is defined inductively by

$$RSS(M_l) = \inf_{j \in M_{l-1}^c} \left(RSS(M_{l-1} \cup \{j\}) \right), \qquad l = 1, \ldots, s,$$

where $M^c = J_s \backslash M$ denotes the complement set of M and $RSS(M_l)$ the residual sum of squares of the model obtained when explaining the dependent variable by those input variables indicated by M_l.

Definition 7 (Backward order)
A set N_l with elements $\{n_1, \ldots, n_l\} \subset \{1, \ldots, s\}$ is called backward order index set, if $N_s = J_s$ and $N_l = \{n_1, \ldots, n_l\}$ is defined inductively by

$$RSS(N_{l-1}) = \inf_{j \in N_l} RSS(N_l \backslash \{j\}), \qquad l = s, s-1, \ldots, 2$$

with $N_0 = \emptyset$ and $RSS(N_{l-1})$ is in an analogous way the residual sum of squares of the model obtained when explaining the dependent variable by those input variables indicated by N_{l-1}.

If we are using M_l instead of J_l, we face the following optimization problem:

$$BIC_F(M_l^*) = \min_{l=0,\ldots,s} BIC_F(M_l) = \min_{l=0,\ldots,s} \log RSS(M_l) + \frac{l \log T}{T}. \tag{4.15}$$

Thus an optimal index set M_l^* is obtained by applying the so called *forward* method as the subscript F in $BIC_F(M_l^*)$ already indicates.

If we use on the other hand N_l instead of J_l, we have

$$BIC_B(N_l^*) = \min_{l=0,\ldots,s} BIC_B(N_l) = \min_{l=0,\ldots,s} \log RSS(N_l) + \frac{l \log T}{T}. \tag{4.16}$$

Finding the optimal index set that minimizes the series $< BIC_B(N_l) >$ over all $l = 0, \ldots, s$ is

called *backward* method accordingly and is marked with a B in the subscript.

Since an analogous procedure can be run by using AIC instead of BIC, we can distinguish between the AIC_F, AIC_B, BIC_F and BIC_B methods.
The advantage of these approaches lies obviously in a considerable reduction of the number of candidate sets that have to be taken into account, namely only $s(s+1)/2$ in comparison to $2^s - 1$ in the case of an exhaustive search, especially for large s[2]. Nevertheless, it has to be mentioned, that the solution will in general only be a suboptimal one, if not all possible subsets of the available inputs are used.

4.3.2 The fast step procedure

Based on the subset J_l^* selected by the forward or the backward search described above the following modifications of this index set are possible in the fast step procedure (*FSP*):

- If $J_l^* \neq J_s$ a variable that has not been chosen yet can be added.
- If $J_l^* \neq \emptyset$ a variable that has already been chosen can be dropped.

The decision of adding a variable to the currently chosen subset or deleting it from it is based on comparing the values of the information criteria AIC resp. BIC of equation (4.11) and (4.12) of the so created subsets.

Thus the following iterative procedure has to be carried out:

1. If in the forward or backward search the optimal subset was found with l elements, set $k = l$.
2. Build the union sets $J_{k+1} = J_k \cup \{k_0\} \quad \forall k_0 \in J_k^c$, where J_k^c denotes the complement set of J_k in the overall index set $J_s = (1, \ldots, s)$. If at least in one case the new index set J_{k+1} leads to an AIC or BIC value less than the one of J_k, find that variable k_0^* and thus that index set J_{k+1}^* that produces the minimal value for the respective information criterion.
3. Build all sets $J_{k-1} = J_k \backslash \{k_0\} \quad \forall k_0 \in J_k$. If at least in one case the new index set J_{k-1} leads to an AIC or BIC value less than the one of J_k, find that variable k_0^* and thus that index set J_{k-1}^* that produces the minimal value for the respective information criterion.
4. If any of the in 2. and 3. calculated subsets J_{k+1} or J_{k-1} yielded a further reduction of the AIC resp. BIC, it's overall minimum defines which variable should be added or dropped. So if the smallest value was achieved by adding a variable k_0^* to the index set

[2]This is exactly the interesting case because for small s no subset selection has to be performed.

J_k, set $k = k + 1$ and $J^*_{k+1} = J_k \cup \{k^*_0\}$. However, if a reduction of the information criterion is obtained by dropping a variable k^*_0 of the index set J_k, set $k = k - 1$ and $J^*_{k-1} = J_k \backslash \{k^*_0\}$.

5. As long as a decrease of the information criterion was reached go to 2. by using the criterion value of J^*_{k+1} resp. J^*_{k-1} as basis of comparison. If no further reduction can be achieved, stop the iteration.

Consistency results of the forward search, the backward search and the *FSP* can be found in An and Gu ([3]).

Chapter 5

Reduced rank regression model

In the present chapter another class of a factor model will be presented, namely the reduced rank regression model. Before going more into detail, a short introduction on multivariate linear regression models will be given, which serve as a basis for the model class of interest.

5.1 The multivariate linear regression model

A multivariate linear regression model seeks to relate a set of N responses y_t, $t = 1, \ldots, T$ to a set of s explanatory variables x_t in a linear way:

$$y_t = Cx_t + \epsilon_t, \tag{5.1}$$

where ϵ_t is a N dimensional random error vector with $E(\epsilon_t) = 0$ and $cov(\epsilon_t) = \Sigma_\epsilon$, which is a $N \times N$ positive definite covariance matrix. An important assumption is the stochastic independence between the errors and the regressors, i.e. $E(\epsilon_t x_t') = 0$. $C \in \mathbb{R}^{N \times s}$ stands for the matrix of regression coefficients, that have to be estimated. Stacking all T observations of y_t and x_t in a matrix, the resulting matrices are $Y = (y_1, \ldots, y_T)' \in \mathbb{R}^{T \times N}$ and $X = (x_1, \ldots, x_T)' \in \mathbb{R}^{T \times s}$. With the help of these matrices, equation (5.1) can be rewritten in a more compact notation as

$$Y = XC' + \epsilon, \tag{5.2}$$

where $\epsilon = (\epsilon_1, \ldots, \epsilon_T)'$.
Moreover, the inequality $N + s \leq T$ should hold and the noise vectors should be independent for different points in time, i.e. $E(\epsilon_s \epsilon_t) = 0$ for $s \neq t$. Further it is assumed that X is of full rank $s < T$, which is a sufficient condition to ensure the uniqueness of the least squares solution.
The unknown parameters of such a multivariate linear regression model, namely C and Σ_ϵ, can be estimated by the least squares (LS) or the maximum likelihood (ML) method. In the

former case the expression

$$\|Y - XC'\|_F^2 = trace[(Y - XC')'(Y - XC')] = trace[\epsilon'\epsilon] \tag{5.3}$$

is minimized which leads to the following least squares estimator for the parameter matrix C:

$$\hat{C} = Y'X(X'X)^{-1}. \tag{5.4}$$

In the case of the maximum likelihood method some distributional assumptions have to be made. The errors ϵ_t are assumed to be multivariate normal distributed, i.e. $\epsilon_t \sim \mathcal{N}(\mathbf{0}, \Sigma_\epsilon)$ and the predictor variables x_t are known vectors.
Maximizing the likelihood

$$L(\epsilon) = (2\pi)^{-\frac{NT}{2}} |\Sigma_\epsilon|^{-\frac{T}{2}} \exp\Big[-\frac{1}{2} trace(\Sigma_\epsilon^{-1}\epsilon'\epsilon)\Big]$$

is equivalent to minimizing

$$trace(\Sigma_\epsilon^{-1}\epsilon\epsilon') = trace\Big(\Sigma_\epsilon^{-\frac{1}{2}}(Y - XC')'(Y - XC')\Sigma_\epsilon^{-\frac{1}{2}}\Big). \tag{5.5}$$

The derivative of the expression in equation (5.5) with respect to C yields the same solution for $\hat{C}$ as obtained in equation (5.4) for the least squares case.

Nevertheless, if no possible relations between the dependent variables are taken into account, there is no difference between estimating the multivariate linear equations jointly or separately. This can be seen from the fact, that the j^{th} column of $\hat{C}$, say $\hat{C}_{(j)}$, is calculated as

$$\hat{C}_{(j)} = Y'_{(j)}X(X'X)^{-1},$$

where $Y_{(j)}$ denotes the j^{th} column of Y. This means, that each dependent variable could be regressed separately on X, and thus the multivariate model contains no new information in comparison with the univariate multiple regression model. Moreover, as already mentioned in the introduction, a more parsimonious model would be more desirable both from the estimation and the interpretation point of view. The number of parameters contained in the matrix C alone is $N \times s$, which can become quite large easily. As a consequence estimation accuracy suffers and inference becomes difficult.

Due to these disadvantages of multivariate linear models it seems reasonable under certain circumstances to set restrictions on the model in order to reduce the number of the parameters and to capture possible correlations between the response variables. One possibility how to do that is presented in the following sections.

5.2 The reduced rank model

An example for a more parsimonious model than the multivariate linear regression model presented in the previous section is the reduced rank regression model. A convenient form for the one step ahead prediction is as follows:

$$y_t = A\underbrace{B'x_{t-1}}_{f_t} + \epsilon_t = Cx_{t-1} + \epsilon_t, \tag{5.6}$$

where, for $t = 1, \ldots, T$, $y_t \in \mathbb{R}^N$ is the dependent variable, $x_{t-1} \in \mathbb{R}^s$ is a vector of exogenous variables, $A \in \mathbb{R}^{N\times k}$, $B \in \mathbb{R}^{s\times k}$ and $C \in \mathbb{R}^{N\times s}$ are matrices of unknown parameters of the model and $\epsilon_t \in \mathbb{R}^N$ is a white noise error process. The coefficient matrices A, B and C are all matrices of rank $k \leq \min(N, s)$. For convenience of notation let us assume that $k < N \leq s$, although the methodology also works for $N > s$. Note, that here the vector of explanatory variables at time $t-1$, x_{t-1}, is already used, which contains just values of variables prior to time t and thus the model incorporates the possibility of calculating forecasts. Moreover, the similarity of equation (5.6) describing a reduced rank model to equation (3.13) stating the properties of a PCA model has to be mentioned. They are distinguished by the fact, that PCA builds linear combinations of the target vector itself and reduced rank analysis approximates the dependent variables by another vector of explanatory variables, whereby in both cases the resulting approximation $\hat{Y} = (\hat{y}_1, \ldots, \hat{y}_T)'$ is of lower rank $k < N$.
Using again a more compact notation, the reduced rank factor model can be written as

$$Y = \underbrace{XB}_{F} A' + \epsilon = XC' + \epsilon, \tag{5.7}$$

where the target matrix $Y = (y_1, \ldots, y_T)'$ is a real matrix of dimension $T \times N$, the matrix of exogenous variables $X = (x_0, \ldots, x_{T-1})'$ is a $T \times s$ matrix and the noise matrix $\epsilon = (\epsilon_1, \ldots, \epsilon_T)'$ is of dimension $T \times N$.
When interpreting A as a factor loadings matrix and defining XB as the factor matrix F, a special type of a factor model as described in section 1.4 is obtained. However, here the error terms are not required to be orthogonal.
In other words, we face a linear model with $N-k$ linear restrictions on the regression coefficient matrix $C = AB'$:

$$l_i'C = 0, \qquad i = 1, \ldots, N-k, \tag{5.8}$$

where $l_1, \ldots, l_{N-k}$ are generally unknown a priori. One of the practical aspects justifying such restrictions is given by the fact, that the number of parameters, that has to be estimated in a linear model, can become quite large for increasing N or s. Thus a more parsimonious structure of the model is often desirable. Moreover, estimation becomes more precise if the number of

parameters is reduced for a fixed sample size and in some situation a reduced rank model may capture the characteristics of the 'true model' in a better way.

5.3 Estimation

In order to estimate a reduced rank model as given in equation (5.7) the parameter values of A, B and Σ_ϵ, the covariance matrix of ϵ, have to be determined. Analogously to the indeterminacy of Γ_1 in section 2.3 the coefficient matrices A and B are not identifiable without further restrictions. This means that for any nonsingular[1] matrix $S \in \mathbb{R}^{k\times k}$ and the linear transformations $\tilde{A} = AS'$ and $\tilde{B} = BS^{-1}$ the equality $\tilde{B}\tilde{A}' = BS^{-1}\ SA' = BA'$ holds. Thus the number of parameters, that have to be estimated in a reduced rank model, is given by $k(N + s - k)$, which is in general much smaller than the Ns parameters of the full rank linear regression model.
In order to derive a unique solution for the estimates of A and B the following lemma is needed, which follows immediately from theorem 2.2.4:

Lemma 5.3.1. *Let A be a symmetric matrix of dimension $N \times N$ and let the eigenvalues of A be arranged in decreasing order of magnitude by $\lambda_1 \geq \ldots \geq \lambda_N$. Let $\gamma_1, \ldots, \gamma_N$ denote the corresponding eigenvectors.*
Then the supremum of $\sum_{i=1}^{k} X_i' A X_i = tr(X'AX)$ over all matrices X with orthogonal columns $(X_1, \ldots, X_k)$ and $k \leq N$ is attained for $X_i = \gamma_i$, $i = 1, \ldots, k$, and is equal to $\sum_{i=1}^{k} \lambda_i$.

By dint of the above theorem the Householder-Young Theorem can be stated, which is a well known result of PCA that has already been mentioned before (see [60]):

Theorem 5.3.1. *Let C be a $N \times s$ matrix of rank N. Then the minimum of $tr[(C-P)(C-P)']$ over all $N \times s$ matrices P with rank $k \leq N$ is attained when $P = \Gamma_1\Gamma_1'C$, where $\Gamma_1 \in \mathbb{R}^{N\times k}$ contains those normalized eigenvectors of CC', that belong to the k largest eigenvalues of CC'.*

Proof. Let $P = QR'$ with $Q \in \mathbb{R}^{N\times k}$ and $R \in \mathbb{R}^{s\times k}$ and, without loss of generality, let us assume that Q is orthonormal which gives $Q'Q = I_k$. Minimizing $tr[(C - QR')(C - QR')']$ over R for a given Q yields the least squares solution $\hat{R} = C'Q(Q'Q)^{-1} = C'Q$. Substituting this expression in the objective function and applying some basic matrix rules gives

$$\begin{aligned} tr[(C-P)(C-P)'] &= tr[(C - QQ'C)(C - QQ'C)'] = \\ & tr[CC'(I_N - QQ')] = tr[CC'] - tr[Q'CC'Q]. \end{aligned} \tag{5.9}$$

[1] In the case of reduced rank models the coefficient matrices need not be orthogonal and thus any regular matrix S can be postmultiplied to get another feasible solution.

Minimizing equation (5.9) with respect to Q is equivalent to

$$\begin{aligned} \max_{Q} \quad & tr[Q'CC'Q] \\ \text{s.t.} \quad & Q'Q = I_k. \end{aligned}$$

By setting $CC' = A$ lemma 5.3.1 can be applied and thus minimization is achieved when choosing the columns of Q as the eigenvectors of the matrix CC' belonging to the k largest eigenvalues.

□

Due to the fact that the positive square roots of the eigenvalues of a matrix CC' are the singular values of the matrix C, the above calculations can be reduced to a singular value decomposition of the matrix C.
In general a matrix $C \in \mathbb{R}^{N\times s}$ of rank N_1 can be decomposed as $V\Lambda U'$, where $V = (v_1, \ldots, v_{N_1})$ is an orthogonal matrix of dimension $N \times N_1$ such that $V'V = I_{N_1}$, $U = (u_1, \ldots, u_{N_1}) \in \mathbb{R}^{s\times N_1}$ is also orthogonal such that $U'U = I_{N_1}$ and $\Lambda = diag(\lambda_1, \ldots, \lambda_{N_1})$ with $\lambda_1^2 \geq \lambda_2^2 \geq \ldots \geq \lambda_{N_1}^2 > 0$ stating the nonnegative and nonzero eigenvalues of CC'. Then for $i = 1, \ldots, N_1$ the columns v_i are normalized eigenvectors of CC' belonging to the eigenvalues λ_i^2 and $u_i = \frac{1}{\lambda_i}C'v_i$.
So when minimizing $tr[(C-P)(C-P)'] = tr[(V\Lambda U' - QR')(V\Lambda U' - QR')']$ over all $N \times s$ matrices P with rank $k < N_1$ in theorem 5.3.1, Q is given by $V_{(k)} = (v_1, \ldots, v_k)$ and R by $C'Q = C'V_{(k)} = U\Lambda V'V_{(k)} = (\lambda_1 u_1, \ldots, \lambda_k u_k) \equiv U_{(k)}\Lambda_{(k)}$ with $U_{(k)} = (u_1, \ldots, u_k)$ and $\Lambda_{(k)} = diag(\lambda_1, \ldots, \lambda_k)$.
Thus the rank k approximation of a $N \times s$ matrix $C = V\Lambda U'$ is given by $P = QR' = V_{(k)}\Lambda_{(k)}U'_{(k)}$ where the index (k) denotes that part of the singular value decomposition that belongs to the k largest singular values of C. This approach will be called the *direct approach*, because the estimators for the coefficient matrices A and B are obtained directly by the singular value decomposition of the (full rank) least squares estimator $\hat{C}$.
Furthermore, the minimum of $tr[(V\Lambda U' - QR')(V\Lambda U' - QR')'] = tr[(\Lambda - V'V_{(k)}\Lambda_{(k)}U'_{(k)}U)(\Lambda - V'V_{(k)}\Lambda_{(k)}U'_{(k)}U)']$ results in $\sum_{i=k+1}^{N}\lambda_i^2$.

A generalization of theorem 5.3.1 is given by the following theorem:

Theorem 5.3.2. *Let $Z = (Y, X)$ be the joint matrix of the target matrix Y and the matrix of explanatory variables X with dimension $T \times (N + s)$. Let the mean vector of Z, $\mu_Z \in \mathbb{R}^{N+s}$, be $\mathbf{0}$ and its covariance matrix be*

$$cov(Z) = \begin{pmatrix} \Sigma & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{pmatrix},$$

where Σ_{XX} is required to be nonsingular.

Then for any positive definite matrix $\Omega \in \mathbb{R}^{N\times N}$*, matrices* $\hat{A}_{(k)} \in \mathbb{R}^{N\times k}$ *and* $\hat{B}_{(k)} \in \mathbb{R}^{s\times k}$ *with* $k \leq min(N, s)$ *exist, which minimize*

$$trace[\Omega^{\frac{1}{2}}(Y - XBA')'(Y - XBA')\Omega^{\frac{1}{2}}]. \tag{5.10}$$

They are given by

$$\begin{aligned}\hat{A}_{(k)} &= \Omega^{-\frac{1}{2}}(v_1, \ldots, v_k) = \Omega^{-\frac{1}{2}}V_{(k)}\\ \hat{B}_{(k)} &= \Sigma_{XX}^{-1}\Sigma_{XY}\Omega^{\frac{1}{2}}V_{(k)},\end{aligned}$$

where $V_{(k)} = (v_1, \ldots, v_k)$ *is the matrix of the* k *largest eigenvectors of the matrix* $\Omega^{\frac{1}{2}}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}\Omega^{\frac{1}{2}}$ *belonging to the eigenvalues* $(\lambda_1^2, \ldots, \lambda_k^2)$*.*

Proof. Equation (5.10) can be rewritten as

$$\begin{aligned}&trace[\Omega^{1/2}(\Sigma - AB'\Sigma_{XY} - \Sigma_{YX}BA' + AB'\Sigma_{XX}BA')\Omega^{1/2}] =\\ &= trace[\Omega^{1/2}(\Sigma - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY})\Omega^{1/2}]+\\ &\quad + trace[\Omega^{1/2}(\Sigma_{YX}\Sigma_{XX}^{-1/2} - AB'\Sigma_{XX}^{1/2})(\Sigma_{YX}\Sigma_{XX}^{-1/2} - AB'\Sigma_{XX}^{1/2})'\Omega^{1/2}].\end{aligned}$$

Minimizing it with respect to A and B means minimizing the last line of the above equation, which can be done easily with the help of the results of theorem 5.3.1. If C is set as $\Omega^{1/2}\Sigma_{YX}\Sigma_{XX}^{-1/2}$ and P as $\Omega^{1/2}AB'\Sigma_{XX}^{1/2}$, the quantities Q and R of the previous theorem are given as

$$Q = \Omega^{1/2}\hat{A}_{(k)} \quad \text{and} \quad R = \Sigma_{XX}^{1/2}\hat{B}_{(k)}.$$

The minimum of the objective function in equation (5.10) is then given by $trace(\Sigma\Omega) - \sum_{i=1}^{k}\lambda_i^2$.

□

Hence, the optimal low rank approximation of C is given by

$$\hat{C}_{(k)} = \hat{A}_{(k)}\hat{B}'_{(k)} = \Omega^{-1/2}V_{(k)}V'_{(k)}\Omega^{1/2}\Sigma_{YX}\Sigma_{XX}^{-1} = P_\Omega\Sigma_{YX}\Sigma_{XX}^{-1},$$

where P_Ω is an idempotent but not necessarily symmetric matrix. The above equation also shows, that for $k = N$ the optimal matrix $\hat{C}_{(N)}$ is equal to the full rank least squares estimator $\hat{C}$.

Nevertheless, it is a well known result, that there is no advantage compared to single linear regression models if a multivariate regression model is estimated by ordinary least squares (OLS) with a coefficient matrix of full rank $k = N$. The reasonability for estimation in a multivariate framework is apparent when for example additional rank restrictions are imposed on the parameter matrix C. It has already been mentioned before, that the decomposition of

C into matrices A and B of rank k is just unique except for transformations with a regular matrix. So the multiplication of $\tilde{A} = AS$ with $\tilde{B}' = S^{-1}B'$ with a regular matrix S of rank k yields the same solution AB'. Moreover, in theorem 5.3.2 the normalization of the eigenvectors has been required, which means that $V'_{(k)}V_{(k)} = I_k$. This last restriction is equivalent to the normalization of the parameter matrices A and B in the following way:

$$B'\Sigma_{XX}B = \begin{pmatrix} \lambda_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_k^2 \end{pmatrix} \quad \text{and} \quad A'\Omega A = I_k. \tag{5.11}$$

Another remark worth noting here is the fact, that in theorem 5.3.1 the optimal matrix R was obtained for a given Q and then the optimal Q was calculated. This is equivalent to deriving first B in terms of A and afterwards the optimal matrix A.
Conversely, one could fix B before, calculate an optimal A based on B and then derive the matrix B. Considering the model stated in equation (5.6), the model could be interpreted in the following way:

$$y_t = A(B'x_{t-1}) + \epsilon_t = Af_t + \epsilon_t, \tag{5.12}$$

where f_t represents a factor process and A can be seen as its matrix of loadings.
Assuming that $f_t = B'x_{t-1}$ is given, the matrix A can be calculated by regressing y_t on f_t:

$$\hat{A} = \Sigma_{YX}B(B'\Sigma_{XX}B)^{-1}. \tag{5.13}$$

Substituting this ordinary least squares estimator in equation (5.10) of theorem 5.3.2 and making use of the equality $trace(UV) = trace(VU)$ for all matrices $U \in \mathbb{R}^{m\times n}$ and $V \in \mathbb{R}^{n\times m}$, the objective function used there simplifies to

$$trace[\Sigma\Omega] - trace[(B'\Sigma_{XX}B)^{-1}B'\Sigma_{XY}\Omega\Sigma_{YX}B]. \tag{5.14}$$

As shown in Reinsel and Velu [60] on page 32, the optimum is achieved when choosing the columns of $\Sigma_{XX}^{1/2}B$ as the eigenvectors corresponding to the k largest eigenvectors of

$$\Sigma_{XX}^{-1/2}\Sigma_{XY}\Omega\Sigma_{YX}\Sigma_{XX}^{-1/2}.$$

Hence here the eigenvectors of CC' are needed for deriving an explicit solution for the component matrices A and B whereas when fixing A first the eigenvectors of $C'C$ are required.

Going back to the original task of estimating the parameters A and B in the reduced rank model

$$y_t = AB'x_{t-1} + \epsilon_t = Cx_{t-1} + \epsilon_t, \qquad t = 1, \ldots, T$$

where the ϵ_t are independent with zero mean vector and positive definite covariance matrix Σ_ϵ, one may consider the methodology described in theorem 5.3.2 similar to the approach used for canonical correlation analysis. With the choice $\Omega = \Sigma^{-1}$ it can be interpreted as follows.
First the above equation will be premultiplied with $\Sigma^{-1/2}$ which leads to a standardized matrix of observations as response variable:

$$\begin{aligned}\Sigma^{-1/2} y_t &= \Sigma^{-1/2} AB' x_{t-1} + \Sigma^{-1/2}\epsilon_t = \\ &= \Sigma^{-1/2} C \Sigma_{XX}^{1/2} \Sigma_{XX}^{-1/2} x_{t-1} + \Sigma^{-1/2}\epsilon_t, \qquad t = 1, \ldots, T.\end{aligned}$$

Rewriting this model in a more compact way gives

$$Y\Sigma^{-1/2} = X\Sigma_{XX}^{-1/2}\Sigma_{XX}^{1/2} C'\Sigma^{-1/2} + \epsilon\Sigma^{-1/2}$$

or

$$Y^{(s)} = X^{(s)}\Sigma_{XX}^{1/2} C'\Sigma^{-1/2} + \epsilon\Sigma^{-1/2},$$

where $Y^{(s)} = Y\Sigma^{-1/2}$ and $X^{(s)} = X\Sigma_{XX}^{-1/2}$ are the standardized response and predictor matrices respectively.
Denoting by $\Sigma_{XX}^{1/2}\hat{C}'\Sigma^{-1/2}$ the least squares estimator of the above regression, this matrix can be decomposed in analogy to the direct approach by means of a singular value decomposition

$$\Sigma_{XX}^{1/2}\hat{C}'\Sigma^{-1/2} = U\Lambda V'.$$

Note, that here U, Λ and V are different from the ones obtained in the direct approach. Again just the k largest singular values $\Lambda_{(k)}$ and the corresponding left and right singular vectors $U_{(k)}$ and $V_{(k)}$ are retained.
Then the final rank k estimator for C is

$$\hat{C}_{(k)} = \Sigma^{1/2} V_{(k)} \Lambda_{(k)} U'_{(k)} \Sigma_{XX}^{-1/2}. \tag{5.15}$$

Because of modifying the principal equation before reducing the rank of its regressor matrix, this methodology is called the *indirect procedure*.

For the previously chosen matrix $\Omega = \Sigma^{-1}$ Rao [58] has shown an even stronger result for the solutions $A_{(k)}$ and $B_{(k)}$ minimizing the objective function in theorem 5.3.2. He proves that for this specific choice of Ω the obtained coefficient matrices minimize even all the eigenvalues of the matrix given in equation (5.10) simultaneously.

Another possible choice for Ω could be $\tilde{\Sigma}_\epsilon^{-1}$, which is the inverse of the maximum likelihood

estimate of the error covariance matrix of the unrestricted model, that is given by

$$\tilde{\Sigma}_\epsilon = \frac{1}{T}(Y - \tilde{C}X)'(Y - \tilde{C}X),$$

where $\tilde{C}$ denotes the full rank estimate for the overall coefficient matrix. Robinson [61] showed that with this choice for Ω the optimal component estimates in equation (5.10) are the maximum likelihood estimates under the assumption, that the noise ϵ_t is Gaussian, i.e. independent and identically normal distributed ($iid\mathcal{N}$) with mean vector zero and covariance Σ_ϵ.
So maximum likelihood estimation is another possibility to calculate estimates for the parameters of a reduced rank model. Therefore the slightly modified log-likelihood function, which is given by

$$\log L(C, \Sigma_\epsilon) = \frac{T}{2}\left[\log|\Sigma_\epsilon^{-1}| - trace\left(\Sigma_\epsilon^{-1}\frac{1}{T}(Y - CX)'(Y - CX)\right)\right], \tag{5.16}$$

has to be maximized. $|.|$ stands for the determinant of the matrix. Irrelevant constants, that do not depend on C or Σ_ϵ, have been removed in equation (5.16) for means of simplicity. If Σ_ϵ is unknown, its maximum likelihood solution is $\hat{\Sigma}_\epsilon = \frac{1}{T}(Y - CX)'(Y - CX)$. When substituting this expression in the above equation and writing C as AB', it can be simplified further to

$$\log L(A, B, \hat{\Sigma}_\epsilon) = -\frac{T}{2}\left[\log\left|\frac{1}{T}(Y - AB'X)'(Y - AB'X)\right| + N\right]. \tag{5.17}$$

Obviously, the maximum of equation (5.17) is obtained if

$$\left|\frac{1}{T}(Y - AB'X)'(Y - AB'X)\right| \tag{5.18}$$

is minimized.
A well known result from algebra is that all the eigenvalues of a positive definite matrix A_1 are positive and therefore the determinant $|A_1|$, which is the product of these eigenvalues, has to be positive too. Taking into account furthermore, that the equality $|A_1A_2| = |A_1|\,|A_2|$ holds for two matrices A_1 and A_2 of appropriate dimension, the objective function

$$\left|\tilde{\Sigma}_\epsilon^{-1}\frac{1}{T}(Y - AB'X)'(Y - AB'X)\right| \tag{5.19}$$

yields the same optimal rank deficient matrices as the expression in equation (5.18). $\tilde{\Sigma}_\epsilon^{-1}$ denotes again the maximum likelihood estimator for the covariance matrix of the innovations in the case of a full rank coefficient matrix C, which is a fixed positive definite matrix, so that $|\tilde{\Sigma}_\epsilon^{-1}|$ is a positive value.

If $\frac{1}{T}(Y - AB'X)'(Y - AB'X)$ is rewritten as

$$\begin{aligned}\frac{1}{T}(Y - AB'X)'(Y - AB'X) &= \frac{1}{T}\left(Y - \tilde{C}X + (\tilde{C} - AB')X\right)'\left(Y - \tilde{C}X + (\tilde{C} - AB')X\right) = \\ &= \frac{1}{T}\left(Y - \tilde{C}X\right)'\left(Y - \tilde{C}X\right) + \frac{1}{T}\left(\tilde{C} - AB'\right)'X'X\left(\tilde{C} - AB'\right) = \\ &= \tilde{\Sigma}_\epsilon + \left(\tilde{C} - AB'\right)'\Sigma_{XX}\left(\tilde{C} - AB'\right)\end{aligned}$$

the expression $\left|\tilde{\Sigma}_\epsilon^{-1}\frac{1}{T}(Y - AB'X)'(Y - AB'X)\right|$ can be modified as

$$\left|I_N + \tilde{\Sigma}_\epsilon^{-1}\left(\tilde{C} - AB'\right)'\Sigma_{XX}\left(\tilde{C} - AB'\right)\right| = \prod_{i=1}^{N}(1 + \delta_i^2),$$

where I_N is the $N \times N$ identity matrix and δ_i^2, $i = 1, \ldots, N$, are the eigenvalues of the matrix $\tilde{\Sigma}_\epsilon^{-1}\left(\tilde{C} - AB'\right)'\Sigma_{XX}\left(\tilde{C} - AB'\right)$. Hence, minimizing the objective function in equation (5.19) is equivalent to minimize simultaneously all the eigenvalues of

$$\tilde{\Sigma}_\epsilon^{-1/2}\left(\tilde{C} - AB'\right)'\Sigma_{XX}\left(\tilde{C} - AB'\right)\tilde{\Sigma}_\epsilon^{-1/2} =: (C^{(*)} - P)'(C^{(*)} - P)$$

with $C^{(*)} = \tilde{\Sigma}_\epsilon^{-1/2}\tilde{C}\Sigma_{XX}^{1/2}$ and $P = \tilde{\Sigma}_\epsilon^{-1/2}AB'\Sigma_{XX}^{1/2}$. In analogy to lemma 2.2.6 a similar result can be stated for singular values instead of eigenvalues in order to derive the minimum of the expression above:

Lemma 5.3.2. *For a rank N matrix $C^{(*)} \in \mathbb{R}^{N\times s}$ and a matrix $P \in \mathbb{R}^{N\times s}$ of rank $k \leq N$ the following inequality holds for any i:*

$$\lambda_i(C^{(*)} - P) \geq \lambda_{k+i}(C^{(*)}),$$

where $\lambda_i(C^{(*)})$ denotes the i^{th} largest singular value of $C^{(*)}$ and $\lambda_{k+i}(C^{(*)}) = 0$ for $k + i \geq N$. The equality is attained iff P is defined as the best rank k approximation of $C^{(*)}$, i.e. for the singular value decomposition of $C^{(*)} = V\Lambda U'$ its approximation P is given as $V_{(k)}\Lambda_{(k)}U'_{(k)}$, where the subscript (k) indicates again that just the first k singular values and their corresponding left and right singular vectors are used.

According to the above lemma the required minimum of $(C^{(*)} - P)'(C^{(*)} - P)$ is achieved if P is chosen as best rank k approximation of $C^{(*)}$, i.e. for $\tilde{\Sigma}_\epsilon^{-1/2}\tilde{C}\Sigma_{XX}^{1/2} = V\Lambda U'$ it is given by

$$P = V_{(k)}\Lambda_{(k)}U'_{(k)} = V_{(k)}V'_{(k)}C^{(*)} = V_{(k)}V'_{(k)}\tilde{\Sigma}_\epsilon^{-1/2}\tilde{C}\Sigma_{XX}^{1/2} =: \tilde{\Sigma}_\epsilon^{-1/2}\tilde{A}_{(k)}\tilde{B}'_{(k)}\Sigma_{XX}^{1/2}.$$

Thus the maximum likelihood estimate $\tilde{C}_{(k)}$ of rank k can be calculated as

$$\tilde{C}_{(k)} = \tilde{A}_{(k)}\tilde{B}'_{(k)} = \tilde{\Sigma}_\epsilon^{1/2} V_{(k)} V'_{(k)} \tilde{\Sigma}_\epsilon^{-1/2} \tilde{C}, \tag{5.20}$$

which gives the same optimal solution as theorem 5.3.2 with $\Omega = \tilde{\Sigma}_\epsilon^{-1}$. Because of the equality of the full rank ML estimator $\tilde{C}$ and the full rank least squares estimator $\hat{C}$, the recently deduced rank k approximation gives also the best approximation of the least squares estimator $\hat{C}$. Under the assumption that the noise ϵ_t is independent and identically normal distributed with mean vector $\mathbf{0}$ and covariance matrix Σ_ϵ, these maximum likelihood estimates $\tilde{C}_{(k)} = \tilde{A}_{(k)}\tilde{B}'_{(k)}$ are proven to be asymptotically efficient.

Note furthermore, that in equation (5.8) $N-k$ (unknown) restrictions $l_i' C_{(k)} = 0$ are defined for $i = 1, \ldots, N-k$, which can be seen as the complementary problem. The estimates above can be used now to write down l_i' explicitly:

$$l_i' = v_i' \Omega^{1/2} \qquad \text{for} \quad i = 1, \ldots, N-k.$$

With the help of this definition equation (5.8) can be restated as

$$l_i' C_{(k)} = v_i' \Omega^{1/2} C_{(k)} = v_i' \Omega^{1/2} \Omega^{-1/2} V_{(k)} B_{(k)} = 0 \qquad \text{for} \quad i = 1, \ldots, N-k,$$

because of the orthogonality of the eigenvectors $\{v_1, \ldots, v_N\}$, what proves the validity of the choice of l_i'.

Another aspect, that should be mentioned, is the fact that the choice of $\Omega = \Sigma^{-1}$ or $\Omega = \Sigma_\epsilon^{-1}$ leads to different parameter estimates $\hat{A}_{(k)}$ and $\hat{B}_{(k)}$ respectively $\tilde{A}_{(k)}$ and $\tilde{B}_{(k)}$. Nevertheless, the final result for the optimal low rank coefficient matrix $C_{(k)}$ stays the same, i.e.

$$\hat{C}_{(k)} = \hat{A}_{(k)}\hat{B}_{(k)} = \tilde{A}_{(k)}\tilde{B}_{(k)} = \tilde{C}_{(k)}.$$

5.4 Further specifications

In literature there exist various ways of generalization or adaption of reduced rank regression models as presented here. One possibility consists of allowing for autoregressive errors. Another example would be the model class of reduced rank autoregressive models that try to find a low rank approximation of the coefficient matrix of a vector autoregressive (VAR) model. Or one may impose rank restrictions on seemingly unrelated regression (SURE) models. All these models are explained in more detail in Reinsel and Velu [60].

However, in this thesis more emphasis will be given again on possible zero restrictions imposed on the parameters of the model in a similar way as presented in chapter 3 for the case of the principal component model. Details concerning this aspect are described in the following chapter.

Chapter 6

Sparse reduced rank regression model

The aim of this chapter consists of defining a sparse reduced rank regression model and proposing an estimation methodology similar to the one explained in chapter 3.4 for PCA models. As principal component models can be seen as a special case of reduced rank models it seems to suggest itself to choose a similar way of proceeding as in the former case.

6.1 The model

A sparse reduced rank regression model is a reduced rank model which has zero restrictions incorporated. The sparseness is defined here in the same sense as in the former explanations, namely by imposing zero restriction on the coefficient matrix L of equation (1.1).
This means that the model of interest is of the form

$$y_t = Cx_{t-1} + \epsilon_t = A\underbrace{B'x_{t-1}}_{f_t} + \epsilon_t = Af_t + \epsilon_t, \tag{6.1}$$
$$s.t. \quad \Psi\, vec(A') = \mathbf{0},$$

where $y_t \in \mathbb{R}^N$, $x_{t-1} \in \mathbb{R}^s$ and C is a rank k matrix that can be expressed as the product of a sparse matrix $A \in \mathbb{R}^{N\times k}$ with a regular matrix $B' \in \mathbb{R}^{k\times s}$ which are both of full rank $k < \min(N, s)$.

6.2 Estimation of the sparse reduced rank model

Taking into account the estimation of the unrestricted model as described in section 5.3, it is obvious that the solution obtained through the singular value decomposition does not lead to

the desired result of obtaining a sparse estimate for the parameter matrix A, that obeys certain optimality conditions. Although the original full rank coefficient matrix C can be estimated with least squares under consideration of additional zero restrictions of certain entries of this matrix, the zeros cannot be retained when approximating C by a lower rank approximation neglecting the smallest singular values in the decomposition. Thus another approach has to be adopted to incorporate additional sparsity constraints.

It can be seen easily that the reduced rank model is nonlinear in the parameter matrices A and B. Nevertheless, according to the remark mentioned on page 65, its structure can be regarded as bilinear which allows for certain computational simplifications.
To induce the idea behind the algorithm that estimates such restricted reduced rank models, an alternative for estimating the unrestricted model is described first. Therefore the objective function

$$trace[\Omega^{\frac{1}{2}}(Y - XBA')'(Y - XBA')\Omega^{\frac{1}{2}}], \tag{6.2}$$

that minimizes the sum of the weighted squared error of the model, will be considered again. The first order equations obtained when building the first partial derivatives of the objective function of equation (6.2) with respect to A and B and setting them equal to zero are given by

$$\Sigma_{YX}B - AB'\Sigma_{XX}B = 0 \tag{6.3}$$

and

$$A'\Omega\Sigma_{YX} - (A'\Omega A)B'\Sigma_{XX} = 0. \tag{6.4}$$

Hence as already previously observed (see equation (5.13)) equation (6.3) states that the solution for A depending on B is given by

$$A = \Sigma_{YX}B(B'\Sigma_{XX}B)^{-1}. \tag{6.5}$$

In the same way equation (6.4) leads to an estimator for B depending on A, namely

$$B = \Sigma_{XX}^{-1}\Sigma_{XY}\Omega A(A'\Omega A)^{-1}. \tag{6.6}$$

As already described in equation (5.11) some normalization conditions have to be imposed on the parameter matrices A and B to ensure the uniqueness of the obtained result. So $A'\Omega A = I_k$ has to be valid and the i^{th} element in the diagonal of $B'\Sigma_{XX}B$ has to be equal to λ_i^2, which denotes the i^{th} eigenvalue of the matrix $\Omega^{\frac{1}{2}}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}\Omega^{\frac{1}{2}}$ for $i = 1, \ldots, k$. The off-diagonal elements of $B'\Sigma_{XX}B$ are zero.
Substituting these restrictions into the equations (6.5) and (6.6), they can be simplified further

to

$$A = \Sigma_{YX} B \begin{pmatrix} \frac{1}{\lambda_1^2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\lambda_k^2} \end{pmatrix} \tag{6.7}$$

and

$$B = \Sigma_{XX}^{-1} \Sigma_{XY} \Omega A. \tag{6.8}$$

The above equations indicate again that A can be calculated in terms of B and vice versa. Thus it is self-evident to estimate these parameter matrices iteratively which leads to a procedure that is similar to the one known as partial least squares estimation (PLS) in literature. The difference between these methodologies lies in the manner of factor extraction. In the case of reduced rank regression the aim is to select factors that account for as much variation of the response variable Y as possible without taking into account the variation of the predictor variables X. However, partial least squares regression selects factors of X and Y that have maximum covariance.

Taking all these considerations into account, additional restrictions on the parameter matrix A will be imposed by applying a similar methodology as described before. Therefore equation (6.1) will be restated in a more compact way as

$$\begin{aligned} Y &= XBA' + \epsilon = FA' + \epsilon \\ s.t. \quad & \Psi \; vec(A') = \mathbf{0}, \end{aligned} \tag{6.9}$$

where the variables have the same meaning as in the previous equations and Ψ is defined in such a way, that the resulting coefficient matrix $\hat{A}$ has zero restrictions on certain predefined positions.
Now let $G \in \mathbb{R}^{m \times n}$ and $H \in \mathbb{R}^{r \times q}$ be two arbitrary matrices. Then $vec(.)$ denotes again the *vec operator*, that stacks the columns of the matrix $G = [g_1, \ldots, g_n]$ with $g_i = (g_{1i}, \ldots, g_{mi})'$ into a vector $vec(G) = (g_{11}, g_{21}, \ldots, g_{m1}, g_{12}, \ldots, g_{mn})'$, and $G \otimes H$ characterizes the *Kronecker product* of two matrices G and H as described on page 47, that results in a matrix of dimension $mr \times nq$. Further matrix rules based on the *vec operator* and the *Kronecker product* can be found in the appendix.

With the help of these two operators equation (6.9) can be reformulated as

$$vec(Y) = (I_N \otimes XB) vec(A') + vec(\epsilon) \tag{6.10}$$

or as

$$vec(Y) = (A \otimes X)vec(B) + vec(\epsilon). \tag{6.11}$$

Now suppose that linear restrictions for the parameter matrix A are given as

$$vec(A') = R_A\beta_A + r_A, \tag{6.12}$$

where the vector β_A denotes an unrestricted vector of unknown parameters and R_A and r_A are predefined by the practitioner and therefore assumed as known.
Note, that an alternative way of notation for defining restrictions for the vector $vec(A')$ is given by $\Psi\ vec(A') = c$ which is equivalent to the one that is defined here. Assuming that the first p columns of Ψ are linearly independent the matrix Ψ and the vector $vec(A')$ can be partitioned in such a way that the equations of the restrictions can be written as

$$[\Psi_1 \quad \Psi_2] \begin{bmatrix} vec(A')_1 \\ vec(A')_2 \end{bmatrix} = \Psi_1 vec(A')_1 + \Psi_2 vec(A')_2 = c,$$

where Ψ_1 contains the first p columns of Ψ. So choosing $R_A = \begin{bmatrix} -\Psi_1^{-1}\Psi_2 \\ I \end{bmatrix}$, $\beta_A = vec(A')_2$ and $r_A = \begin{bmatrix} \Psi_1^{-1}c \\ 0 \end{bmatrix}$ this approach leads to the same equations of restrictions.
For the purpose of defining zero restrictions in A the vector c and thus r_A are both vectors of zeros and thus equation (6.12) simplifies to

$$vec(A') = R_A\beta_A, \tag{6.13}$$

where β_A contains exactly those elements of $vec(A')$, that are not zero. The optimization problem of interest for the estimation of A and a known matrix B can now be restated as

$$\begin{aligned} vec(Y) &= (I_N \otimes XB)vec(A') + vec(\epsilon) \\ s.t. \quad vec(A') &= R_A\beta_A, \end{aligned}$$

or as

$$vec(Y) = (I_N \otimes XB)R_A\beta_A + vec(\epsilon) = \tilde{X}\beta_A + vec(\epsilon). \tag{6.14}$$

Then the ordinary least squares estimate of β_A for given B and R_A is obtained by

$$\begin{aligned} \hat{\beta}_A(B) = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'vec(Y) &= \left[R_A'(I_N \otimes B'X'XB)R_A\right]^{-1} R_A'(I_N \otimes B'X')vec(Y) \\ &= \left[R_A'(I_N \otimes B'\Sigma_{XX}B)R_A\right]^{-1} R_A'vec(B'\Sigma_{XY}). \end{aligned}$$

Substituting this estimate $\hat{\beta}_A$ in equation (6.13) gives then the restricted estimator for $vec(A')$ resp. A:

$$\widehat{vec(A')}(B) = R_A \hat{\beta}_A(B). \tag{6.15}$$

On the other hand, if A is known, an estimate for B can be obtained due to the following considerations. Equation (6.11) shows that $vec(B)$ can be estimated by a simple least squares estimate, i.e.

$$\begin{aligned}\widehat{vec(B)}(A) &= \left[(A \otimes X)'(A \otimes X)\right]^{-1} (A \otimes X)' vec(Y) \\ &= \left[(A'A)^{-1} \otimes (X'X)^{-1}\right] vec(X'YA) = vec\left((X'X)^{-1}X'YA(A'A)^{-1}\right). \end{aligned} \tag{6.16}$$

Thus,

$$\hat{B}(A) = (X'X)^{-1}X'YA(A'A)^{-1} = \Sigma_{XX}^{-1}\Sigma_{XY}\left[A^{+}\right]'. \tag{6.17}$$

So basically a similar result as in the case of PCA is found. When setting $X := Y$ as it is the case in the principal component model, the same estimator $\hat{B}$ as in section 3.4 will be obtained. Note, that the above result will also be obtained, when equation (6.9) is postmultiplied with the transpose of the Moore Penrose Pseudoinverse $A^{+} = (A'A)^{-1}A'$ and then the coefficient matrix B of the resulting equation is estimated by the method of ordinary least squares.

Instead of estimating the above equation with ordinary least squares one may prefer a weighted or generalized least squares estimator, which has in general a smaller asymptotic covariance matrix in the sense of the ordering of positive semidefinite matrices[1]. So instead of optimizing the sum of squared errors given by

$$f^{(1)}(A, B) = trace[(Y - XBA')'(Y - XBA')]$$

an objective function as in equation (6.2) could be considered:

$$f^{(2)}(A, B) = trace[\Omega^{\frac{1}{2}}(Y - XBA')'(Y - XBA')\Omega^{\frac{1}{2}}], \tag{6.18}$$

which is equivalent to a system of equations given by

$$Y\Omega^{\frac{1}{2}} = XBA'\Omega^{\frac{1}{2}} + \tilde{\epsilon}. \tag{6.19}$$

The optimal estimator of B has already been given in equation (6.6) due to the fact, that no additional restrictions have been added. Solely the factor $A'\Omega A$ cannot be assumed to be

[1]For two given positive semidefinite matrices $A \geq 0$ and $B \geq 0$ of the same dimension A has the property to be smaller then B, i.e. $B \geq A$ if $B - A \geq 0$.

equal to the $k \times k$ identity matrix I_k as in the unrestricted case, because further restrictions are imposed on A, and thus the last term cannot be dropped.
Restating equation (6.19) with the help of the *vec* operator and adding the restriction $vec(A') = R_A\beta_A$ gives

$$(\Omega^{\frac{1}{2}} \otimes I_N)vec(Y) = (\Omega^{\frac{1}{2}} \otimes XB)R_A\beta_A + vec(\tilde{\epsilon}). \tag{6.20}$$

The optimal parameter estimate $\hat{\beta}_A(B)$ is then given by

$$\hat{\beta}_A(B) = \left[R_A'(\Omega \otimes B'\Sigma_{XX}B)R_A\right]^{-1} R_A'vec(B'\Sigma_{XY}\Omega). \tag{6.21}$$

Premultiplying this estimate for β_A with R_A gives the final weighted least squares estimate for $vec(A')$ resp. after resizing for A, which will be called $\hat{A}(B)$.

Based on these two estimates $\hat{A}(B)$ and $\hat{B}(A)$ an iterative procedure can be applied for obtaining the final estimates. So an arbitrary matrix of starting values for $\hat{B}^{(1)}$ has to be defined, which can for example be the unrestricted estimate of the reduced rank regression model. Next, for $i \geq 2$

$$vec(\hat{A}^{(i)}) = R_A\hat{\beta}_A\left(\hat{B}^{(i-1)}\right)$$

and

$$\hat{B}^{(i)} = \Sigma_{XX}^{-1}\Sigma_{XY}\Omega\hat{A}^{(i)}\left[\left(\hat{A}^{(i)}\right)'\Omega\hat{A}^{(i)}\right]^{-1} = \tag{6.22}$$

$$= \Sigma_{XX}^{-1}\Sigma_{XY}\left[\left(\hat{A}^{(i)}\right)^{+}\right]' \tag{6.23}$$

are calculated iteratively. Note, that $\left[\left(\hat{A}^{(i)}\right)'\Omega\hat{A}^{(i)}\right]^{-1}\left(\hat{A}^{(i)}\right)'\Omega$ can also be regarded as pseudoinverse $\left(\hat{A}^{(i)}\right)^{+}$ of $\hat{A}^{(i)}$ as the main property $\hat{A}^{(i)}\left(\hat{A}^{(i)}\right)^{+}\hat{A}^{(i)} = \hat{A}^{(i)}$ is fulfilled.
Furthermore, in each step of the iteration the estimators have to be rescaled in an appropriate way, whereby the normalization conditions of the unrestricted model (see page 65) are not suitable anymore. For the same reasons as in the case of the restricted PCA model the orthogonality of the loadings matrix A can not be required anymore, if additional zero restrictions on this matrix of coefficients are present. So the same restrictions as for restricted PCA models are defined, namely that the columns of the factor matrix $F = XB$ have length 1.

Again the question of identifiability arises. Similar to the arguments given on page 48 for the case of restricted principal component models, conditions can be given for a regular matrix S, that have to be met in order to guarantee the optimality of $\tilde{A}' = SA'$. Therefore, the

transformation S has to fulfill for a given matrix of restrictions Ψ the following equations:

$$\Psi(A \otimes I_k)vec(S) = 0 \tag{6.24}$$

or

$$\Psi(I_N \otimes S)vec(A') = 0. \tag{6.25}$$

Finally, the iteration stops when the relative change of the objective function is beyond a certain threshold τ. If $f_k^{(j)}$ denotes for $j \in 1, 2$ the value of $f^{(1)}$ or $f^{(2)}$ in the k^{th} iteration, a stopping criterion for the algorithm proposed here, is given by:

$$\frac{f_k^{(j)} - f_{k-1}^{(j)}}{f_{k-1}^{(j)}} < \tau.$$

This type of iteration again leads in the case of ordinary least squares estimation as well as in the case of generalized least squares estimation to monotone convergence, which means that

$$f^{(j)}(\hat{A}^{(2)}, \hat{B}^{(1)}) \geq f^{(j)}(\hat{A}^{(2)}, \hat{B}^{(2)}) \geq f^{(j)}(\hat{A}^{(3)}, \hat{B}^{(2)}) \geq f^{(j)}(\hat{A}^{(3)}, \hat{B}^{(3)}) \geq \ldots, \quad j = 1, 2.$$

This property ensures, that the above defined alternating least squares algorithm converges, because $f^{(1)}$ resp. $f^{(2)}$ are bounded below by the values of the (weighted) sum of squared errors of the unrestricted reduced rank regression model. Nevertheless, it could not be proofed, whether the obtained solution is even a local minimum or not.

Chapter 7

Forecasting in reduced rank regression models

As already mentioned in the introduction, the main aim of this thesis is to propose forecasting models relying on restricted PCA and reduced rank models. The way how to proceed in the latter case is obvious. As the equation of interest is already stated in a dynamic way as

$$y_t = AB'x_{t-1} + \epsilon_t, \qquad t = 1, \ldots, T$$

one may define the predictor $\hat{y}_{\tilde{t}+1}|\tilde{t}$ for instance in time $\tilde{t}+1$ based on data available until $\tilde{t}$ as

$$\hat{y}_{\tilde{t}+1}|\tilde{t} = \hat{A}_{\tilde{t}+1}|\tilde{t}\ \hat{B}'_{\tilde{t}+1}|\tilde{t}\ x_{\tilde{t}}.$$

Here again the same settings on the dimensionality of the parameters as in chapter 5 are made. When assuming that the forecasts for the parameter matrices $\hat{A}_{\tilde{t}+1}|\tilde{t}$ resp. $\hat{B}_{\tilde{t}+1}|\tilde{t}$ at time $\tilde{t}+1$ are the naive forecasts $\hat{A}_{\tilde{t}}$ resp. $\hat{B}_{\tilde{t}}$, the final estimate for $\hat{y}_{\tilde{t}+1}|\tilde{t}$ is given immediately by

$$\hat{y}_{\tilde{t}+1}|\tilde{t} = \hat{A}_{\tilde{t}}\hat{B}'_{\tilde{t}}x_{\tilde{t}}. \tag{7.1}$$

Although the number of parameters is already reduced by imposing additional zero restriction in the reduced rank forecasting model, one may try to reduce them even further by doing input selection on the s-dimensional vector x_t. So if one variable is skipped in x_t the number of parameters to estimate in B is reduced by k. This means that a significant reduction in the number of parameters to be estimated can still be achieved by carrying out additionally variable selection. In section 4.3 a methodology proposed by An and Gu [3] was already described that selects a subset of possible candidates of inputs due to information criteria such as the *AIC* or the *BIC*. They measure the tradeoff between the mean square error of the model and the number of free parameters used in the estimation of AB' in relation to the sample

size. In the case of unrestricted reduced rank models this number of parameters $n_u(N, k, p)$ is equal to $nk + kp - k^2$ because of the possible rotation of the loadings matrix with an orthogonal matrix. When estimating a restricted reduced rank model, this number is given by $n_r(N, k, p) = Nk - a + kp - k$, where a denotes the overall number of zero restrictions in the matrix of loadings and here just k has to be subtracted because an orthogonal rotation of the loadings matrix is not possible anymore since the structure of zeros in A would be destroyed. As already mentioned earlier, this property of reduced rank models has to be reduced in the restricted case to requiring the length of the columns of the factors $\|f_t\| = \|B'x_t\| = 1$. If in the case of input selection the number of input variables is reduced from p to a subset of cardinality p_1, the above formulas for $n_u(N, k, p)$ and $n_r(N, k, p)$ have to be updated accordingly.
Now the way to incorporate the methodology of An and Gu [3] in this framework is straightforward. For every possible k, which should be larger than 2 in the restricted case to ensure that every dependent variable is explained by at least one factor, calculate a reduced rank model with the predefined zero restrictions.

Chapter 8

Empirics

Factor models are a standard tool in financial econometrics. As two popular examples the capital asset pricing model (CAPM) and the arbitrage pricing theory (APT) can be mentioned. In this thesis a PCA model and a sparse PCA model as described in the chapters 2 and 3 are implemented and tested with financial time series, namely with equities.
The question that arises is how to measure the goodness of fit of the restricted factor models and how to compare these models with the unrestricted ones. Therefore two definitions will be given before.

Definition 8 (In-sample Period)
Concerning parameter estimation the in-sample period is the historical time span in which the data used for creating and calibrating the econometrical models are observed.

Definition 9 (Out-of-sample Period)
The out-of-sample period is the time span following the in-sample period until the present, in which forecasts are generated based on the parameter estimates obtained in the in-sample period.

Naturally, it is impossible to improve the in-sample results of the unrestricted models when imposing additional restrictions. Nevertheless, out-of-sample an outperformance of the unrestricted model can be expected if, for example, the *'true model'* has zeros on certain positions of its loadings matrix. In the following two sections two possibilities will be given for measuring the out-of-sample goodness of fit of the forecasting models that can be used to compare the results of the unrestricted models with those of the restricted ones. Firstly, *a posteriori model statistics* can be calculated and secondly, a *portfolio evaluation* can be done in order to carry out model selection or model evaluation.

8.1 A posteriori analysis of the model

In order to calculate such model statistics the relative differences of the targets of a model, $y_{\tilde{t}+1} = (y_{\tilde{t}+1,1}, \ldots, y_{t\tilde{+}1,N})'$, have to be compared with the out-of-sample forecasts $\hat{y}_{\tilde{t}+1}|\tilde{t} = (\hat{y}_{\tilde{t}+1,1}|\tilde{t}, \ldots, \hat{y}_{\tilde{t}+1,N}|\tilde{t})'$ with $\tilde{t} < T$. The former vector contains the returns of the target price time series as entries which are calculated as $y_{\tilde{t},i} = \frac{p_{\tilde{t},i}}{p_{\tilde{t}-1,i}} - 1$ with close prices $p_{\tilde{t},i}$ for target i at instant in time $\tilde{t}$. Choosing a window length of T_1 for the estimation of the parameters, forecasts can be generated for the time period between $T_1 + 1$ and $T + 1$. In a next step the forecasts for the instants in time from $T_1 + 1$ to T can be compared with the observations of the target for this time span.
The statistics taken into account in this thesis for model evaluation are the following:

Hit: An out-of-sample hit can be defined as $hit_{t+1,i} = sign(y_{t+1,i}\ \hat{y}_{t+1,i}|t)$, whereby 1 means that the forecasts shows the same direction as the target and -1 vice versa.

Hitrate: The hitrate measures the average number of hits in a certain period of time. Thus it can be stated as $hitrate_i = \frac{1}{T-T_1} \sum_{t=T_1+1}^{T} hit_{t,i}$.

R^2: In analogy to the in-sample coefficient of determination the out-of-sample coefficient of determination can be expressed as $R_i^2 = cor(y_i, \hat{y}_i)^2 * sign(cor(y_i, \hat{y}_i))$ where $y_i = (y_{T_1+1,i}, \ldots, y_{T,i})'$ and $\hat{y}_i = (\hat{y}_{T_1+1,i}|T_1, \ldots, \hat{y}_{T,i}|T-1)'$ are defined as target resp. forecast vector for the i^{th} security and $cor(.)$ stands for the Pearson's coefficient of correlation. Note, that the out-of-sample R^2 need not be in the interval $[0; 1]$ because geometrically speaking no orthogonality between the forecast and the error vector can be assumed. In order to account for the possibility, that the angle between the target and the forecast vector can also be larger than 90 degrees, the squared coefficient of correlation is multiplied additionally with its sign.

8.2 Portfolio evaluation

The three criteria described in the previous section are all based on a certain loss function and thus may not be adequate in this context. Another possibility for evaluating out of sample forecasts of a financial forecasting model, which may be more meaningful, consists in calculating a portfolio evaluation. Therefore the possibility of considering a single- or a multi-asset portfolio exists. A single asset portfolio can be evaluated for each target separately by defining the following investment rules. If the forecast for the next day has a positive sign, a long position is taken. On the other hand, if next days forecast is negative, one may hold a short position. This strategy allows the portfolio value to increase although the value of the underlying financial instrument is falling.
One of the famous approaches for a multiple portfolio optimization is based on the portfolio

theory proposed by Markowitz [53] in 1952. For deriving optimal portfolio weights for the individual financial instruments, the following objective function has to be minimized:

$$\begin{aligned} \min_{w_t \in \mathbb{R}^N} \quad & -w_t' \hat{y}_t | t-1 + \alpha \; w_t' \hat{\Sigma}_t w_t \\ s.t. \quad & \sum_{i=1}^{N} w_{t,i} = 0, \end{aligned}$$

where w_t is a N-dimensional vector of portfolio weights for instant in time t, α is the so called risk aversion factor, a coefficient punishing risky assets in the optimization. $\hat{\Sigma}_t$ is a risk matrix predicted for time t that can, for example, be chosen as historic covariance matrix of the errors of the forecast model or alternatively it could be modeled by a generalized autoregressive conditional heteroscedasticity (GARCH) model.

Because of the fact, that the forecasting accuracy of point forecasts of financial forecasting models in general is quite poor and multivariate portfolio optimization as defined above includes additional tuning or uncertainty parameters, namely the choice of the risk aversion factor α and of the predicted risk matrix $\hat{\Sigma}_t$, within the framework of this thesis just single asset portfolios will be considered as model evaluation criterion.

Furthermore, portfolio statistics can be calculated in order to evaluate different performance curves, which are the graphs of the portfolio values over time. Therefore measures such as *total return, annualized return, annualized volatility, Sharpe ratio or maximum drawdown* are famous criteria for analyzing the performance of financial products.

8.3 World equities

This data set contains 14 of the leading world indices from 2005-07-29 to 2008-09-12. For the empirical research Bloomberg is chosen as a data provider and the Bloomberg Tickers of the targets and their explanations are given in table 8.1[1].

Their discrete weekly returns calculated as $y_{i,t} = \frac{p_{i,t}}{p_{i,t-1}} - 1$ with close prices $p_{i,t}$ for target i at instant in time t are shown in figure 8.1. The volatility of the returns increases a lot after the news about the bankruptcy of Lehman Brothers on September 15^{th} in 2008 spread around, which contradicts the desired assumption of homoscedasticity of econometric time series. Such extraordinary events are far off predictability and therefore the period after September 12^{th}, 2008 will not be included in further calculations.
In table 8.2 the summary statistics of the equities data are listed. Descriptive statistics such as the quartiles, the mean and distributional measures can be found there. This statistics as

[1]The data were provided by C-Quadrat, Vienna.

	Bloomberg Ticker	Field	Description
1	DAX Index	PX_LAST	German Stock Index (30 selected German blue chip stocks)
2	SPX Index	PX_LAST	Standard and Poor's (S&P) 500 Index (capitalization-weighted index of 500 stocks representing all major industries)
3	SMI Index	PX_LAST	Swiss Market Index (capitalization-weighted index of the 20 largest and most liquid stocks of the SPI universe)
4	NDX Index	PX_LAST	NASDAQ 100 Index (modified capitalization-weighted index of the 100 largest and most active non-financial domestic and international issues listed on the NASDAQ)
5	SX5E Index	PX_LAST	EURO STOXX 50 Price Index (free-float market capitalization-weighted index of 50 European blue-chip stocks from those countries participating in the EMU)
6	UKX Index	PX_LAST	FTSE 100 Index (capitalization-weighted index of the 100 most highly capitalized companies traded on the London Stock Exchange)
7	CAC Index	PX_LAST	CAC 40 Index (narrow-based, modified capitalization-weighted index of 40 companies listed on the Paris Bourse)
8	AEX Index	PX_LAST	AEX Index (free-float adjusted market capitalization-weighted index of the leading Dutch stocks traded on the Amsterdam Exchange)
9	INDU Index	PX_LAST	Dow Jones Industrial Average Index (price-weighted average of 30 blue-chip stocks that are generally the leaders in their industry)
10	IBEX Index	PX_LAST	IBEX 35 Index (official index of the Spanish Continuous Market comprised of the 35 most liquid stocks traded on the Continuous market)
11	E100 Index	PX_LAST	FTSE Eurotop 100 Index (modified capitalization-weighted index of the most actively traded and highly capitalized stocks in the pan-European markets)
12	BEL20 Index	PX_LAST	BEL 20 Index (modified capitalization-weighted index of the 20 most capitalized and liquid Belgian stocks that are traded on the Brussels Stock Exchange)
13	SPTSX60 Index	PX_LAST	S&P/Toronto Stock Exchange 60 Index(capitalization-weighted index consisting of 60 of the largest and most liquid stocks listed on the Toronto Stock Exchange)
14	RTY Index	PX_LAST	Russell 2000 Index (is comprised of the smallest 2000 companies in the Russell 3000 Index, representing approximately 8% of the Russell 3000 total market capitalization)

Table 8.1: Bloomberg Tickers, Fields and Description of some of the most important world equities used in this empirical application.

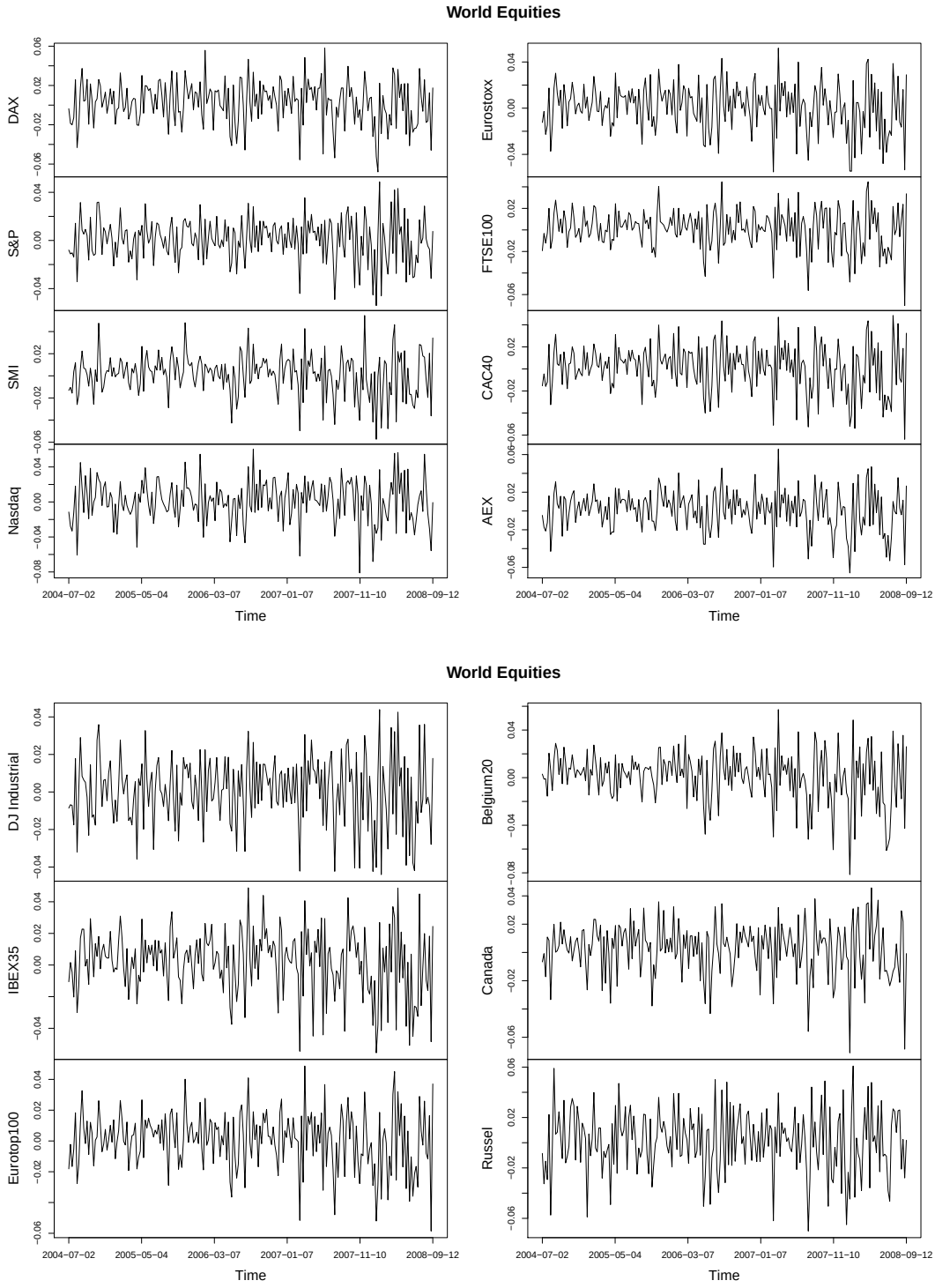

Figure 8.1: weekly returns of world equities from 2005-07-29 to 2008-09-12.

	DAX	S&P	SMI	Nasdaq	Eurostoxx	FTSE100	CAC40
min	−0.0680	−0.0541	−0.0573	−0.0811	−0.0552	−0.0702	−0.0638
1^{st} quantile	−0.0105	−0.0100	−0.0101	−0.0116	−0.0117	−0.0099	−0.0122
median	0.0048	0.0015	0.0037	0.0017	0.0023	0.0015	0.0025
3^{rd} quantile	0.0172	0.0122	0.0123	0.0164	0.0153	0.0136	0.0156
max	0.0580	0.0487	0.0545	0.0603	0.0520	0.0447	0.0485
mean	0.0022	0.0006	0.0013	0.0010	0.0009	0.0010	0.0009
skewness	−0.3863	−0.3303	−0.3576	−0.3299	−0.3983	−0.5312	−0.3830
kurtosis	3.2026	3.3715	3.7185	3.5074	3.0789	3.8820	3.0568

	AEX	DJ Indust.	IBEX35	Eurotop100	Belgium20	Canada	Russel
min	−0.0659	−0.0440	−0.0555	−0.0585	−0.0815	−0.0710	−0.0701
1^{st} quantile	−0.0122	−0.0093	−0.0077	−0.0102	−0.0106	−0.0056	−0.0138
median	0.0023	0.0018	0.0044	0.0014	0.0030	0.0046	0.0025
3^{rd} quantile	0.0141	0.0133	0.0150	0.0122	0.0148	0.0141	0.0189
max	0.0657	0.0439	0.0488	0.0487	0.0572	0.0457	0.0608
mean	0.0009	0.0006	0.0018	0.0007	0.0013	0.0024	0.0012
skewness	−0.3319	−0.3203	−0.5003	−0.3692	−0.6946	−0.7840	−0.2462
kurtosis	3.4800	2.9222	3.2443	3.3607	4.1322	4.4949	2.9886

Table 8.2: Descriptive statistics of the equities data on a weekly basis from 2005-07-29 to 2008-09-12

well as the histograms in figure 8.2 indicate, that one has to be careful when working with financial data because the often required assumption of normal distribution is not always met. The data often show a leptokurtic distribution which means that in comparison with a normal distribution it has higher peaks and so called *fat tails*.

Another problematic characteristic of financial data consists in the presence of a unit root. Therefore the autocorrelation functions of the data are given in figure 8.3, which show no severe problems in the data analyzed here. Moreover, the Augmented Dickey Fuller Test (ADF Test) rejects for all targets the null hypothesis of the presence of a unit root.

In order to estimate restricted factor models, as explained in the previous chapters, a pattern matrix has to be defined a priori, that marks the restricted positions of the loadings matrix with zeros. Here the matrix given in table 8.3 is used, which interprets the first factor as European market and the second one as American market. Therefore the European indices load (mainly) on the first factor and the others on the second one. Solely, *FTSE* 100 loads on both factors because it shows a slightly different behavior than the other European indices and contains partly also assets from other non European countries. Thus, the pattern matrix shows the required structure of not being decomposed entirely in block matrices and of restricting at least $k = 2$ elements in each column to zero, which can not be reached by simple orthogonal rotation.

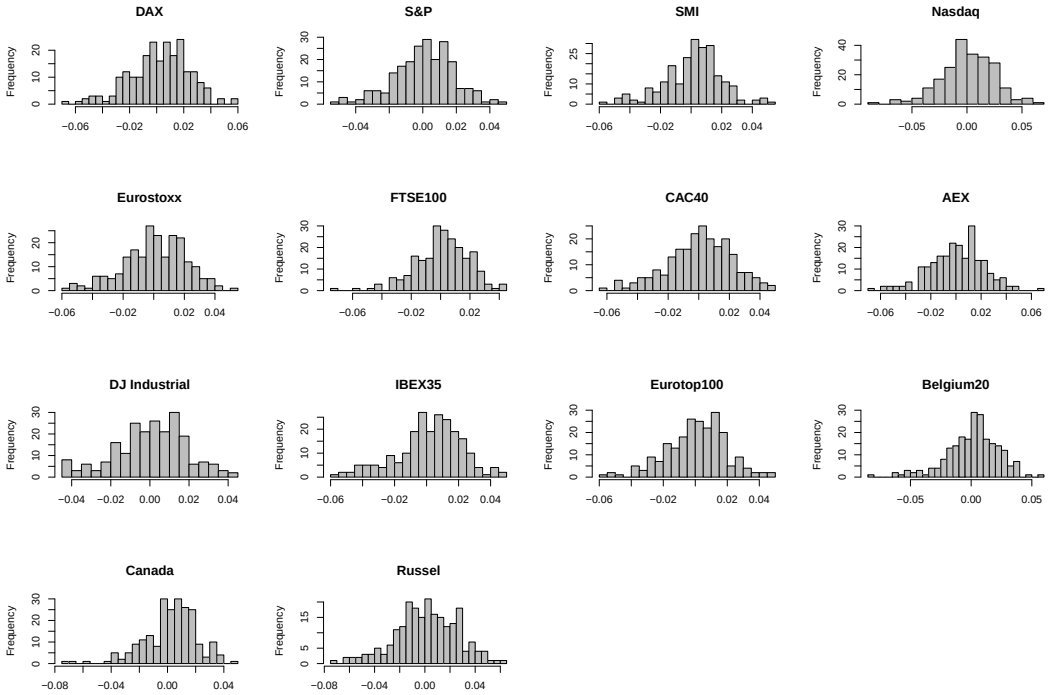

Figure 8.2: histograms of the weekly returns of the equities data from 2005-07-29 to 2008-09-12

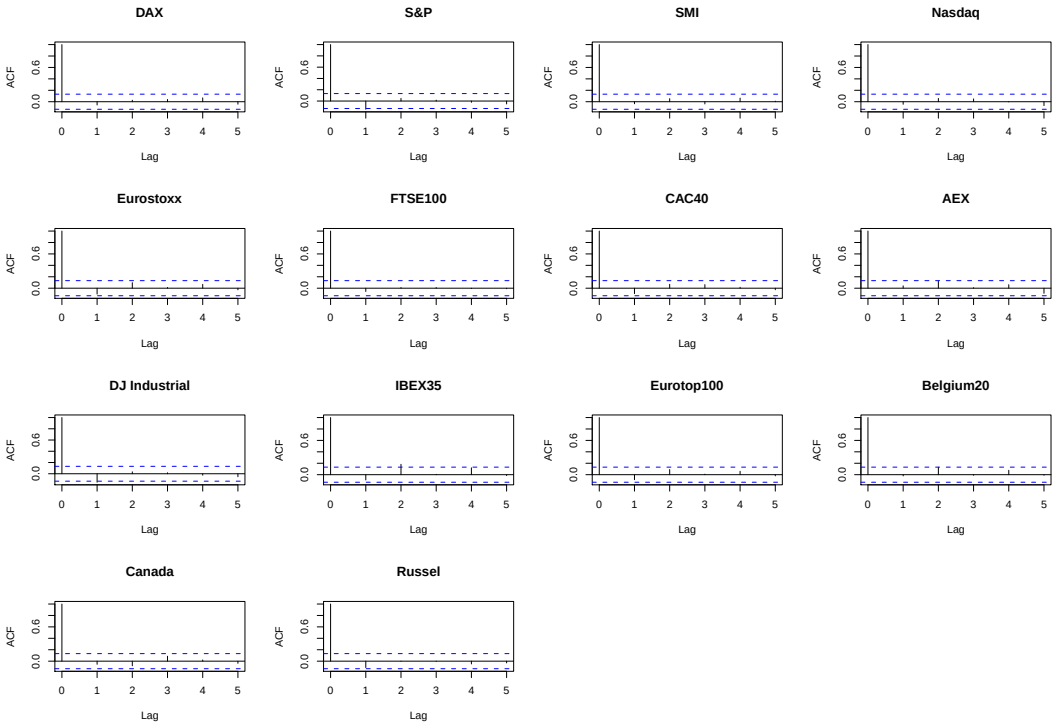

Figure 8.3: autocorrelation function of the weekly returns of the equities data from 2005-07-29 to 2008-09-12

	EU	US
DAX	1	0
S&P	0	1
SMI	1	0
Nasdaq	0	1
Eurostoxx	1	0
FTSE100	1	1
CAC40	1	0
AEX	1	0
DJ Industrial	0	1
IBEX35	1	0
Eurotop100	1	0
Belgium20	1	0
Canada	0	1
Russel	0	1

Table 8.3: Pattern matrix for the world equities data defining the positions of the loadings matrix which are restricted to be zero in the estimation.

Apart from the dependent variables described above also input variables have to be selected and assigned to the different factors. Therefore the variables which can be seen in table 8.4 have been chosen and attributed to the European and American market respectively, which will also be the explanation of the factors later on. The original list of inputs has been reduced to this 17 final variables by means of a cluster and correlation analysis and variables with extreme outliers have been skipped. So the list of possible explanatory variables consists of an intercept, lags 1 to 4 of the lagged dependent variable (4 autoregressive variables) and lags 1 to 4 of the 17 exogenous variables. But not all of these variables are used for calculating the forecast. As described in section 4.3 a subset selection algorithm is applied to reduce the number of inputs further.

8.3.1 Results

Based on the 14 indices and the 17 inputs described in this section an unrestricted and a restricted principal components model have been estimated. As rolling window size 80 observations have been chosen in the estimation. The number of selected inputs in each estimation step has been forced to be between 2 and 10 and is shown in figure 8.4 and 8.5 for both model types.

In table 8.5 an example for an unrestricted (first two columns) versus a restricted (columns 3 and 4) loadings matrix is presented. In the restricted case exact zeros are obtained on the

	Bloomberg Ticker	EU	US	Description
1	USDJPY Curncy	1	1	USD-JPY exchange rate (amount of Japanese Yen for 1 US Dollar)
2	EURUSD Curncy	1	1	EUR-USD exchange rate (amount of US Dollars for 1 Euro)
3	SX8P Index	1	0	DJ Stoxx sector index technology
4	SX4P Index	1	0	DJ Stoxx sector index chemicals
5	SX6P Index	1	0	DJ Stoxx sector index utilities
6	SX7P Index	1	0	DJ Stoxx sector index banks
7	EUR001M Index	1	0	EU 1-month yield curve
8	EUR012M Index	1	0	EU 12-months yield curve
9	RX1 Comdty	1	0	Eurobund future with a 10-year maturity
10	CL1 Comdty	1	1	crude oil future
11	GC1 Comdty	1	1	gold future
12	VDAX Index	1	0	German volatility index
13	TY1 Comdty	0	1	US 10-years treasury note
14	MOODCAVG Index	0	1	Moody's rating and risk analysis index (lagged 1 day)
15	US0012M Index	0	1	US 12-months yield curve
16	USSWAP2 CMPL Curncy	0	1	2-year vanilla interest rate swap
17	USSWAP5 CMPL Curncy	0	1	5-year vanilla interest rate swap

Table 8.4: List of exogenous inputs used for forecasting with their assignment to European and US-based indices. A '1' in the columns 'EU' or 'US' means, that the corresponding input may have predictive power for forecasting the behavior of the European resp. US market and a '0' vice versa. The data is available from 1999-01-01 up to the present.

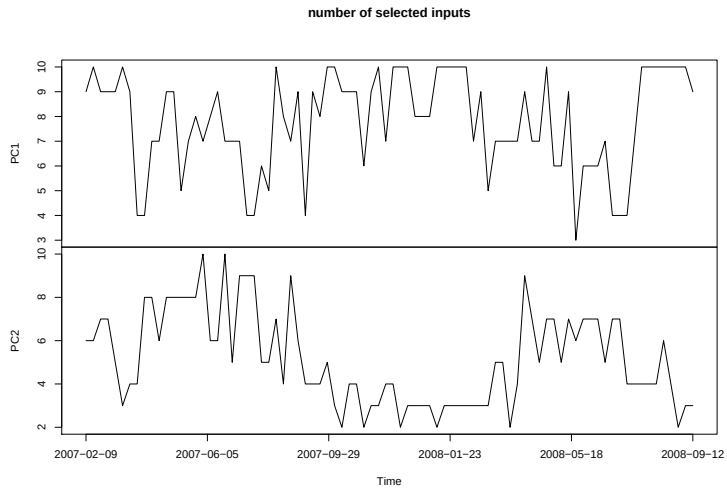

Figure 8.4: Number of selected inputs over time for each principal component for the (unrestricted) principal component forecast model.

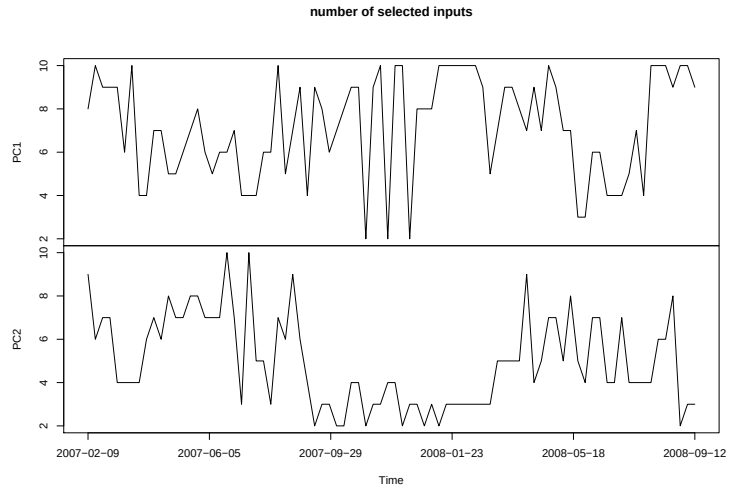

Figure 8.5: Number of selected inputs over time for each modified principal component for the restricted principal component forecast model.

	PC 1	PC 2	PC 1_{restr}	PC 2_{restr}
DAX	0.3413	−0.0377	0.2080	0.0000
S&P	0.0279	0.4221	0.0000	−0.1950
SMI	0.2500	0.0830	0.1923	0.0000
Nasdaq	0.0163	0.5028	0.0000	−0.2229
Eurostoxx	0.3422	−0.0199	0.2163	0.0000
FTSE100	0.3004	0.0062	0.1822	−0.0161
CAC40	0.3497	0.0158	0.2334	0.0000
AEX	0.3199	0.0393	0.2203	0.0000
DJ Industrial	0.0450	0.3861	0.0000	−0.1885
IBEX35	0.3400	−0.0630	0.1996	0.0000
Eurotop100	0.3193	−0.0067	0.2056	0.0000
Belgium20	0.3420	0.0272	0.2334	0.0000
Canada	0.2123	0.0326	0.0000	−0.1399
Russel	−0.0805	0.6353	0.0000	−0.2234

Table 8.5: Example for an unrestricted and a restricted loadings matrix on 2008-09-12.

specified positions of the loadings matrix whereas the unrestricted loadings matrix has just small values in the according positions. To enhance comparability, the loadings matrix of the unrestricted model is rotated by an orthogonal *varimax* rotation as described in section 2.3.

The final out of sample statistics of this analysis can be found in the tables 8.6 and 8.7. There it can be seen that on average the restricted PCA model outperforms the unrestricted one in the sense of a higher average portfolio value starting from a value of 100 on $2007-02-02$(140.96 compared to 133.11). The mean of the R^2 statistics is in both cases

	DAX	S&P	SMI	Nasdaq	Eurostoxx	FTSE100	CAC40
R2	0.0301	−0.0002	0.0083	0.00	0.0265	0.0197	0.0177
Skewness	0.31	−0.05	0.11	−0.25	0.32	0.10	0.25
Kurtosis	2.37	2.25	2.33	2.38	2.35	2.64	2.40
Jarque Bera	0.26	0.37	0.42	0.33	0.23	0.74	0.35
ADF	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Hitrate	0.52	0.51	0.58	0.56	0.51	0.60	0.50
Portfolio value	139.66	118.93	120.01	150.47	134.74	139.22	127.23

	AEX	DJ Industrial	IBEX35	Eurotop100	Belgium20	Canada	Russel
R2	0.029	0.0002	0.0134	0.0231	0.0104	0.0257	−0.001
Skewness	0.23	−0.12	0.29	0.29	0.32	−0.35	−0.09
Kurtosis	2.52	2.18	2.93	2.29	2.52	2.9	2.78
Jarque Bera	0.47	0.28	0.55	0.23	0.32	0.41	0.87
ADF	0.01	0.02	0.01	0.01	0.01	0.02	0.01
Hitrate	0.48	0.56	0.56	0.51	0.52	0.65	0.48
Portfolio value	146.67	115.3	123.58	136.3	131.89	171.6	107.91

Table 8.6: Out-of-sample model statistics of the unrestricted PCA model based on a window length of 80 weekly datapoints for generating 1-step ahead forecasts from 2007-02-09 to 2008-09-12.

similar (0.0121 in the restricted case vs. 0.0145 in the unrestricted one). The restricted model also has a slightly higher average hitrate of 55.87% in comparison to 53.91% for the unrestricted PCA model. For both model types the null hypothesis of normality of the residuals, tested by the Jarque Bera test, cannot be rejected on a confidence level of $\alpha = 0.05$, whereas the Augmented Dickey Fuller (ADF) Test rejects the null hypothesis of the presence of a unit root of the residuals in all cases, if the same confidence level of 0.05 is assumed.

In figure 8.6 the performance curves for all 14 indices from 2007-02-02 to 2008-09-12 can be found. The increase in performance of the European indices is quite promising at the beginning whereas the American ones start performing well in October 2007. The graphic also shows that is quite a difficult issue calculating real out-of-sample econometrical forecasting models that perform well also on the short run at every instant in time. Nevertheless, in the author's opinion it is possible to obtain good results on a long-term basis, that can outperform actively managed portfolios.

To round up the results obtained for the restricted and the unrestricted model of the world equities models, another comparison is given here, that takes just the last 30 weeks before 2008-09-12 into account. There the average performance of the restricted model is again clearly better than the one of the unrestricted PCA model (122.10 vs. 109.72), if 100 is chosen as a starting value on $2008-02-15$ for all securities. Also the hitrate of the restricted model indicates with

	DAX	S&P	SMI	Nasdaq	Eurostoxx	FTSE100	CAC40
R2	0.0180	0.0040	0.0100	0.0104	0.0167	0.0216	0.0086
Skewness	0.35	-0.06	0.27	-0.36	0.40	-0.02	0.36
Kurtosis	2.77	2.34	2.36	2.74	2.57	2.50	2.68
Jarque Bera	0.38	0.45	0.30	0.36	0.24	0.65	0.33
ADF	0.01	0.02	0.01	0.01	0.01	0.01	0.01
Hitrate	0.56	0.54	0.60	0.58	0.55	0.63	0.54
Portfolio value	141.24	132.11	123.78	172.47	141.07	149.68	136.09

	AEX	DJ Industrial	IBEX35	Eurotop100	Belgium20	Canada	Russel
R2	0.0235	0.0093	0.0109	0.0198	0.0078	0.0079	0.0009
Skewness	0.23	-0.02	0.28	0.31	0.32	-0.46	-0.11
Kurtosis	2.61	2.13	2.78	2.35	2.43	3.31	2.76
Jarque Bera	0.54	0.26	0.53	0.25	0.28	0.19	0.83
ADF	0.01	0.02	0.01	0.01	0.01	0.01	0.01
Hitrate	0.51	0.56	0.57	0.55	0.52	0.61	0.51
Portfolio value	143.33	133.17	128.28	140.57	139.82	156.17	135.68

Table 8.7: Out-of-sample model statistics of the restricted PCA model based on a window length of 80 weekly datapoints for generating 1-step ahead forecasts from 2007-02-09 to 2008-09-12.

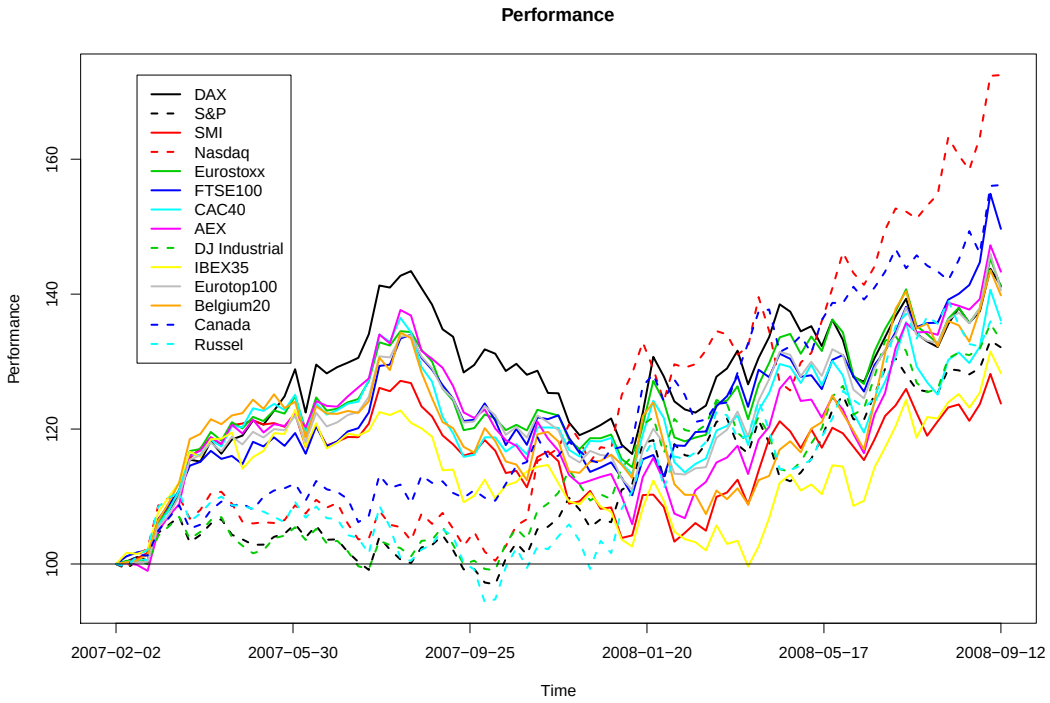

Figure 8.6: Performance curves for all 14 indices from 2007-02-02 to 2008-09-12 based on forecasts calculated with a restricted principal component forecast model. For the European indices solid lines are used and for the American ones dashed lines.

	DAX	S&P	SMI	Nasdaq	Eurostoxx	FTSE100	CAC40
R2	0.0448	-0.0204	0.0031	-0.0220	0.0381	0.0947	0.0410
Skewness	0.53	0.03	0.09	-0.05	0.34	-0.08	0.31
Kurtosis	2.29	2.05	1.91	2.17	1.93	2.63	2.08
Jarque Bera	0.37	0.57	0.47	0.65	0.37	0.90	0.47
ADF	0.13	0.52	0.21	0.37	0.15	0.35	0.30
Hitrate	0.50	0.33	0.53	0.50	0.57	0.67	0.53
Portfolio value	110.91	87.42	102.05	100.84	111.93	124.37	113.46

	AEX	DJ Industrial	IBEX35	Eurotop100	Belgium20	Canada	Russel
R2	0.1338	-0.0210	0.0782	0.0554	0.0373	0.1097	-0.0495
Skewness	0.11	-0.09	0.33	0.30	0.11	-0.24	-0.10
Kurtosis	1.93	1.97	2.05	1.89	2.11	3.25	2.28
Jarque Bera	0.48	0.51	0.43	0.37	0.59	0.83	0.71
ADF	0.42	0.53	0.19	0.26	0.34	0.30	0.45
Hitrate	0.57	0.40	0.53	0.53	0.50	0.77	0.47
Portfolio value	130.63	85.13	114.87	115.94	115.31	133.38	89.83

Table 8.8: Out-of-sample model statistics of the unrestricted PCA model based on a window length of 80 weekly datapoints for generating 1-step ahead forecasts from 2008-02-22 to 2008-09-12 (a period of 30 weeks).

a mean of 60.24% a considerable improvement against the unrestricted one (52.86%). What seems a bit surprising here is the fact, that the average R^2 of the restricted model is worse than the one of the unrestricted model (0.0187 vs. 0.0374). Nevertheless, it has to be taken into account, that no point forecasts are considered in the portfolio evaluation for several reasons mentioned before, and that's why less importance may be given to this statistical measure in this context.

Last but not least some performance statistics of the performance curves of the restricted principal component models and of the indices as a benchmark are summarized in the tables 8.10 and 8.11, respectively. In the sense of generated returns the long/short strategy of the restricted PCA forecasts outperforms clearly the indices themselves whereby the annualized volatility of these two groups of variables are very similar. Smaller drawdowns as well as much higher Sharpe ratios underline furthermore the meaningfulness of the obtained restricted forecasts in combination with the proposed portfolio strategy.

	DAX	S&P	SMI	Nasdaq	Eurostoxx	FTSE100	CAC40
R2	0.0084	0.0015	0.0008	0.0036	0.0079	0.0531	0.0081
Skewness	0.39	0.03	0.22	-0.09	0.27	-0.11	0.29
Kurtosis	2.44	2.16	2.03	2.41	2.12	2.79	2.23
Jarque Bera	0.56	0.64	0.49	0.79	0.51	0.95	0.56
ADF	0.13	0.59	0.16	0.44	0.15	0.35	0.25
Hitrate	0.57	0.50	0.60	0.63	0.63	0.70	0.60
Portfolio value	114.91	113.76	118.32	133.58	119.37	127.38	119.82

	AEX	DJ Industrial	IBEX35	Eurotop100	Belgium20	Canada	Russel
R2	0.0917	0.0017	0.0250	0.0228	0.0141	0.0236	0.0001
Skewness	0.07	0.10	0.23	0.21	0.07	-0.16	-0.44
Kurtosis	2.35	1.94	2.19	2.02	1.98	3.23	2.60
Jarque Bera	0.76	0.49	0.58	0.49	0.51	0.90	0.56
ADF	0.31	0.62	0.31	0.22	0.38	0.46	0.52
Hitrate	0.63	0.57	0.60	0.60	0.57	0.67	0.57
Portfolio value	134.21	111.11	123.72	124.10	126.82	124.87	117.45

Table 8.9: Out-of-sample model statistics of the restricted PCA model based on a window length of 80 weekly datapoints for generating 1-step ahead forecasts from 2008-02-22 to 2008-09-12 (a period of 30 weeks).

	DAX	S&P	SMI	Nasdaq	Eurostoxx	FTSE100	CAC40
Total return %	41.24	32.11	23.78	72.47	41.07	49.68	36.09
Total return p.a. %	23.86	18.83	14.14	40.18	23.77	28.39	21.04
Volatility p.a. %	18.16	16.19	17.29	19.25	17.85	16.67	19.32
Sharpe ratio	1.2	1.04	0.7	1.98	1.22	1.58	0.99
Max. % loss 1 week	-4.97	-4.87	-5.45	-5.65	-4.8	-3.6	-5.39
Max. % loss 5 weeks	-10.47	-7.16	-8.48	-6.58	-10.82	-9.55	-13.74
Max. % loss 20 weeks	-17.95	-7.19	-18.1	-5.66	-14.01	-15.83	-14.48
Max. drawdown %	18.9	9.35	18.74	9.92	15	17.77	16.98

	AEX	DJ Indust.	IBEX35	Eurotop100	Belgium20	Canada	Russel
Total return %	43.33	33.17	28.28	40.57	39.82	56.17	35.68
Total return p.a. %	24.99	19.42	16.69	23.49	23.09	31.82	20.81
Volatility p.a. %	19.19	16.25	18.64	16.84	20.43	15.88	19.93
Sharpe ratio	1.2	1.07	0.79	1.28	1.03	1.88	0.94
Max. % loss 1 week	-4.32	-4.39	-5.08	-3.92	-5.19	-4.57	-6.08
Max. % loss 5 weeks	-10.49	-7.81	-9.73	-9.66	-13.35	-5.26	-9.86
Max. % loss 20 weeks	-19.59	-4.92	-14.57	-15.86	-14.03	-1.66	-11.82
Max. drawdown %	23.08	8.55	18.89	17.11	20.04	5.53	14.77

Table 8.10: Performance statistics of the performance curves obtained of the restricted PCA model in combination with a simple one asset long/short strategy based on data from 2007-02-02 to 2008-09-12.

	DAX	S&P	SMI	Nasdaq	Eurostoxx	FTSE100	CAC40
Total return %	-9.45	-13.58	-22.08	-1.72	-22.48	-14.17	-23.68
Ttotal return p.a. %	-5.97	-8.65	-14.33	-1.07	-14.59	-9.03	-15.42
Volatility p.a. %	18.44	16.36	17.3	19.9	18.03	17.04	19.43
Sharpe ratio	-0.43	-0.65	-0.94	-0.15	-0.92	-0.65	-0.9
Max. % loss 1 week	-6.8	-5.41	-5.73	-8.11	-5.52	-7.02	-6.38
Max. % loss 5 weeks	-14.82	-10.36	-11.34	-15.28	-13.86	-10.92	-14.92
Max. % loss 20 weeks	-19.49	-16.1	-20.43	-21.91	-20.5	-15.85	-20.75
Max. drawdown %	24.29	20.64	30.35	22.87	30.09	22.16	33.52

	AEX	DJ Indust.	IBEX35	Eurotop100	Belgium20	Canada	Russel
Total return %	-21.16	-9.73	-22	-23.03	-31.65	1.73	-11.02
Ttotal return p.a. %	-13.7	-6.15	-14.27	-14.98	-21.01	1.07	-6.98
Volatility p.a. %	19.4	16.46	18.7	17.01	20.46	16.39	20.13
Sharpe ratio	-0.81	-0.5	-0.87	-1	-1.12	-0.06	-0.45
Max. % loss 1 week	-6.59	-4.4	-5.55	-5.85	-8.15	-7.1	-7.01
Max. % loss 5 weeks	-16.92	-10.68	-14.05	-11.96	-19.59	-9.07	-12.35
Max. % loss 20 weeks	-21.04	-13.44	-20	-21.22	-23.67	-9.91	-19.3
Max. drawdown %	30.61	21.23	29.6	29.75	37.69	14.58	22.86

Table 8.11: Performance statistics of the indices themselves as a benchmark from 2007-02-02 to 2008-09-12.

Chapter 9

Conclusion and extensions

The main parts of this thesis are devoted to the development of sparse *principal components* and *reduced rank regression models*. Therefore the unrestricted model classes are presented first and then similar objective functions as in the classical case are defined in order to estimate the unknown parameters, whereby restrictions are imposed on the corresponding matrix of loadings. Based on this specifications an adaptive least squares algorithm is presented as a solution to this optimization problems that works for both model types.
These sparse factor models are used further as forecasting models, whereby for restricted *PCA* a two-step procedure is necessary and for restricted *RRRA* a direct approach can be chosen. The problematic of inputselection for the choice of exogenous or autoregressive variables is done with the help of an algorithm similar to the one proposed by An and Gu [3], which is based on information criteria such as *AIC* and/or *BIC*.
Finally, the directional forecasts of a sparse principal component model for financial instruments are employed in an empirical study in a simple single asset portfolio long/short strategy. The obtained results show that the restricted forecasting model for the 14 indices

- enhances interpretability of the factors
- outperforms the unrestricted model in terms of better out-of-sample model statistics for most of the analyzed targets
- produces higher portfolio values than the forecasts of the unrestricted models.

It is more or less surprising, that post - statistics such as the R^2 give no reliable hint about the quality of the financial forecasts for usage in a portfolio, as even models with not so good R^2 values can bring out a good performance. Nevertheless, the ability of econometric models of generating good point forecasts in finance is limited and therefore one should not impose too much weight on this criterion.
Furthermore, it is shown that the out-of-sample Hitrate contributes in a positive way to the

performance. However, some examples in chapter 8 demonstrate, that even with a Hitrate of 50 percent, which comes close to throwing a coin, one can generate persistently a good performance, if the timing of the signals is right.
Comparing finally the portfolio statistics of the proposed portfolios with their targets, a manifest improvement over the indices themselves can be observed, and therefore utilizations of such restricted forecasts in some areas of the wide range of financial products such as e.g. exchange traded funds (ETFs), which are basically index trackers, can be suggested.

Besides the topics analyzed in the framework of this thesis, there are still a number of open problems or questions which are a matter of future research. In the sequel some of them, which are of interest to the author, are pointed out. Firstly, the procedure gives no indication about the correctness of the assumptions of sparseness as preknowledge is postulated. So the development of statistical tests regarding the meaningfulness of the determined structure of the loadings matrix is up to future research.
Next, modifications of existing sparse principal components techniques explained in section 3.2 such as *SPCA* in order to obtain a sparse loadings matrix A instead of B (see equation (3.13)) would be interesting. To obtain comparability between my technique and others based on the *LASSO*, the penalty coefficients of the *LASSO* components have to be set individually for each element in the loadings matrix separately with an accordingly high value for certain position, where zeros should be enforced.
Theoretically one may also consider new optimization technologies solving the nonlinear optimization problem in equation (3.17), which is neither convex nor concave. But this proceeding with a so called 'black box' as a solver was not within the scope of this thesis.

Appendix A

Vector and matrix algebra

As several operators and derivatives applied to vectors or matrices are used in the framework of this thesis, a few well known definitions and results will be summarized in the following sections.

A.1 Derivatives

Let A, B, C and D be matrices of appropriate dimension where the entries of $A = (a_{ij})$ may be functions of a real value t where indicated. Defining y and x as vectors of appropriate lengths, then

$$\frac{\partial x'y}{\partial x} = y \qquad\qquad \frac{\partial x'Ax}{\partial x} = (A + A')x$$

The derivatives of a matrix A with respect to t or its entries a_{ij} are

$$\frac{\partial A}{\partial t} = \left(\frac{\partial a_{ij}}{\partial t}\right) \qquad \text{resp.} \qquad \frac{\partial A}{\partial a_{ij}} = e_i e_j',$$

where $e_i = (0, \ldots, 0, \underbrace{1}_{i}, 0, \ldots, 0)'$ resp. $e_j = (0, \ldots, 0, \underbrace{1}_{j}, 0, \ldots, 0)'$ are the i^{th} resp. j^{th} canonical basis vectors.
Similarly,

$$\frac{\partial A'}{\partial a_{ij}} = e_j e_i'$$

and the product rule is given by

$$\frac{\partial AB}{\partial t} = \frac{\partial A}{\partial t}B + A\frac{\partial B}{\partial t}.$$

Let $f(A)$ be a differentiable, real valued function of the entries a_{ij} of A. Then the differentiation of f with respect to the matrix A can be stated as

$$\frac{\partial f}{\partial A} = \left(\frac{\partial f}{\partial a_{ij}}\right).$$

The chain rule for a function $g(U) = g(f(A))$ is of the form

$$\frac{\partial g(U)}{\partial a_{ij}} = \frac{\partial g(f(A))}{\partial a_{ij}} = trace\left[\left(\frac{\partial g(U)}{\partial U}\right)' \frac{\partial f(A)}{\partial a_{ij}}\right].$$

For square matrices A the following equalities hold:

$$\begin{aligned}
\frac{\partial trace(A)}{\partial t} &= trace\left(\frac{\partial A}{\partial t}\right) && \text{for the trace}\\
\frac{\partial A^{-1}}{\partial t} &= -A^{-1}\frac{\partial A}{\partial t}A^{-1} && \text{for the inverse}\\
\frac{\partial log(|A|)}{\partial t} &= trace\left(A^{-1}\frac{\partial A}{\partial t}\right) && \text{for the logarithmic determinant}
\end{aligned}$$

of a matrix.
For first, second and higher order derivatives with respect to a matrix A the following rules are valid:

$$\begin{aligned}
\frac{\partial trace(A)}{\partial A} &= I & \frac{\partial trace(BA)}{\partial A} &= B'\\
\frac{\partial trace(BAC)}{\partial A} &= B'C' & \frac{\partial trace(BA'C)}{\partial A} &= CB\\
\frac{\partial trace(B \otimes A)}{\partial A} &= trace(B)I & \frac{\partial trace(A \otimes A)}{\partial A} &= 2trace(A)I\\
\frac{\partial trace(A'BA)}{\partial A} &= (B + B')A & \frac{\partial trace(ABA)}{\partial A} &= A'B' + B'A'\\
\frac{\partial trace(BACA)}{\partial A} &= B'A'C' + C'A'B' & \frac{\partial trace(C'A'DAC)}{\partial A} &= D'ACC' + DACC'\\
\frac{\partial trace(BACA'D)}{\partial A} &= B'D'AC' + DBAC & \frac{\partial trace(A^k)}{\partial A} &= k(A^{k-1})'
\end{aligned}$$

Other useful matrix derivatives are

$$\begin{aligned}
\frac{\partial |A|}{\partial A} &= |A|\left(A^{-1}\right)' & \frac{\partial |CAD|}{\partial A} &= |CAD|\left(A^{-1}\right)'\\
\frac{\partial trace(BA^{-1}C)}{\partial A} &= -(A^{-1}CBA^{-1})' & \frac{\partial ||A||_F^2}{\partial A} &= 2A
\end{aligned}$$

$$\frac{\partial x'Ay}{\partial A} = xy' \qquad \frac{\partial x'A'y}{\partial A} = yx'$$
$$\frac{\partial x'A'BAy}{\partial A} = B'Axy' + BAyx' \qquad \frac{\partial (Ax+y)'B(Ax+y)}{\partial A} = (B+B')(Ax+y)x'$$

A.2 Kronecker and vec Operator

As already stated on page 47, the symbol $\otimes$ is known as Kronecker product, that concatenates a rectangular matrix $G = \begin{pmatrix} g_{11} & \cdots & g_{1n} \\ \vdots & & \vdots \\ g_{m1} & \cdots & g_{mn} \end{pmatrix}$ of dimension $m \times n$ and a $r \times q$ matrix $H = \begin{pmatrix} h_{11} & \cdots & h_{1q} \\ \vdots & & \vdots \\ g_{r1} & \cdots & g_{rq} \end{pmatrix}$ to a matrix of dimension $mr \times nq$ in the following way:

$$G \otimes H = \begin{pmatrix} g_{11}H & \cdots & g_{1n}H \\ \vdots & & \vdots \\ g_{m1}H & \cdots & g_{mn}H \end{pmatrix}.$$

Let A, B, C and D be matrices of appropriate dimension and α and β are constants. Then the Kronecker product can be characterized by the following properties:

$$A \otimes B \neq B \otimes A \quad \text{in general} \qquad rk(A \otimes B) = rk(A)rk(B)$$
$$A \otimes (B + C) = A \otimes B + A \otimes C \qquad A \otimes (B \otimes C) = (A \otimes B) \otimes C$$
$$\alpha A \otimes \beta B = \alpha\beta(A \otimes B) \qquad (A \otimes B)' = A' \otimes B'$$
$$(A \otimes B)(C \otimes D) = AC \otimes BD \qquad (A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$$
$$(A \otimes B)^{+} = A^{+} \otimes B^{+} \qquad trace(A \otimes B) = trace(A)trace(B)$$

Another operator used frequently in this thesis is the vec operator. Applied to a matrix $A = (a_1, \ldots, a_N)$ it stacks the columns of A into a vector $vec(A) = (a_1', \ldots, a_N')'$. For matrices A, B and C and a constant α the properties of the vec operator include

$$vec(BAC) = (C' \otimes B)vec(A) \qquad vec(A + B) = vec(A) + vec(B)$$
$$trace(A'B) = vec(A)'vec(B) \qquad vec(\alpha A) = \alpha vec(A)$$

Bibliography

[1] S. K. Ahn and G. C. Reinsel. Nested reduced rank autoregressive models for multiple time series. *J. of the American Statistical Association*, 83:849–856, 1988.

[2] M. A. Aizerman, E. A. Braverman, and L. Rozonoer. Theoretical foundations of the potential function method in pattern recognition learning. In *Automation and Remote Control,*, number 25, pages 821–837, 1964.

[3] H. An and L. Gu. On the selection of regression variables. *Acta Mathematicae Applicatae Sinica*, 2(1):27–36, June 1985.

[4] T. W. Anderson. Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Annals of Mathematical Statistics*, 22:327–351, 1951.

[5] T. W. Anderson. Estimating linear statistical relationships. *The Annals of Statistics*, 12(1):1–45, March 1984. The 1982 Wald Memorial Lectures.

[6] T. W. Anderson. Specification and misspecification in reduced rank regression. In *San Antonio Conference: selected articles*, volume 64 of *Series A*, pages 193–205, 2002.

[7] T. W. Anderson and H. Rubin. Statistical inference in factor analysis. In *Third Berkeley Symposium on mathematical statistics and probability*, volume V, pages 111–150. University of California Press, 1956.

[8] J. Bai and S. Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, January 2002.

[9] C. Becker, R. Fried, and U. Gather. Applying sliced inverse regression to dynamical data. November 2000.

[10] D. R. Brillinger. *Time Series: Data Analysis and Theory.* Holden-Day, San Francisco, CA, expanded edition, 1981.

[11] C. Burt. Experimental tests of general intelligence. *British Journal of Psychology*, 3:94–177, 1909.

[12] R. B. Cattell. The scree test for the number of factors. *Multivariate Behavioral Research*, 2(1):245–276, 1966.

[13] G. Chamberlain. Funds, factors and diversification in arbitrage pricing models. *Econometrica*, 51(5):1305–1324, 1983.

[14] G. Chamberlain and M. Rothschild. Arbitrage, factor structure and mean-variance analysis on large asset markets. *Econometrica*, 51(5):1281–1304, 1983.

[15] J. N. Darroch. An optimal property of principal components. *The Annals of Mathematical Statistics*, 36, October 1965.

[16] A. d'Aspremont, F. R. Bach, and L. El Ghaoui. Optimal solutions for sparse principal component analysis. *J. of Machine Learning Research*, 9:1269–1294, 2008.

[17] A. d'Aspremont, L. El Ghaoui, M. Jordan, and G. Lanckriet. A direct formulation for sparse PCA using semidefinite programming. *SIAM Review*, 49(3):434–448, 2007.

[18] P. T. Davies and M. K-S. Tso. Procedures for reduced-rank regression. *Applied Statistics*, 31:244–255, 1982.

[19] M. Deistler and E. Hamann. Identification of factor models for forecasting returns. *Journal of Financial Econometrics*, 3(2):256–281, 2005.

[20] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, September 1936.

[21] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. In *Annals of Statistics (with discussion)*, volume 32, pages 407–499. 2004.

[22] R. F. Engle and M. W. Watson. A one-factor multivariate time series model of metropolitan wage rates. *J. of the American Statistical Association*, 76:774–781, 1981.

[23] J. Fan and R. Li. Variable slection via nonconcave penalized likelihood and its oracle properties. *J. of the American Statistical Association*, 96:1348:1360, 2001.

[24] M. Forni, D. Giannone, M. Lippi, and L. Reichlin. Opening the black box: structural factor models versus structural *VAR*s. April 2004.

[25] M. Forni, M. Hallin, M. Lippi, and L. Reichlin. The generalized dynamic factor model: identification and estimation. *The Review of Economics and Statistics*, 82(4):540–554, November 2000.

[26] M. Forni, M. Hallin, M. Lippi, and L. Reichlin. The generalized dynamic factor model: one-sided estimation and forecasting. February 2003.

[27] M. Forni, M. Hallin, M. Lippi, and L. Reichlin. The generalized dynamic factor model: consistency and rates. *Journal of Ecnometrics*, 119:231–255, 2004.

[28] M. Forni and M. Lippi. The generalized dynamic factor model: representation theory. *Econometric Theory*, 17:113–1141, 2001.

[29] K. Gabriel and S. Zamir. Lower rank approximation of matrices by least squares with any choice of weights. *Technometrics*, 21:489–498, 1979.

[30] U. Gather, R. Fried, V. Lanius, and M. Imhoff. Online monitoring of high dimensional physiological time series - a case-study.

[31] J. Geweke. The dynamic factor analysis of economic time series. In D. Aigner and A. Goldberger, editors, *Latent variables in socio-economic models*, pages 365–383. Amsterdam, North Holland, 1977.

[32] R. Guidorzi, R. Diversi, and U. Soverini. Errors-in-variables filtering in behavioural and state-space contexts.

[33] E. Hannan and M. Deistler. *The Statistical Theory of Linear Systems*. Wiley, 1988.

[34] P. R. Hansen. A reality check for data snooping: A comment on White. Brown University.

[35] P. R. Hansen. Generalized reduced rank regression. Technical report, 2002.

[36] P. R. Hansen. On the estimation of reduced rank regression. March 2002. working paper 2002-08.

[37] Ch. Heaton and V. Solo. Asymptotic principal components estimation of large factor models. Research Papers 0303, Macquarie University, Department of Economics, June 2003.

[38] Ch. Heaton and V. Solo. Estimation of approximate factor models: Is it important to have a large number of variables? Research Papers 0605, Macquarie University, Department of Economics, September 2006.

[39] A. E. Hendrickson and P. O White. Promax: a quick method for rotation to orthogonal oblique structure. *British Journal of Statistical Psychology*, 17:6570, 1964.

[40] J. G. Hirschberg and D. J. Slottje. The reparametrization of linear models subject to exact linear restrictions. Department of Economics - Working Papers Series 702, The University of Melbourne, 1999. Available at http://ideas.repec.org/p/mlb/wpaper/702.html.

[41] P. Horst. *Factor Analysis of Data Matrices*, chapter 10. Holt, Rinehart and Winston, 1965.

[42] H. Hotelling. Analysis of a complex of statistical variables with principal components. *J. of Educational Psychology*, 24:417–441, 1933.

[43] A. J. Izenman. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5:248–264, 1975.

[44] S. Johanson. Likelihood-based interference in cointegrated vector autoregressive models. *Oxford University Press*, 1995.

[45] I. T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer, 2002.

[46] I. T. Jolliffe, N. T. Trendafilov, and M. Uddin. A modified principal component technique based on the LASSO. *Journal of Computational and Graphical Statistics*, 12(3):531–547, 2003.

[47] I.T. Jolliffe and M. Uddin. The simplified component technique - an alternative to rotated principal components. *Journal of Computational and Graphical Statistics*, 9:689–710, 2000.

[48] M. Journée, Y. Nesterov, P. Richtarik, and R. Sepulchre. Generalized power method for sparse principal component analysis. *ArXiv*, 2008. http://arxiv.org/abs/0811.4724.

[49] H. F. Kaiser. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23:187–200, 1958.

[50] D. N. Lawley and A. E. Maxwell. *Factor Analysis as a statistical method*. Butterworths, 1988.

[51] C. Leng and H. Wang. On general adaptive sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 18:201–215, 2009.

[52] H. Lütkepohl. *Introduction to Multiple Time Series Analysis*. Second Edition. Springer Verlag, 1993.

[53] H. M. Markowitz. Portfolio selection. *J. of Finance*, 7(1):77–91, March 1952.

[54] M. Okamoto. Optimality of principal components. *Multivariate Analysis 2 (P. R. Krishnaiah, ed.)*, pages 673–685, 1969.

[55] M. Okamoto and M. Kanazawa. Minimization of eigenvalues of a matrix and optimality of principal components. *The Annals of Mathematical Statistics*, 39(3):859–863, 1968.

[56] K. Pearson. On lines and planes of closest fit to system of points in space. *Philosophical Magazine*, 2:559–572, 1901.

[57] C. R. Rao. The use and interpretation of principal component analysis in applied research. *Sankhya*, 26:329–359, 1964.

[58] C. R. Rao. Separation theorems for singular values of matrices and their applications in multivariate analysis. *Journal of Multivariate Analysis*, 9(3):362–377, 1979.

[59] L. Reichlin. Extracting business cycle indexes from large data sets: aggregation, estimation, identification. November 2000. Paper prepared for the World Congress of the Econometric Society, Seattle, August 2000. Visit www.dynfactors.org.

[60] G. C. Reinsel and R. P. Velu. *Multivariate Reduced Rank Regression, Theory and Aplli-cations*. Lecture Notes in Statistics 136. Springer Verlag New York, Inc., 1998.

[61] P. M. Robinson. Identification, estimation and large-sample theory for regressions containing unobservable variables. *International Economic Review*, 15(3):680–92, October 1974.

[62] T. J. Sargent and C. A. Sims. Business cycle modelling without pretending to have too much a priori economic theory. In C. A. Sims, editor, *New Methods in Business Cycle Research*. Minneapolis, 1977.

[63] W. Scherrer and M. Deistler. A strucure theory for linear dynamic errors-in-variables models. *SIAM J. Control Optim.*, 36(6):2148–2175, November 1998. AMS subject classification: 93B30, 93B15, 62H25. PII: S0363012994262464.

[64] H. Shen and J. Z. Huang. Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, 99(6):1015–1034, 2008.

[65] T. Söderström, U. Soverini, and K. Mahata. Perspectives on errors-in-variables estimation for dynamic systems.

[66] C. Spearman. General intelligence, objectively determined and measured. *American Journal of Psychology*, 15:201–203, 1904.

[67] N. Srebro and T. Jaakkola. Weighted low-rank approximations. In *Proceedings of the 20th International Conference on Machine Learning, ICML-2003*, pages 720–727. AAAI Press, 2003.

[68] J. H. Stock and M. W. Watson. Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economic Statistics*, 20(2):147–162, 2002.

[69] J. H. Stock and M. W. Watson. Forecasting with many predictors. *Handbook of economic forecasting*, 2004.

[70] P. Stoica and M. Viberg. Reduced-rank linear regression. *Signal Processing Workshop on Statistical Signal and Array Processing, IEEE*, page 542, 1996.

[71] Y. Takane and M. A. Hunter. Constrained principal component analysis: A comprehensive theory. *Applicable Algebra in Engineering, Communication and Computing*, 12(5):391–419, 2001.

[72] Y. Takane, H. Kiers, and J. Leeuw. Component analysis with different sets of constraints on different dimensions. *Psychometrika*, 60(2):259–280, June 1995.

[73] Y. Takane, H. Yanai, and S. Mayekawa. Relationships among several methods of linearly constrained correspondence analysis. *Psychometrika*, 56(4):667–684, December 1991.

[74] L. L. Thurstone. The vectors of mind. *Psychological Review*, 41:1–32, 1932.

[75] L. L Thurstone. Multiple-factor analysis. *University of Chicago Press*, 1947.

[76] G. C. Tiao and R. S. Tsay. Model specification in multivariate time series (with discussion). *J. of the Royal Statistical Society*, B 51:157–213, 1989.

[77] R. Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B*, 58(2):267–288, 1996.

[78] K. D. West. Asymptotic inference about predictive ability. *Econometrica*, 64(5):1067–1084, September 1996.

[79] H. White. A reality check for data snooping. *Econometrica*, 68(5):1097–1126, September 2000.

[80] J. Ye. On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, 93(441):120–130, March 1998.

[81] P. A. Zadrozny. Estimating a *VARMA* model with mixed-frequency and temporally-aggregated data: an application to forecasting u.s. gnp at monthly intervals. February 2000.

[82] Ch. Zinner. *Modeling of high dimensional time series by generalized dynamic factor models.* PhD thesis, TU Wien, 2008.

[83] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, December 2006.

[84] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. R. Statistical Society B*, 67:301–320, 2005.

[85] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286, June 2006.

Index

Curriculum Vitae

- **Personal Information**
  - Title: Dipl.-Ing.
  - Date and place of birth: September 16, 1978, Krems/Donau, Austria
  - Citizenship: Austria
  - Marital status: single
  - Children: Nico (2004) and Lena (2007)
  - Languages: German (mother tongue), English (fluent), Spanish (fluent), French (basic)
- **Education**
  - **since March 2002:** Ph.D. studies in Technical Mathematics, with emphasis on Mathematical Economics, at the Faculty of Financial Mathematics, Vienna University of Technology, Austria.
    Ph.D. Thesis: Estimation of Constrained Factor Models with application to Financial Time Series
    Supervisor: O.Univ.Prof. Dipl.-Ing. Dr.techn. Manfred Deistler
  - **1996-2001:** Master of Science (MSc) in Technical Mathematics, with emphasis on Mathematical Economics, Vienna University of Technology
    First and second diploma passed with distinction
    Master Thesis: Modellierung der Phillipskurve mittels Smooth Transition Regression Modellen
    Supervisor: Ao.Univ.Prof. Dr.iur. Bernhard Böhm
  - **1988-1996:** High school *(BRG Krems/Donau)*, Austria, graduation (Matura) passed with distinction
  - **1984-1988:** Primary school, Langenlois, Austria.

- **Professional Career**
  Internships:
  - **07/2001 - 09/2001 :** FSC Financial Soft Computing GmbH in Vienna, Austria
  - **03/2001 - 05/2001:** Research Assistant at the Institute for Mathematical Methods in Economics, Research Unit Econometrics and System Theory, Vienna University of Technology, Austria
    Main Project: EU-Project (IDEE) in cooperation with French and Italian researchers on the development of European economies
  - **03/2000 - 04/2000:** Voluntary job at the commercial section of the Austrian Embassy in Santiago, Chile

  Long Term Positions:
  - **since 11/2009:** Senior Financial Mathematician at **C-QUADRAT Kapitalanlage AG** in Vienna, Austria
    Main Activities: Development of proprietary indices based on quantitative forecasting models in order to create Exchange Traded Funds (ETFs) and other structured financial products.
  - **10/2003 - 01/2009: FSC Financial Soft Computing GmbH** in Vienna, Austria
    Main Activities: Development, estimation and selection of econometric models using the open source statistical program 'R' in order to forecast financial time series; main focus on estimating and developing factor models to forecast financial time series as FX, equities, commodities and assets and on estimation of volatilities with GARCH models
  - **03/2002 - 09/2003:** Research Assistant at the **Institute for Mathematical Methods in Economics**, Research Unit Econometrics and System Theory, Vienna University of Technology, Austria in cooperation with FSC Financial Soft Computing GmbH in Vienna, Austria
    Main Project: Development, estimation and selection of econometric models using the open source statistical program 'R' in order to forecast financial time series; main focus on estimating and developing factor models to forecast financial time series as FX, equities, commodities and assets