Turun yliopisto
University of Turku

ON INDEPENDENT COMPONENT ANALYSIS AND SUPERVISED DIMENSION REDUCTION FOR TIME SERIES

Markus Matilainen

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
Sarja - ser. AI osa - tom. 573 | Astronomica - Chemica - Physica - Mathematica | Turku 2018

Turun yliopisto
University of Turku

ON INDEPENDENT COMPONENT ANALYSIS AND SUPERVISED DIMENSION REDUCTION FOR TIME SERIES

Markus Matilainen

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
Sarja - ser. A I osa - tom. 573 | Astronomica - Chemica - Physica - Mathematica | Turku 2018

University of Turku

Faculty of Science and Engineering
Department of Mathematics and Statistics
Doctoral Programme in Mathematics and Computer Sciences

Supervised by

Professor Hannu Oja
Department of Mathematics and Statistics
University of Turku, Turku, Finland

Assistant Professor Klaus Nordhausen
CSTAT - Computational Statistics
Institute of Statistics & Mathematical Methods in Economics
Vienna University of Technology, Vienna, Austria

Reviewed by

Associate Professor Luke Anthony Prendergast
Department of Mathematics and Statistics
La Trobe University, Melbourne, Australia

Assistant Professor Sarah Gelper
Department of Industrial Engineering & Innovation Sciences
Eindhoven University of Technology, Eindhoven, Netherlands

Opponent

Professor Efstathia Bura
ASTAT - Applied Statistics
Institute of Statistics and Mathematical Methods in Economics
Vienna University of Technology, Vienna, Austria

ISBN 978-951-29-7105-3 (PRINT)
ISBN 978-951-29-7106-0 (PDF)
ISSN 0082-7002 (Print)
ISSN 2343-3175 (Online)
Painosalama Oy - Turku, Finland 2018

Abstract

The main goal of this thesis work has been to develop tools to recover hidden structures, latent variables, or latent subspaces for multivariate and dependent time series data. The secondary goal has been to write computationally efficient algorithms for the methods to an R-package.

In Blind Source Separation (BSS) the goal is to find uncorrelated latent sources by transforming the observed data in an appropriate way. In Independent Component Analysis (ICA) the latent sources are assumed to be independent. The well-known ICA methods FOBI and JADE are generalized to work with multivariate time series, where the latent components exhibit stochastic volatility. In such time series the volatility cannot be regarded as a constant in time, as often there are periods of high and periods of low volatility. The new methods are called gFOBI and gJADE. Also SOBI, a classic method which works well once the volatility is assumed to be constant, is given a variant called vSOBI, that also works with time series with stochastic volatility.

In dimension reduction the idea is to transform the data into a new coordinate system, where the components are uncorrelated or even independent, and then keep only some of the transformed variables in such way that we do not lose too much of the important information of the data. The aforementioned BSS methods can be used in unsupervised dimension reduction; all the variables or time series have the same role.

In supervised dimension reduction the relationship between a response and predictor variables needs to be considered as well. Well-known supervised dimension reduction methods for independent and identically distributed data, SIR and SAVE, are generalized to work for time series data. The methods TSIR and TSAVE are introduced and shown to work well for time series, as they also use the information on the past values of the predictor time series. Also TSSH, a hybrid version of TSIR and TSAVE, is introduced.

All the methods that have been developed in this thesis have also been implemented in R package tsBSS.

Tiivistelmä

Tämän väitöskirjatyön tavoitteena on ollut kehittää teoreettisia työkaluja moniulotteisten aikasarjojen piilevien rakenteiden ja kiinnostavien latenttien aikasarjojen etsimiseen ja niiden määrän supistamiseen. Tavoite on myös ollut kehittää tätä tarkoitusta varten tehokkaita laskenta-algoritmeja ja koota niiden pohjalta käytännön tarpeita varten R-ohjelmistopaketti.

Sokeassa signaalien erottelussa riippumattomien moniulotteisten havaintojen tapauksessa tavoitteena on löytää korreloimattomia latentteja muuttujia siirtämällä alkuperäiset havainnot uuteen koordinaatistoon. Riippumattomien komponenttien analyysissä oletetaan, että näin saatavat uudet muuttujat ovat riippumattomia. Tässä työssä perinteiset riippumattomien komponenttien analyysin menetelmät FOBI ja JADE yleistetään toimimaan myös moniulotteisten aikasarjojen, s.o. riippuvien moniulotteisten havaintojen, tapauksessa. Nämä yleistykset gFOBI ja gJADE toimivat myös silloin, kun aikasarjoilla on stokastista volatiliteettia. Stokastisen volatiliteetin tapauksessa esiintyy satunnaisia ajanjaksoja, jolloin havaintojen vaihtelu on pientä, ja ajanjaksoja, jolloin vaihtelu on suurta. Myöskään klassinen menetelmä SOBI ei välttämättä löydä komponenttisarjoja stokastisen volatiliteetin tapauksessa ja sille kehitetään tällaisessakin tapauksessa toimiva vaihtoehtoinen vSOBI.

Havaintoaineiston dimension supistamisessa pyritään korvaamaan alkuperäiset muuttujat huomattavasti pienemmällä määrällä muuttujia, jotka pitävät sisällään kaiken tai lähes kaiken aineiston sisältämän informaation. Tämän voi toteuttaa esimerkiksi sokean signaalien erottelun avulla hyväksymällä vain informatiivisimman muuttujajoukon uudessa koordinaatistossa sopivan kriteerin mielessä. Niin sanotussa ohjatussa dimension supistamisessa käytetään ulkopuolista muuttujaa, vastetta, ja valikoidun muuttujajoukon ajatellaan olevan riittävä selittämään vasteen ja alkuperäisen muuttujajoukon välisen riippuvuuden mahdollisimman tyhjentävästi. Tässä työssä perinteiset ohjatut menetelmät riippumattomille moniulotteisille havainnoille, SIR ja SAVE, yleistetään aikasarjoille. Toisin kuin SIR ja SAVE, nämä uudet aikasarjamenetelmät, TSIR ja TSAVE ja niiden hybridimenetelmä TSSH, hyödyntävät myös havaintojen aikariippuvuuden luonnollisella tavalla.

R-ohjelmistopaketti tsBSS sisältää algoritmit kaikille tässä väitöskirjatyössä kehitetyille menetelmille.

Acknowledgements

The whole journey of the PhD research has been amazing. When I was starting, I had no idea what kind of people I meet and collaborate with or things I experience. My life has changed a lot in the past few years. Of course this is not all due to the PhD research, but it has had a significant positive effect.

Firstly I wish to thank my amazing supervisors, Professor Hannu Oja and Assistant Professor Klaus Nordhausen. Without them none of this would have been possible. I am glad I managed to find my way into their research group. I cannot even list all the things I have learned from them, as there were so many. Hannu's infinite knowledge on everything kept me amazed. Klaus always had an answer when I had troubles and he also made me a better programmer. They are awesome both as supervisors and as persons. I feel very lucky to have had such guidance.

I wish to thank Professor Christophe Croux for the cooperation with my research and for the hospitality during my research visit to KU Leuven. I am very glad that I had an opportunity to work with him. I also want to thank PhD Sara Taskinen and PhD Jari Miettinen for the cooperation in research and for the hospitality during my research visits to University of Jyväskylä.

I would also like to express my gratitude to the pre-examiners of this thesis, Associate Professor Luke Anthony Prendergast and Assistant Professor Sarah Gelper, for their careful reading of my thesis and their comments that made the thesis still better.

Also my special thanks to Assistant Professor Pauliina Ilmonen and Professor Jaakko Nevalainen, who were the ones that encouraged me to do a PhD.

This research was carried out in the Department of Mathematics and Statistics in University of Turku. The research was funded by the Academy of Finland, Jenny and Antti Wihuri Foundation and the Doctoral Programme in Mathematics and Computer Sciences (MATTI). I also had financial support through Oskar Öflunds Stiftelse and Turun yliopistosäätiö foundations.

Big thanks to all the statistics staff members in our department. Research would be nothing without awesome co-workers. I thank them all for making the work environment great and lunch breaks so interesting. And all those board game evenings! Especially I wish to thank MSc (soon to be PhD) Joni Virta for all the discussions related to research as well as other things. I can never forget all the

Wikipedia searches, Finnish Population Register Centre statistics or the list of literal translations of Finnish idioms to English that Joni helped me to expand. I also thank Ilmari and Henri for the numerous sports related discussions. Also the supporting staff members in our department and faculty have done a great job – thank you for making my life easier!

A big thank you also for all the researchers in the conferences and workshops I have attended to. I have managed to get lots of new ideas and connections.

I wish to acknowledge some friends I have had during this period. All the people related in any way to Statistika, ISTU, Tenants Committee of Student Village East as well as Haliskylä, Finnish language circle, Ingress and FinBuffy – you are all amazing. I also had a nice time with many different groups of board game and sauna lovers. I also wish to thank my friends in Agora for all the late afternoon tea breaks. If you are my friend and not affiliated to aforementioned people or groups of people in any way, you are also awesome. I especially wish to mention Ifti, who gave support just when I needed, and Roni for all the countless conversations about life.

Finally, I especially wish to thank my parents for always supporting me and believing in me. I wouldn't be the person I am without them! Thanks also to all my siblings for being there for me!

December 21, 2017

Markus Matilainen

Contents

Abbreviations and notation

a.s.	almost surely
iid	independent and identically distributed
wlog	without loss of generality
$f_x(x)$	density function of a random variable $\mathbf{x}$
$\mathbf{A}$	real $p \times q$ matrix with a_{ij} as element (i,j).
$\mathbf{A}'$	transpose of a matrix $\mathbf{A}$
$\mathbf{A}^{-1}$	inverse of a $p \times p$ matrix $\mathbf{A}$
$\mathbf{A}^{-1/2}$	symmetric inverse square root of a $p \times p$ matrix $\mathbf{A}$
$\mathrm{diag}(\mathbf{A})$	diagonal matrix with same diagonal elements as $\mathbf{A}$
$\mathrm{off}(\mathbf{A})$	$\mathbf{A} - \mathrm{diag}(\mathbf{A})$
$\det(\mathbf{A})$	determinant of a $p \times p$ matrix $\mathbf{A}$
$\mathrm{trace}(\mathbf{A})$	trace of $p \times p$ matrix $\mathbf{A}$: $\sum_{i=1}^{p} a_{ii}$
$\mathrm{vec}(\mathbf{A})$	operator that stacks the columns of a matrix $\mathbf{A}$ into a single column
$\mathrm{vech}(\mathbf{A})$	operator that stacks column-wise the on and below diagonal elements of a $p \times p$ matrix $\mathbf{A}$ into a single column
$\lVert \cdot \rVert$	Frobenius (matrix) norm: $\lVert \mathbf{A} \rVert = \left(\sum_{i=1}^{p} \sum_{j=1}^{q} a_{ij}^2 \right)^{1/2}$
$\mathbf{E}^{jk}$	$p \times p$ matrix where the element (j,k) equals to 1 and other elements equal to 0
$\mathbf{I}_p$	$p \times p$ identity matrix
$\mathbf{J}$	sign-change matrix: a diagonal matrix with diagonal values ± 1
$\mathbf{K}$	permutation matrix: a matrix where the rows and/or columns of $\mathbf{I}_p$ are permuted
$\mathbf{P}$	projection matrix, for which $\mathbf{P}^2 = \mathbf{P}$ and $\mathbf{P}' = \mathbf{P}$
$\mathbf{U}$	orthogonal matrix: invertible matrix such that $\mathbf{U}' = \mathbf{U}^{-1}$
$\mathbf{\Omega}$	$p \times p$ mixing matrix
$\mathbf{\Gamma}$	unmixing (or a signal separation) matrix
$\mathbf{x}$	random variable or a stochastic process
$\boldsymbol{\mu}$	expected value $E(\mathbf{x})$
$\mathbf{\Sigma}$	covariance $\mathrm{Cov}(\mathbf{x})$
$\mathbf{x}^{(s)}$	standardized $\mathbf{x}$: $\mathbf{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$
$\mathcal{O}^{k \times p}$	set of $k \times p$ matrices with orthonormal rows
$\mathbb{Z}_+$	set of positive integers (excluding zero)
$\mathbb{R}^p$	set of real-valued vectors
$\mathbb{R}^{p \times q}$	set of real-valued matrices
τ	time series lag in $\mathbb{Z}_+$
T	set of lags τ (unless specified differently)
$y \perp\!\!\!\perp x$	y is independent of x

List of publications

The thesis consists of the introduction and the following original publications:

I Matilainen M, Nordhausen K & Oja H (2015), *New Independent Component Analysis Tools for Time Series*, Statistics & Probability Letters, 105, 80–87.

II Matilainen M, Miettinen J, Nordhausen K, Oja H & Taskinen S (2017), *On Independent Component Analysis with Stochastic Volatility Models*, Austrian Journal of Statistics, 46(3–4), 57–66.

III Matilainen M, Croux C, Nordhausen K & Oja H (2017): *Supervised Dimension Reduction for Multivariate Time Series*, Econometrics and Statistics, 4, 57–69.

IV Matilainen M, Croux C, Nordhausen K & Oja H, *Sliced Average Variance Estimation for Multivariate Time Series*, submitted.

Part I

Introduction

1 Background

Multivariate statistics deals with datasets with large number of variables that depend on each other. There are many different kind of techniques to extract meaningful information from such large datasets, such as clustering or classification of observations (see for example Rencher and Christensen, 2012). A popular way to handle such data is to create new uncorrelated or even independent variables based on the original ones and then possibly reduce the amount of them, without losing too much of the important information.

Principal Component Analysis (PCA) uses some orthogonal transformation to create a set of uncorrelated variables from the original correlated variables; the new variables are ordered according to their variances and then usually only a few first ones would be used in further analysis. Pearson (1901) has given early ideas for PCA and then Hotelling (1933) has developed it further independently.

In Blind Source Separation (BSS) we assume that the observed variables are actually a linear combination (linear mixture) of some latent unknown sources. These sources then need to be uncovered. There are different types of BSS methods, such as Independent Component Analysis. For the details of the early development of BSS, see Jutten and Taleb (2000).

In Independent Component Analysis (ICA) we transform variables into a new coordinate system by forming a set of independent variables. The idea is to maximize some measure of non-Gaussianity of the independent variables (sources). Such measures include skewness (a measure of asymmetry) and high or low kurtosis (heavy and light tailedness) of the density distributions of the sources. According to the central limit theorem, these independent variables are more non-Gaussian than their linear combinations that we observe.

For the early development of the ICA concept, see Comon (1994). For an early overview of ICA, see for example Hyvärinen (2001) and references therein.

Sometimes when reducing the number of variables, there might be one or more response variable(s), i.e. one or more variable(s) that depend on another set of variables. In this supervised dimension reduction the relationship between the response and the explaining variable needs to be considered as well. For this there are methods such as Sliced Inverse Regression (SIR) (Li, 1991) and Sliced Average Variance Estimation (SAVE) (Cook, 2000). Also well-known Canonical Correlation Analysis (CCA) (Hotelling, 1936) can be thought as a supervised dimension reduction method. CCA first creates linear combinations of two separate sets of variables such that the linear combinations of the sets have a maximum correlation. Then the rest of the linear combinations are created in the same way such that they are uncorrelated to all the previous linear combinations. In the end the number of the most correlated linear combinations are chosen in an appropriate way.

Analysing time series data is more complex compared to independent and identically distributed data, as also temporal dependence needs to be acknowledged as well. For BSS there exists Second Order Source Separation methods such as Second Order Blind Identification (SOBI), which aims to transform the observations to another coordinate system in order to find latent uncorrelated stationary source time series (Belouchrani et al., 1997). The method jointly diagonalizes a set of lagged covariance matrices to extract the components.

ICA methods not designed for time series have also been used in time series context for example in financial applications, see for example Broda and Paolella (2009) and García-Ferrer et al. (2012). However, the methods used in these papers do not take the temporal dependence into account, so the methods do not utilize all the information available in data.

The outline of the thesis is as follows. In Chapter 2 known methods for independent and identically distributed (iid) methods are reviewed. Chapter 3 first discusses existing BSS methods and then extensions of ICA methods for time series are considered. In the last part of the chapter popular supervised dimension reduction methods are generalized to work with time series. In Chapter 4 the R package tsBSS is discussed with examples and finally Chapter 5 summarizes what has been done and what still needs to be investigated.

2 Independent and identically distributed observations

When dealing with data with a large number of variables, the number of parameters in the models may be very large, which may cause computational issues in their estimation. For more on this, known as the *curse of dimensionality*, see for example Scott (2015).

One way to deal with this issue is to transform the data into a new coordinate system, where the components are uncorrelated or even independent. Then we can deal separately with each transformed variable. In dimension reduction the key is to use only some of these transformed variables in such way that we do not lose too much of the important information of the data.

Sometimes we have datasets where one of the variables is treated as a response, i.e. one of the variables is to be explained by other variables. Naturally there can also be more than one response. Then the relationship between the response(s) and the other variables needs to be considered when reducing the dimension. Such a scenario is called *supervised* dimension reduction. If no such variable exists, i.e. all the variables are treated equally, then dimension reduction is called *unsupervised*.

In Sections 2.1 we first introduce notation and give some preliminary results. In Section 2.2, moments and cumulants are reviewed and in Section 2.3 we give a general multivariate model for iid data. Then we review two widely used procedures, Principal Component Analysis (PCA) in Section 2.4 and Independent Component Analysis (ICA) in Section 2.5. In the last part of the chapter, in Section 2.6, we discuss some supervised dimension reduction methods for iid observations.

2.1 Notation and preliminary results

Throughout this thesis the following notation is used. Let $\mathbf{A}$ be a $p \times q$ matrix with elements $a_{ij}, i = 1, \dots, p$ and $j = 1, \dots, q$, and $\mathbf{A}'$ its transpose. In this thesis a p-vector $\mathbf{x}$ is denoted as $\mathbf{x} \in \mathbb{R}^p$ and $p \times q$ matrix $\mathbf{A}$ is denoted as $\mathbf{A} \in \mathbb{R}^{p \times q}$.

A square matrix $\mathbf{A} \in \mathbb{R}^{p \times p}$ is symmetric if $a_{ij} = a_{ji}$ for all $i, j = 1, \dots, p$. The determinant of a square matrix $\mathbf{A}$ is $\det(\mathbf{A})$, and the trace of a square matrix $\mathbf{A}$ is $\mathrm{trace}(\mathbf{A}) = \sum_{i=1}^{p} a_{ii}$. In addition, $\mathrm{diag}(\mathbf{A})$ is a diagonal matrix with the same diagonal elements than a square matrix $\mathbf{A}$ and $\mathrm{off}(\mathbf{A}) = \mathbf{A} - \mathrm{diag}(\mathbf{A})$. A matrix norm of $\mathbf{A} \in \mathbb{R}^{p \times q}$ is $\|\mathbf{A}\| = \left(\sum_{i=1}^{p} \sum_{j=1}^{q} a_{ij}^2 \right)^{1/2}$.

For a matrix $\mathbf{A} = \left(\mathbf{a}_1, \dots, \mathbf{a}_q \right) \in \mathbb{R}^{p \times q}$, a vectorization operator $\mathrm{vec}(\cdot)$ from $\mathbb{R}^{p \times q}$ to $\mathbb{R}^{pq}$ is defined as

$$\mathrm{vec}(\mathbf{A}) = \begin{pmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_q \end{pmatrix}.$$

While the vec operator stacks all the vectors to a single column, a vech operator only stacks the elements on and below the diagonal of a square matrix to a vector. The $\mathrm{vech}(\cdot)$ operator from a matrix $\mathbf{A} \in \mathbb{R}^{p \times p}$ to a vector is $\mathrm{vech}(\mathbf{A}) \in \mathbb{R}^{p(p+1)/2}$.

Some special matrices are also used in this thesis. $\mathbf{I}_p \in \mathbb{R}^{p \times p}$ is an identity matrix, $\mathbf{J}$ is a sign-change matrix, a diagonal matrix with diagonal values ±1, and $\mathbf{K}$ is a permutation matrix, a matrix where the rows and/or columns of $\mathbf{I}_p$ are permuted.

In addition, $\mathbf{U} \in \mathbb{R}^{p \times p}$ is an orthogonal matrix if it is invertible and $\mathbf{U}' = \mathbf{U}^{-1}$. A matrix has orthonormal rows if it is the first $k \leq p$ rows of an orthogonal matrix. Denote $\mathcal{O}^{k \times p}$ for a set of $k \times p$ matrices with orthonormal rows. $\mathbf{P}$ is a projection matrix if $\mathbf{P}^2 = \mathbf{P}$ and $\mathbf{P}' = \mathbf{P}$. In a matrix $\mathbf{E}^{jk}$, $j, k = 1, \dots, p$, the element (j, k) equals to 1 and other elements equal to zero.

$\mathbf{A}^{-1}$ is the inverse of a matrix $\mathbf{A} \in \mathbb{R}^{p \times p}$; $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_p$. A square matrix is called singular, if it does not have an inverse matrix, i.e. if $\det(\mathbf{A}) = 0$. A symmetric matrix $\mathbf{A}$ is positive definite if $\mathbf{b}'\mathbf{A}\mathbf{b} > 0$ for any non-zero vector $\mathbf{b}$. $\mathbf{B} = \mathbf{A}^{1/2}$ is the symmetric square root of a $p \times p$ matrix $\mathbf{A}$, if $\mathbf{B}$ is a symmetric matrix such that $\mathbf{B}^2 = \mathbf{A}$. $\mathbf{A}^{-1/2}$ is then the symmetric inverse square root of a $p \times p$ matrix $\mathbf{A}$.

For more details on basic matrix notation and theory related to multivariate statistics, see for example Chapter 1 in Kollo and von Rosen (2005).

A singular value decomposition of a positive-definite matrix $\mathbf{A} \in \mathbb{R}^{p \times p}$ is

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}',$$

where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices and $\mathbf{D}$ a diagonal matrix such that the diagonal elements are all positive. The eigenvalue-eigendecomposition of a positive-definite matrix $\mathbf{A} \in \mathbb{R}^{p \times p}$ is

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}',$$

where $\mathbf{U} \in \mathbb{R}^{p \times p}$ is an orthogonal matrix and $\mathbf{D} \in \mathbb{R}^{p \times p}$ a diagonal matrix such that the eigenvalues as the diagonal elements are all positive.

For a symmetric matrix $\mathbf{A} \in \mathbb{R}^{p \times p}$ maximizing $\left\| \mathrm{diag}\left(\mathbf{U}\mathbf{A}\mathbf{U}'\right) \right\|^2$ produces an orthogonal matrix $\mathbf{U} \in \mathbb{R}^{p \times p}$, where the rows are the eigenvectors of $\mathbf{A}$. Also

$$\left\| \mathrm{diag}\left(\mathbf{U}\mathbf{A}\mathbf{U}'\right) \right\|^2 + \left\| \mathrm{off}\left(\mathbf{U}\mathbf{A}\mathbf{U}'\right) \right\|^2 = \|\mathbf{A}\|^2. \tag{2.1}$$

Thus a maximization $\left\| \mathrm{diag}\left(\mathbf{U}\mathbf{A}\mathbf{U}'\right) \right\|^2$ for an orthogonal matrix $\mathbf{U}$ is equivalent to minimizing $\left\| \mathrm{off}\left(\mathbf{U}\mathbf{A}\mathbf{U}'\right) \right\|^2$.

2.2 Moments and cumulants

Moments and cumulants are quantitative measures that have been long used to describe characteristics of probability distributions, such as its location, scale, skewness and kurtosis. For an early history of moments and cumulants, see for example Hald (2000). In this section we define moments and cumulants of a random variable and random vector. We focus here only on continuous distributions.

Consider a continuous random variable x with a density function $f_x(x)$. Then the j:th moment of a random variable x is defined as

$$\mu_j = \mathrm{E}\left(x^j\right) = \int_{-\infty}^{\infty} t^j f_t(t) dt,$$

for $j = 1, 2, \ldots$. The first moment $\mu := \mathrm{E}(x)$ is called the expected value of a random variable. The j:th central moment is

$$\mu_j^{(c)} = \mathrm{E}\left(\left(x - \mu\right)^j\right) = \int_{-\infty}^{\infty} \left(t - \mu\right)^j f_t(t) dt.$$

The second central moment $\sigma^2 := \mathrm{Var}(x) = \mathrm{E}\left(\left(x-\mu\right)^2\right)$ is called the variance of a random variable.

The moment generating function of a random variable is

$$m_x(t) := \mathrm{E}\left(\exp(tx)\right) = \sum_{j=0}^{\infty} \frac{t^j \mathrm{E}\left(x^j\right)}{j!} = \sum_{j=0}^{\infty} \frac{t^j \mu_j}{j!},$$

for $t \in \mathbb{R}$ and with $\mu_0 = 1$. Then the j:th moment is

$$\mu_j = m_x^{(j)}(t)\,|_{t=0},$$

where $m_x^{(j)}(t)$ is the j:th derivative of $m_x(t)$ with respect to t.

The cumulant generating function is

$$c_x(t) := \log\left(m_x(t)\right) = \sum_{j=0}^{\infty} \frac{t^j \kappa_j}{j!}. \tag{2.2}$$

Then the j:th cumulant

$$\kappa_j = c_x^{(j)}(t)\,|_{t=0},$$

where $c_x^{(j)}(t)$ is the j:th derivative of $c_x(t)$ with respect to t. Based on (2.2) we can write the cumulants in terms of the moments (and vice versa). The first few are:

$$\kappa_1 = \mu_1, \quad \kappa_2 = \mu_2 - \mu_1^2, \quad \kappa_3 = \mu_3 - 3\mu_1\mu_2 + 2\mu_1^3 \text{ and}$$
$$\kappa_4 = \mu_4 - 4\mu_3\mu_1 - 3\mu_2^2 + 12\mu_2\mu_1^2 - 6\mu_1^4.$$

For a standardized variable $x^{(s)} = \frac{x-\mu}{\sigma}$ then

$$\kappa_1 = 0, \quad \kappa_2 = 1, \quad \kappa_3 = \mu_3 \text{ and } \kappa_4 = \mu_4 - 3.$$

Skewness measures how much a distribution deviates from a symmetric distribution and it can be defined as the third standardized moment of a random variable x, i.e.

$$\gamma_1 := \frac{\mu_3^{(c)}}{\left(\sigma^2\right)^{3/2}} = \frac{\kappa_3}{\kappa_2^{3/2}}.$$

The kurtosis of the random variable x can be defined as the fourth standardized moment

$$\mu_4^{(c)*} := \frac{\mu_4^{(c)}}{\left(\sigma^2\right)^2}.$$

Excess kurtosis measures how heavy the tails of the distribution of a random variable x are compared to a normal distribution, and it can be defined as

$$\gamma_2 = \frac{\kappa_4}{\kappa_2^2} = \mu_4^{(c)*} - 3.$$

For $\mathrm{x}^{(s)}$ skewness is just $\mathrm{E}\left(\left(\mathrm{x}^{(s)}\right)^3\right)$, kurtosis $\mathrm{E}\left(\left(\mathrm{x}^{(s)}\right)^4\right)$ and the excess kurtosis $\mathrm{E}\left(\left(\mathrm{x}^{(s)}\right)^4\right) - 3$. For more on moments and cumulants, see for example Shynk (2012).

Moments and cumulants can be used to characterize a normal distribution as follows.

Definition 2.1 *A random variable* x *has a normal (Gaussian) distribution, i.e.* $\mathrm{x} \sim \mathrm{N}\left(\mu, \sigma^2\right)$, *if*

$$f_x(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(x-\mu\right)^2}{2\sigma^2}\right).$$

The normal distribution is fully specified by its mean and variance, and for the cumulants it is true that $\kappa_j = 0$, when $j \geq 3$. Thus it can be seen that the skewness γ_1 for a normal distribution is zero, since the normal distribution is symmetric around its mean. Also the kurtosis of a normal distribution equals to 3 and excess kurtosis γ_2 naturally 0. Any distribution is said to be *heavy-tailed* if its kurtosis is above 3 and *light-tailed* if its kurtosis is below 3.

Now similarly, for p-variate random vector $\mathbf{x} = \left(x_1, \ldots, x_p\right)'$, the joint moment generating function is

$$m_{\mathbf{x}}(\mathbf{t}) := \mathrm{E}\left(\exp\left(\mathbf{t}'\mathbf{x}\right)\right),$$

where $\mathbf{t} = \left(t_1, \ldots, t_p\right)'$, and the joint cumulant generating function is

$$c_{\mathbf{x}}(\mathbf{t}) := \log\left(m_{\mathbf{x}}(\mathbf{t})\right).$$

For more on joint moments and cumulants, see for example Mittelhammer (1996) and the references therein.

For simplicity, let us assume that $\mathrm{E}(\mathbf{x}) = \mathbf{0}$. The second-order cumulants of $\mathbf{x}$ are then

$$\kappa\left(x_i, x_j\right) := \mathrm{E}\left(x_i x_j\right),$$

which is just the covariance $\mathsf{Cov}\left(x_i, x_j\right)$, $i, j = 1, \ldots, p$. Similarly, the third-order cumulants are

$$\kappa\left(x_i, x_j, x_k\right) := \mathrm{E}\left(x_i x_j x_k\right),$$

and fourth-order cumulants are

$$\begin{aligned}&\kappa\left(x_i, x_j, x_k, x_l\right) := \mathrm{E}\left(x_i x_j x_k x_l\right)\\ &-\mathrm{E}\left(x_i x_j\right)\mathrm{E}\left(x_k x_l\right) - \mathrm{E}\left(x_i x_k\right)\mathrm{E}\left(x_j x_l\right) - \mathrm{E}\left(x_i x_l\right)\mathrm{E}\left(x_j x_k\right),\end{aligned} \tag{2.3}$$

$i, j, k, l = 1, \ldots, p$. For more details, see for example Mittelhammer (1996); Hyvärinen (2001).

The classic measures of multivariate skewness and kurtosis by Mardia (1970) for a standardized random vector $\mathbf{x}^{(s)}$ are

$$\mathrm{E}\left(\left(\mathbf{x}^{(s)\prime}\mathbf{x}^{*(s)}\right)^3\right) \text{ and } \mathrm{E}\left(\left(\mathbf{x}^{(s)\prime}\mathbf{x}^{(s)}\right)^4\right),$$

respectively, where $\mathbf{x}^{(s)}$ and $\mathbf{x}^{*(s)}$ are independent copies of $\mathbf{x}^{(s)}$.

2.3 Models for iid data

Let $\mathbf{x} = \left(x_1, \ldots, x_p\right)' \in \mathbb{R}^p$ be a random vector. Consider a general multivariate location-scatter model

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Omega}\mathbf{z}, \tag{2.4}$$

where $\boldsymbol{\mu} \in \mathbb{R}^p$ is a location vector, $\boldsymbol{\Omega} \in \mathbb{R}^{p \times p}$ a full-rank transformation matrix and $\mathbf{z} = \left(z_1, \ldots, z_p\right)' \in \mathbb{R}^p$ an unknown random vector. In this section we discuss about the assumptions for $\mathbf{z}$ in different models.

For a $\mathbf{x} \in \mathbb{R}^p$ a mean vector $\boldsymbol{\mu} := \mathrm{E}(\mathbf{x}) = \left(\mathrm{E}\left(x_1\right), \ldots, \mathrm{E}\left(x_p\right)\right)' \in \mathbb{R}^p$ and a covariance matrix $\boldsymbol{\Sigma} = \mathsf{Cov}(\mathbf{x}) = \mathsf{Cov}\left(x_i, x_j\right)_{i,j=1,\ldots,p}$. The correlation matrix is then $\boldsymbol{\rho} = \mathsf{diag}(\boldsymbol{\Sigma})^{-1/2}\boldsymbol{\Sigma}\ \mathsf{diag}(\boldsymbol{\Sigma})^{-1/2}$.

Definition 2.2 *A random vector* $\mathbf{x} \in \mathbb{R}^p$ *has a multivariate normal distribution, i.e.* $\mathbf{x} \sim \mathrm{N}_p\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$, *if*

$$f_{\mathbf{x}}\left(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = \frac{1}{(2\pi)^{p/2}\det\left(\boldsymbol{\Sigma}\right)}\exp\left(-\frac{1}{2}\left(\mathbf{x} - \boldsymbol{\mu}\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{x} - \boldsymbol{\mu}\right)\right).$$

In the classic multivariate normal model $\mathbf{z} \sim \mathrm{N}_p\left(\mathbf{0}, \mathbf{I}_p\right)$ *and the covariance matrix* $\boldsymbol{\Sigma} = \boldsymbol{\Omega}\boldsymbol{\Omega}'$.

The assumptions for $\mathbf{z}$ can also be relaxed in many ways.

Definition 2.3 *A random vector $\mathbf{x} \in \mathbb{R}^p$ has an elliptical distribution if*

$$f_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\det(\boldsymbol{\Sigma})} \exp\left(-g\left((\mathbf{x}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})\right)\right),$$

where $g(\cdot)$ is a function that does not depend on μ and $\boldsymbol{\Sigma}$. It is assumed that $\mathbf{U}\mathbf{z} \sim \mathbf{z}$ for an orthogonal matrix $\mathbf{U}$, i.e. that $\mathbf{z}$ has a spherical distribution around the origin.

With $g(r) = \frac{r}{2} + \frac{p}{2}\log(2\pi)$ this reduces to the multivariate normal distribution. Also the multivariate t distribution (see e.g. Kotz and Nadarajah, 2004) and power-exponential distribution (Gómez et al., 1998) are among the elliptical distributions.

Elliptical distribution can be seen as an extension of the multivariate normal distribution. The first two moments are

$$E(\mathbf{x}) = \boldsymbol{\mu} \text{ and } \mathrm{Cov}(\mathbf{x}) = c\boldsymbol{\Sigma},$$

if they exist. The constant c depends on the function g. For the multivariate normal distribution $c = 1$. For more on the elliptical distributions, see for example Fang and Zhang (1990); Kollo and von Rosen (2005).

Definition 2.4 *In Independent Component (IC) model we assume that the components of a random vector $\mathbf{z} \in \mathbb{R}^p$ are independent and $\mathbf{z}$ is standardized such that*

$$E(\mathbf{z}) = \mathbf{0} \text{ and}$$
$$\mathrm{Cov}(\mathbf{z}) = \mathbf{I}_p.$$

In the IC model for the random variable $\mathbf{x}$ then $E(\mathbf{x}) = \boldsymbol{\mu}$ and $\mathrm{Cov}(\mathbf{x}) = \boldsymbol{\Sigma} = \boldsymbol{\Omega}\boldsymbol{\Omega}'$. The IC model is another extension of the multivariate normal model. In Section 2.5 we discuss about Independent Component Analysis (ICA) that is based on the IC model.

For a more detailed overview of the location-scatter models, see for example Chapter 2 in Oja (2010).

Sometimes some of the latent variables are not of interest and they can be regarded as noise. Assume that the first k variables are important and the last $p-k$ are not. Then we can divide $\mathbf{z}$ into two subvectors, to $\mathbf{z}_{(1)} \in \mathbb{R}^k$ and to $\mathbf{z}_{(2)} \in \mathbb{R}^{p-k}$. Now $\mathbf{z}_{(1)}$ is the meaningful part and $\mathbf{z}_{(2)}$ is just noise. Write $\mathbf{\Omega} = \left(\mathbf{\Omega}_1, \mathbf{\Omega}_2\right)$. Now the model (2.4) can be written as

$$\mathbf{x} = \boldsymbol{\mu} + \mathbf{\Omega}_1 \mathbf{z}_{(1)} + \mathbf{\Omega}_2 \mathbf{z}_{(2)}. \tag{2.5}$$

Often we are also interested in a variable or a vector that depends on a p-variate $\mathbf{x}$. Such dependent variable is called a response and we denote it here by y, if it is a scalar, and $\mathbf{y}$, if it is a vector. In such case we can assume that there is an unknown relationship between the response y and the explaining variables $\mathbf{x}$, i.e.

$$y = f(\mathbf{x}, \epsilon), \tag{2.6}$$

where $f(\cdot)$ is an unspecified function and ϵ is an unobserved random variable or vector independent of $\mathbf{x}$.

2.4 Principal Component Analysis

The main idea in Principal Component Analysis (PCA) is to create uncorrelated linear combinations of the original variables and then usually keeping only some of the uncorrelated variables for further analysis. This is done in a way that as much variability in the data is retained as possible, while still keeping the number of chosen linear combinations low enough. The linear combinations are called principal components and they are required to be uncorrelated and ordered in a descending order of their variances (Jolliffe, 2002).

Assume a model of the form (2.4). Consider the eigenvalue-eigendecomposition of $\mathbf{\Sigma}$, that is,

$$\mathbf{\Sigma} = \mathbf{U}\mathbf{D}\mathbf{U}',$$

where the diagonal elements of the diagonal matrix $\mathbf{D}$ are in descending order. Generally PCA does not have any specific distributional assumptions, other than the existence of the second moments.

The columns of $\mathbf{U}$ contain the eigenvectors and the diagonal elements of $\mathbf{D}$ are the eigenvalues. The principal components are then

$$\mathbf{z} = \mathbf{U}'\left(\mathbf{x} - \boldsymbol{\mu}\right)$$

with $\mathrm{Cov}(\mathbf{z}) = \mathbf{D}$. The principal components $\mathbf{z}$ are uncorrelated, but not necessarily independent, unless we assume multivariate normality.

An issue with PCA is that, if the measurement unit of a component of $\mathbf{x}$ is changed (e.g. from cm to mm), the resulting principal components will be different. Also if a variable has a larger or smaller scale than others, it might just form a principal component only by itself. This does not really serve the purpose of dimension reduction. In order to avoid such situations, the variables $\mathbf{x}$ are often transformed first. Let $\mathrm{diag}(\boldsymbol{\Sigma})$ be a diagonal matrix, where the diagonal elements are the variances of the components of $\mathbf{x}$. Then we can use the scaled variables $\mathbf{x}^{(sc)} = \mathrm{diag}(\boldsymbol{\Sigma})^{-1/2}(\mathbf{x} - \mathrm{E}(\mathbf{x}))$ instead of $\mathbf{x}$. Thus we perform the analysis with the correlation matrix $\boldsymbol{\rho}$ instead of the covariance matrix $\boldsymbol{\Sigma}$.

Using some appropriate rules we can assess how many of these components are enough in further analysis. This can be done e.g. by setting a threshold value to the proportion of the original sum of the variances of the components, $\mathrm{trace}(\mathrm{Cov}(\mathbf{x}))$, that these principal components need to explain at least. If we choose to take k components, those are the first k components, as they are ordered according to their variances. The screeplots can be used to choose the value of k graphically and there are tests that can be used to determine if the last $p-k$ values are zero, see for example Jolliffe (2002) and the references therein.

For more on PCA, including robust versions of PCA which are not sensitive to outliers, and the differences between PCA and another dimension reduction technique *factor analysis*, see e.g. Jolliffe (2002). For a recent review on robust PCA methods, see Bouwmans and Zahzah (2014).

2.5 Independent Component Analysis

Consider an IC model of the form (2.4), i.e.

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Omega}\mathbf{z},$$

where $\boldsymbol{\mu} \in \mathbb{R}^p$ is a location vector, $\boldsymbol{\Omega} \in \mathbb{R}^{p \times p}$ a full-rank matrix and $\mathbf{z} \in \mathbb{R}^p$ a latent random vector with independent components. Matrix $\boldsymbol{\Omega}$ can be called here a *mixing* matrix.

As stated earlier, we assume that the latent sources $\mathbf{z}$ are standardized. We also assume that at most one of the components of $\mathbf{z}$ has a Gaussian distribution.

The goal in Independent Component Analysis (ICA) is to find an unmixing matrix $\boldsymbol{\Gamma} = \boldsymbol{\Omega}^{-1}$ in such way that $\boldsymbol{\Gamma}\left(\mathbf{x} - \boldsymbol{\mu}\right)$ has independent components. However, $\mathbf{z}$ can only be found up to the signs and order of the components. This is easy to see using a sign change matrix $\mathbf{J}$ and a permutation matrix $\mathbf{K}$, as

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Omega}\mathbf{z} = \boldsymbol{\mu} + \left(\boldsymbol{\Omega}\mathbf{K}'\mathbf{J}\right)(\mathbf{J}\mathbf{K}\mathbf{z}) = \boldsymbol{\mu} + \boldsymbol{\Omega}^*\mathbf{z}^*,$$

where also $\mathbf{z}^*$ fulfills the assumptions for $\mathbf{z}$.

There exists an orthogonal matrix $\mathbf{U}_0$ such that

$$\mathbf{z} = \mathbf{U}_0\mathbf{x}^{(s)}, \tag{2.7}$$

where $\mathbf{x}^{(s)} = \boldsymbol{\Sigma}^{-1/2}\left(\mathbf{x} - \boldsymbol{\mu}\right)$ (see for example Miettinen et al., 2015b). This can be seen as follows. Consider a singular value decomposition $\boldsymbol{\Omega} = \mathbf{U}\mathbf{D}\mathbf{V}'$, where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices and $\mathbf{D}$ a diagonal matrix such that the diagonal elements are all positive. Then $\boldsymbol{\Sigma} = \boldsymbol{\Omega}\boldsymbol{\Omega}' = \mathbf{U}\mathbf{D}^2\mathbf{U}'$ and hence $\boldsymbol{\Sigma}^{-1/2} = \mathbf{U}\mathbf{D}^{-1}\mathbf{U}'$. From this it follows that

$$\mathbf{x}^{(s)} = \boldsymbol{\Sigma}^{-1/2}\left(\mathbf{x} - \boldsymbol{\mu}\right) = \mathbf{U}\mathbf{D}^{-1}\mathbf{U}'\mathbf{U}\mathbf{D}\mathbf{V}'\mathbf{z} = \mathbf{U}\mathbf{V}'\mathbf{z} = \mathbf{U}_0'\mathbf{z},$$

where $\mathbf{U}_0 := \mathbf{V}\mathbf{U}'$ is an orthogonal matrix. It follows from this result that we can solve this Independent Component Analysis problem by only finding an orthogonal matrix $\mathbf{U}_0$.

Therefore the general procedure that is applied in this thesis is

1. Standardize $\mathbf{x}$: $\mathbf{x}^{(s)} = \boldsymbol{\Sigma}^{-1/2}\left(\mathbf{x} - \boldsymbol{\mu}\right)$.

2. Find an orthogonal matrix $\mathbf{U}$ that maximizes a *criterion function* $G\left(\mathbf{U}, \mathbf{x}^{(s)}\right)$. Two main approaches here for finding the latent components are deflation-based (one by one) and symmetric (simultaneously).

3. Calculate an unmixing matrix functional: $\boldsymbol{\Gamma} := \mathbf{U}\boldsymbol{\Sigma}^{-1/2}$.

Remark 2.1 *Note that while* $\boldsymbol{\Sigma}^{-1/2}$ *only depends on the distribution of* $\mathbf{x}$*, the orthogonal matrix* $\mathbf{U}$*, and thus also* $\boldsymbol{\Gamma}$*, depends also on the criterion function used. In practice also the used algorithm affects the properties of* $\mathbf{U}$*.*

For observed iid data $\mathbf{x}_1, \ldots, \mathbf{x}_n$ the procedure goes a follows. Let $\bar{\mathbf{x}}$ be the sample mean vector and $\mathbf{S}$ the sample covariance matrix. First calculate the standardized observations $\hat{\mathbf{x}}_i^{(s)} = \mathbf{S}^{-1/2}\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)$, $i = 1, \ldots, n$. Then for calculating a criterion function the theoretical quantities are replaced by their sample counterparts. The estimate

$\hat{\mathbf{U}}$ is obtained by maximizing $G\left(\hat{\mathbf{U}}, \hat{\mathbf{x}}^{(s)}\right)$. The unmixing matrix estimate $\hat{\boldsymbol{\Gamma}} = \hat{\mathbf{U}}\mathbf{S}^{-1/2}$. This procedure applies to all the methods presented in this thesis.

An important requirement for an unmixing matrix functional $\boldsymbol{\Gamma} = \boldsymbol{\Omega}^{-1}$ is that it should be *affine equivariant*. This means the following. Let $\mathbf{x}^* = \mathbf{A}\mathbf{x} + \mathbf{b}$, where $\mathbf{A} \in \mathbb{R}^{p \times p}$ is a non-singular transformation matrix and $\mathbf{b} \in \mathbb{R}^p$ a location shift vector. If $\boldsymbol{\Gamma}$ is affine equivariant, then $\boldsymbol{\Gamma}^* = \boldsymbol{\Gamma}\mathbf{A}^{-1}$ and thus $\boldsymbol{\Gamma}^*\mathbf{x}^* = \boldsymbol{\Gamma}\mathbf{x}$, up to the location shifts, signs and order of the components. This property ensures that transforming the observed data does not change the independent components in any fundamental way.

Denote by $\mathbf{E}^{jk}$ a matrix, where the element (j,k) equals to 1 and all others equal to 0. Then for a p-variate random vector $\mathbf{x}$ all the possible fourth moments are included in the matrices

$$\mathbf{B}^{jk}(\mathbf{x}) = E\left(\mathbf{x}\mathbf{x}'\mathbf{E}^{jk}\mathbf{x}\mathbf{x}'\right), \; j,k = 1, \ldots, p. \tag{2.8}$$

Also all the possible fourth order cumulants (2.3) for a standardized random vector $\mathbf{x}^{(s)}$ are obtained from the matrices

$$\mathbf{C}^{jk}\left(\mathbf{x}^{(s)}\right) = \mathbf{B}^{jk}\left(\mathbf{x}^{(s)}\right) - \mathbf{E}^{jk} - \mathbf{E}^{kj} - \mathrm{trace}\left(\mathbf{E}_{jk}\right)\mathbf{I}_p, \; j,k = 1, \ldots, p. \tag{2.9}$$

Note that the only non-zero elements of $\mathbf{C}^{jk}(\mathbf{z})$, $j,k = 1, \ldots, p$, are $\left(\mathbf{C}^{jj}(\mathbf{z})\right)_{jj}$ (see e.g. Miettinen et al., 2015b).

Fourth Order Blind Identification (FOBI) Cardoso (1989) introduces FOBI, which uses a matrix of fourth moments,

$$\mathbf{B}\left(\mathbf{x}^{(s)}\right) = \sum_{j=1}^{p} \mathbf{B}^{jj}\left(\mathbf{x}^{(s)}\right) = E\left(\mathbf{x}^{(s)}\mathbf{x}^{(s)\prime}\mathbf{x}^{(s)}\mathbf{x}^{(s)\prime}\right), \tag{2.10}$$

which is a measure of multivariate kurtosis. Note that $\mathbf{B}(\mathbf{z}) = \sum_{j=1}^{p}\left(\kappa_{4,j} + p + 2\right)\mathbf{E}^{jj}$, where $\kappa_{4,j} = E\left(z_j^4\right) - 3$ (see e.g. Miettinen et al., 2015b).

Then FOBI uses (2.10) and searches for an orthogonal matrix $\mathbf{U} = \left(\mathbf{u}_1, \ldots, \mathbf{u}_p\right)' \in \mathbb{R}^{p \times p}$ that maximizes the *criterion function*

$$\left\|\mathrm{diag}\left(\mathbf{U}\mathbf{B}\left(\mathbf{x}^{(s)}\right)\mathbf{U}'\right)\right\|^2 = \sum_{i=1}^{p}\left(\mathbf{u}_i'\mathbf{B}\left(\mathbf{x}^{(s)}\right)\mathbf{u}_i\right)^2, \tag{2.11}$$

and then finds an unmixing matrix $\mathbf{\Gamma} = \mathbf{U\Sigma}^{-1/2}$. This maximization produces an orthogonal matrix $\mathbf{U}$, where the rows are the eigenvectors of $\mathbf{B}\left(\mathbf{x}^{(s)}\right)$ and also $\mathbf{B}(\mathbf{z})$ is diagonal. (Miettinen et al., 2015b)

The independent components are ordered according to the diagonal values of $\mathbf{B}(\mathbf{z})$. If the diagonal values are the same, then the corresponding rows of $\mathbf{\Gamma}$ are not uniquely defined. The diagonal values are the kurtosis values $\kappa_{4,j}, j = 1, \ldots, p$ of the independent components.

In order to make some inference on the efficiency of the estimate $\hat{\mathbf{\Gamma}}$ in large samples and compare it to others, we need to assess its asymptotic behaviour. As the estimate $\hat{\mathbf{\Gamma}}$ is affine equivariant, it is not a restriction to use $\mathbf{\Omega} = \mathbf{I}_p$ for asymptotic considerations, and thus $\mathbf{\Gamma} = \mathbf{I}_p$. If the fourth moments of the components of $\mathbf{z}$ are distinct and the eight moments bounded, then $\hat{\mathbf{\Gamma}} \to_P \mathbf{I}_p$ (i.e. the unmixing matrix estimate is consistent, since it converges in probability to the identity matrix). The limiting distribution of $\sqrt{n}\,\mathrm{vec}\left(\hat{\mathbf{\Gamma}} - \mathbf{I}_p\right)$ is a multivariate normal distribution with a zero mean vector. For more details, see Ilmonen et al. (2010b) and Miettinen et al. (2015b).

Joint Approximate Diagonalization of Eigen-matrices (JADE) As having distinct diagonal elements in $\mathbf{B}(\mathbf{z})$ is a restrictive assumption, Cardoso and Souloumiac (1993) have proposed JADE, where such an assumption is not needed. JADE uses the fourth cumulant matrices (2.9) for the standardized observations. The *criterion function* to be maximized here for an orthogonal matrix $\mathbf{U} = \left(\mathbf{u}_1, \ldots, \mathbf{u}_p\right)'$ is

$$\sum_{j=1}^{p}\sum_{k=1}^{p}\left\|\mathrm{diag}\left(\mathbf{U}\mathbf{C}^{jk}\left(\mathbf{x}^{(s)}\right)\mathbf{U}'\right)\right\|^2 = \sum_{i=1}^{p}\sum_{j=1}^{p}\sum_{k=1}^{p}\left(\mathbf{u}_i'\mathbf{C}^{jk}\left(\mathbf{x}^{(s)}\right)\mathbf{u}_i\right)^2. \tag{2.12}$$

Remark 2.2 *According to* (2.1), *the maximization of* (2.11) *and* (2.12) *can be written as a minimization of*

$$\left\|\mathrm{off}\left(\mathbf{U}\mathbf{B}\left(\mathbf{x}^{(s)}\right)\mathbf{U}'\right)\right\|^2$$

and

$$\sum_{j=1}^{p}\sum_{k=1}^{p}\left\|\mathrm{off}\left(\mathbf{U}\mathbf{C}^{jk}\left(\mathbf{x}^{(s)}\right)\mathbf{U}'\right)\right\|^2$$

for an orthogonal matrix $\mathbf{U}$, *respectively. Then* $\mathbf{\Gamma} = \mathbf{U\Sigma}^{-1/2}$.

JADE aims to maximize the sum of the squared diagonal elements of the matrices $\mathbf{U}\mathbf{C}^{jk}\left(\mathbf{x}^{(s)}\right)\mathbf{U}'$ to make the matrices 'as diagonal as possible'. There are several methods available for the approximate joint diagonalization of symmetric matrices. Miettinen et al. (2015b) use a fixed-point algorithm for JADE that is based on Lagrangian multiplier technique and in their simulations a popular JADE algorithm based on Jacobi rotations by Clarkson (1988) seems to provide the same solution.

The unmixing matrix estimate is uniquely defined (up to the order and signs of the rows) if and only if at most one $\mathbf{C}^{jj}(\mathbf{z}) = \mathbf{0}$ (see e.g. Miettinen et al., 2015b). In other words, if at most one of the values $\left(\mathbf{C}^{jj}(\mathbf{z})\right)_{jj} = \kappa_{4,j} = 0$, $j = 1, \ldots, p$, then JADE works. This means that two or more components may have same kurtosis values as long as they are not zero.

As the estimate $\hat{\boldsymbol{\Gamma}}$ is affine equivariant, we can consider the case where $\boldsymbol{\Omega} = \mathbf{I}_p$ and thus $\boldsymbol{\Gamma} = \mathbf{I}_p$. If the eight moments of the components of $\mathbf{z}$ are bounded and at most one $\mathbf{C}^{jj}(\mathbf{z}) = \mathbf{0}$, then $\hat{\boldsymbol{\Gamma}} \rightarrow_P \mathbf{I}_p$. The limiting distribution of $\sqrt{n}\ \mathrm{vec}\left(\hat{\boldsymbol{\Gamma}} - \mathbf{I}_p\right)$ is also a multivariate normal distribution with a zero mean vector. For more details, see Bonhomme and Robin (2009) and Miettinen et al. (2015b).

Projection pursuit In projection pursuit we search for directions that maximize a criterion function (see for example Huber, 1985). FastICA, proposed by Hyvärinen (1999), is a method that uses the projection pursuit approach for ICA and maximizes, for an orthogonal matrix $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_p)'$, a criterion function

$$\sum_{i=1}^{p} \left| E\left(G\left(\mathbf{u}_i'\mathbf{x}^{(s)}\right)\right)\right|, \tag{2.13}$$

where the function G is chosen to be nonlinear, nonquadratic and twice continuously differentiable, such that for a normally distributed variable z it is required that $E(G(z)) = 0$. FastICA has a deflation-based version, where the components are found one by one, and a symmetric version, where the components are found simultaneously.

An estimate is consistent, if it converges in probability to the true value of the parameter when the number of observations grows infinitely large. To ensure the consistency of the estimation procedure, for the function G there are some conditions. Assume that wlog that the components are ordered in such way that $\left|E\left(G\left(z_1\right)\right)\right| \geq \ldots \geq \left|E\left(G\left(z_p\right)\right)\right|$. We require that

- $\left|E\left(G\left(\mathbf{u}_i'\mathbf{z}\right)\right)\right| \leq \left|E\left(G\left(z_i\right)\right)\right|$ for all $i = 1, \ldots, p$, when $\mathbf{u}_i'\mathbf{e}_j = 0$ for all $j = 1, \ldots, i-1$ (deflation-based).

- $\sum_{i=1}^{p} \left|E\left(G\left(\mathbf{u}_i'\mathbf{z}\right)\right)\right| \leq \sum_{i=1}^{p} \left|E\left(G\left(z_i\right)\right)\right|$ (symmetric).

Note that, as $\mathbf{x}^{(s)} = \mathbf{U}_0\mathbf{z}$, for an orthogonal matrix $\mathbf{U}_0$, then $E\left(G\left(\mathbf{u}_i'\mathbf{z}\right)\right) := E\left(G\left(\left(\mathbf{u}_i^{*\prime}\mathbf{U}_0\right)\mathbf{z}\right)\right) = E\left(G\left(\mathbf{u}_i^{*\prime}\mathbf{x}^{(s)}\right)\right)$, which is in the same format than the G function in (2.13).

The function $g(z) = G'(z)$, i.e. the derivative of $G(z)$, is the so-called nonlinearity function. The term 'nonlinearity' comes from the fact that G is nonquadratic and therefore its derivative g is nonlinear.

Nonlinearity functions that are used in the literature include the following:

- With $g(z) = z^2$ and z^3 we are searching for directions, where the skewness and kurtosis deviates from the skewness and kurtosis of the normal distribution, respectively. These have been proven to satisfy the consistency conditions, see Comon (1994); Miettinen et al. (2015b); Virta et al. (2016).

- $g(z) = \tanh(az)$ (hyperbolic tangent), where $a \in (0, \infty)$ is a constant. This is good in general and is preferred with heavy-tailed sources. For the densities of the distributions, for which $\tanh(az)$ is optimal with different values of a, see Virta and Nordhausen (2017b).

- $g(z) = z \cdot \exp\left(-az^2/2\right)$, where $a \approx 1$ is a constant. This works well when the sources are heavy-tailed.

The last two do not fulfill the consistency conditions, as seen in Wei (2014).

For the limiting distribution of the general deflation-based fastICA unmixing matrix estimator, see for example Ollila (2010), Nordhausen et al. (2011) and Miettinen et al. (2014a). See also Tichavský et al. (2006), Dermoune and Wei (2013) and Miettinen et al. (2015b).

The deflation-based fastICA with $g(z) = z^3$ was already suggested in Hyvärinen and Oja (1997). The criterion function, for an orthogonal matrix $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_p)'$, is

$$\sum_{i=1}^{p} \left|E\left(\left(\mathbf{u}_i'\mathbf{x}^{(s)}\right)^4\right) - 3\right|.$$

For the symmetric and deflation-based fastICA with $g(z) = z^3$, the limiting distribution of $\sqrt{n}\, \mathrm{vec}\left(\hat{\boldsymbol{\Gamma}} - \mathbf{I}_p\right)$ is a multivariate normal distribution with a zero mean vector. For more details, see Miettinen et al. (2015b).

Miettinen et al. (2015b) also compare the asymptotic efficiencies of the estimates of $\boldsymbol{\Gamma}$ using FOBI, JADE and both symmetric and deflation based fastICA with $g(z) = z^3$. If the independent components are identically distributed, JADE and symmetric fastICA estimates are asymptotically equivalent, while FOBI fails due to the fact that the elements of $B(\mathbf{z})$ are not distinct.

Nordhausen et al. (2011) suggest a reloaded version of the deflation-based fastICA. It uses a known estimator, such as FOBI, as an initial value for an unmixing matrix estimate. Then the extraction order of the components is optimized in such way that the trace of the limiting covariance matrix of $\mathrm{vec}(\hat{\boldsymbol{\Gamma}})$ is minimized. This leads to a faster convergence and algorithm becomes more stable in small sample sizes. Also the limiting distribution of the estimate corresponds to the regular limiting distribution of fastICA estimate that extracts the components in the (same) optimal order.

Miettinen et al. (2014a) have proposed a deflation-based method with adaptive choices for nonlinearity functions. This method generalizes the reloaded version by allowing different nonlinearity functions to be used for each component. This again makes the algorithm to converge faster and to become more stable in very small samples (e.g. when $n = 100$).

Koldovský et al. (2006) have proposed a symmetric fastICA method, where different nonlinearities can be used for different sources. For the limiting distribution of the estimate, see Wei (2015); Tichavský et al. (2006).

Miettinen et al. (2017b) introduce a squared version of symmetric fastICA, where a criterion function is

$$\sum_{i=1}^{p} \left(E\left(G\left(\mathbf{u}_i' \mathbf{x}^{(s)}\right)\right)\right)^2,$$

in which for function G it is required that

$$\sum_{i=1}^{p} \left(E\left(G\left(\mathbf{u}_i' \mathbf{z}\right)\right)\right)^2 \leq \sum_{i=1}^{p} \left(E\left(G\left(z_i\right)\right)\right)^2.$$

Miettinen et al. (2017b) show that when they examine the performance of the algorithms, in most of the cases the symmetric fastICA is more efficient (the asymptotic variance is lower) than the deflation based fastICA. Also the squared symmetric version is shown to be more efficient in cases where the separation of the components is difficult. According to all the asymptotic and finite sample estimates, the overall performance of the squared symmetric version seems to be the best, even though it does not beat others in every case.

As the ICA methods only allow one Gaussian component, Blanchard et al. (2006) has suggested Non-Gaussian Component Analysis (NGCA), which allows to find subspaces $\mathbf{z}_{(1)} \in \mathbb{R}^{\mathrm{p}}$ and $\mathbf{z}_{(2)} \in \mathbb{R}^{\mathrm{p-k}}$, where the components are non-Gaussian and Gaussian, respectively. In such case the model can be written as (2.5), i.e. in the form of

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Omega}_1 \mathbf{z}_{(1)} + \boldsymbol{\Omega}_2 \mathbf{z}_{(2)},$$

where $\boldsymbol{\Omega} = \left(\boldsymbol{\Omega}_1, \boldsymbol{\Omega}_2\right)$. Blanchard et al. (2006) has also provided an algorithm for finding the subspace using fastICA algorithm. Recently Nordhausen et al. (2017c) has suggested a test statistic and bootstrap test for finding the subspace dimension in the context of FOBI.

For robust considerations of Independent Component Analysis, see for example Nordhausen et al. (2008); Ilmonen and Paindaveine (2011); Hallin and Mehta (2015).

Applications ICA methods have been used for example in biomedical applications. Such applications include analysis and classification of heartbeats, functional magnetic resonance imaging (fMRI) and brain, EEG (electroencephalogram) and MEG (Magnetoencephalography) studies (see Naik and Wang (2014) and the references therein).

Assume that we have for example four microphones in a room and they record voices from four people. The original signals $\mathbf{z} = \left(z_1, z_2, z_3, z_4\right)'$ are captured partly by different microphones. What the microphones capture is $\mathbf{x} = \left(x_1, x_2, x_3, x_4\right)'$. This is known as the ‘cocktail party problem’. ICA methods can then be use to extract the original voices from the microphone records. (Bell and Sejnowski, 1995)

Other applications include for example financial data (e.g. currency exchange rates, stock market prices), noise reduction in images as well as face recognition (see Hyvärinen and Oja (2000) and Stone (2004) and the references therein).

Differences between ICA and PCA Independent Component Analysis and Principal Component Analysis both can be used in dimension reduction, but they have several differences.

In ICA the goal is to maximize a measure of non-Gaussianity, such as skewness or kurtosis, by finding independent components $\mathbf{z}$ that have a unit variance. Then the most interesting components are the ones that are non-Gaussian.

In PCA the main goal is dimension reduction. First we create new uncorrelated components $\mathbf{z}$ that are ordered in a descending order of their variances. Then usually the first few components are kept. Also PCA has no model assumptions unlike ICA.

In PCA the latent components are

$$\mathbf{z} = \mathbf{U}'\left(\mathbf{x} - \boldsymbol{\mu}\right),$$

while in ICA they are

$$\mathbf{z} = \mathbf{V}\boldsymbol{\Sigma}^{-1/2}\left(\mathbf{x} - \boldsymbol{\mu}\right) = \mathbf{VUD}^{-1/2}\mathbf{U}'\left(\mathbf{x} - \boldsymbol{\mu}\right),$$

where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices. Thus in both we first find uncorrelated variables. While PCA stops here, in ICA we still standardize the variables and then rotate them with an orthogonal matrix $\mathbf{VU}$.

For example, let z_1, z_2 and z_3 be independent $N(0,1)$-distributed variables. Let the distribution of z_4 be

$$\frac{1}{3}N(-5,1) + \frac{1}{3}N(0,1) + \frac{1}{3}N(5,1).$$

The distribution of z_4 is thus a mixture of three normal distributions with different expected values. In order to fulfill the assumptions of ICA models, z_4 is also standardized to have a unit variance. Thus $E(\mathbf{z}) = \mathbf{0}$ and $Cov(\mathbf{z}) = \mathbf{I}$. The excess kurtosis $\kappa_4 \approx -1.335$ for the component z_4 and zero for others.

We simulate 1000 observations from the distribution of $\mathbf{z}$. Figure 2.1 shows the ‘observed’ $\mathbf{x} = \boldsymbol{\Omega}\mathbf{z}$, where $\boldsymbol{\Omega}$ is a random mixing matrix. Figures 2.2 and 2.3 show the resulting principal components from PCA and independent components from ICA method JADE, respectively.

The mixed components in Figure 2.1 do not show anything special. From Figure 2.2 we can see that PCA is not able to find anything, as all the components seem to be just noise. However, JADE finds one independent component (IC4) that clearly has three different groups as in z_4 (Figure 2.3). JADE is able to find the groups

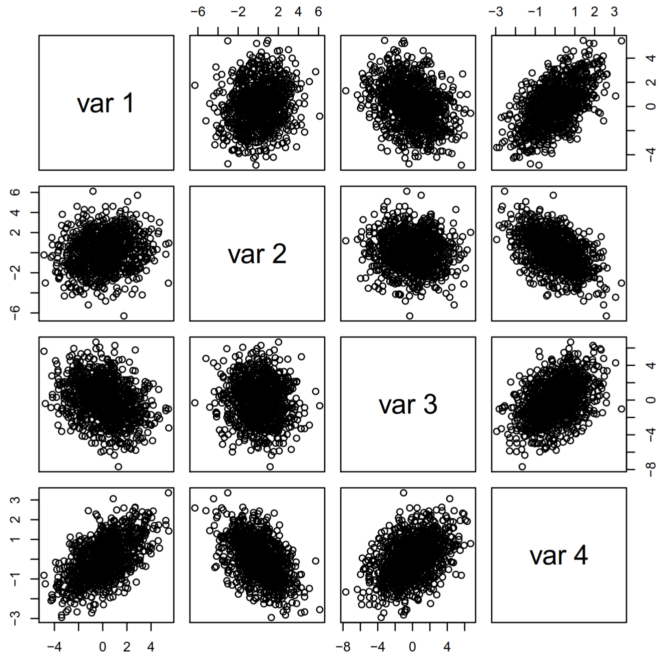


Figure 2.1: Scatterplot matrix of **x** *of a sample of size 1000*

since the excess kurtosis of z_4 is clearly lower compared to others. The other components cannot be differentiated.

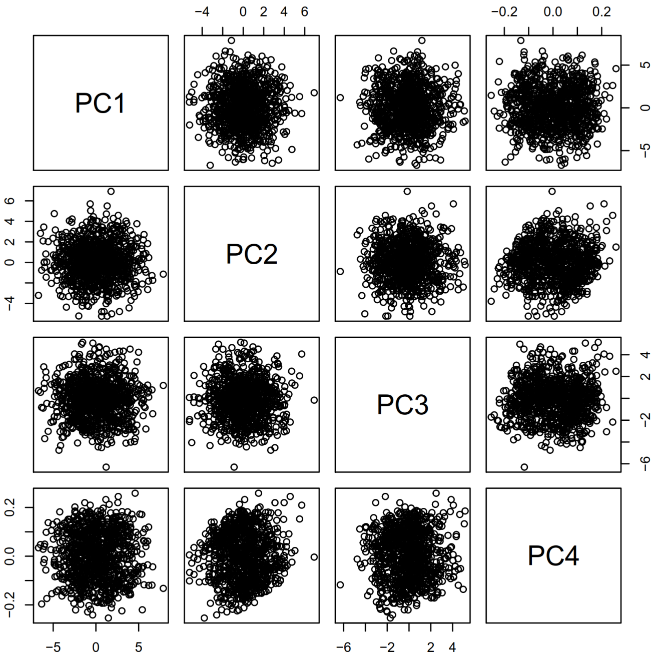


Figure 2.2: Scatterplot matrix of the principal components of a sample of size 1000

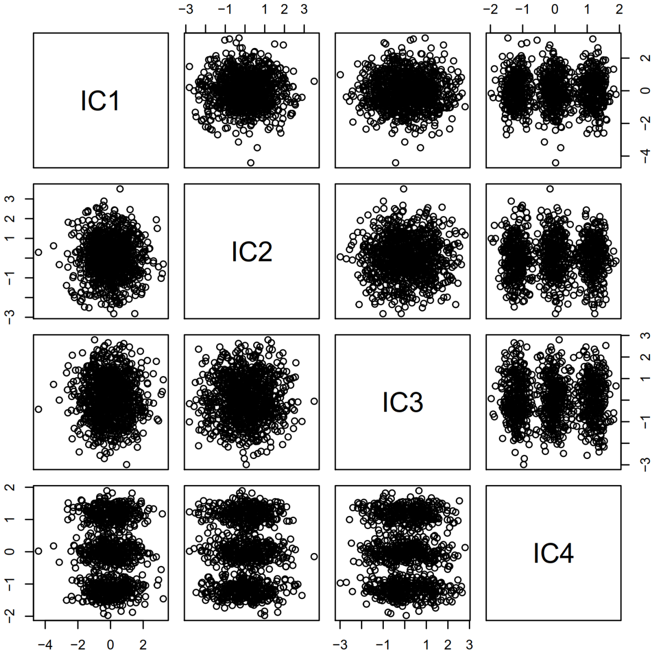


Figure 2.3: Scatterplot matrix of the independent components of a sample of size 1000 using JADE

2.6 Supervised dimension reduction

Now we assume that we have a response y that depends on a random vector $\mathbf{x} \in \mathbb{R}^p$ via some unknown function f as in (2.6). When the dimension p gets higher, modelling such data can become challenging and computationally intensive. We can first reduce the dimension of $\mathbf{x}$ and then model y with the these new variables. However, we might then lose some important information on the relationship between the response and the explaining variables. The main purpose in supervised dimension reduction is to find a k-dimensional subspace of $\mathbf{x}$ that captures as much of the relationship between y and $\mathbf{x}$ as possible.

Definition 2.5 *A Blind Source Separation model for the joint distribution of a random vector* $\mathbf{x}$ *and a response* y *is*

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Omega}\mathbf{z},$$

where $\boldsymbol{\mu} \in \mathbb{R}^p$ *is location vector and* $\boldsymbol{\Omega} \in \mathbb{R}^{p\times p}$ *a mixing matrix, as before. For the latent (unknown) random vector* $\mathbf{z} \in \mathbb{R}^p$ *the assumptions are*

$$\mathrm{E}(\mathbf{z}) = \mathbf{0}, \mathrm{Cov}(\mathbf{z}) = \mathbf{I}_p$$

and

$$\left(y, \mathbf{z}_{(1)}\right) \perp\!\!\!\perp \mathbf{z}_{(2)}, \tag{2.14}$$

where $\mathbf{z} = \left(\mathbf{z}_{(1)}', \mathbf{z}_{(2)}'\right)'$ *is divided into vectors* $\mathbf{z}_{(1)} \in \mathbb{R}^k$ *and* $\mathbf{z}_{(2)} \in \mathbb{R}^{p-k}$.

As only $\mathbf{z}_{(1)}$ contributes to y and $\mathbf{z}_{(2)}$ is just noise in which we are not interested in, we can write the model in the form (2.5).

The goal here is to find an estimate for $\boldsymbol{\Gamma}$, which equals to the first k rows of $\boldsymbol{\Omega}^{-1}$ and therefore $\boldsymbol{\Gamma}\left(\mathbf{x} - \boldsymbol{\mu}\right) = \mathbf{z}_{(1)}$. To be more precise, we estimate a k-variate subspace spanned by the rows of $\boldsymbol{\Gamma}$ and given by its projection matrix $\mathbf{P}_{\boldsymbol{\Gamma}}$. Instead of using a projection matrix in the form

$$\mathbf{P}_{\boldsymbol{\Gamma}} = \boldsymbol{\Gamma}\left(\boldsymbol{\Gamma}'\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\Gamma}',$$

we can utilize standardization. As $\boldsymbol{\Gamma} = \mathbf{U}\boldsymbol{\Sigma}^{-1/2}$ for an orthogonal matrix $\mathbf{U} = \left(\mathbf{U}_1', \mathbf{U}_2'\right)' \in \mathbb{R}^{p\times p}$, we can write

$$\mathbf{P}_{\boldsymbol{\Gamma}} = \boldsymbol{\Sigma}^{1/2}\Big(\mathbf{U}_1' \underbrace{\left(\mathbf{U}_1\mathbf{U}_1'\right)^{-1}}_{=\mathbf{I}_k} \mathbf{U}_1\Big)\boldsymbol{\Sigma}^{-1/2} = \boldsymbol{\Sigma}^{1/2}\left(\mathbf{U}_1'\mathbf{U}_1\right)\boldsymbol{\Sigma}^{-1/2},$$

for $\mathbf{U}_1 \in \mathcal{O}^{k\times p}$. This is a projection with respect to Mahalanobis inner product. It is easy to see that if $k = p$, then $\mathbf{P}_\Gamma = \mathbf{I}$, as no projection would be needed. We can also write

$$\mathbf{Q}_\Gamma = \mathbf{\Sigma}^{1/2}\Big(\mathbf{U}_2' \underbrace{\left(\mathbf{U}_2\mathbf{U}_2'\right)^{-1}}_{=\mathbf{I}_{p-k}} \mathbf{U}_2\Big)\mathbf{\Sigma}^{-1/2} = \mathbf{\Sigma}^{1/2}\left(\mathbf{U}_2'\mathbf{U}_2\right)\mathbf{\Sigma}^{-1/2},$$

for $\mathbf{U}_2 \in \mathcal{O}^{(p-k)\times p}$. Then $\mathbf{P}_\Gamma + \mathbf{Q}_\Gamma = \mathbf{I}_p$. Also $\mathbf{P}_\Gamma\mathbf{\Sigma}\mathbf{Q}_\Gamma = \mathbf{0}_p$, i.e. $\mathbf{P}_\Gamma\mathbf{x}$ and $\mathbf{Q}_\Gamma\mathbf{x}$ are uncorrelated.

Finally we can write

$$\mathbf{x} = \underbrace{\mathbf{P}_\Gamma\mathbf{x}}_{\text{signal}} + \underbrace{\mathbf{Q}_\Gamma\mathbf{x}}_{\text{noise}}.$$

Let $\mathbf{W}_1 \in \mathbb{R}^{k\times k}$ and $\mathbf{W}_2 \in \mathbb{R}^{(p-k)\times(p-k)}$ be some orthogonal matrices. The assumptions for $\mathbf{z}$ are also true for $\mathbf{z}^*_{(1)} = \mathbf{W}_1\mathbf{z}_{(1)}$ and $\mathbf{z}^*_{(2)} = \mathbf{W}_2\mathbf{z}_{(2)}$ with $\mathbf{\Omega}^*_1 = \mathbf{\Omega}_1\mathbf{W}_1'$ and $\mathbf{\Omega}^*_2 = \mathbf{\Omega}_2\mathbf{W}_2'$, as from (2.5) we get

$$\begin{aligned}\mathbf{x} = \boldsymbol{\mu} + \mathbf{\Omega}_1\mathbf{z}_{(1)} + \mathbf{\Omega}_2\mathbf{z}_{(2)} &= \boldsymbol{\mu} + \mathbf{\Omega}_1\mathbf{W}_1'\mathbf{W}_1\mathbf{z}_{(1)} + \mathbf{\Omega}_2\mathbf{W}_2'\mathbf{W}_2\mathbf{z}_{(2)}\\ &= \boldsymbol{\mu} + \mathbf{\Omega}^*_1\mathbf{z}^*_{(1)} + \mathbf{\Omega}^*_2\mathbf{z}^*_{(2)}.\end{aligned} \tag{2.15}$$

However, as we are estimating the subspace, this is not an issue here. In order to avoid any ambiguity in the model, we also assume that the value of k is the smallest such value that assumptions for $\mathbf{z}$ are valid.

Now the regression model (2.6) can be reduced to

$$y = f\left(\mathbf{z}_{(1)}, \epsilon\right),$$

where $f(\cdot)$ is an unspecified function, that may be different than in (2.6), and ϵ is an unknown random variable.

Sliced Inverse Regression Li (1991) has first introduced Sliced Inverse Regression (SIR). In SIR, instead of regressing y on $\mathbf{x}$, Li (1991) uses inverse regression, where $\mathbf{x}$ is regressed against y.

The assumption (2.14) implies the assumptions

$$\mathbf{z}_{(2)} \perp\!\!\!\perp y | \mathbf{z}_{(1)} \text{ and} \tag{2.16}$$

$$E\left(\mathbf{z}_{(2)} | \mathbf{z}_{(1)}\right) = \mathbf{0} \tag{2.17}$$

in the original SIR article (Li, 1991). The equation (2.17) is known as the 'linearity assumption' (see e.g. Li, 1991; Cook and Weisberg, 1991). From (2.16) and (2.17) it follows that

$$\mathrm{Cov}(\mathrm{E}(\mathbf{z}|\mathrm{y})) = \begin{pmatrix} \mathrm{Cov}(\mathrm{E}(\mathbf{z}_{(1)}|\mathrm{y})) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}. \tag{2.18}$$

In order to find the estimate for $\boldsymbol{\Gamma}$, according (2.7) it is enough to start from the standardized variables $\mathbf{x}^{(s)}$ and search for a matrix $\mathbf{U} = \left(\mathbf{u}_1, \ldots, \mathbf{u}_k\right)' \in \mathcal{O}^{k\times p}$ that maximizes

$$\begin{aligned} &\left\| \mathrm{diag}\left(\mathbf{U}\mathrm{Cov}\left(\mathrm{E}\left(\mathbf{x}^{(s)}|\mathrm{y}\right)\right)\mathbf{U}'\right)\right\|^2 \\ &= \sum_{i=1}^{k} \left(\mathbf{u}_i' \mathrm{Cov}\left(\mathrm{E}\left(\mathbf{x}^{(s)}|\mathrm{y}\right)\right)\mathbf{u}_i\right)^2. \end{aligned} \tag{2.19}$$

Thus we get an estimate for $\boldsymbol{\Gamma} = \mathbf{U}\boldsymbol{\Sigma}^{-1/2}$. Just like 'true' latent components $\mathbf{z}$, the components $\boldsymbol{\Gamma}\left(\mathbf{x} - \boldsymbol{\mu}\right)$ are standardized. Also the components of $\mathrm{E}\left(\boldsymbol{\Gamma}\left(\mathbf{x} - \boldsymbol{\mu}\right)|\mathrm{y}\right)$ are uncorrelated and ordered according to their variances. This means that the elements of the diagonal matrices $\mathrm{Cov}\left(\mathrm{E}\left(\boldsymbol{\Gamma}\left(\mathbf{x} - \boldsymbol{\mu}\right)|\mathrm{y}\right)\right)$ are ordered according to their value. A larger value for the variance $\lambda_i = \mathrm{Cov}\left(\mathrm{E}\left(\boldsymbol{\Gamma}\left(\mathbf{x} - \boldsymbol{\mu}\right)|\mathrm{y}\right)_i\right), i = 1, \ldots, k$, means the stronger dependence between $\mathrm{E}\left(\boldsymbol{\Gamma}\left(\mathbf{x} - \boldsymbol{\mu}\right)|\mathrm{y}\right)_i$ and the response y. Note also that if some of the λ_i's are the same, then the corresponding rows of $\boldsymbol{\Gamma}$ are not uniquely defined.

Also, similar to FOBI, the eigenvectors of $\mathrm{Cov}\left(\mathrm{E}\left(\mathbf{x}^{(s)}|\mathrm{y}\right)\right)$ are the rows of the matrix $\mathbf{U} \in \mathcal{O}^{k\times p}$ and λ_i's are the eigenvalues associated to them.

In practice the response y needs to be sliced into some H disjoint intervals $S_h, h = 1, \ldots, H$ in order to calculate the estimate of $\mathrm{Cov}\left(\mathrm{E}\left(\mathbf{x}^{(s)}|\mathrm{y}\right)\right)$. The new discretized variable can be defined as a classificatory variable

$$\mathrm{y}^* := \sum_{h=1}^{H} h\, I\left(\mathrm{y} \in S_h\right),$$

where $I(\cdot)$ is an indicator function that equals to 1 if $\mathrm{y} \in S_h$ and zero otherwise. Thus y^* has H unique values and the conditional covariance matrices $\mathrm{Cov}\left(\mathrm{E}\left(\mathbf{x}^{(s)}|\mathrm{y}^*\right)\right)$ are used as an approximation for $\mathrm{Cov}\left(\mathrm{E}\left(\mathbf{x}^{(s)}|\mathrm{y}\right)\right)$. For observed data we calculate first the average value of $\mathbf{x}$ for each slice S_h and denote them by $\boldsymbol{\mu}_h$, $h = 1, \ldots, H$. Then we calculate

$$\widehat{\mathrm{Cov}}\left(E\left(\mathbf{x}^{(s)} | y^*\right)\right) = \frac{1}{H} \sum_{h=1}^{H} \boldsymbol{\mu}_h \boldsymbol{\mu}_h'$$

to estimate $\mathrm{Cov}\left(E\left(\mathbf{x}^{(s)} | y^*\right)\right)$. Li (1991) suggested that $H = 10$ is an appropriate value for SIR. Also $H > k + 1$ is required. Often the slices are chosen to have the same size, e.g. H quantiles of the response y, or as close to that as possible.

For observed data the value of k needs to be estimated. First we search for an for an orthogonal matrix estimate $\hat{\mathbf{U}} = \left(\mathbf{u}_1, \ldots, \mathbf{u}_p\right)' \in \mathbb{R}^{p \times p}$ that maximizes

$$\begin{aligned} &\left\| \mathrm{diag}\left(\hat{\mathbf{U}}\widehat{\mathrm{Cov}}\left(E\left(\mathbf{x}^{(s)} | y^*\right)\right)\hat{\mathbf{U}}'\right)\right\|^2 \\ &= \sum_{i=1}^{p} \left(\mathbf{u}_i' \widehat{\mathrm{Cov}}\left(E\left(\mathbf{x}^{(s)} | y^*\right)\right)\mathbf{u}_i\right)^2 \end{aligned}$$

and using then some appropriate rule to decide which of the p components are important. If the first k components are chosen, the estimate $\hat{\boldsymbol{\Gamma}}$ is then the first k rows of $\hat{\mathbf{U}}\mathbf{S}^{-1/2}$.

Li (1991) has proposed a Chi Squared test to estimate the subspace dimension k assuming multivariate normality for $\mathbf{x}$. The estimated subspace dimension is $\hat{k}$ if the average of the last $p - \hat{k}$ eigenvalues $\hat{\lambda}_i$ can be considered zero. Bura and Cook (2001a) have proposed a weighted Chi-Squared test, which also relies on asymptotics and sequential testing strategies can be used in estimating the subspace dimension k. The normality assumptions are relaxed and there are restrictions only for the conditional covariance structure of the predictors $\mathbf{x}$. An additional assumption

$$\mathrm{Cov}\left(\mathbf{z}_{(2)} | \mathbf{z}_{(1)}\right) = \mathbf{I}_{p-k} \text{ (a.s.)}, \tag{2.20}$$

which also follows from (2.14), gives a simpler Chi-squared test.

Also BIC (Bayesian Information Criterion) type criterion functions have been suggested for the subspace estimation, see for example Zhu et al. (2006) and Zhu et al. (2010).

Liquet and Saracco (2012) have suggested a practical bootstrap based criterion to estimate the subspace dimension k as well as the amount of slices H.

Nordhausen et al. (2017b) have used Assumption (2.14) in asymptotic and bootstrap tests for the estimation of the subspace dimension k. It could be argued that the difference between the assumption (2.14) and the assumptions (2.16), (2.17) and (2.20) together is not very big. Finding a distribution that fulfills the aforementioned three weaker assumptions but not (2.14) could be hard. However, this needs more research. Another recent contribution to the estimation of k is Luo and Li (2016).

Remark 2.3 *There are things to consider when estimating the value* $\hat{k}$*. The slicing may affect the value of* k*, as seen for example in Bura and Cook (2001b). The estimate* $\hat{k}$ *may be different with different values of* H*. Also methods may not find the whole subspace, as seen in SIR when* $E(\mathbf{x}|y^*) = \mathbf{0}$ *(see for example Cook and Weisberg (1991)). Thus it may be that* $\hat{k} < k$*.*

Remark 2.4 *Note that when slicing a variable, the matrix* (2.18) *naturally changes, but it still keeps the original block-diagonal structure. Also when slicing affects the estimate of* k*, the sizes of the blocks may change, but not the type of the structure.*

Zhu and Ng (1995) have shown that, assuming certain conditions, $\sqrt{n}\text{vech}\left(\widehat{\text{Cov}}(\mathbf{x}|y^*) - \text{Cov}(\mathbf{x}|y^*)\right)$ has a multivariate normal limiting distribution with zero mean vector and a bounded covariance matrix. See also Li and Zhu (2007) and the references therein.

Sliced Average Variance Estimation The drawback of SIR is that it fails to work when there are e.g. symmetric relationship between the explaining variable and the response variable. (Cook and Weisberg, 1991) Consider a regression model

$$y_i = 1 + x_{1i}^2 + x_{2i}^2 + \epsilon_i, \text{ for } i = 1, \dots, n,$$

where x_{1i} and x_{2i} have standardized normal distributions and ϵ is a $N(0, 0.1)$-distributed random variable.

Now $E\left(\mathbf{x}|y^*\right) = \mathbf{0}$ and thus $\text{Cov}\left(E\left(\mathbf{x}|y^*\right)\right) = \mathbf{0}$ and SIR fails. This is illustrated in Figure 2.4 using $n = 5000$ simulated observations based on the regression model. The scatterplot of y and x_1 is shown with 10 regions showing the slicing of y (10 quantiles). Each region contains 500 values. In each region the estimated value of $E\left(\mathbf{x}_1|y^*\right)$ is quite close to zero (the large points in the figure).

Therefore SAVE (Sliced Average Variance Estimation) is suggested (Cook and Weisberg, 1991; Cook, 2000). SAVE needs the

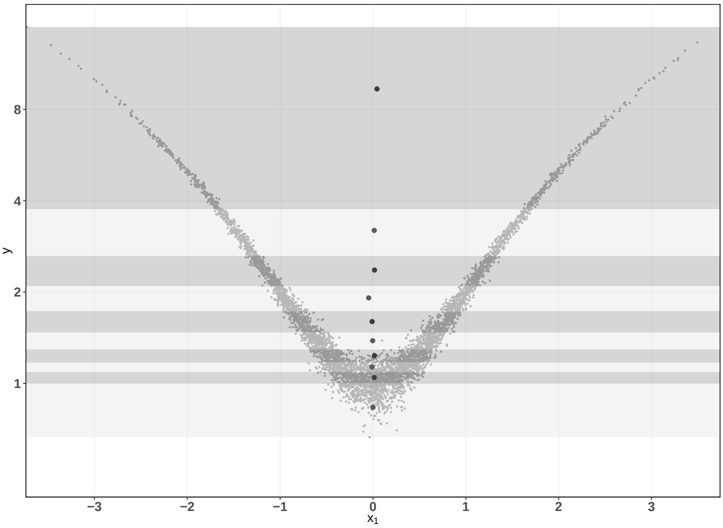


Figure 2.4: Scatterplot of y *and* x_1 *with the slices of* y *as the shaded areas. The large black point in each slice is an estimate of* $\mathrm{E}\left(x_1 \mid y^*\right)$ *in that slice. The* y*-axis is in a logarithmic scale.*

assumption (2.16) and the constant covariance assumption (2.20) and then

$$\mathrm{E}\left(\left(\mathbf{I}_p - \mathrm{Cov}(\mathbf{z} \mid y)\right)^2\right) = \begin{pmatrix} \mathrm{E}\left(\left(\mathbf{I}_k - \mathrm{Cov}\left(\mathbf{z}_{(1)} \mid y\right)\right)^2\right) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \tag{2.21}$$

a structure that holds also when the response y is sliced, but possibly with different block sizes, if slicing affects the value of k as in SIR. Again (2.14) implies both (2.16) and (2.20).

From Figure 2.4 it is rather easy to see that $\mathrm{E}\left(\mathbf{I}_p - \mathrm{Cov}\left(\mathbf{x}^{(s)} \mid y\right)\right)$ is non-zero, as variation grows when y grows. In order to use SAVE, we first calculate the standardized observations $\mathbf{x}^{(s)}$, as before. Then we search for a matrix $\mathbf{U} = \left(\mathbf{u}_1, \ldots, \mathbf{u}_k\right)' \in \mathscr{O}^{k \times p}$ that maximizes

$$\begin{aligned} &\left\| \mathrm{diag}\left(\mathbf{U}\mathrm{E}\left(\mathbf{I}_p - \mathrm{Cov}\left(\mathbf{x}^{(s)} \mid y\right)\right)^2 \mathbf{U}'\right) \right\|^2 \\ &= \sum_{i=1}^{k} \left(\mathbf{u}_i' \mathrm{E}\left(\left(\mathbf{I}_p - \mathrm{Cov}\left(\mathbf{x}^{(s)} \mid y\right)\right)^2\right) \mathbf{u}_i\right)^2. \end{aligned}$$

Finally we can estimate the value of $\mathbf{\Gamma} = \mathbf{U}\mathbf{\Sigma}^{-1/2}$. As in SIR, the value of k may also need to be estimated and $\mathrm{Cov}\left(\mathbf{x}^{(s)} \mid y\right)$ is approximated using the sliced y, i.e. by $\mathrm{Cov}\left(\mathbf{x}^{(s)} \mid y^*\right)$.

Let $x_{h1}^{(s)}, \dots, x_{hn_h}^{(s)}$ be the observations in slice S_h. For observed data we calculate

$$\frac{1}{H}\sum_{h=1}^{H}\left(\mathbf{I}_p - \widehat{\mathrm{Cov}}\left(\mathbf{x}^{(s)} \,|\, y^* = h\right)\right)^2$$
$$= \frac{1}{H}\sum_{h=1}^{H}\left(\mathbf{I}_p - \frac{1}{n_h - 1}\sum_{i=1}^{n_h}\left(x_{hi}^{(s)} - \bar{x}_h^{(s)}\right)\left(x_{hi}^{(s)} - \bar{x}_h^{(s)}\right)'\right)^2$$

to estimate $\mathrm{Cov}\left(\mathbf{x}^{(s)} \,|\, y^* = h\right)$. BIC type criterions have been suggested for the estimation of the subspace dimension k for SAVE, see for example Zhu et al. (2006) and Zhu and Zhu (2007). See also Zhu et al. (2007); Shao et al. (2007). The bootstrap based criterion by Liquet and Saracco (2012) can be used to estimate k as well as H also in SAVE.

In addition to being able to handle aforementioned symmetric regressions, Cook and Critchley (2000) have shown that SAVE is more comprehensive than SIR, as it is able to capture generally a larger part of the central subspace. However, SAVE is less efficient than SIR, because it needs more observations than SIR to work equally well. Another drawback of SAVE, compared to SIR, is its sensitivity to the choice of H, which has been discussed e.g. in Cook (2000) and Li and Zhu (2007). Cook (2000) states that SAVE does not work if the number of observations per slice, c, is too small, i.e. the amount of slices $H = n/c$ is too large.

Li and Zhu (2007) have shown that $\frac{1}{H}\sum_{h=1}^{H}\left(\mathbf{I}_p - \widehat{\mathrm{Cov}}\left(\mathbf{x} \,|\, y^* = h\right)\right)^2$ is $\sqrt{n}$ has a multivariate normal limiting distribution with a zero mean vector only if the response y is discrete-valued.

For the affine equivariance of SIR and SAVE, see for example Liski et al. (2014). As SIR and SAVE are moment based methods and not robust, outlying observations may affect the results. Robustness issues of SIR have been studied for example in Gather et al. (2001, 2002); Prendergast (2005, 2006, 2007); Prendergast (2007) has also investigated SAVE.

Hybrid of SIR and SAVE Due to drawbacks of both methods, Ye and Weiss (2003) has proposed and Zhu et al. (2007) discussed more in detail about a hybrid version of SIR and SAVE. The idea behind this hybrid is to combine the best parts of both methods to uncover the structures as efficiently as possible with as little sensitivity as possible. The hybrid uses the convex combination

$$\mathbf{H}_{1b} = b \cdot \mathrm{E}\left(\left(\mathbf{I}_p - \mathrm{Cov}\left(\mathbf{x}^{(s)} | y\right)\right)^2\right) + (1-b) \cdot \mathrm{Cov}\left(\mathrm{E}\left(\mathbf{x}^{(s)} | y\right)\right),$$

where $b \in [0, 1]$. With $b = 0$ we get SIR and with $b = 1$ we get SAVE. In order to estimate $\mathbf{\Gamma} = \mathbf{U}\mathbf{\Sigma}^{-1/2}$, we need to find a matrix $\mathbf{U} = \left(\mathbf{u}_1, \ldots, \mathbf{u}_k\right)' \in \mathcal{O}^{k \times p}$ that maximizes

$$\left\|\mathrm{diag}\left(\mathbf{U}\mathbf{H}_{1b}\mathbf{U}'\right)\right\|^2 = \sum_{i=1}^{k} \left(\mathbf{u}_i' \mathbf{H}_{1b} \mathbf{u}_i\right)^2.$$

In practice we can again approximate y by a sliced variable y^*. For the discussion on choosing the value b, see Zhu et al. (2007). They also show that if the predictors of y have both even and odd functions, e.g. $y_i = z_{1i}^2 + z_{2i}^3 + \epsilon_i$, the hybrid version works better than SIR and SAVE alone. This is due to SIR contributing more to find the odd one and SAVE contributing more to find the even one. For the estimation of the subspace dimension k in the hybrid methods of SIR and SAVE, see for example Zhu and Zhu (2007).

Shaker and Prendergast (2011) have also suggested SAVE|SIR, another combination of SIR and SAVE. As SIR is efficient in finding linear relationships, it can be used to find efficiently a partial dimension reduction subspace. Then SAVE would be used to find the remainder of the subspace that SIR was unable to find.

Other methods In SIR, SAVE and their hybrid, the criterion functions are of the form

$$\left\|\mathrm{diag}\left(\mathbf{U}\mathbf{M}\mathbf{U}'\right)\right\|^2,$$

where $\mathbf{M} \in \mathbb{R}^{p \times p}$ is a *kernel matrix*.

Li (1992) has introduced Principal Hessian Directions (PHD) method for supervised dimension reduction. Here $\mathbf{M} = \mathrm{E}\left(\mathrm{H}\left(\mathbf{x}^{(s)}\right)\right)$, where

$$\mathrm{H}\left(\mathbf{x}^{(s)}\right) = \frac{\partial^2}{\partial \mathbf{x}^{(s)} \partial \mathbf{x}^{(s)\prime}} \mathrm{E}\left(y | \mathbf{x}^{(s)}\right)$$

is the so-called Hessian matrix.

Li and Wang (2007) have introduced Directional Regression (DR). In DR we first assume that $\tilde{y}$ and $\tilde{\mathbf{x}}^{(s)}$ are independent copies of y and $\mathbf{x}^{(s)}$, respectively. Write $\mathbf{A}\left(y, \tilde{y}\right) := \mathrm{E}\left(\left(\mathbf{x}^{(s)} - \tilde{\mathbf{x}}^{(s)}\right)\left(\mathbf{x}^{(s)} - \tilde{\mathbf{x}}^{(s)}\right)' | y, \tilde{y}\right)$. Then

$$\mathbf{M} = \mathrm{E}\left(2\mathbf{I}_p - \mathbf{A}\left(y, \tilde{y}\right)\right)^2.$$

PHD and DR are both inverse regression methods just like SIR and SAVE and for them we assume also the linearity and constant covariance assumptions (2.17) and (2.20).

There are also some non-parametric methods, such as Minimum Variance Average Estimation (MAVE) (Xia et al., 2002) and its variants, and semiparametric methods. For a review on these and other methods for supervised dimension reduction, see for example Ma and Zhu (2013).

Bura and Yang (2011) have also proposed asymptotic Chi Squared tests for the rank of an asymptotically normal random matrix, such as an estimate of a kernel matrix $\mathbf{M}$. Such tests can be used to determine the dimension k of the subspace in dimension reduction methods based on the kernel matrices, including the ones described in this section.

Example: Supervised dimension reduction with SIR and SAVE Consider a random vector $\mathbf{z} = \left(z_1, \ldots, z_5\right)'$, with a distribution $N\left(\mathbf{0}, \mathbf{I}_5\right)$, and a random variable $\epsilon \sim N(0, 0.1)$ that is independent of $\mathbf{z}$. Consider then the following two models:

M1: $y = z_1 - z_2 + \epsilon$ and

M2: $y = z_1^2 - z_2^2 + \epsilon$.

We simulate 1000 observations from $\mathbf{z}$ and ϵ and then calculate the values of the response y based on the aforementioned models. We then use a random mixing matrix $\mathbf{\Omega}$ to create the ‘observed’ predictors $\mathbf{x} = \mathbf{\Omega z}$.

Figure 2.5 has the response y of the model M1 with the observed predictors $\mathbf{x}$, while Figure 2.6 has the directions found using the SIR method with $H = 10$. We can see that Figure 2.5 does not give us a clear idea of the relationship between y and $\mathbf{x}$. On the other hand, Figure 2.6 shows clearly one direction with a linear relationship with y, while the other directions seem to be just noise.

Figure 2.7 has the observed values based on the model M2. This figure does not seem to reveal anything interesting. Figure 2.8 then has the directions found using the SAVE method with $H = 5$ and it shows clearly two directions with quadratic relationships between them and y. The other directions seem to be just noise.

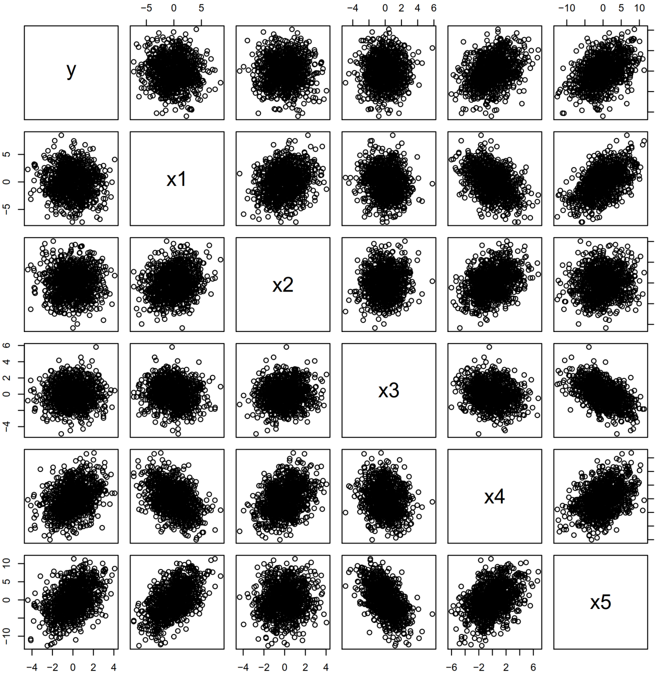


Figure 2.5: Scatterplot of y *and* **x** *based on model* M1*.*

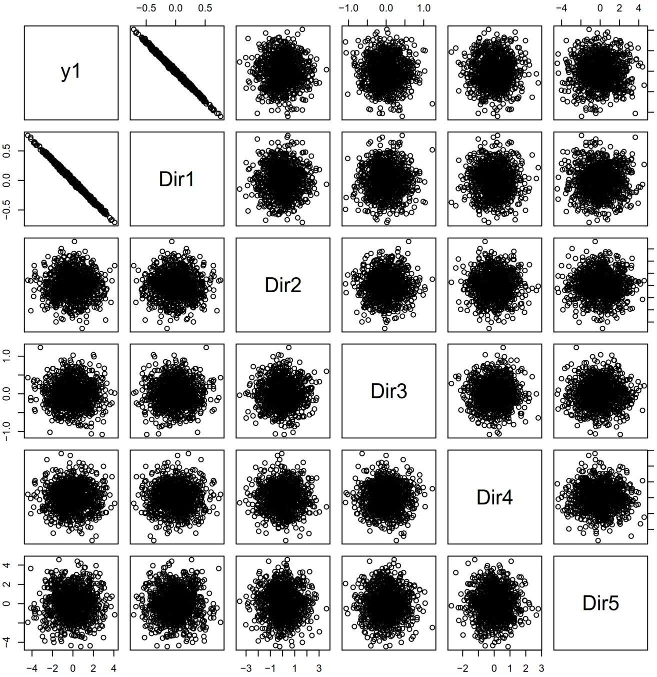


Figure 2.6: Scatterplot of y *and the directions uncovered using the SIR method with* H = 10 *based on model* M1*.*

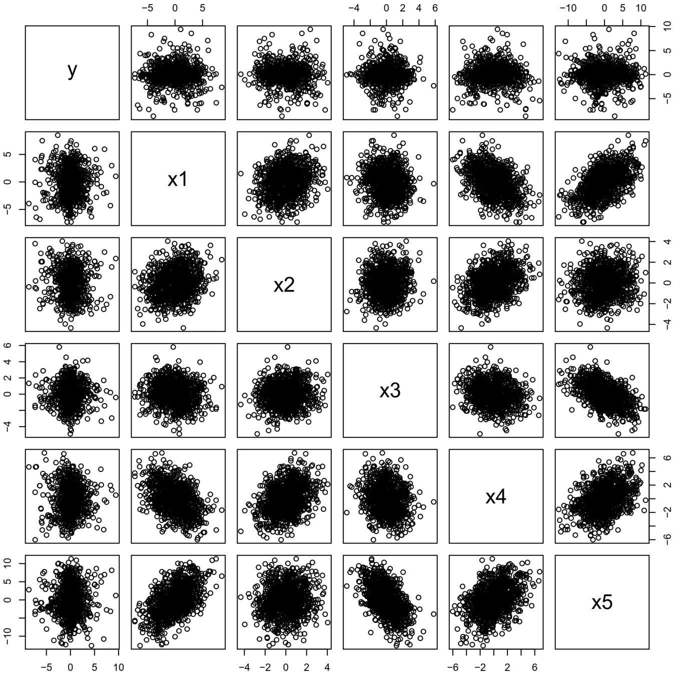


Figure 2.7: Scatterplot of y *and* **x** *based on model* M2.

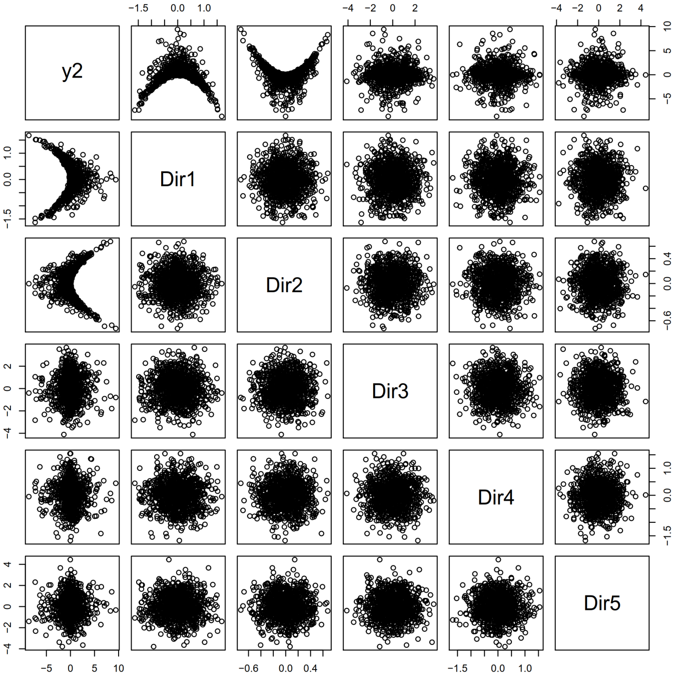


Figure 2.8: Scatterplot of y *and the directions uncovered using the SAVE method with* H = 5 *based on model* M2.

3 Time series

Time series data differ greatly from iid data, as observations at time t may depend on observations on the previous time points $\mathsf{t}-1, \mathsf{t}-2, \ldots$. This presents a challenge for data analysis. Methods for iid observations can often be used in time series context, but they do not utilize the information on temporal dependence and therefore cannot extract all the information available. Sometimes this can be circumvented by introducing lagged values of variables $\mathbf{x}_\mathsf{t}$ as new variables $\mathbf{x}^*_\mathsf{t} = \left(\mathbf{x}'_\mathsf{t}, \mathbf{x}'_{\mathsf{t}-1}, \ldots, \mathbf{x}'_{\mathsf{t}-\mathsf{s}}\right)'$ and then use the original iid methods for the augmented $\mathbf{x}^*_\mathsf{t}$.

Principal Component Analysis has been used as a tool also in time series applications, see for example Stock and Watson (2002). Ku et al. (1995) have also proposed to apply PCA for augmented variables $\mathbf{x}^*_\mathsf{t}$. For an overview of PCA for time series, see Jolliffe (2002). For some recent contributions on methods related to the use of PCA for time series data as well as dynamic PCA and factor models that are designed for dimension reduction for multivariate time series, see for example Matteson and Tsay (2011); Barbarino and Bura (2015); Forni et al. (2015); Peña and Yohai (2016).

In this chapter we first discuss some time series models in Section 3.1. In Section 3.2 we review Second Order Source Separation (SOS) models, where the latent components are assumed to be uncorrelated, as in PCA. In Section 3.3 we generalize Independent Component Analysis (ICA) for time series. Both SOS and ICA models are submodels of the Blind Source Separation (BSS) model. The methods in Sections 3.2 and 3.3 utilize the information on temporal dependence without needing to use augmented sets of variables. In Section 3.4 we discuss supervised dimension reduction for time series.

3.1 Some time series models

In this section we consider various time series models. The simplest univariate time series is a white noise process $\left(\epsilon_\mathsf{t}\right)_{\mathsf{t}\in\mathbb{Z}}$, which is a series of uncorrelated random variables that have a zero mean and a finite variance. The process ϵ_t is called a Gaussian white noise, if ϵ_t is a series of independent random variables with a distribution $N(0,\sigma^2)$, i.e. $\epsilon_t \sim \text{iid } N(0,\sigma^2)$ (Shumway and Stoffer, 2011).

Similarly, a multivariate iid time series $(\boldsymbol{\epsilon}_t)_{t\in\mathbb{Z}}$ is a white noise process if $E(\boldsymbol{\epsilon}_t) = \mathbf{0}$, $\text{Cov}(\boldsymbol{\epsilon}_t) = \boldsymbol{\Sigma}_\epsilon$ is a non-singular matrix and $\text{Cov}(\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_s') = \mathbf{0}$ for $s \neq t$. If the distribution is multivariate normal, then $\boldsymbol{\epsilon}_t$ is Gaussian white noise (Lütkepohl, 2005).

For a univariate process $(x_t)_{t\in\mathbb{Z}}$ the autocovariance is defined as

$$\sigma_{x;s,t} = E\big((x_s - \mu_s)(x_t - \mu_t)\big), \tag{3.1}$$

for all time points s and t, measures a linear dependence between two values x_s and x_t. The values μ_s and μ_t denote the theoretical means of the process at time s and t, respectively. Note that the variance of the process $\sigma^2_{x;t} = \sigma_{x;t,t} = E\big((x_t - \mu_t)^2\big)$. The autocorrelation

$$\rho_{x;s,t} = \frac{\sigma_{x;s,t}}{\sqrt{\sigma^2_{x;s}\sigma^2_{x;t}}}$$

measures how well a value x_t of a time series can be predicted by using only the value x_s.

The cross-covariance between the two time series y_t and x_t can be defined as

$$\sigma_{xy;s,t} = E\big((x_s - \mu_{x;s})(y_t - \mu_{y;t})\big), \tag{3.2}$$

where $\mu_{x;t}$ is the theoretical means of the process x_t at time t and $\mu_{y;t}$ the theoretical mean of the process y_t at time t. The cross-autocorrelation is

$$\rho_{xy;s,t} = \frac{\sigma_{xy;s,t}}{\sqrt{\sigma^2_{x;s}\sigma^2_{y;t}}}.$$

For many time series methods stationarity is an important property. For a strict stationarity of a time series x_t it is assumed that

$$P\left(x_{s_1} \leq a_1, \ldots, x_{s_k} \leq a_k\right) = P\left(x_{s_1+t} \leq a_1, \ldots, x_{s_k+t} \leq a_k\right),$$

for all $t \in \mathbb{Z}$, all $k = 1, 2, \ldots$, and all $a_1, \ldots, a_k$ and $s_1, \ldots, s_k$. The joint strict stationarity of the time series x_t and y_t can be defined similarly. In practice, however, weaker assumptions are generally enough. For a weak (or second-order) stationarity it is assumed that the time series x_t has a finite variance and its

- mean value μ_t does not depend on time t.
- autocovariance (3.1) (and hence the autocorrelation) depends only on the absolute difference of the time points, i.e. $\tau = |s-t|$.

In this thesis we only need weak stationarity and the series is then called briefly stationary. The aforementioned difference $\tau \in \mathbb{Z}_+$ is called here *lag*.

Assume now a stationary time series x_t that has a mean value μ. Then (3.1) simplifies to

$$\sigma_\tau = \mathrm{E}\big((x_t-\mu)(x_{t+\tau}-\mu)\big).$$

Time series x_t and y_t are jointly stationary if they both are stationary and the cross-autocovariance (3.2) does not depend on time, i.e.

$$\sigma_{xy;\tau} = \mathrm{E}\big((x_t-\mu_x)(y_{t+\tau}-\mu_y)\big). \tag{3.3}$$

For an overview on characteristics of time series, see for example Shumway and Stoffer (2011).

Autoregressive Moving Average models

Box and Jenkins (1970) introduce the AutoRegressive Moving Average (ARMA) process for time series. An ARMA(p,q) process is defined as

$$x_t = \sum_{i=1}^{p} \phi_i x_{t-i} + \epsilon_t + \sum_{j=1}^{q} \theta_j \epsilon_{t-j}, \tag{3.4}$$

where the coefficients ϕ_i, $i = 1,\dots,p$ are autoregressive coefficients, the coefficients θ_j, $j = 1,\dots,q$, moving average coefficients and ϵ_t an unobserved Gaussian white noise process with a variance σ^2 (Shumway and Stoffer, 2011). To simplify notation, we assume here wlog that the mean value of the process $\mu = 0$. If $q = 0$, then (3.4) is called an AR(p) process and if $p = 0$, then it is an MA(q) process.

We also assume that the ARMA processes are causal and invertible. An ARMA process is causal if it can be written as a linear MA(∞) process

$$x_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}, \ t = 0, \pm1, \pm2, \dots,$$

where $\sum_{j=0}^{\infty} |\psi_j| < \infty$ and ψ_0 is set to 1, and invertible if it can be written as

$$\epsilon_t = \sum_{j=0}^{\infty} \pi_j x_{t-j}, \ t = 0, \pm 1, \pm 2, \ldots,$$

where $\sum_{j=0}^{\infty} |\pi_j| < \infty$ and π_0 is set to 1. Causality ensures that the process does not depend on its future values and invertibility ensures that the model is uniquely defined.

For an MA(∞) process here $\mathrm{Var}(x_t) = \sigma^2 \sum_{j=0}^{\infty} \psi_j^2$ for all t. Also the autocovariance for a lag $\tau > 0$ is

$$\sigma_\tau = \mathrm{E}\left(x_t x_{t+\tau}\right) = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+\tau}.$$

This process is stationary in the sense that the mean value does not depend on the time point and the autocovariance only depends on the time difference.

ARMA model parameters can be estimated for example using a maximum likelihood method, where the likelihood can be written, using a conditional distribution of x_t given its past values, as $\prod_{t=1}^{T} f\left(x_t | x_{t-1}, \ldots, x_1\right)$. For more on ARMA-processes, including different estimation methods of the model parameters, see for example Shumway and Stoffer (2011).

For a p-variate time series $\left(\mathbf{x}_t\right)_{t \in \mathbb{Z}}$ the VARMA(p, q) (Vector ARMA) process is

$$\mathbf{x}_t = \sum_{i=1}^{p} \mathbf{\Phi}_i \mathbf{x}_{t-i} + \boldsymbol{\epsilon}_t + \sum_{j=1}^{q} \mathbf{\Theta}_j \boldsymbol{\epsilon}_{t-j}, \ t = 0, \pm 1, \pm 2, \ldots,$$

where $\boldsymbol{\epsilon}_t \in \mathbb{R}^p$ is a white noise process with a covariance matrix $\mathbf{\Sigma}$, $\mathbf{\Phi}_i \in \mathbb{R}^{p \times p}$ is a matrix of VAR coefficients and $\mathbf{\Theta}_j \in \mathbb{R}^{p \times p}$ is a matrix of vector MA coefficients. We assume here wlog that $\boldsymbol{\mu} = \mathrm{E}\left(\mathbf{x}_t\right) = \mathbf{0}$.

Similar to the univariate case, VARMA process is causal if it can be written as a vector MA(∞) process

$$\mathbf{x}_t = \sum_{j=0}^{\infty} \mathbf{\Psi}_j \boldsymbol{\epsilon}_{t-j}, t = 0, \pm 1, \pm 2, \ldots,$$

where $\sum_{j=0}^{\infty} \sum_{k=1}^{p} \sum_{l=1}^{p} \left| \left(\mathbf{\Psi}_j\right)_{kl} \right| < \infty$ and $\mathbf{\Psi}_j \in \mathbb{R}^{p \times p}$ is a coefficient matrix. Invertibility is also analogous to the univariate case (Shumway and Stoffer, 2011).

For a vector MA(∞) process $\mathrm{Var}\left(\mathbf{x}_t\right) = \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{\Sigma} \boldsymbol{\Psi}_j'$ for all t. Also the autocovariance for a lag $\tau > 0$ is

$$\boldsymbol{\Sigma}_\tau = \mathrm{E}\left(\mathbf{x}_t \mathbf{x}_{t+\tau}\right) = \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{\Sigma} \boldsymbol{\Psi}_{j+\tau}'.$$

This process is again stationary in the sense that the mean value does not change over time and the autocovariances of the process do not depend on the time t, but only on the absolute time difference $\tau = |s - t|$. The maximum likelihood estimation can be used for estimating the VARMA model parameters. For more on the multivariate ARMA models, including the parameter estimation, see for example Lütkepohl (2005).

Assume that we have two zero mean ARMA(1,1) processes. If we estimate their parameters separately, four AR and MA coefficients need to be estimated; for both ARMA processes parameters ϕ_1 and θ_1. If we use a bivariate VARMA(1,1) model, eight VAR and vector MA coefficients need to be estimated: 2×2-parameter matrices $\boldsymbol{\Phi}_1$ and $\boldsymbol{\Theta}_1$.

With p zero mean ARMA(1,1) processes we need to estimate $2p$ AR and MA coefficients, and with a p-variate VARMA(1,1) process we need to estimate $2p^2$ VAR and vector MA coefficients. It is easy to see that when p gets larger, VARMA modelling becomes harder. Hence if the p-variate process can be transformed into p uncorrelated univariate processes, much less parameters would need to be estimated.

Stochastic volatility models

In stochastic volatility processes the variance of the process is a random process itself. For such models the variance, also called volatility, has its own process and own parameters related to it. Such processes are used widely in finance, where often there are periods of low volatility followed by the periods of high volatility. For example, in Lütkepohl (2005) an example of the monthly returns of German Stock Index (DAX) between 1965 and 1995 is shown to have no significant linear autocorrelation. However, some autocorrelations of the squared process is shown to be significant. This indicates that an ARMA process is not enough to capture the autocorrelation structure of the returns, as the autocorrelation is non-linear.

In stochastic volatility process the main interest is usually the volatility process and the actual value of the time series is of less interest. For time series with stochastic volatility, some of the most popular models for such time series are ARCH, GARCH and SV models. Assume that $\mathscr{F}_t = \left(x_s\right)_{s \leq t}$, i.e. it contains all the information on the process x_t until time t.

ARCH and GARCH models Engle (1982) first introduces ARCH(p) (AutoRegressive Conditional Heteroskedasticity) processes, where the individual time series process can be written as

$$x_t = \sigma_t \epsilon_t,$$

where the conditional variance process $\left(\sigma_t^2\right)_{t \in \mathbb{Z}}$, given $\mathscr{F}_{t-1}$, is

$$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i x_{t-i}^2 \tag{3.5}$$

and ϵ_t a process with $\mathrm{E}\left(\epsilon_t\right) = 0$ and $\mathrm{Var}\left(\epsilon_t\right) = 1$. Originally Engle (1982) assumed that $\epsilon_t \sim \text{iid } \mathrm{N}(0,1)$, but also other distributions may be used (Lütkepohl, 2005). Bollerslev (1986) then generalizes this and introduces GARCH(p,q) processes, where (3.5) is replaced by

$$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i x_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2, \tag{3.6}$$

where $\omega > 0$ and $\alpha_i, \beta_j \geq 0$, for all i and j. The current value of the conditional variance $\sigma_t^2 = \mathrm{Var}\left(x_t \,|\, \mathscr{F}_{t-1}\right)$ thus depends on the previous values of the series and the previous values of the conditional variance itself, and ω is the constant part of the conditional variance.

For the second order stationarity of a GARCH(p,q) process it is required that

$$\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j < 1. \tag{3.7}$$

With these requirements the unconditional variance is

$$\sigma^2 = \frac{\omega}{1 - \sum_{i=1}^{p} \alpha_i - \sum_{j=1}^{q} \beta_j} > 0.$$

Let us now define a process $\upsilon_t := x_t^2 - \sigma_{t|t-1}^2$ and then substitute $\sigma_{t|t-1}^2$ in (3.6) by $x_t^2 - \upsilon_t$. We get

$$x_t^2 = \omega + \sum_{i=1}^{p} \left(\alpha_i + \beta_i\right) x_{t-i}^2 + \upsilon_t + \sum_{j=1}^{q} \beta_j \upsilon_j,$$

where wlog we assume that $p \leq q$ and β_j's are set to zero when $j > m$. This is now in the form of an ARMA(p, q) process for x_t^2. Hence calculating the autocorrelations for the squared process can reveal the presence of stochastic volatility, as in the German Stock Index example. (Lütkepohl, 2005)

Also in GARCH(1,1) process, the constraint (3.7) for α_1 and β_1 are sufficient to ensure the positivity of the conditional variance. (see e.g. Teräsvirta, 2009). However, in higher-order models the necessary and sufficient conditions for this are more complicated (see Nelson and Cao, 1992). Some time series methods designed for the stochastic volatility processes also require the existence of higher order moments of the processes. To assess the finiteness of the moments, see e.g. Lindner (2009). For example, finite eight moments exist for a GARCH(1,1) with a Gaussian white noise process ϵ_t if and only if

$$\beta_1^4 + 4\beta_1^3\alpha_1 + 18\beta_1^2\alpha_1^2 + 60\beta_1\alpha_1^3 + 105\alpha_1^4 < 1.$$

For GARCH models there also exists many different versions, see for example Teräsvirta (2009).

Just like with ARMA models, also GARCH model has multivariate extensions. Engle et al. (1986) first generalize ARCH model to a bivariate case. Then Diebold and Nerlove (1989) suggest a multivariate ARCH type model and Bollerslev et al. (1988) propose the first multivariate GARCH type model. Assume a multivariate time series process

$$\mathbf{x}_t = \mathbf{\Sigma}_t^{1/2} \boldsymbol{\epsilon}_t,$$

where $\boldsymbol{\epsilon}_t \in \mathbb{R}^p$ is a white noise process, with a mean $\mathbf{0}$ and a variance $\mathbf{I}_p$, and $\mathbf{\Sigma}_t^{-1/2}$, the symmetric positive definite square root of $\mathbf{\Sigma}_t$, is the conditional covariance matrix of $\mathbf{x}_t$, given $\mathscr{F}_{t-1}$, i.e. the history of the process $\mathbf{x}_t$. For $\boldsymbol{\epsilon}_t$ for example the multivariate normal distribution can be considered.

For a multivariate GARCH (MGARCH) the conditional variance process can be written as

$$\operatorname{vech}\left(\boldsymbol{\Sigma}_t\right) = \boldsymbol{\omega} + \sum_{i=1}^{p} \mathbf{A}_i \operatorname{vech}\left(\mathbf{x}_{t-i}\mathbf{x}'_{t-i}\right) + \sum_{j=1}^{q} \mathbf{B}_j \operatorname{vech}\left(\boldsymbol{\Sigma}_{t-j|t-j-1}\right),$$

where $\boldsymbol{\omega} \in \mathbb{R}^{p(p+1)/2}$ is a base level of the variance process, and $\mathbf{A}_i \in \mathbb{R}^{(p(p+1)/2)\times(p(p+1)/2)}$ and $\mathbf{B}_j \in \mathbb{R}^{(p(p+1)/2)\times(p(p+1)/2)}$ are coefficient matrices.

As a GARCH model can be written as an ARMA model of a squared process, also MGARCH model can be similarly written as a VARMA model of a p-variate squared process. MGARCH model is stationary if for a matrix

$$\mathbf{C} = \sum_{i=1}^{p} \mathbf{A}_i + \sum_{j=1}^{q} \mathbf{B}_j$$

it is true that $\det\left(\mathbf{I}_p - \mathbf{C}z\right) \neq 0$ for all $z \in \mathbb{C}$ and $|z| \leq 1$. (Lütkepohl, 2005)

The number of the parameters to be estimated in multivariate GARCH model is very large. Bollerslev et al. (1988) has discussed MGARCH models, where the matrices $\mathbf{A}_i$ and $\mathbf{B}_j$ are diagonal, i.e. the covariances only depend on their own past. A simpler way to reduce the number of parameters to be estimated would be to transform the p-variate processes to a different coordinate system, where the processes are independent and therefore the parameters of the different processes could estimated separately.

Parameter estimation in GARCH models, in both univariate and multivariate case, can be done for example via maximum likelihood or quasi maximum likelihood estimation. For more on GARCH models and different multivariate GARCH models and their parameter estimation, see for example Lütkepohl (2005); Matteson and Ruppert (2011).

SV models Taylor (1982) has introduced an SV (Stochastic Volatility) model, where the process is

$$x_t = e^{h_t/2}\epsilon_t, \tag{3.8}$$

where

$$h_t = \mu + \phi\left(h_{t-1} - \mu\right) + \sigma\eta_t.$$

The form (3.8) can also be written in a 'linearized' form as

$$\log\left(x_t^2\right) = h_t + \log\left(\epsilon_t^2\right).$$

The volatility process h_t has a stationary distribution with an initial state $h_0 \,|\, \left(\mu, \phi, \sigma\right) \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{1-\phi^2}\right)$. Requirement for stationarity of the process x_t is $|\phi| < 1$. The processes ϵ_t and η_t are independent $N(0,1)$ processes. For the finiteness of the moments, see e.g. Andersen (1994).

Melino and Turnbull (1990) have used the generalized method of moments and Harvey et al. (1994) the quasi-maximum likelihood method for the estimation of the SV parameters. Using the maximum likelihood method in SV parameter estimation is not straightforward, as the conditional likelihood cannot be expressed analytically (see e.g. Bauwens et al., 2012a); numerical methods are needed. For a review on different estimation methods, see for example Broto and Ruiz (2004). For a recent contribution on the parameter estimation of SV models, see Kastner and Frühwirth-Schnatter (2014).

SV model has also been generalized to a multivariate case in Harvey et al. (1994), where they use the quasi-maximum likelihood method for the parameter estimation. For an overview of different stochastic volatility models, see for example Shephard and Andersen (2009); Bauwens et al. (2012b).

In general the SV models have not been used in applications as much as GARCH models, likely due to a large scale of different estimation methods and the lack of appropriate software packages (Bos, 2012). Kastner and Frühwirth-Schnatter (2014) has recently proposed an efficient estimation procedure and Kastner (2016) a software package.

Examples of ARMA, GARCH and SV models are in Figure 3.1. The models are

- ARMA(1,1) process with $\phi_1 = 0.8$ and $\theta_1 = -0.2$.
- GARCH(1,1) process with $\omega = 0.1$, $\alpha_1 = 0.2$ and $\beta_1 = 0.7$.
- SV process with $\mu = -1.5$, $\phi = 0.95$ and $\sigma = 0.4$.

For GARCH and SV processes we can see that there are periods of high and periods of low volatility.

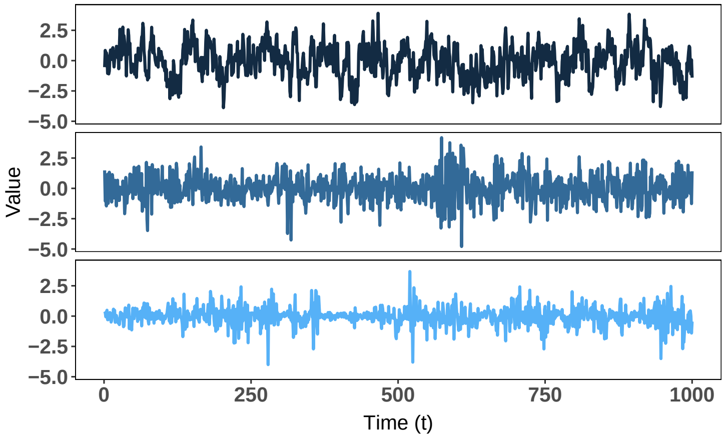


Figure 3.1: Simulated ARMA (top panel), GARCH (middle panel) and SV (bottom panel) processes.

3.2 Second Order Source Separation

From this section onwards we assume that $\mathbf{x} = (\mathrm{x}_t)_{t\in\mathbb{Z}}$ is a p-variate time series. The term time series here means both the process that creates them and the observed time series, as from the context it is clear which one is meant. For a matrix $\mathbf{A} \in \mathbb{R}^{p\times p}$ and a vector $\mathbf{b} \in \mathbb{R}^p$, $\mathbf{Ax} + \mathbf{b}$ is a time series $(\mathbf{Ax}_t + \mathbf{b})_{t\in\mathbb{Z}}$.

Definition 3.1 *Time series* x *and* y *are said to be uncorrelated if* x_t *and* y_s *are uncorrelated for all* t *and* s*.*

Definition 3.2 *A Second Order Source Separation (SOS) model, a submodel of the BSS model, is written as*

$$\mathbf{x}_t = \boldsymbol{\mu} + \boldsymbol{\Omega}\mathbf{z}_t, \ t = 0, \pm 1, \pm 2, \ldots, \tag{3.9}$$

where $\boldsymbol{\mu} \in \mathbb{R}^p$ *is a location vector and* $\boldsymbol{\Omega} \in \mathbb{R}^{p\times p}$ *a mixing matrix, sometimes called also a signal separation matrix. For the* p*-variate time series* $\mathbf{z} = (\mathbf{z}_t)_{t\in\mathbb{Z}}$ *we assume that*

$$\mathrm{E}(\mathbf{z}_t) = \mathbf{0}, \ \mathrm{Cov}(\mathbf{z}_t) = \mathbf{I}_p \tag{3.10}$$

and the p *processes in* $\mathbf{z}$ *are assumed to be jointly stationary and uncorrelated.*

The general procedure to find an unmixing (or signal separation) matrix $\boldsymbol{\Gamma}$ here is the same than in i.i.d. case. Now the goal is to uncover the latent uncorrelated time series $\mathbf{z}$ up to sign changes and the order of the components. For time series $\mathbf{z}$ and a standardized time series $\mathbf{x}^{(s)}$,

$$\mathbf{z} = \mathbf{U}_0 \mathbf{x}^{(s)},$$

for an orthogonal matrix $\mathbf{U}_0$. This is analogous to (2.7) in iid data. This means that after calculating the standardized variables $\mathbf{x}^{(s)} = \boldsymbol{\Sigma}^{-1/2}\left(\mathbf{x} - \boldsymbol{\mu}\right)$, a BSS problem can be solved by only finding an orthogonal matrix $\mathbf{U}$.

In the SOS model the latent time series $\mathbf{z}$ are only assumed to be uncorrelated, so the assumption on independence does not need to be fulfilled. This means that

$$\mathrm{E}\left(\mathbf{z}_t \mathbf{z}'_{t+\tau}\right) = \mathrm{E}\left(\mathbf{z}_{t+\tau} \mathbf{z}'_t\right) = \mathbf{D}_\tau, \tag{3.11}$$

for $\tau > 0$, where $\mathbf{D}_\tau$ is a diagonal $p \times p$ matrix. For a chosen $\tau > 0$ the diagonal values $\mathbf{D}_\tau$ are assumed to be distinct.

Tong et al. (1990) has introduced AMUSE (Algorithm for Multiple Unknown Signals Extraction) method. AMUSE uses the cross-autocovariance matrix, a matrix version of (3.3),

$$\boldsymbol{\Sigma}_\tau\left(\mathbf{x}\right) = \mathrm{E}\left(\left(\mathbf{x}_t - \boldsymbol{\mu}\right)\left(\mathbf{x}_{t+\tau} - \boldsymbol{\mu}\right)'\right), \tag{3.12}$$

for a lag $\tau > 0$, and maximizes, for an orthogonal matrix $\mathbf{U}$, the *criterion function*

$$\left\|\mathrm{diag}\left(\mathbf{U}\boldsymbol{\Sigma}_\tau\left(\mathbf{x}^{(s)}\right)\mathbf{U}'\right)\right\|^2 = \sum_{i=1}^{p}\left(\mathbf{u}'_i \boldsymbol{\Sigma}_\tau\left(\mathbf{x}^{(s)}\right)\mathbf{u}_i\right)^2, \tag{3.13}$$

for a selected lag τ.

Similar to FOBI, also for AMUSE the eigenvectors of $\boldsymbol{\Sigma}_\tau\left(\mathbf{x}^{(s)}\right)$ are the rows of an orthogonal matrix $\mathbf{U}$ and $\boldsymbol{\Gamma} = \mathbf{U}\boldsymbol{\Sigma}^{-1/2}$ is uniquely defined only if the eigenvalues of $\boldsymbol{\Sigma}_\tau\left(\mathbf{x}^{(s)}\right)$ are distinct. The affine equivariance of $\boldsymbol{\Gamma}$ has been shown in Miettinen et al. (2012).

Now $\boldsymbol{\Gamma} \in \mathbb{R}^{p \times p}$ satisfies

$$\boldsymbol{\Gamma}\boldsymbol{\Sigma}\boldsymbol{\Gamma}' = \mathbf{I}_p \text{ and } \boldsymbol{\Gamma}\boldsymbol{\Sigma}_\tau(\mathbf{x})\boldsymbol{\Gamma}' = \boldsymbol{\Lambda}_\tau,$$

where $\boldsymbol{\Lambda}_\tau$ is a diagonal matrix in which the diagonal elements are in decreasing order. The matrices $\boldsymbol{\Lambda}_\tau$ and $\mathbf{D}_\tau$ are the same up to the order of the diagonal elements. Note that as $\boldsymbol{\Gamma}\boldsymbol{\Sigma}\boldsymbol{\Gamma}' = \mathbf{I}_p$, $\boldsymbol{\Gamma} = \mathbf{U}\boldsymbol{\Sigma}^{-1/2}$

for an orthogonal matrix $\mathbf{U}$. Now $\mathbf{\Sigma}_\tau\left(\mathbf{x}^{(s)}\right) = \mathbf{\Sigma}^{-1/2}\mathbf{\Sigma}_\tau(\mathbf{x})\mathbf{\Sigma}^{-1/2}$ is the autocorrelation matrix and thus the criterion function can be written in the form of (3.13) (Miettinen et al., 2012).

Let $\mathbf{x}_1, \dots, \mathbf{x}_T$ be the observed time series. An estimate for $\mathbf{\Sigma}_\tau(\mathbf{x})$ can be achieved using a symmetrized version of the sample autocovariance matrices

$$\hat{\mathbf{\Sigma}}_\tau = \frac{1}{T-\tau}\sum_{t=1}^{T-\tau} \mathbf{x}_t \mathbf{x}_{t-\tau}', \quad \tau = 0, 1, 2, \dots.$$

As the matrices (3.11) are assumed to be symmetric under SOS model, it is natural to symmetrize the sample autocovariaces. The symmetrized version of this matrix

$$\hat{\mathbf{\Sigma}}_\tau{}^S = \frac{1}{2}\left(\hat{\mathbf{\Sigma}}_\tau + \hat{\mathbf{\Sigma}}_\tau{}'\right).$$

Thus the estimate $\hat{\mathbf{\Gamma}}$ is a $p \times p$ matrix that satisfies

$$\hat{\mathbf{\Gamma}}\hat{\mathbf{\Sigma}}\hat{\mathbf{\Gamma}}' = \mathbf{I}_p \text{ and}$$
$$\hat{\mathbf{\Gamma}}\hat{\mathbf{\Sigma}}_\tau^S\hat{\mathbf{\Gamma}}' = \hat{\mathbf{\Lambda}}_\tau.$$

In order to find the joint limiting distribution, we consider the model (3.9) with $\boldsymbol{\mu} = \mathbf{0}$ (wlog). Firstly we assume that $\mathbf{z}$ is a multivariate MA(∞) process that fulfills the assumptions (3.10) and (3.11). We also assume that the components of $\mathbf{z}$ have finite fourth moments and they are exchangeable and marginally symmetric, i.e.

$$\mathbf{JKz}_t \sim \mathbf{z}_t,$$

for all sign change matrices $\mathbf{J}$ and permutation matrices $\mathbf{K}$.

As $\hat{\mathbf{\Gamma}}$ is affine equivariant, we can concentrate on the case where $\mathbf{\Gamma} = \mathbf{I}_p$. The limiting distribution for $\sqrt{T}\mathrm{vec}\left(\hat{\mathbf{\Gamma}} - \mathbf{I}_p\right)$ is a p^2-variate normal distribution with a zero mean vector and the covariance matrix as in Miettinen et al. (2012).

The drawback with AMUSE is that the choice τ is crucial, as only one lag is used. Belouchrani et al. (1997) has generalized this to use of a set of lags. In symmetric SOBI (Second Order Blind Identification) the covariance matrices of all lags are jointly diagonalized. The *criterion function* to be maximized for an orthogonal matrix $\mathbf{U}$ is

$$\sum_{\tau \in T}\left\|\mathrm{diag}\left(\mathbf{U}\mathbf{\Sigma}_\tau\left(\mathbf{x}^{(s)}\right)\mathbf{U}'\right)\right\|^2 = \sum_{\tau \in T}\sum_{i=1}^{p}\left(\mathbf{u}_i'\mathbf{\Sigma}_\tau\left(\mathbf{x}^{(s)}\right)\mathbf{u}_i\right)^2, \qquad (3.14)$$

where T is a set on lags in $\mathbb{Z}_+$. Then $\mathbf{\Gamma} = \mathbf{U}\mathbf{\Sigma}^{-1/2}$. In deflation-based SOBI (Miettinen et al., 2014b), on the other hand, the uncorrelated components are found one by one.

Assume that we have a multivariate MA(∞) process that fulfills all the assumptions mentioned in the context of AMUSE, except for (3.11) that now is assumed for all K lags $\tau_1, \dots, \tau_K$. In addition, we assume that the diagonal elements of $\sum_{i=1}^{K} \mathbf{\Lambda}_{\tau_i}^2$ are distinct and in decreasing order. Due to affine equivariant we use $\mathbf{\Gamma} = \mathbf{I}_p$. Then

$$\sqrt{T} \mathrm{vec}\left(\hat{\mathbf{\Gamma}} - \mathbf{I}_p\right)$$

is a p^2-variate normal with a zero mean vector. For more details, see Miettinen et al. (2014b) for the deflation-based SOBI and Miettinen et al. (2016) for the symmetric SOBI.

Applications of SOBI SOBI method has been widely used for example with EEG, MEG and fMRI data, see e.g. Tang et al. (2005); Tang (2010) and the references therein. For example, Joyce et al. (2004) have used SOBI for EEG data in the automatic removal of eye movement and blink artifacts.

SOBI has also been used in separating the vibrations caused by the underground traffic from other vibration sources (Popescu and Manolescu, 2007), in the operational modal analysis of civil structures (Rainieri, 2014) and in the forecasting of wind speed (Firat et al., 2010), among others.

Some other versions of SOBI SOBI is a popular algorithm and there are several versions and extensions available. Taskinen et al. (2016) have proposed a method that allows the user to use several lag combinations and chooses the combination that leads to the lowest sum of the limiting variances of the off-diagonal elements of $\sqrt{T} \mathrm{vec}\left(\hat{\mathbf{\Gamma}}\mathbf{\Omega} - \mathbf{I}_p\right)$. Miettinen (2015) has suggested a method that uses $\sum_{\tau \in T} \sum_{i=1}^{p} \left| \mathbf{u}_i' \mathbf{\Sigma}_\tau \left(\mathbf{x}^{(s)}\right) \mathbf{u}_i \right|^a$, where $a \in (1, \infty)$, as a criterion function. With $a = 2$ the regular symmetric SOBI is obtained.

For the SOBI method it should be noted that the sample mean vector, the sample covariance matrix and the sample autocovariance matrices are highly non-robust, i.e. they are sensitive to outliers. In order to make SOBI robust, these population quantities should be replaced by their robust counterparts. Theis et al. (2010) have proposed a robustified version of SOBI, where the sample mean is replaced by spatial median (see for example Haldane

(1948)) and the sample covariance matrix are replaced by the spatial sign covariance matrix (see Visuri et al. (2000) and the references therein). Similarly the autocovariance matrices are replaced by the spatial sign autocovariance matrices. Ilmonen et al. (2015) have proposed an affine equivariant robust version of SOBI, where the sample mean vector and the sample covariance are replaced instead by the Hettmansperger-Randles estimates of location and scatter (Hettmansperger and Randles, 2002), which are robust and affine equivariant.

For other recent extensions of SOBI, see for example Theis et al. (2004), Lietzén et al. (2017) and Virta and Nordhausen (2017a).

3.3 Independent Component Analysis

The ICA methods in Section 2.5 can be used for time series, but as they ignore the temporal dependence, they may not utilize all the information available in data. In addition to the EEG, MEG and fMRI applications mentioned in Section 2.5, these ICA methods have been popular also in financial applications. ICA methods not designed for time series have been used for financial time series for example in Back and Weigend (1997); Chen et al. (2007); Broda and Paolella (2009); Lu et al. (2009); Kumiega et al. (2011); García-Ferrer et al. (2011, 2012). Next we present some methods that are designed for time series with stochastic volatility, a phenomenon that is common for example in financial data.

We know that SOBI works very well with the second order stationary time series, such as ARMA models, but what happens when there are individual time series with stochastic volatility, i.e. time series where the variance is changing over time?

Assume that the components of $\mathbf{z}$ are all ARCH and GARCH processes. As $\mathbf{z}$ is a standardized process, the cross-autocovariance matrices (3.12) are

$$\boldsymbol{\Sigma}\left(\mathbf{z}_t\right) = \mathrm{E}\left(\mathbf{z}_t \mathbf{z}'_{t+\tau}\right) = \mathrm{E}\left(\boldsymbol{\sigma}_t \boldsymbol{\epsilon}_t \boldsymbol{\epsilon}'_{t+\tau} \boldsymbol{\sigma}'_{t+\tau}\right) = \mathbf{0}_p$$

for every lag $\tau > 0$. As $\boldsymbol{\Sigma}\left(\mathbf{z}_t\right) = \mathbf{0}_p$ for all lags τ, the assumption (3.11) for SOBI (and, of course, for AMUSE) is violated.

The same is clearly true for the SV processes. To overcome this issue we can use for example fourth cross-moment and fourth cumulant matrices instead of the cross-autocovariance matrices. First we define independence for time series

Definition 3.3 *Time series* x *and* y *are said to be independent, written as* $\mathrm{x} \perp\!\!\!\perp \mathrm{y}$, *if* $\left(\mathrm{x}_{\mathrm{t}_1},\dots,\mathrm{x}_{\mathrm{t}_\mathrm{n}}\right)$ *and* $\left(\mathrm{y}_{\mathrm{s}_1},\dots,\mathrm{y}_{\mathrm{s}_\mathrm{m}}\right)$ *are independent for all* $\mathrm{t}_1,\dots,\mathrm{t}_\mathrm{n}$ *and* $\mathrm{s}_1,\dots,\mathrm{s}_\mathrm{m}$.

Independent Component (IC) model for time series is defined as follows.

Definition 3.4 *In IC model for time series we assume that*

$$\mathbf{x}_\mathrm{t} = \boldsymbol{\mu} + \boldsymbol{\Omega}\mathbf{z}_\mathrm{t}, \ \mathrm{t} = 0, \pm 1, \pm 2, \dots,$$

where $\boldsymbol{\mu} \in \mathbb{R}^\mathrm{p}$ *is a location vector and* $\boldsymbol{\Omega} \in \mathbb{R}^{\mathrm{p}\times\mathrm{p}}$ *a mixing matrix. For the* p*-variate time series* $\mathbf{z} = \left(\mathbf{z}_\mathrm{t}\right)_{\mathrm{t}\in\mathbb{Z}}$ *we assume that*

$$\mathrm{E}\left(\mathbf{z}_\mathrm{t}\right) = \mathbf{0}, \ \mathrm{Cov}\left(\mathbf{z}_\mathrm{t}\right) = \mathbf{I}_\mathrm{p}$$

and the p *processes in* $\mathbf{z}$ *are assumed to be jointly stationary and independent.*

Similar to (2.8) and (2.9), consider the fourth cross-moment matrices

$$\mathbf{B}_\tau^{\mathrm{jk}}(\mathbf{x}) = \mathrm{E}\left(\mathbf{x}_{\mathrm{t}+\tau}\mathbf{x}_\mathrm{t}'\mathbf{E}^{\mathrm{jk}}\mathbf{x}_\mathrm{t}\mathbf{x}_{\mathrm{t}+\tau}'\right) \tag{3.15}$$

and the cross-cumulant matrices

$$\mathbf{C}_\tau^{\mathrm{jk}}(\mathbf{x}) = \mathbf{B}_\tau^{\mathrm{jk}}(\mathbf{x}) - \boldsymbol{\Sigma}_\tau(\mathbf{x})\left(\mathbf{E}^{\mathrm{jk}} + \mathbf{E}^{\mathrm{kj}}\right)\boldsymbol{\Sigma}_\tau(\mathbf{x})' - \mathrm{trace}\left(\mathbf{E}_{\mathrm{jk}}\right)\mathbf{I}_\mathrm{p}, \tag{3.16}$$

for all $\mathrm{j}, \mathrm{k} = 1, \dots, \mathrm{p}$. For the latent process $\mathbf{z}$,

$$\mathbf{B}_\tau^{\mathrm{jk}}(\mathbf{z}) = \mathrm{E}\left(\mathbf{x}_{\mathrm{t}+\tau}\mathbf{x}_\mathrm{t}'\mathbf{E}^{\mathrm{jk}}\mathbf{x}_\mathrm{t}\mathbf{x}_{\mathrm{t}+\tau}'\right).$$

and the cross-cumulant matrices

$$\mathbf{C}_\tau^{\mathrm{jk}}(\mathbf{z}) = \mathbf{B}_\tau^{\mathrm{jk}}(\mathbf{x}) - \boldsymbol{\Sigma}_\tau(\mathbf{x})\left(\mathbf{E}^{\mathrm{jk}} + \mathbf{E}^{\mathrm{kj}}\right)\boldsymbol{\Sigma}_\tau(\mathbf{x})' - \mathrm{trace}\left(\mathbf{E}_{\mathrm{jk}}\right)\mathbf{I}_\mathrm{p}.$$

Generalized FOBI (gFOBI) Matilainen et al. (2015) have proposed gFOBI, a generalized version of FOBI, where we use the fourth cross-moment matrices. From (3.15) it follows that

$$\mathbf{B}_\tau(\mathbf{x}) = \sum_{\mathrm{j}=1}^{\mathrm{p}} \mathbf{B}_\tau^{\mathrm{jj}}(\mathbf{x}) = \mathrm{E}\left(\mathbf{x}_{\mathrm{t}+\tau}\mathbf{x}_\mathrm{t}'\mathbf{x}_\mathrm{t}\mathbf{x}_{\mathrm{t}+\tau}'\right).$$

Then the *criterion function* to be maximized for an orthogonal matrix $\mathbf{U} = \left(\mathbf{u}_1', \ldots, \mathbf{u}_p'\right)$ is

$$\sum_{\tau \in \mathsf{T}} \left\| \mathrm{diag}\left(\mathbf{U}\mathbf{B}_\tau\left(\mathbf{x}^{(s)}\right)\mathbf{U}'\right)\right\|^2 = \sum_{\tau \in \mathsf{T}} \sum_{i=1}^{p} \left(\mathbf{u}_i'\mathbf{B}_\tau\left(\mathbf{x}^{(s)}\right)\mathbf{u}_i\right)^2, \qquad (3.17)$$

where T is a set of lags τ in $\mathbb{Z}_{0,+}$. Thus the unmixing matrix is $\boldsymbol{\Gamma} = \mathbf{U}\boldsymbol{\Sigma}^{-1/2}$. The solution for gFOBI is unique (up to the order and signs of the rows) if, for all $j \neq k$, $j, k = 1, \ldots, p$, there exists a lag $\tau > 0$, such that the j:th and the k:th diagonal values of $\mathbf{B}_\tau(\mathbf{z}) = \sum_{j=1}^{p} \left(\mathrm{E}\left((z_j)_t^2\right)\mathrm{E}\left((z_j)_{t+\tau}^2\right) + p - 1\right)\mathrm{E}^{jj}$ are different.

Generalized JADE (gJADE) Matilainen et al. (2015) have also proposed gJADE, a generalized version of JADE. The gJADE method uses the fourth cross-cumulant matrices (3.16). The *criterion function* to be maximized for an orthogonal matrix $\mathbf{U} = \left(\mathbf{u}_1', \ldots, \mathbf{u}_p'\right)$ is

$$\begin{aligned}
&\sum_{\tau \in \mathsf{T}} \sum_{j=1}^{p} \sum_{k=1}^{p} \left\| \mathrm{diag}\left(\mathbf{U}\mathbf{C}_\tau^{jk}\left(\mathbf{x}^{(s)}\right)\mathbf{U}'\right)\right\|^2 \\
&= \sum_{\tau \in \mathsf{T}} \sum_{i=1}^{p} \sum_{j=1}^{p} \sum_{k=1}^{p} \left(\mathbf{u}_i'\mathbf{C}_\tau^{jk}\left(\mathbf{x}^{(s)}\right)\mathbf{u}_i\right)^2, \qquad (3.18)
\end{aligned}$$

where T is a set of lags τ in $\mathbb{Z}_{0,+}$. Now an unmixing matrix for gJADE is $\boldsymbol{\Gamma} = \mathbf{U}\boldsymbol{\Sigma}^{-1/2}$. The gJADE solution is unique (up to the order and signs of the rows) if, for at least $p - 1$ components, there is a lag $\tau \in \mathsf{T}$ such that $\mathbf{C}_\tau^{jj}(\mathbf{z}) \neq \mathbf{0}$

For both methods also more general lag combinations are possible. Write $\boldsymbol{\tau}_{1:4} := \left(\tau_1, \tau_2, \tau_3, \tau_4\right)$. Then

$$\mathbf{B}_{\boldsymbol{\tau}_{1:4}}(\mathbf{x}) = \mathrm{E}\left(\mathbf{x}_{t+\tau_1}\mathbf{x}_{t+\tau_2}'\mathbf{x}_{t+\tau_3}\mathbf{x}_{t+\tau_4}'\right)$$

and

$$\begin{aligned}
\mathbf{C}_{\boldsymbol{\tau}_{1:4}}^{jk}(\mathbf{x}) =& \mathbf{B}_{\boldsymbol{\tau}_{1:4}}^{jk}(\mathbf{x}) - \boldsymbol{\Sigma}_{12}(\mathbf{x})\mathbf{E}^{jk}\boldsymbol{\Sigma}_{43}(\mathbf{x})' \\
& - \boldsymbol{\Sigma}_{13}(\mathbf{x})\mathbf{E}^{kj}\boldsymbol{\Sigma}_{42}(\mathbf{x})' - \boldsymbol{\Sigma}_{14}(\mathbf{x})\boldsymbol{\Sigma}_{23}(\mathbf{x})',
\end{aligned}$$

where $\boldsymbol{\Sigma}_{jk}(\mathbf{x}) := \boldsymbol{\Sigma}_{\tau_j - \tau_k}(\mathbf{x})$. Then $\mathbf{B}_{\boldsymbol{\tau}_{1:4}}\left(\mathbf{x}^{(s)}\right)$ would replace $\mathbf{B}_\tau\left(\mathbf{x}^{(s)}\right)$ in (3.17) and $\mathbf{C}_{\boldsymbol{\tau}_{1:4}}^{jk}\left(\mathbf{x}^{(s)}\right)$ would replace $\mathbf{C}_\tau^{jk}\left(\mathbf{x}^{(s)}\right)$ in (3.18).

The gJADE with more general lag combinations has also been proposed in the PhD thesis of González Prieto (2011) under the name FOTBI (Fourth Order Temporal Blind Identification); see also García-Ferrer et al. (2011).

According to the limited simulations in Matilainen et al. (2015), the more general combinations in gFOBI and gJADE do not seem to produce better separation of the components. These combinations are also computationally more intensive than the basic combinations in (3.17) and (3.18).

For both gFOBI and gJADE the joint diagonalization methods are needed to solve the optimization problem. The affine equivariance properties of the unmixing matrix functionals of gFOBI and gJADE have been proved in Matilainen et al. (2015).

Remark 3.1 *If we choose* $\tau = 0$ *in* (3.17) *and* (3.18)*, we have the classic FOBI and JADE methods designed for iid observations.*

Use of different nonlinearity functions There are also several methods for time series that use nonlinearity functions. The criterion functions mentioned in this part search for an orthogonal matrix $\mathbf{U} = \left(\mathbf{u}_1, \ldots, \mathbf{u}_p\right)'$, which is then used to calculate $\mathbf{\Sigma}$, as before.

Hyvärinen (2001) has proposed the *criterion function*

$$\sum_{i=1}^{p} \left| E\left(\left(\mathbf{u}_i'\mathbf{x}_{t+\tau}\right)^2 \left(\mathbf{u}_i'\mathbf{x}_t\right)^2\right) - E\left(\left(\mathbf{u}_i'\mathbf{x}_t\right)^2\right) E\left(\left(\mathbf{u}_i'\mathbf{x}_{t+\tau}\right)^2\right) \right| \tag{3.19}$$

for a lag $\tau > 0$.

Shi et al. (2009) introduce FixNA (Fixed-point algorithm for maximizing the Nonlinear Autocorrelation), where the *criterion function* is

$$\sum_{\tau \in T} \sum_{i=1}^{p} E\left(G\left(\mathbf{u}_i'\mathbf{x}_{t+\tau}\right) G\left(\mathbf{u}_i'\mathbf{x}_t\right)\right).$$

The choices mentioned for $G(y)$ are y^2 and $\log(\cosh(y))$. The sources can be estimated one by one (deflation-based) or all at once (symmetric; joint diagonalization). Shi et al. (2009) also discuss some theoretical properties of the method. Matilainen et al. (2017d) generalize (3.19) to FixNA2 method with the *criterion function*

$$\sum_{\tau \in T} \sum_{i=1}^{p} \left| E\left(G\left(\mathbf{u}_i'\mathbf{x}_{t+\tau}\right) G\left(\mathbf{u}_i'\mathbf{x}_t\right)\right) - E\left(G\left(\mathbf{u}_i'\mathbf{x}_t\right)\right) E\left(G\left(\mathbf{u}_i'\mathbf{x}_{t+\tau}\right)\right) \right|, \tag{3.20}$$

where the function $G(y)$ is as in FixNA method. Both FixNA and FixNA2 are also implemented in the R package tsBSS (See Section 4).

Matilainen et al. (2017d) also propose an alternative version to (3.20) called vSOBI (variant of SOBI), where the *criterion function* is of the form

$$\sum_{\tau \in T} \sum_{i=1}^{p} \left(E\left(G\left(\mathbf{u}_i' \mathbf{x}_{t+\tau} \right) G\left(\mathbf{u}_i' \mathbf{x}_t \right) \right) - E\left(G\left(\mathbf{u}_i' \mathbf{x}_t \right) \right) E\left(G\left(\mathbf{u}_i' \mathbf{x}_{t+\tau} \right) \right) \right)^2 .$$

Remark 3.2 *The name for vSOBI comes from the SOBI method, for which the criterion function* (3.14) *can be written as*

$$\sum_{\tau \in T} \sum_{i=1}^{p} \left(\mathbf{u}_i' E\left(\mathbf{x}_t^{st} \mathbf{x}_{t+\tau}^{st}{}' \right) \mathbf{u}_i \right)^2 = \sum_{\tau \in T} \sum_{i=1}^{p} \left(E\left(\left(\mathbf{u}_i' \mathbf{x}_t^{st} \right) \left(\mathbf{u}_i' \mathbf{x}_{t+\tau}^{st} \right)' \right) \right)^2 .$$

Remark 3.3 *As the methods described in this section and in Section 3.2 transform the* p*-variate time series into* p *uncorrelated (or mutually independent) time series, the univariate time series models can be used for the estimation instead the multivariate time series models.*

The main benefit here is that less parameters need to be estimated even with a large p*. Also the computational burden is lower, as the large residual covariance matrices do not need to be calculated.*

Illustration To illustrate how for example vSOBI works, consider a time series $\mathbf{z} = \left(z_1, z_2, z_3 \right)'$, where the components are GARCH(1,1) processes with $\left(\omega, \alpha_1, \beta_1 \right)$ parameter vectors $(1, 0.2, 0.7)$, $(1, 0.1, 0.8)$ and $(1, 0.05, 0.9)$. The components are also standardized to meet the requirements of the ICA methods. The components are then mixed using a random full-rank mixing matrix $\boldsymbol{\Omega}$. The vSOBI method is then used to uncover $\mathbf{z}$, up to the signs and order of the components, with lags $\tau = 1, \ldots, 10$.

We simulate 5000 observations based on $\mathbf{z}$. The first 1000 values of the latent $\mathbf{z}$ are in Figure 3.2.

Figure 3.3 has the mixed components and we can see that there are two very similar time series and one that is clearly in a different scale. The results using the vSOBI method can be seen in Figure 3.4. The uncovered time series are coloured to match the latent time series in order to see clearly which of the mutually independent time series corresponds to which of the latent time series. It is rather easy

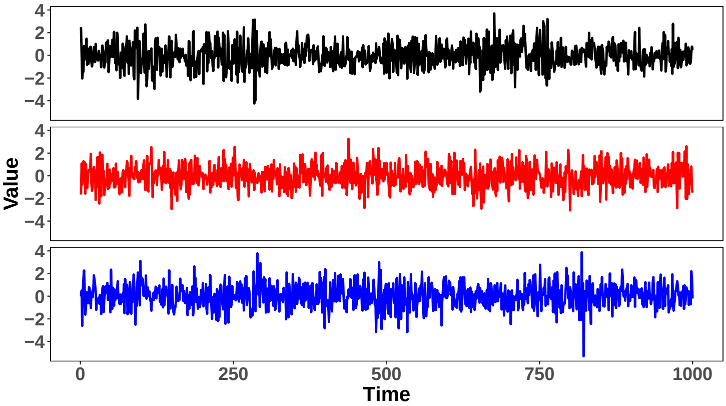


Figure 3.2: The simulated GARCH(1,1) processes based on $\mathbf{z}$.

to see that vSOBI results match very well to the latent time series, up to the signs and order of the components.

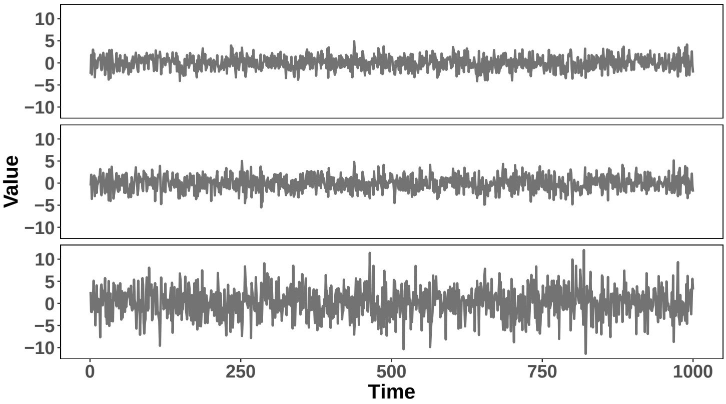


Figure 3.3: The mixed components $\mathbf{x} = \boldsymbol{\Omega}\mathbf{z}$.

3.4 Supervised dimension reduction

Assume that we have a univariate response time series $\mathrm{y} = \left(\mathrm{y}_t\right)_{t \in \mathbb{Z}}$, that is dependent on the p-variate time series $\mathbf{x} = \left(\mathbf{x}_{t-s}\right)_{t \in \mathbb{Z};\ s=0,1,\ldots}$. As p may be large, it would be advantageous if we could explain the response y adequately with some $\mathrm{k} < \mathrm{p}$ time series. The supervised dimension reduction methods designed for iid observations in Section 2.6 can be used for also time series, but they do not utilize any information on temporal dependence. This is a drawback, as now

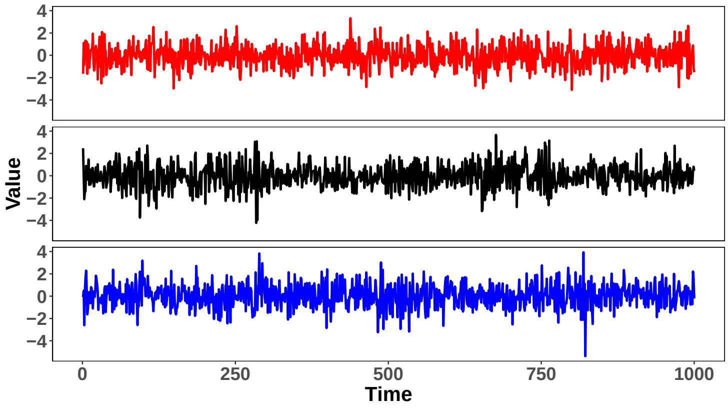


Figure 3.4: The mutually independent time series uncovered by vSOBI.

the response y might depend also on the previous value of the time series $\mathbf{x}$.

Becker and Fried (2003) apply the original SIR method to time series data by using the lagged values of y and $\mathbf{x}$ as the predictors, i.e.

$$\mathbf{x}_t^* = \left(\mathbf{x}_t', y_{t-1}, \mathbf{x}_{t-1}', \dots, y_{t-\tau_{\max}}, \mathbf{x}_{t-\tau_{\max}}'\right)'.$$

Then the standardized vector $\mathbf{x}^{*(s)}$ is used in (2.19) instead of $\mathbf{x}^{(s)}$. This approach can also be used with other iid supervised dimension reduction methods such as SAVE. With these methods in applications $\tau_{\max}$ can be very large. This may lead to a large number of variables and reduce the sample size, which are clear drawbacks.

Barbarino and Bura (2015) propose RSIR (Regularized SIR) method for time series, see also Barbarino and Bura (2017). First PCA is applied to predictors $\mathbf{x}$ to create the principal components $z_1, \dots, z_p$. Then often m of them are kept using an appropriate rule for PCA. Finally they compile a set of predictors

$$\mathbf{x}_t^{\#} = \left(z_{1t}, \dots, z_{mt}, y_t, y_{t-1}, \dots, y_{t-\tau_{\max}}\right)'$$

and then apply the regular SIR method to $\mathbf{x}^{\#(s)}$. Here the number of predictor time series depends on the number of lags used as well as how many principal components are chosen.

Another approach, discussed next, would be to directly reduce the number of predictor time series using the joint distribution of the time series y and $\mathbf{x} \in \mathbb{R}^p$. This way it is possible to also find out with which lag(s) the chosen directions contribute to the response time series y.

Assume that a univariate response time series y and a time series $\mathbf{x} \in \mathbb{R}^p$ are jointly (second-order) stationary and they have a relationship of the form

$$y_{t+1} = f\big(\mathbf{x}_t, \mathbf{x}_{t-1}, \ldots, \epsilon_t, \epsilon_{t-1}, \ldots\big), \tag{3.21}$$

where $f(\cdot)$ is an unspecified function and $\epsilon = \big(\epsilon_t\big)_{t\in\mathbb{Z}}$ is an unobserved stochastic process independent of $\mathbf{x}$.

Definition 3.5 *A Blind Source Separation model for the joint distribution of a time series* $\mathbf{x}$ *and a response time series* y *is*

$$\mathbf{x}_t = \boldsymbol{\mu} + \boldsymbol{\Omega}\mathbf{z}_t, \;\; t = 0, \pm 1, \pm 2, \ldots,$$

where, as before, $\boldsymbol{\Omega} \in \mathbb{R}^{p\times p}$ *is a full-rank mixing matrix and* $\boldsymbol{\mu} \in \mathbb{R}^p$ *is a location vector. For the stationary latent* p*-variate process* $\mathbf{z} = \big(\mathbf{z}_t\big)_{t\in\mathbb{Z}}$ *the assumptions are*

$$E\big(\mathbf{z}_t\big) = \mathbf{0}, \;\; \mathrm{Cov}\big(\mathbf{z}_t\big) = \mathbf{I}_p \tag{3.22}$$

and

$$\big(y, \mathbf{z}'_{(1)}\big)' \perp\!\!\!\perp \mathbf{z}_{(2)}, \tag{3.23}$$

where $\mathbf{z} = \big(\mathbf{z}'_{(1)}, \mathbf{z}'_{(2)}\big)'$ *can be divided into two subseries* $\mathbf{z}_{(1)} \in \mathbb{R}^k$ *and* $\mathbf{z}_{(2)} \in \mathbb{R}^{p-k}$*, as in Section 2.6.*

The subseries $\mathbf{z}_{(1)} \in \mathbb{R}^k$ is again of interest, while $\mathbf{z}_{(2)} \in \mathbb{R}^{p-k}$ can be considered noise. The assumption (3.23) implies that for all $t_1, t_2, \tau \in \mathbb{Z}$ the following is true:

$$\big(y_{t_1+\tau}, \mathbf{z}'_{(1),t_1}\big)' \perp\!\!\!\perp \mathbf{z}_{(2),t_2}.$$

Again the value of k is chosen to be the smallest value for which the assumptions (3.22) and (3.23) are true. All this leads to a prediction model

$$y_{t+1} = f\big(\mathbf{z}_{(1),t}, \mathbf{z}_{(1),t-1}, \ldots, \epsilon_t, \epsilon_{t-1}, \ldots\big),$$

where $f(\cdot)$ is an unspecified function, that may be different than in (3.21), and ϵ is an unobserved stochastic process independent of $\mathbf{x}$.

As in Section 2.6, the goal is to find an estimate for $\boldsymbol{\Gamma}$, such that $\boldsymbol{\Gamma}\big(\mathbf{x} - \boldsymbol{\mu}\big) = \mathbf{z}_{(1)}$. Also, even though the model is again not directly well-defined as in (2.15), we note that we are only looking for the subspace spanned by the rows of $\boldsymbol{\Gamma}$. Note that here also the important lags corresponding to the components of $\mathbf{z}_{(1)}$ are also uncovered.

Time series SIR Matilainen et al. (2017b) proposes TSIR, a time series version of SIR. For TSIR we need two additional assumptions for $\mathbf{z}$,

$$\mathbf{z}_{(2)} \perp\!\!\!\perp y | \mathbf{z}_{(1)} \text{ and} \tag{3.24}$$

$$\mathrm{E}\left(\mathbf{z}_{(2),t+\tau} | \mathbf{z}_{(1),t}\right) = \mathbf{0}, \tag{3.25}$$

which are time series versions of (2.16) and (2.17), respectively. These assumptions follow from (3.23).

From (3.24) and (3.25) it follows that

$$\mathrm{Cov}\left(\mathrm{E}\left(\mathbf{z}_t | y_{t+\tau}\right)\right) = \begin{pmatrix} \mathrm{Cov}\left(\mathrm{E}\left(\mathbf{z}_{(1),t} | y_{t+\tau}\right)\right) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}. \tag{3.26}$$

Note that the main difference between (2.21) and (3.26) is that in the latter one we calculate also the *lagged* supervised covariance matrices.

Starting from the standardized variables $\mathbf{x}^{(s)}$, we search for a matrix $\mathbf{U} = \left(\mathbf{u}_1, \ldots, \mathbf{u}_k\right)' \in \mathcal{O}^{k \times p}$ that maximizes

$$\begin{aligned} &\sum_{\tau \in T} \left\| \mathrm{diag}\left(\mathbf{U}\mathrm{Cov}\left(\mathrm{E}\left(\mathbf{x}_t^{(s)} | y_{t+\tau}\right)\right)\mathbf{U}'\right) \right\|^2 \\ &= \sum_{\tau \in T} \sum_{i=1}^{k} \left(\mathbf{u}_i' \mathrm{Cov}\left(\mathrm{E}\left(\mathbf{x}_t^{(s)} | y_{t+\tau}\right)\right)\mathbf{u}_i\right)^2 =: \sum_{\tau \in T} \sum_{i=1}^{k} \left(\lambda_{i\tau}\right)^2. \end{aligned}$$

Now we can estimate $\mathbf{\Gamma} = \mathbf{U}\mathbf{\Sigma}^{-1/2}$ and then the standardized components $\mathbf{\Gamma x}$. As this leads to a joint diagonalization of several scatter matrices, the ordering of the components is more complicated. Write $\lambda_{i\cdot} := \sum_{\tau \in T} \lambda_{i\tau}$, $i = 1, \ldots, k$. Then the components of $\mathrm{E}\left(\mathbf{\Gamma x} | y\right)$ are uncorrelated and ordered according to $\lambda_{i\cdot}$'s (directions).

TSIR is shown to work generally better than the vectorized SIR by Becker and Fried (2003) in simulations. It also has more desirable properties such as producing a table of $\lambda_{i\tau}$'s, from where it is easy to see which latent sources contribute to the response at which lag(s). Also in a real data example it shows better performance (Matilainen et al., 2017b).

Time series SAVE TSIR has the same kind of drawbacks as the original SIR. It works well when the relationship is linear, no matter what is the lag, but for example when squared relationships, i.e. relationships of the form $y = z^2$, are used, then it fails. This can be seen in the illustrations and simulations in Matilainen et al. (2017c), where also TSAVE, a time series version of SAVE, is proposed.

TSAVE needs the assumptions needed in TSIR, and in addition

$$\mathrm{Cov}\left(\mathbf{z}_{(2),t+\tau} \,|\, \mathbf{z}_{(1),t}\right) = \mathbf{I}_{p-k} \text{ (a.s.), for all } \tau \in \mathsf{T},$$

similar to Section 2.6. This assumption also follows from (3.23). Eventually we get

$$\mathrm{E}\left(\left(\mathbf{I}_p - \mathrm{Cov}\left(\mathbf{z}_t \,|\, y_{t+\tau}\right)\right)^2\right) = \begin{pmatrix} \mathrm{E}\left(\left(\mathbf{I}_k - \mathrm{Cov}\left(\mathbf{z}_{(1),t} \,|\, y_{t+\tau}\right)\right)^2\right) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}. \tag{3.27}$$

For TSAVE, starting from the standardized observations $\mathbf{x}^{(s)}$, we search for a matrix $\mathbf{U} = \left(\mathbf{u}_1, \ldots, \mathbf{u}_k\right)' \in \mathcal{O}^{k\times p}$ that maximizes

$$\begin{aligned}
&\sum_{\tau\in\mathsf{T}} \left\| \mathrm{diag}\left(\mathbf{U}\mathrm{E}\left(\left(\mathbf{I}_p - \mathrm{Cov}\left(\mathbf{x}_t^{(s)} \,|\, y_{t+\tau}\right)\right)^2\right)\mathbf{U}'\right)\right\|^2 \\
&= \sum_{\tau\in\mathsf{T}} \sum_{i=1}^{k} \left(\mathbf{u}_i'\mathrm{E}\left(\left(\mathbf{I}_p - \mathrm{Cov}\left(\mathbf{x}_t^{(s)} \,|\, y_{t+\tau}\right)\right)^2\right)\mathbf{u}_i\right)^2 =: \sum_{\tau\in\mathsf{T}} \sum_{i=1}^{k} \left(\lambda_{i\tau}\right)^2.
\end{aligned}$$

From this we get an estimate for $\mathbf{\Gamma} = \mathbf{U}\mathbf{\Sigma}^{-1/2}$ and then the standardized components $\mathbf{\Gamma x}$. The components of $\mathrm{E}\left(\mathbf{\Gamma x} | y\right)$ are uncorrelated and ordered according to the values of $\lambda_{i\cdot} = \sum_{\tau\in\mathsf{T}} \lambda_{i\tau}$, $i = 1, \ldots, k$.

According to the simulations in Matilainen et al. (2017c), TSAVE produces generally better results than a vectorized SAVE, where the SAVE method is used to a set of predictors that include also some lagged variables.

The lagged supervised covariance matrices are approximated by slicing the response y, as in iid case. The slicing affects the values of the covariance matrices, but not their block diagonal structure (3.26) or (3.27). The number of slices may also affect the sizes of the blocks in (3.26) and (3.27), as it is possible that $\hat{k}$, the estimated dimension of the subspace, maybe smaller than k. Also the used method may affect the value $\hat{k}$, as shown with TSIR in Matilainen et al. (2017b). For some ideas on how to estimate k in TSIR and TSAVE, see Matilainen et al. (2017b).

TSAVE also has the same kind of drawbacks than the original SAVE, as it is not as efficient as TSIR in examples where they both work, and is more affected by the the choice of number of slices H; TSAVE needs more observations per slice than TSIR, i.e. the value of H for TSAVE should be smaller than for TSIR. Matilainen et al. (2017c) concluded that $\mathsf{H}=10$ is the best for TSIR and $\mathsf{H}=2$ and 5 are the best for TSAVE.

Hybrid of TSIR and TSAVE Generalizing the SAVE|SIR method by Shaker and Prendergast (2011) (see Section 2.6) for time series would be difficult. It is not clear how we would choose the partial dimension reduction subspace, as we do not check just the directions but the combinations of directions and lags. Also the threshold value would have a significant impact on the results.

However, in the same way as Zhu et al. (2007) for iid data, Matilainen et al. (2017c) proposes TSSH, a hybrid which is a convex combination of TSIR and TSAVE. The TSSH method is shown to work more efficiently when the response y depends on the explaining variables with odd and even powers. Write

$$\mathbf{H}_{2b} = b \cdot E\left(\left(\mathbf{I}_p - \mathrm{Cov}\left(\mathbf{x}_t^{(s)}\,|y_{t+\tau}\right)\right)^2\right) + (1-b)\cdot \mathrm{Cov}\left(E\left(\mathbf{x}_t^{(s)}\,|y_{t+\tau}\right)\right),$$

where $b \in [0,1]$. Similar to Section 2.6, with $b=0$ we get TSIR and with $b=1$ we get TSAVE.

In order to estimate $\mathbf{\Gamma} = \mathbf{U}\mathbf{\Sigma}^{-1/2}$, we need to find a matrix $\mathbf{U} = \left(\mathbf{u}_1,\dots,\mathbf{u}_k\right)' \in \mathcal{O}^{k\times p}$ that maximizes

$$\sum_{\tau\in T} \left\|\mathrm{diag}\left(\mathbf{U}\mathbf{H}_{2b}\mathbf{U}'\right)\right\|^2 = \sum_{\tau\in T}\sum_{i=1}^{k}\left(\mathbf{u}_i'\mathbf{H}_{2b}\mathbf{u}_i\right)^2.$$

According to Matilainen et al. (2017c) the values of b around 0.5 and 0.6 are generally preferable.

Matilainen et al. (2017c) also suggest to use different values of H for TSIR and TSAVE part of the method, as the methods work best with different H's. The suggested values are $\mathsf{H}=10$ for the TSIR part and $\mathsf{H}=2$ (or $\mathsf{H}=5$) for the TSAVE part.

For TSIR, TSAVE and TSSH examples, see Chapter 4.

4 R package *tsBSS* and examples

For the methods developed in this work there is an R (R Core Team, 2017) package tsBSS (version 0.3.1; Matilainen et al., 2017a) available on CRAN. The version 0.3.1 described here includes the following main functions:

- gFOBI
- gJADE
- vSOBI
- FixNA with an option to choose between the original FixNA by Shi et al. (2009) and FixNA2
- tssdr with an option to choose between TSIR, TSAVE and TSSH methods

All these functions, except for gFOBI, are partly implemented in C++ in order to reduce the computation time. In addition, gFOBI, gJADE and tssdr use the joint diagonalization algorithm from the JADE package (Miettinen et al., 2017c), which is implemented in C++.

Using the package In each function the user is allowed to input the data $\mathbf{x}$ as a multivariate time series ('ts' object) or as a matrix, and in function tssdr also the response y can be input as a time series ('ts' object) or as a vector. The lags used in the algorithms are given as a vector.

The method for the joint diagonalization along with the maximum number of iterations and the convergence tolerance are passed on to the JADE package in gFOBI, gJADE and tssdr. In vSOBI and FixNA the optimization is implemented using the Lagrangian multiplier technique, naturally with options to choose the maximum number of iterations and the convergence tolerance. Also for vSOBI and FixNA there are currently two non-linearity functions G available, $G(z) = z^2$ and $G(z) = \log(\cosh(z))$.

For tssdr also the number of slices H is given by a user. For TSSH it is given as a 2-vector, where the first value corresponds to the TSIR part and the second value to the TSAVE part. The function creates a class 'tssdr' for printing, plotting and extracting the latent sources of 'tssdr' objects.

In addition, in the summary function summary.tssdr there are different strategies available for choosing the number of lags and the amount of directions. The summary function lets the user choose the desired strategy along with the threshold value for how much of the dependence needs to be retained at least. This function also creates its own class, 'summary.tssdr', which is used for printing, plotting as well as extracting the coefficients and the chosen latent directions of 'summary.tssdr' objects. The package uses the components method from the ICtest package (Nordhausen et al., 2017a) for the extraction of the source components for the 'tssdr' and 'summary.tssdr' objects.

Other packages related to package tsBSS Currently the tsBSS package depends also on the JADE package (Miettinen et al., 2017c), in which several BSS methods have been implemented, along with joint diagonalization methods as well as other utility functions.

The ICA methods FOBI and JADE have been implemented in the JADE package, as well as AMUSE and SOBI for the second order stationary time series. The package also has a joint diagonalization algorithm for both the deflation-based and the symmetric approaches and tsBSS package uses the symmetric one. For the symmetric approach it uses Jacobi angles (Cardoso and Souloumiac, 1996).

JADE package also includes the Minimum Distance Index (MDI) by Ilmonen et al. (2010a), which is used in simulations to assess how well different BSS methods perform. The MDI for an unmixing matrix estimate $\hat{\boldsymbol{\Gamma}}$ can be defined as

$$\hat{D} = \frac{1}{\sqrt{p-1}} \inf_{C \in \mathscr{C}} \left\| \mathbf{C}\hat{\boldsymbol{\Gamma}}\boldsymbol{\Omega} - \mathbf{I}_p \right\|,$$

where $\mathscr{C}$ is a set of matrices with exactly one non-zero element in each row and column. Clearly $0 \leq \hat{D} \leq 1$ and $\hat{D} = 0$ only if $\mathbf{C}\hat{\boldsymbol{\Gamma}} = \boldsymbol{\Omega}^{-1}$. The lower the value the better it separates the components. The index does not depend on the model specification and hence it is affine invariant.

Package tsBSS also uses the class 'bss' from the JADE package. It is used by gFOBI, gJADE and vSOBI functions for printing, plotting as well as extracting the coefficients and the latent sources of 'bss' objects.

The package tensorBSS (Virta et al., 2017) uses tsBSS functions gFOBI and gJADE in their tensor-valued versions of the methods.

Other packages related to time series, ICA and supervised dimension reduction For GARCH models there exists packages such as fGarch (Wuertz and Rmetrics Core Team, 2016) and rugarch (Ghalanos, 2015). Package tseries (Trapletti and Hornik, 2017) includes tools for example for ARMA and GARCH models. Tools for forecasting using different models, such as ARMA models, see package forecast (Hyndman, 2017). For SV models there is the package stochvol (Kastner, 2016).

For multivariate time series analysis package MTS (Tsay, 2015) includes functions for VARMA models, multivariate GARCH type models, among others. BigVAR package (Nicholson et al., 2017) includes tools for dimension reduction for multivariate time series.

For Independent Component Analysis, other than JADE package, there are several packages, such as ica (Helwig, 2015) and steadyICA (Risk et al., 2015) as well as fastICA (Marchini et al., 2017) and fICA (Miettinen et al., 2015a) for fastICA, to name a few. Also ICtest package (Nordhausen et al., 2017a) includes tools for ICA (and PCA).

The package BSSasymp (Miettinen et al., 2017c) can be used to compute the asymptotic covariance matrices of the mixing and unmixing matrix estimates of several BSS methods.

For supervised dimension reduction there are for example the packages dr (Weisberg, 2002), which has a function for SIR and SAVE for example, edrGraphicalTools (Coudret et al., 2017) and MAVE (Weiqiang and Yingcun, 2017).

Example usage of tsBSS package We assume that the user has installed the package tsBSS along with its dependencies JADE and ICtest. First load tsBSS.

```
> library(tsBSS)
```

The stochvol package is used for simulating SV models and the fGarch package for simulating GARCH models.

```
> library(stochvol); library(fGarch)
```

Assume we have four source time series of length 10000. Two are based on SV models and two on GARCH models. The first and the second source are simulated from an SV model with parameter vectors ($\mu = -0.04$, $\phi = 0.8$, $\sigma = 0.1$) and ($\mu = -0.12$, $\phi = 0.9$, $\sigma = 0.2$), respectively. The third and the fourth source are simulated from GARCH(1,1) models with parameter vectors ($\omega = 0.255$, $\alpha = 0.05$, $\beta = 0.7$) and ($\omega = 0.101$, $\alpha = 0.1$, $\beta = 0.8$), respectively.

```
> set.seed(2)
> n <- 10000
> s1 <- svsim(n, mu = -0.04, phi = 0.8, sigma = 0.1)$y
> s2 <- svsim(n, mu = -0.12, phi = 0.9, sigma = 0.2)$y
> s3 <- garchSim(garchSpec(model = list(omega = 0.255,
+                   alpha = 0.05, beta = 0.7)), n = n)
> s4 <- garchSim(garchSpec(model = list(omega = 0.101,
+                   alpha = 0.1, beta = 0.8)), n = n)
```

These four source time series are then mixed with a mixing matrix in order to get the 'observed' time series **x**. The mixing matrix can be any full-rank matrix of size 4×4.

```
> A <- matrix(rnorm(16), 4, 4)
> X <- cbind(s1, s2, s3, s4) %*% t(A)
```

Then different methods can be used to uncover the latent source time series **z**. Here the methods gFOBI, gJADE, vSOBI and FixNA methods are used with the default lags $0, \ldots, 12$.

```
> res1 <- gFOBI(X)
> res2 <- gJADE(X)
> res3 <- FixNA(X)
> res4 <- vSOBI(X)
```

The comparison of these unmixing matrix estimates can be done using the Minimum Distance Index. Values of the index are between 0 and 1 and the lower the value the better the separation of the components.

```
> MD(coef(res1), A)

[1] 0.3257535
```

```
> MD(coef(res2), A)

[1] 0.07726117

> MD(coef(res3), A)

[1] 0.1027123

> MD(coef(res4), A)

[1] 0.07107797
```

From all the methods gFOBI is clearly the worst and it seems that vSOBI gives here the best results. Its unmixing (signal separation) matrix estimate is the following.

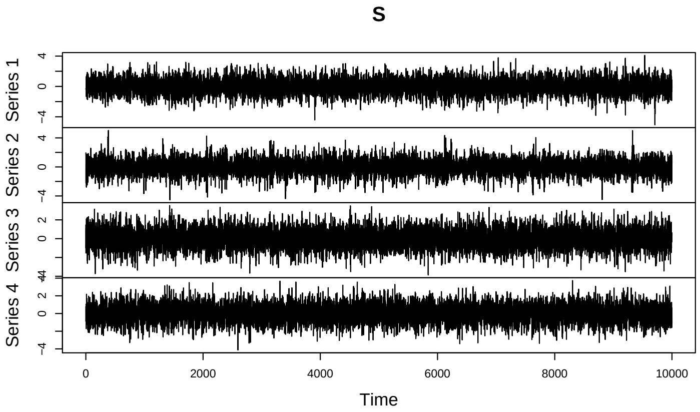
```
> coef(res4)

            [,1]       [,2]      [,3]        [,4]
[1,]  0.19748665  0.3461760 0.9652557 -0.13391310
[2,] -0.41172965  0.9126516 1.0750717 -0.02615942
[3,] -0.48743988 -0.2346775 1.4604355  0.23368565
[4,]  0.04665906  0.6246627 0.4970670  0.41666496
```

If this matrix is multiplied by the mixing matrix **A**, the resulting matrix should be close to an identity matrix up to the signs and order of the rows.

```
> coef(res4) %*% A

            [,1]         [,2]        [,3]         [,4]
[1,]  0.05615189  0.017184664 -0.01094783  0.997069916
[2,]  0.01705792 -0.997650043 -0.01042140  0.004608566
[3,]  0.99300526  0.005295150  0.06422456 -0.037852962
[4,] -0.07736697 -0.001452764  0.99505378  0.009100759
```

For the chosen method vSOBI the estimated sources are plotted.

```
> plot(res4)
```



The first few observations of the latent sources can also be printed.

```
> head(bss.components(res4))

       Series 1   Series 2   Series 3   Series 4
[1,] -0.3217990  1.9448108 -0.9915257  0.7659239
[2,] -0.3023507  1.4110913  1.4826359 -0.8927301
[3,] -0.3589765 -2.2087150 -1.5360480 -0.3917388
[4,] -0.3169541  0.3187540  1.0166698  1.0637498
[5,]  0.1666706  0.8522360 -0.5873536 -1.0398834
[6,]  0.2977453  0.9505946 -0.7337683 -1.7155187
```

Let y be a response time series that depends on the predictor **x** in such way that $y_t = z_{1,t-1}^2 + 3z_{2,t-3} + \epsilon_t$, where the process $\epsilon_t \sim N(0,1)$ is independent of **x**.

```
> eps <- rnorm(n - 3)
> y <- s1[3:(n - 1)]^2 + s2[1:(n - 3)] + eps
> X <- (cbind(s1, s2, s3, s4)[4:n, ]) %*% t(A)
```

The dimension of the subspace is known to be $k = 2$ and especially the lags 1 and 3 are important. The TSIR, TSAVE and TSSH with $b = 0.5$ are used with lags $\tau = 1, \ldots, 5$. For TSIR and the TSIR part of TSSH $H = 10$ and for TSAVE and the TSAVE part of TSSH $H = 2$.

```
> res1 <- tssdr(y, X, algorithm = "TSIR", k = 1:5, H = 10)
> res2 <- tssdr(y, X, algorithm = "TSAVE", k = 1:5, H = 2)
> res3 <- tssdr(y, X, algorithm = "TSSH", k = 1:5,
+                H = c(10, 2), weight = 0.5)
```

The summary method is used to find the dimension of the subspace and the important lags for each method. Threshold value 0.8 is used along with the rectangle method to choose lags and the number of directions.

```
> summ1 <- summary(res1, type="rectangle", thres = 0.8)
> summ2 <- summary(res2, type="rectangle", thres = 0.8)
> summ3 <- summary(res3, type="rectangle", thres = 0.8)
> summ1

        Summary of TSIR for response y and predictors X

The signal separation matrix W is:

       [,1]  [,2] [,3]    [,4]
[1,] -0.406 0.906 1.02 -0.0301

The L matrix is:

        Dir.1   Dir.2   Dir.3    Dir.4
Lag 1 0.00125 0.00107 0.00370 0.000745
Lag 2 0.00598 0.00310 0.00196 0.002475
Lag 3 0.94038 0.00359 0.00115 0.002655
Lag 4 0.00395 0.00684 0.00290 0.003269
Lag 5 0.00224 0.00664 0.00466 0.001424

Using the rectangle method:

The first direction and the first 3 lags are relevant.

> summ2


        Summary of TSAVE for response y and predictors X

The signal separation matrix W is:

       [,1]   [,2] [,3] [,4]
[1,] -0.476 -0.238 1.49 0.21

The L matrix is:
```

```
         Dir.1   Dir.2    Dir.3    Dir.4
Lag 1 0.866762 0.00201 0.000453 0.000460
Lag 2 0.000452 0.00221 0.004916 0.002937
Lag 3 0.000689 0.10748 0.000343 0.001997
Lag 4 0.000989 0.00043 0.001573 0.000891
Lag 5 0.000373 0.00209 0.002434 0.000512

Using the rectangle method:

The first direction and the first lag are relevant.
```

```
> summ3
```

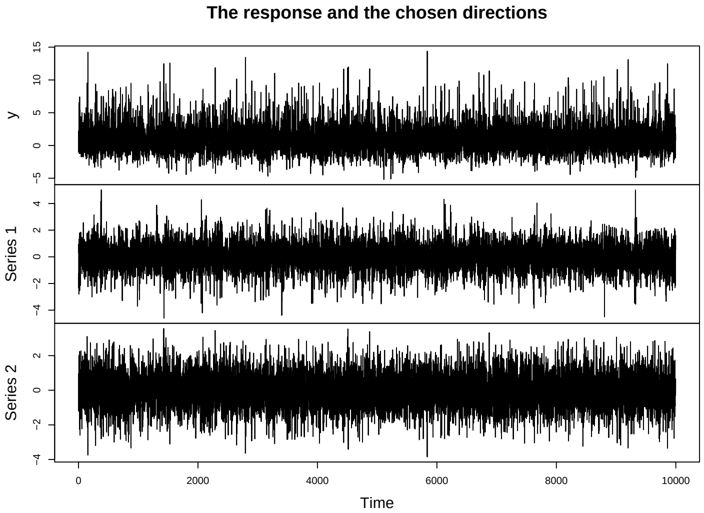
```
        Summary of TSSH for response y and predictors X

The signal separation matrix W is:

       [,1]   [,2] [,3]    [,4]
[1,] -0.412  0.904 1.03 -0.0269
[2,] -0.481 -0.226 1.51  0.2092

The L matrix is:

        Dir.1   Dir.2    Dir.3    Dir.4
Lag 1 0.00173 0.45986 0.002062 0.000625
Lag 2 0.00401 0.00197 0.002798 0.003143
Lag 3 0.49876 0.00147 0.000995 0.002651
Lag 4 0.00209 0.00247 0.003113 0.002344
Lag 5 0.00216 0.00208 0.004752 0.000915

Using the rectangle method:

The first 2 directions and the first 3 lags are relevant.
```

TSIR is able to find only one direction and completely misses the direction corresponding to the quadratic term. On the other hand, TSAVE finds easily the direction corresponding to the quadratic term, but with the used threshold value the direction corresponding to the linear term is not chosen. The hybrid method TSSH finds easily both directions. The signal separation matrix is now a 2×4-matrix, as from the four time series two are chosen.

```
> coef(summ3)
```

```
             [,1]       [,2]     [,3]        [,4]
[1,] -0.4116843  0.9042441 1.034378 -0.02690265
[2,] -0.4813061 -0.2261688 1.508242  0.20922292
```

```
> plot(summ3)
> head(components(summ3))
```



```
       Series 1   Series 2
[1,]  0.3120054  0.9592157
[2,]  0.8583688 -0.5268373
[3,]  0.9576241 -0.6385566
[4,] -2.3195483 -1.2101137
[5,] -1.2353125  0.7278113
[6,] -1.4249150 -0.6469661
```

The chosen two directions was plotted along with the response. Also the first few values of them were printed.

5 Conclusions

This work has dealt with Blind Source Separation, with a focus on Independent Component Analysis, and dimension reduction in multivariate time series. The ICA methods gFOBI and gJADE, time series versions of well-known ICA methods FOBI and JADE, were introduced. These methods work well with the components of the multivariate time series exhibit stochastic volatility. Another method called vSOBI, a variant of SOBI, was also introduced. While SOBI works well with the second order stationary time series, vSOBI method works well with the time series with stochastic volatility. According to Matilainen et al. (2017d) vSOBI seems to work even better than gFOBI and gJADE. As FixNA, a method similar to vSOBI, has already been introduced earlier (Shi et al., 2009), also FixNA2 method in its general form was introduced. The methods gFOBI, gJADE, vSOBI and FixNA2, introduced in this thesis, are included in the R-package tsBSS. Also an implementation for FixNA is included in the package, as it has not been done earlier, to the best of our knowledge.

The aforementioned methods can be used for dimension reduction, when all the time series have the same role, i.e. none is used as a response. For supervised dimension reduction, when there is also a response y that depends on some predictor time series $\mathbf{x}$, two new methods TSIR and TSAVE, as well as their hybrid TSSH, were introduced. These methods are generalizations of SIR, SAVE and their hybrid. A function tssdr for the time series supervised dimension reduction with options for TSIR, TSAVE and TSSH has also been included in the tsBSS package, with its own class with methods for printing, plotting etc.

FixNA, FixNA2 and vSOBI methods all work with time series with stochastic volatility, whereas SOBI does not. But what if the latent time series $\mathbf{z}$ includes components with stochastic volatility and components without stochastic volatility? Currently gSOBI, a generalized version of SOBI which combines SOBI and vSOBI with $\mathsf{g}(\mathsf{z}) = \mathsf{z}^2$, is being investigated along with ways to order the latent time series by their 'volatilitiness' (Miettinen et al., 2017a). The purpose for the method would be to find efficiently different types of autocorrelation (linear and squared) in the data.

Currently the order of the latent components obtained using gFOBI, gJADE, FixNA and vSOBI methods do not have a specific order. The ordering of the components according to their volatility is also being investigated.

Theoretical properties of FOBI and JADE as well as SIR and SAVE have been investigated. For the time series versions of these methods, theoretical properties still need to be investigated more, including the limiting distributions of the estimators. For SOBI, Miettinen et al. (2016) have derived the limiting distribution for $\hat{\mathbf{\Gamma}}$ in a case of uncorrelated multivariate linear processes; such linear processes include ARMA(p, q) processes. Also for gSOBI (and for vSOBI as its special case) some results are already available in Miettinen et al. (2017a).

In location-scatter models only the covariance matrix $\mathbf{\Sigma}$ have been used in this thesis. These matrices can also be replaced by other scatter matrices, given appropriate additional assumptions. For example, when robustifying methods, the covariance matrix can also be replaced, under certain circumstances, with a robust counterpart of the covariance, one that is not sensitive to outliers in the data. Such scatter matrix needs to have the so-called independence property, see for example Taskinen et al. (2007). Robustifying the whole methods proposed here is still to be investigated, as different kind of outliers, such as level shifts, may affect the results in ways that are not desirable.

In this thesis we have assumed that the source time series are stationary. This assumption is relaxed in Nonstationary Source Separation (NSS), where the variances of the source time series are allowed to change over time. There are several second-order NSS models, see for example Nordhausen (2014) and the references therein.

The methods gFOBI and gJADE have also been generalized to tensor-valued time series in Virta and Nordhausen (2017a).

For SOBI Taskinen et al. (2016) have proposed a way to choose the best lag combinations by utilizing the asymptotic distribution of the estimator. For the time series methods presented in this thesis such methods are not yet available. Thus the current guideline for applications would be to take enough of them, for example lags $1, \dots, 12$, as it is safer to take too many than too few.

It may be unreasonable to always assume that all the components are independent. Independent Subspace Analysis is an extension of ICA, where instead of the component-wise independence assumption, groups of components are assumed to be independent from each other, see for example Theis (2007) and Nordhausen and Oja (2011).

Future work also includes extending the idea of Non-Gaussian Component Analysis (NGCA) (Blanchard et al., 2006) to time series context. Only one Gaussian component is allowed in ICA, including gFOBI, gJADE and vSOBI. In a time series version of NGCA one could assume that the signal part $\mathsf{z}_{(1)}$ is a k-variate possibly dependent time series with non-trivial linear autocorrelations and $\mathsf{z}_{(2)}$ is $\mathsf{p}-\mathsf{k}$-dimensional Gaussian white noise.

Summaries of original publications

I There are several methods available for finding latent structures in high-dimensional data such as Independent Component Analysis (ICA) methods FOBI (Cardoso, 1989) and JADE (Cardoso and Souloumiac, 1993). Such methods are not very efficient when dealing with time series data, as temporal dependence needs to be considered as well. SOBI (Belouchrani et al., 1997) is a second-order Blind Source Separation (BSS) method for time series. However, SOBI cannot handle time series with stochastic volatility, i.e. time series where the volatility may change over time. Paper I focuses on extending the ICA methods FOBI and JADE for time series, especially to handle time series with stochastic volatility. Methods, called gFOBI and gJADE, are compared to classic methods to show that not only they work well with time series with stochastic volatility, but also that utilizing information on temporal dependence is important.

II In paper II, SOBI, a popular BSS methods for time series, and two methods developed in paper I, namely gFOBI and gJADE, are first reviewed. Also fastICA (Hyvärinen, 1999), a method designed for iid data, but widely used also in time series context, is discussed. Then a new family of vSOBI methods are proposed and compared to the aforementioned methods. According to the simulation results vSOBI methods seems to work better than fastICA and the methods in paper I.

III In supervised dimension reduction one (or more) variable(s) depends on another set of variables. Li (1991) has introduced such method by proposing Sliced Inverse Regression (SIR). SIR method has been used for time series before (see e.g. Becker and Fried (2003)), but only in a way that all the lagged values of the predictors are also used as predictors in the original SIR method. This type of vectorizing does not produce any easy way to examine which of the lags of which latent variables contribute to the response variable. Therefore in paper III the original SIR method is generalized to work directly with time series data. This TSIR method can now be easily used to assess the relationships between the response and the latent variables and their lags. Then, by using some appropriate criterion, we can choose the most important lags and number of directions and use them for prediction. Simulations and a real data example are used to show that this method works usually as good as or even better than the vectorized version, where also the lagged values are used as the predictors.

IV SIR method has been found to have issues when it comes to symmetric relationships around zero between a set of predictors and the response. Therefore Cook and Weisberg (1991) have introduced Sliced Average Variance Estimate (SAVE), which works in such situations. Also Zhu et al. (2007) has proposed a hybrid version of SIR and SAVE methods. This method uses jointly the efficiency of SIR and the comprehensiveness of SAVE. In Paper IV a time series version of SAVE is proposed. Also a hybrid of this TSAVE and TSIR is introduced along with a proposition on how to choose the weights for the convex combination. A simulation study to find the optimal values for the number of slices used in TSIR and TSAVE is conducted. TSAVE is also shown to perform better than the vectorised version of SAVE as well as TSIR, using different types of models, levels of autocorrelation and threshold values.

Bibliography

Andersen, T. (1994), *Stochastic Autoregressive Volatility: A Framework for Volatility Modeling*, Mathematical Finance, 4(2), 75–102.

Back, A. D. & Weigend, A. S. (1997), *A First Application of Independent Component Analysis to Extracting Structure from Stock Returns*, International Journal of Neural Systems, 8(4), 473–484.

Barbarino, A. & Bura, E., (2015), *Forecasting with Sufficient Dimension Reductions*, Finance and economics discussion series 2015-074, Washington: Board of Governors of the Federal Reserve System.

Barbarino, A. & Bura, E., (2017), *A Unified Framework for Dimension Reduction in Forecasting*, Feds working paper no. 2017–004, Washington: Board of Governors of the Federal Reserve System.

Bauwens, L., Hafner, C. & Laurent, S. (2012a), *Volatility Models* in Handbook of Volatility Models and Their Applications, eds. Bauwens, L., Hafner, C. & Laurent, S., pp. 1–45. John Wiley & Sons, Inc.

Bauwens, L., Hafner, C. M. & Laurent, S. (2012b) *Handbook of Volatility Models and Their Applications.* John Wiley & Sons, Inc.

Becker, C. & Fried, R. (2003), *Sliced Inverse Regression for High-dimensional Time Series* in Exploratory Data Analysis in Empirical Research, eds. Schwaiger, M. & Opitz, O., pp. 3–11. Springer.

Bell, A. J. & Sejnowski, T. J. (1995), *An Information-Maximization Approach to Blind Separation and Blind Deconvolution*, Neural Computation, 7(6), 1129–1159.

Belouchrani, A., Abed Meraim, K., Cardoso, J.-F. & Moulines, E. (1997), *A Blind Source Separation Technique Based on Second Order Statistics*, IEEE Transactions on Signal Processing, 45, 434–444.

Blanchard, G., Kawanabe, M., Sugiyama, M., Spokoiny, V. & Müller, K.-R. (2006), *In Search of Non-Gaussian Components of a High-Dimensional Distribution*, Journal of Machine Learning Research, 7, 247–282.

Bollerslev, T. (1986), *Generalized Autoregressive Conditional Heteroskedasticity*, Journal of Econometrics, 31(3), 307–327.

Bollerslev, T., Engle, R. F. & Wooldridge, J. M. (1988), *A Capital Asset Pricing Model with Time-Varying Covariances*, Journal of Political Economy, 96, 116–131.

Bonhomme, S. & Robin, J.-M. (2009), *Consistent Noisy Independent Component Analysis*, Journal of Econometrics, 149(1), 12–25.

Bos, C. S. (2012), *Relating Stochastic Volatility Estimation Methods* in Handbook of Volatility Models and Their Applications, eds. Bauwens, L., Hafner, C. & Laurent, S., pp. 147–174. John Wiley & Sons, Inc.

Bouwmans, T. & Zahzah, E. H. (2014), *Robust PCA via Principal Component Pursuit: A Review for a Comparative Evaluation in Video Surveillance*, Computer Vision and Image Understanding, 122, 22–34.

Box, G. E. P. & Jenkins, G. M. (1970) *Time Series Analysis: Forecasting and Control.* Holden Day, San Francisco, first edition.

Broda, S. A. & Paolella, M. S. (2009), *CHICAGO: A Fast and Accurate Method for Portfolio Risk Calculation*, Journal of Financial Econometrics, 7(4), 412–436.

Broto, C. & Ruiz, E. (2004), *Estimation Methods for Stochastic Volatility Models: A Survey*, Journal of Economic Surveys, pp. 613–649.

Bura, E. & Cook, R. (2001a), *Extending Sliced Inverse Regression: The Weighted Chi-squared Test*, Journal of the American Statistical Association, 96, 996–1003.

Bura, E. & Cook, R. (2001b), *Estimating the Structural Dimension of Regressions via Parametric Inverse Regression*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63, 393–410.

Bura, E. & Yang, J. (2011), *Dimension Estimation in Sufficient Dimension Reduction: A Unifying Approach*, Journal of Multivariate Analysis, 102, 130–142.

Cardoso, J.-F. (1989), *Source Separation Using Higher Order Moments* in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, pp. 2109–2112.

Cardoso, J.-F. & Souloumiac, A. (1993), *Blind Beamforming for Non-Gaussian Signals*, IEE-Proceedings-F, 140(6), 362–370.

Cardoso, J.-F. & Souloumiac, A. (1996), *Jacobi Angles for Simultaneous Diagonalization*, SIAM Journal on Matrix Analysis and Applications, 17, 161–164.

Chen, Y., Härdle, W. & Spokoiny, V. (2007), *Portfolio Value at Risk Based on Independent Component Analysis*, Journal of Computational and Applied Mathematics, 205, 594–607.

Clarkson, D. B. (1988), *Remark AS R74: A Least Squares Version of Algorithm AS 211: The F-G Diagonalization Algorithm*, Journal of the Royal Statistical Society. Series C (Applied Statistics), 37 (2), 317–321.

Comon, P. (1994), *Independent Component Analysis, A New Concept?*, Signal Processing, 36, 287–314.

Cook, R. D. & Critchley, F. (2000), *Identifying Regression Outliers and Mixtures Graphically*, Journal of the American Statistical Association, 95(451), 781–794.

Cook, R. (2000), *SAVE: A Method for Dimension Reduction and Graphics in Regression*, Communications in Statistics – Theory and Methods, 29, 2109–2121.

Cook, R. & Weisberg, S. (1991), *Sliced Inverse Regression for Dimension Reduction: Comment*, Journal of the American Statistical Association, 86, 328–332.

Coudret, R., Liquet, B. & Saracco, J. (2017) *edrGraphicalTools: Provides Tools for Dimension Reduction Methods.* URL `https://CRAN.R-project.org/package=edrGraphicalTools`. R package version 2.2.

Dermoune, A. & Wei, T. (2013), *FastICA Algorithm: Five Criteria for the Optimal Choice of the Nonlinearity Function*, IEEE Transactions on Signal Processing, 61(8), 2078–2087.

Diebold, F. X. & Nerlove, M. (1989), *Dynamic Exchange Rate Volatility: A Multivariate Latent Factor ARCH Model*, Journal of Applied Econometrics, 4, 1–21.

Engle, R. F., Granger, C. W. J. & Kraft, D. (1986), *Combining Competing Forecasts of Inflation Using a Bivariate ARCH Model*, Journal of Economic Dynamics and Control, 8, 151–165.

Engle, R. (1982), *Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of U.K. Inflation*, Econometrica, 50, 987–1008.

Fang, K. & Zhang, Y. (1990) *Generalized Multivariate Analysis.* Science Press.

Firat, U., Engin, S. N., Saraclar, M. & Ertuzun, A. B. (2010), *Wind Speed Forecasting Based on Second Order Blind Identification and Autoregressive Model* in Ninth International Conference on Machine Learning and Applications. IEEE, pp. 686–691.

Forni, M., Hallin, M., Lippi, M. & Zaffaroni, P. (2015), *Dynamic Factor Models with Infinite-Dimensional Factor Spaces: One-sided Representations*, Journal of Econometrics, 185(2), 359–371.

García-Ferrer, A., González-Prieto, E. & Peña, D., (2011), *Exploring ICA for Time Series Decomposition*, UC3M Working Papers. Statistics and Econometrics 11 [online], Universidad Carlos III de Madrid. URL `http://hdl.handle.net/10016/11285`.

García-Ferrer, A., González-Prieto, E. & Peña, D. (2012), *A Conditionally Heteroskedastic Independent Factor Model with an Application to Financial Stock Returns*, International Journal of Forecasting, 28(1), 70–93.

Gather, U., Hilker, T. & Becker, C. (2001), *A Robustified Version of Sliced Inverse Regression* in Statistics in Genetics and in the Environmental Sciences, eds. Fernholz, L. T., Morgenthaler, S. & Stahel, W., pp. 147–157. Birkhäuser, Basel.

Gather, U., Hilker, T. & Becker, C. (2002), *A Note on Outlier Sensitivity of Sliced Inverse Regression*, Statistics, 36(4), 271–281.

Ghalanos, A. (2015) *rugarch: Univariate GARCH Models.* URL https://CRAN.R-project.org/package=rugarch. R package version 1.3-6.

Gómez, E., Gomez-Viilegas, M. & Marín, J. (1998), *A Multivariate Generalization of the Power Exponential Family of Distributions*, Communications in Statistics - Theory and Methods, 27(3), 589–600.

González Prieto, E. (2011), "Independent Component Analysis for Time Series" PhD thesis, Universidad Carlos III de Madrid. URL http://hdl.handle.net/10016/11958.

Hald, A. (2000), *The Early History of the Cumulants and the Gram-Charlier Series*, International Statistical Review, 68(2), 137–153.

Haldane, J. B. S. (1948), *Note on the Median of a Multivariate Distribution*, Biometrika, 35(3–4), 414–417.

Hallin, M. & Mehta, C. (2015), *R-Estimation for Asymmetric Independent Component Analysis*, Journal of the American Statistical Association, 110(509), 218–232.

Harvey, A., Ruiz, E. & Shephard, N. (1994), *Multivariate Stochastic Variance Models*, Review of Economic Studies, 61(2), 247–264.

Helwig, N. E. (2015) *ica: Independent Component Analysis.* URL https://CRAN.R-project.org/package=ica. R package version 1.0-1.

Hettmansperger, T. P. & Randles, R. H. (2002), *A Practical Affine Equivariant Multivariate Median*, Biometrika, 89(4), 851–860.

Hotelling, H. (1933), *Analysis of a Complex of Statistical Variables into Principal Components*, Journal of Educational Psychology, 24, 417–441 and 498–520.

Hotelling, H. (1936), *Relations Between Two Sets of Variates*, Biometrika, 28(3/4), 321–377.

Huber, P. J. (1985), *Projection Pursuit*, The Annals of Statistics, 13(2), 435–475.

Hyndman, R. J. (2017) *forecast: Forecasting Functions for Time Series and Linear Models.* URL `http://github.com/robjhyndman/forecast`. R package version 8.1.

Hyvärinen, A. (1999), *Fast and Robust Fixed-point Algorithms for Independent Component Analysis*, IEEE Transactions on Neural Networks, 10, 626–634.

Hyvärinen, A. (2001), *Blind Source Separation by Nonstationarity of Variance: A Cumulant-based Approach*, IEEE Transactions on Neural Networks, 12(6), 1471–1474.

Hyvärinen, A. & Oja, E. (1997), *A Fast Fixed-point Algorithm for Independent Component Analysis*, Neural Computation, 9, 1483–1492.

Hyvärinen, A. & Oja, E. (2000), *Independent Component Analysis: Algorithms and Applications*, Neural Networks, 13(4–5), 411–430.

Ilmonen, P., Nordhausen, K., Oja, H. & Ollila, E. (2010a), *A New Performance Index for ICA: Properties Computation and Asymptotic Analysis* in Latent Variable Analysis and Signal Separation, eds. Vigneron, V., Zarzoso, V., Moreau, E., Gribonval, R. & Vincent, E. Springer, pp. 229–236.

Ilmonen, P. & Paindaveine, D. (2011), *Semiparametrically Efficient Inference Based on Signed Ranks in Symmetric Independent Component Models*, Annals of Statistics, 39(5), 2448–2476.

Ilmonen, P., Nevalainen, J. & Oja, H. (2010b), *Characteristics of Multivariate Distributions and the Invariant Coordinate System*, Statistics & Probability Letters, 80(23), 1844–1853.

Ilmonen, P., Nordhausen, K., Oja, H. & Theis, F. (2015), *An Affine Equivariant Robust Second-Order BSS Method* in Latent Variable Analysis and Signal Separation: 12th International Conference LVA/ICA 2015, eds. Vincent, E., Yeredor, A., Koldovský, Z. & Tichavský, P. Springer International Publishing, pp. 328–335.

Jolliffe, I. (2002) *Principal Component Analysis.* Springer-Verlag.

Joyce, C. A., Gorodnitsky, I. F. & Kutas, M. (2004), *Automatic Removal of Eye Movement and Blink Artifacts from EEG Data Using Blind Component Separation*, Psychophysiology, 41(2), 313–325.

Jutten, C. & Taleb, A. (2000), *Source Separation: From Dusk Till Dawn* in Proceedings of the International Symposium on Independent Component Analysis and Blind Signal Separation. pp. 15–26.

Kastner, G. (2016), *Dealing with Stochastic Volatility in Time Series Using the R Package stochvol*, Journal of Statistical Software, 69 (5), 1–30.

Kastner, G. & Frühwirth-Schnatter, S. (2014), *Ancillarity-Sufficiency Interweaving Strategy (ASIS) for Boosting MCMC Estimation of Stochastic Volatility Models*, Computational Statistics & Data Analysis, 76, 408 – 423.

Koldovský, Z., Tichavský, P. & Oja, E. (2006), *Efficient Variant of Algorithm FastICA for Independent Component Analysis Attaining the Cramér-Rao Lower Bound*, IEEE Transactions on Neural Networks, 17(5), 1265–1277.

Kollo, T. & von Rosen, D. (2005) *Advanced Multivariate Statistics with Matrices.* Springer.

Kotz, S. & Nadarajah, S. (2004) *Multivariate T-Distributions and Their Applications.* Cambridge University Press.

Ku, W., Storer, R. H. & Georgakis, C. (1995), *Disturbance Detection and Isolation by Dynamic Principal Component Analysis*, Chemometrics and Intelligent Laboratory Systems, 30(1), 179–196.

Kumiega, A., Neururer, T. & Vliet, B. V. (2011), *Independent Component Analysis for Realized Volatility: Analysis of the Stock Market Crash of 2008*, The Quarterly Review of Economics and Finance, 51(3), 292–302.

Li, B. & Wang, S. (2007), *On Directional Regression for Dimension Reduction*, Journal of the American Statistical Association, 102 (479), 997–1008.

Li, K.-C. (1991), *Sliced Inverse Regression for Dimension Reduction*, Journal of the American Statistical Association, 86(414), 316–327.

Li, K.-C. (1992), *On Principal Hessian Directions for Data Visualization and Dimension Reduction: Another Application of Stein's*

Lemma, Journal of the American Statistical Association, 87(420), 1025–1039.

Li, Y. & Zhu, L.-X. (2007), *Asymptotics for Sliced Average Variance Estimation*, The Annals of Statistics, 35(1), 41–69.

Lietzén, N., Nordhausen, K. & Ilmonen, P. (2017), *Complex Valued Robust Multidimensional SOBI* in Latent Variable Analysis and Signal Separation: 13th International Conference LVA/ICA 2017, eds. Tichavský, P., Babaie-Zadeh, M., Michel, O. J. & Thirion-Moreau, N. Springer International Publishing, pp. 131–140.

Lindner, A. M. (2009), *Stationarity, Mixing, Distributional Properties and Moments of GARCH*(p,q)*–Process* in Handbook of Financial Time Series, eds. Mikosch, T., Kreiß, J.-P., Davis, R. A. & Andersen, T. G., pp. 43–69. Springer.

Liquet, B. & Saracco, J. (2012), *A Graphical Tool for Selecting the Number of Slices and the Dimension of the Model in SIR and SAVE Approaches*, Computational Statistics, 27(1), 103–125.

Liski, E., Nordhausen, K. & Oja, H. (2014), *Supervised Invariant Coordinate Selection*, Statistics: A Journal of Theoretical and Applied Statistics, 4, 711–731.

Lu, C.-J., Wu, J.-Y. & Lee, T.-S. (2009), *Application of Independent Component Analysis Preprocessing and Support Vector Regression in Time Series Prediction* in Proceedings of the Second International Joint Conference on Computational Sciences and Optimization, volume 1. IEEE, pp. 468–471.

Luo, W. & Li, B. (2016), *Combining Eigenvalues and Variation of Eigenvectors for Order Determination*, Biometrika, 103(4), 875–887.

Lütkepohl, H. (2005) *New Introduction to Multiple Time Series Analysis*. Springer.

Ma, Y. & Zhu, L. (2013), *A Review on Dimension Reduction*, International Statistics Review, 81, 134–150.

Marchini, J. L., Heaton, C. & Ripley, B. D. (2017) *fastICA: FastICA Algorithms to Perform ICA and Projection Pursuit*. URL `https://CRAN.R-project.org/package=fastICA`. R package version 1.2-1.

Mardia, K. V. (1970), *Measures of Multivariate Skewness and Kurtosis with Applications*, Biometrika, 57(3), 519–530.

Matilainen, M., Nordhausen, K. & Oja, H. (2015), *New Independent Component Analysis Tools for Time Series*, Statistics & Probability Letters, 105, 80–87.

Matilainen, M., Croux, C., Miettinen, J., Nordhausen, K., Oja, H. & Taskinen, S. (2017a) *tsBSS: Tools for Blind Source Separation and Supervised Dimension Reduction for Time Series.* URL `https://CRAN.R-project.org/package=tsBSS`. R package version 0.3.1.

Matilainen, M., Croux, C., Nordhausen, K. & Oja, H. (2017b), *Supervised Dimension Reduction for Multivariate Time Series*, Econometrics and Statistics, 4, 57–69.

Matilainen, M., Croux, C., Nordhausen, K. & Oja, H. (2017c). *Sliced Average Variance Estimation for Multivariate Time Series*, Submitted.

Matilainen, M., Miettinen, J., Nordhausen, K., Oja, H. & Taskinen, S. (2017d), *On Independent Component Analysis with Stochastic Volatility Models*, Austrian Journal of Statistics, 46(3–4), 57–66.

Matteson, D. S. & Tsay, R. S. (2011), *Dynamic Orthogonal Components for Multivariate Time Series*, Journal of the American Statistical Association, 106(496), 1450–1463.

Matteson, D. & Ruppert, D. (2011), *Time-series Models of Dynamic Volatility and Correlation*, IEEE Signal Processing Magazine, 28 (5), 72–82.

Melino, A. & Turnbull, S. M. (1990), *Pricing Foreign Currency Options with Stochastic Volatility*, Journal of Econometrics, 45 (1–2), 239–265.

Miettinen, J., Nordhausen, K., Oja, H. & Taskinen, S. (2012), *Statistical Properties of a Blind Source Separation Estimator for Stationary Time Series*, Statistics & Probability Letters, 82, 1865–1873.

Miettinen, J., Nordhausen, K., Oja, H. & Taskinen, S. (2014a), *Deflation-based FastICA With Adaptive Choices of Nonlinearities*, IEEE Transactions on Signal Processing, 62(21), 5716–5724.

Miettinen, J., Nordhausen, K., Oja, H. & Taskinen, S. (2014b), *Deflation-based Separation of Uncorrelated Stationary Time Series*, Journal of Multivariate Analysis, 123, 214–227.

Miettinen, J., Nordhausen, K., Oja, H. & Taskinen, S. (2015a) *fICA: Classical, Reloaded and Adaptive FastICA Algorithms.* URL `https://CRAN.R-project.org/package=fICA`. R package version 1.0-3.

Miettinen, J., Taskinen, S., Nordhausen, K. & Oja, H. (2015b), *Fourth Moments and Independent Component Analysis*, Statistical Science, 30, 372–390.

Miettinen, J., Illner, K., Nordhausen, K., Oja, H., Taskinen, S. & Theis, F. (2016), *Separation of Uncorrelated Stationary Time Series Using Autocovariance Matrices*, Journal of Time Series Analysis, 37(3), 337–354.

Miettinen, J., Matilainen, M., Nordhausen, K. & Taskinen, S. (2017a). *Extracting Conditionally Heteroscedastic Components Using ICA*, Submitted.

Miettinen, J., Nordhausen, K., Oja, H., Taskinen, S. & Virta, J. (2017b), *The Squared Symmetric FastICA Estimator*, Signal Processing, 131, 402–411.

Miettinen, J., Nordhausen, K. & Taskinen, S. (2017c), *Blind Source Separation Based on Joint Diagonalization in R: The Packages JADE and BSSasymp*, Journal of Statistical Software, 76(2), 1–31.

Miettinen, J. (2015), *Alternative Diagonality Criteria for SOBI* in Modern Nonparametric, Robust and Multivariate Methods: Festschrift in Honour of Hannu Oja, eds. Nordhausen, K. & Taskinen, S., pp. 455–469. Springer International Publishing.

Mittelhammer, R. C. (1996) *Mathematical Statistics for Economics and Business.* Springer-Verlag, first edition.

Naik, G. R. & Wang, W. (2014) *Blind Source Separation: Advances in Theory, Algorithms and Applications.* Springer.

Nelson, D. B. & Cao, C. Q. (1992), *Inequality Constraints in the Univariate GARCH Model*, Journal of Business & Economic Statistics, 10(2), 229–235.

Nicholson, W., Matteson, D. & Bien, J. (2017) *BigVAR: Dimension Reduction Methods for Multivariate Time Series.* URL `https://CRAN.R-project.org/package=BigVAR`. R package version 1.0.2.

Nordhausen, K. & Oja, H. (2011), *Independent Subspace Analysis using Three Scatter Matrices*, Austrian Journal of Statistics, 40 (1–2), 93–101.

Nordhausen, K., Oja, H. & Ollila, E. (2008), *Robust Independent Component Analysis Based on Two Scatter Matrices*, Austrian Journal of Statistics, 37, 91–100.

Nordhausen, K., Ilmonen, P., Mandal, A., Oja, H. & Ollila, E. (2011), *Deflation-based FastICA Reloaded* in Proceedings of 19th European Signal Processing Conference 2011 (EUSIPCO 2011). IEEE, pp. 1854–1858.

Nordhausen, K. (2014), *On Robustifying Some Second Order Blind Source Separation Methods for Nonstationary Time Series*, Statistical Papers, 55(1), 141–156.

Nordhausen, K., Oja, H., Tyler, D. E. & Virta, J. (2017a) *ICtest: Estimating and Testing the Number of Interesting Components in Linear Dimension Reduction.* URL `https://CRAN.R-project.org/package=ICtest`. R package version 0.3.

Nordhausen, K., Oja, H. & Tyler, D. (2017b). *Asymptotic and Bootstrap Tests for Subspace Dimension*, Submitted. URL `https://arxiv.org/abs/1611.04908v2`.

Nordhausen, K., Oja, H., Tyler, D. & Virta, J. (2017c), *Asymptotic and Bootstrap Tests for the Dimension of the Non-Gaussian Subspace*, IEEE Signal Processing Letters, 24(6), 887–891.

Oja, H. (2010) *Multivariate Nonparametric Methods with R.* Springer.

Ollila, E. (2010), *The Deflation-based FastICA Estimator: Statistical Analysis Revisited*, IEEE Transactions on Signal Processing, 58(3), 1527–1541.

Pearson, K. (1901), *On Lines and Planes of Closest Fit to Systems of Points in Space*, Philosophical Magazine, 2(11), 559–572.

Peña, D. & Yohai, V. J. (2016), *Generalized Dynamic Principal Components*, Journal of the American Statistical Association, 111 (515), 1121–1131.

Popescu, T. D. & Manolescu, M. (2007), *Blind Source Separation of Traffic-Induced Vibrations in Building Monitoring* in IEEE International Conference on Control and Automation. IEEE, pp. 2101–2106.

Prendergast, L. A. (2005), *Influence Functions for Sliced Inverse Regression*, Scandinavian Journal of Statistics, 32(3), 385–404.

Prendergast, L. A. (2006), *Detecting Influential Observations in Sliced Inverse Regression Analysis*, Australian & New Zealand Journal of Statistics, 48(3), 285–304.

Prendergast, L. A. (2007), *Implications of Influence Function Analysis for Sliced Inverse Regression and Sliced Average Variance Estimation*, Biometrika, 94(3), 585–601.

R Core Team (2017) *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`. R version 3.4.1.

Rainieri, C. (2014), *Perspectives of Second-Order Blind Identification for Operational Modal Analysis of Civil Structures*, Shock and Vibration, 2014, 1–9.

Rencher, A. C. & Christensen, W. F. (2012) *Methods of Multivariate Analysis.* John Wiley & Sons, Inc., third edition.

Risk, B. B., James, N. A. & Matteson, D. S. (2015) *steadyICA: ICA and Tests of Independence via Multivariate Distance Covariance.* URL `https://CRAN.R-project.org/package=steadyICA`. R package version 1.0.

Scott, D. W. (2015), *The Curse of Dimensionality and Dimension Reduction* in Multivariate Density Estimation: Theory, Practice, and Visualization. John Wiley & Sons, Inc, second edition.

Shaker, A. J. & Prendergast, L. A. (2011), *Iterative Application of Dimension Reduction Methods*, Electronic Journal of Statistics, 5, 1471–1494.

Shao, Y., Cook, R. D. & Weisberg, S. (2007), *Marginal Tests with Sliced Average Variance Estimation*, Biometrika, 94(2), 285–296.

Shephard, N. & Andersen, T. G. (2009), *Stochastic Volatility: Origins and Overview* in Handbook of Financial Time Series, eds. Mikosch, T., Kreiß, J.-P., Davis, R. A. & Andersen, T. G., pp. 233–254. Springer.

Shi, Z., Jiang, Z. & Zhou, F. (2009), *Blind Source Separation with Nonlinear Autocorrelation and Non-Gaussianity*, Journal of Computational and Applied Mathematics, 223(1), 908–915.

Shumway, R. H. & Stoffer, D. S. (2011) *Time Series Analysis and Its Applications: With R Examples.* Springer, third edition.

Shynk, J. J. (2012) *Probability, Random Variables, and Random Processes: Theory and Signal Processing Applications.* John Wiley & Sons, Inc., first edition.

Stock, J. H. & Watson, M. W. (2002), *Forecasting Using Principal Components from a Large Number of Predictors*, Journal of the American Statistical Association, 97(460), 1167–1179.

Stone, J. V. (2004) *Independent Component Analysis: A Tutorial Introduction.* A Bradford Book.

Tang, A. C., Sutherland, M. T. & McKinney, C. J. (2005), *Validation of SOBI Components from High-Density EEG*, Neuroimage, 25(2), 539–553.

Tang, A. (2010), *Applications of Second Order Blind Identification to High-Density EEG-Based Brain Imaging: A Review* in Advances in Neural Networks - ISNN 2010: 7th International Symposium on Neural Networks, Part II, eds. Zhang, L., Lu, B.-L. & Kwok, J. Springer, pp. 368–377.

Taskinen, S., Sirkiä, S. & Oja, H. (2007), *Independent Component Analysis Based on Symmetrised Scatter Matrices*, Computational Statistics & Data Analysis, 51(10), 5103–5111.

Taskinen, S., Miettinen, J. & Nordhausen, K. (2016), *A More Efficient Second Order Blind Identification Method for Separation of Uncorrelated Stationary Time Series*, Statistics & Probability Letters, 116, 21–26.

Taylor, S. J. (1982), *Financial Returns Modelled by the Product of Two Stochastic Processes – A Study of Daily Sugar Prices 1961–79* in Time Series Analysis: Theory and Practice 1, ed. Anderson, O. D., pp. 203–216. Springer.

Teräsvirta, T. (2009), *An Introduction to Univariate GARCH Models* in Handbook of Financial Time Series, eds. Mikosch, T., Kreiß, J.-P., Davis, R. A. & Andersen, T. G., pp. 17–42. Springer.

Theis, F. J. (2007), *Towards a General Independent Subspace Analysis* in Advances in Neural Information Processing Systems 19, eds. Schölkopf, B., Platt, J. C. & Hoffman, T., pp. 1361–1368. MIT Press.

Theis, F. J., Meyer-Bäse, A. & Lang, E. W. (2004), *Second-Order Blind Source Separation Based on Multi-dimensional Autocovariances* in Independent Component Analysis and Blind Signal Separation: Fifth International Conference, eds. Puntonet, C. G. & Prieto, A. Springer, pp. 726–733.

Theis, F. J., Müller, N. S., Plant, C. & Böhm, C. (2010), *Robust Second-Order Source Separation Identifies Experimental Responses in Biomedical Imaging* in International Conference on Latent Variable Analysis and Signal Separation, eds. Vigneron, V., Zarzoso, V., Moreau, E., Gribonval, R. & Vincent, E. Springer, pp. 466–473.

Tichavský, P., Koldovský, Z. & Oja, E. (2006), *Performance Analysis of the FastICA Algorithm and Cramer-Rao Bounds for Linear Independent Component Analysis*, IEEE Transactions on Signal Processing, 54(4), 1189–1203.

Tong, L., Soon, V., Huang, Y. & Liu, R. (1990), *AMUSE: A New Blind Identification Algorithm* in Proceedings of IEEE International Symposium on Circuits and Systems. IEEE, pp. 1784–1787.

Trapletti, A. & Hornik, K. (2017) *tseries: Time Series Analysis and Computational Finance.* URL `https://CRAN.R-project.org/package=tseries`. R package version 0.10-42.

Tsay, R. S. (2015) *MTS: All-Purpose Toolkit for Analyzing Multivariate Time Series (MTS) and Estimating Multivariate Volatility Models.* URL `https://CRAN.R-project.org/package=MTS`. R package version 0.33.

Virta, J. & Nordhausen, K. (2017a), *Blind Source Separation of Tensor-valued Time Series*, Signal Processing, 141, 204–216.

Virta, J. & Nordhausen, K. (2017b), *On the Optimal Non-linearities for Gaussian Mixtures in FastICA* in Proceedings of Latent Variable Analysis and Signal Separation: 13th International Conference, LVA/ICA 2017, eds. Tichavský, P., Babaie-Zadeh, M., Michel, O. J. & Thirion-Moreau, N. Springer International Publishing, pp. 427–437.

Virta, J., Nordhausen, K. & Oja, H. (2016). *Projection Pursuit for non-Gaussian Independent Components*, Submitted. URL `https://arxiv.org/abs/1612.05445v1`.

Virta, J., Li, B., Nordhausen, K. & Oja, H. (2017) *tensorBSS: Blind Source Separation Methods for Tensor-Valued Observations.* URL `https://CRAN.R-project.org/package=tensorBSS`. R package version 0.3.

Visuri, S., Koivunen, V. & Oja, H. (2000), *Sign and Rank Covariance Matrices*, Journal of Statistical Planning and Inference, 91, 557–575.

Wei, T. (2014), *On the Spurious Solutions of the FastICA Algorithm* in 2014 IEEE Workshop on Statistical Signal Processing (SSP). IEEE, pp. 161–164.

Wei, T. (2015), *A Convergence and Asymptotic Analysis of the Generalized Symmetric FastICA Algorithm*, IEEE Transactions on Signal Processing, 63(24), 6445–6458.

Weiqiang, H. & Yingcun, X. (2017) *MAVE: Methods for Dimension Reduction.* URL `https://CRAN.R-project.org/package=MAVE`. R package version 1.2.9.

Weisberg, S. (2002), *Dimension Reduction Regression in R*, Journal of Statistical Software, 7(1), 1–22.

Wuertz, D. & Rmetrics Core Team (2016) *fGarch: Rmetrics – Autoregressive Conditional Heteroskedastic Modelling.* URL `https://CRAN.R-project.org/package=fGarch`. R package version 3010.82.1.

Xia, Y., Tong, H., Li, W. K. & Zhu, L.-X. (2002), *An Adaptive Estimation of Dimension Reduction Space*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64(3), 363–410.

Ye, Z. & Weiss, R. E. (2003), *Using the Bootstrap to Select One of a New Class of Dimension Reduction Methods*, Journal of the American Statistical Association, 98(464), 968–979.

Zhu, L. P., Wang, T., Zhu, L. & Ferré, L. (2010), *Sufficient Dimension Reduction Through Discretization-Expectation Estimation*, Biometrika, 97, 295–304.

Zhu, L.-X. & Ng, K. W. (1995), *Asymptotics of Sliced Inverse Regression*, Statistica Sinica, 5, 727–736.

Zhu, L.-P. & Zhu, L.-X. (2007), *On Kernel Method for Sliced Average Variance Estimation*, Journal of Multivariate Analysis, 98 (5), 970–991.

Zhu, L.-X., Ohtaki, M. & Li, Y. (2007), *On Hybrid Methods of Inverse Regression-based Algorithms*, Computational Statistics & Data Analysis, 51(5), 2621–2635.

Zhu, L., Miao, B. & Peng, H. (2006), *On Sliced Inverse Regression With High-Dimensional Covariates*, Journal of the American Statistical Association, 101(474), 630–643.

Annales Universitatis Turkuensis

Turun yliopisto
University of Turku

Painosalama Oy, Turku, Finland 2018

ISBN 978-951-29-7105-3 (PRINT)
ISBN 978-951-29-7106-0 (PDF)
ISSN 0082-7002 (PRINT) | ISSN 2343-3175 (PDF)