Turun yliopisto
University of Turku

INDEPENDENT COMPONENT ANALYSIS FOR NON-STANDARD DATA STRUCTURES

Joni Virta

TURUN YLIOPISTON JULKAISUJA - ANNALES UNIVERSITATIS TURKUENSIS
Sarja - ser. AI osa - tom. 577 | Astronomica - Chemica - Physica - Mathematica | Turku 2018

Turun yliopisto
University of Turku

INDEPENDENT COMPONENT ANALYSIS FOR NON-STANDARD DATA STRUCTURES

Joni Virta

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
Sarja - ser. A I osa - tom. 577 | 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

Professor Bing Li
Department of Statistics
Pennsylvania State University, State College, PA, USA

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

Reviewed by

Professor Tõnu Kollo
Institute of Mathematics and Statistics
University of Tartu, Tartu, Estonia

Associate Professor Lexin Li
Division of Biostatistics
University of California, Berkeley, CA, USA

Opponent

Professor Davy Paindaveine
Solvay Brussels School of Economics and Management
Université Libre de Bruxelles, Brussels, Belgium

The originality of this thesis has been checked in accordance with the University of Turku quality assurance system using the Turnitin OriginalityCheck service.

ISBN 978-951-29-7148-0 (PRINT)
ISBN 978-951-29-7149-7 (PDF)
ISSN 0082-7002 (Print)
ISSN 2343-3175 (Online)
Painosalama Oy - Turku, Finland 2018

Abstract

Independent component analysis is a classical multivariate tool used for estimating independent sources among collections of mixed signals. However, modern forms of data are typically too complex for the basic theory to adequately handle. In this thesis extensions of independent component analysis to three cases of non-standard data structures are developed: noisy multivariate data, tensor-valued data and multivariate functional data.

In each case we define the corresponding independent component model along with the related assumptions and implications. The proposed estimators are mostly based on the use of kurtosis and its analogues for the considered structures, resulting into functionals of rather unified form, regardless of the type of the data. We prove the Fisher consistencies of the estimators and particular weight is given to their limiting distributions, using which comparisons between the methods are also made.

Tiivistelmä

Riippumattomien komponenttien analyysi on moniulotteisen tilastotieteen työkalu, jota käytetään estimoimaan riippumattomia lähdesignaaleja sekoitettujen signaalien joukosta. Modernit havaintoaineistot ovat kuitenkin tyypillisesti rakenteeltaan liian monimutkaisia, jotta niitä voitaisiin lähestyä alan perinteisillä menetelmillä. Tässä väitöskirjassa esitellään laajennukset riippumattomien komponenttien analyysin teoriasta kolmelle epästandardille aineiston muodolle: kohinaiselle moniulotteiselle datalle, tensoriarvoiselle datalle ja moniulotteiselle funktionaaliselle datalle.

Kaikissa tapauksissa määritellään vastaava riippumattomien komponenttien malli oletuksineen ja seurauksineen. Esitellyt estimaattorit pohjautuvat enimmäkseen huipukkuuden ja sen laajennuksien käyttöön ja saatavat funktionaalit ovat analyyttisesti varsin yhtenäisen muotoisia riippumatta aineiston tyypistä. Kaikille estimaattoreille näytetään niiden Fisher-konsistenttisuus ja painotettuna on erityisesti estimaattoreiden rajajakaumat, jotka mahdollistavat teoreettiset vertailut eri menetelmien välillä.

Acknowledgements

Although the thesis cover can hold only a single name, a great number of people have shaped the outcome of this work through their actions and words, whether deliberate or not. The following list tries to do justice to the indispensable support of these people.

First of all, I wish to express my sincere gratitude to my three supervisors: Professor Hannu Oja who showed me that statistics does not lose a single a bit in elegance and aesthetics to mathematics and who never backed down from an invitation to discuss all matters theoretical; Assistant Professor Klaus Nordhausen who always knew the right direction to take no matter what I was struggling with, and who also acquainted me with the bizarre ways of the academia, not once failing to accompany his guidance with the relevant anecdote; and Professor Bing Li, who not only invited me to a semester-long research visit to Pennsylvania State University, but without whom the word "non" could be dropped from the title of this thesis.

I am thankful to Professor Tõnu Kollo and Associate Professor Lexin Li, pre-examiners of the thesis, for their careful reviews and comments that helped me make the end result more complete. Similarly, this research would not have been possible if not for the funding provided by the Academy of Finland, the Universty of Turku Doctoral Programme in Mathematics and Computer Sciences (MATTI), Turku University Foundation, Oskar Öflunds Stiftelse and Emil Aaltonen Foundation.

The majority of the thesis research was done by being surrounded by the statistics staff members of the University of Turku and I wish to express my gratitude to all of them. There was always someone to ask when things did not go as planned and a wonderful group of people with whom to apply our extensive knowledge of probability theory to the noble sport of board gaming. I especially want to thank PhD Markus Matilainen for sharing both a room and all the joys and sorrows of being a PhD student with me for the last years. Most memorable projects combining statistics with all things imaginable were undertaken on those lazy Friday afternoons in Quantum 279.

I wish to thank PhD Sara Taskinen and PhD Jari Miettinen for our joint research and also for their company during the numerous conference visits during the last years. Similarly, I am grateful to Assistant Professor Pauliina Ilmonen and MSc Niko Lietzén not only for our current (and future) work together but also for taking great measures to ensure that I feel myself absolutely welcome and at home in Aalto University. I am also grateful to Professor Anne Ruiz-Gazen who hosted me during my research visit to the Toulouse School of Economics.

I would like to thank my parents, Marinella and Jari, whose door and schedule has always been open for a weekend visit and whose constant encouragement to pursue the things I find interesting has been, and still continues to be, invaluable in life. Equally irreplaceable have been the countless hours spent with my two brothers, Miro and Jaro, playing around with projects related to games, art, music, crafts and all things in between.

And finally, I wish to thank Eveliina. No part of my life in the last years has been as important to the whole process of writing the thesis and staying sane throughout as the time spent with her. Every single hour of sitting in cafés, watching movies or just living everyday life has had a role in getting me where I am now.

February 21, 2018

Joni Virta

Contents

List of symbols

x, y, z	random variables
$\mathbf{x}, \mathbf{y}, \mathbf{z}$	random vectors
$\mathbf{X}, \mathbf{Y}, \mathbf{Z}$	random matrices
$\mathcal{X}, \mathcal{Y}, \mathcal{Z}$	random tensors
$\mathcal{D}$	a set of distributions
F_x	the distribution (function) of a random variable x
$\rightsquigarrow$	convergence in distribution
$c_m(\cdot)$	mth marginal cumulant
$\mathrm{tr}^\circ(\mathbf{\Psi})$	the trace of the matrix obtained by replacing the $k + (k-1)p$, $k \in \{1, \ldots, p\}$, diagonal elements of the $p^2 \times p^2$ matrix $\mathbf{\Psi}$ by zeroes
$\mathbf{e}_i$	a vector with a single one as its ith element and other elements zero
$\mathbf{E}^{ij}$	a matrix with a single one as its (i,j)th element and other elements zero
δ_{ij}	the Kronecker delta
$\|\cdot\|_F$	Frobenius norm
$\mathrm{diag}(\mathbf{A})$	diagonal matrix with the same diagonal elements as $\mathbf{A}$
$\mathrm{off}(\mathbf{A})$	$\mathbf{A} - \mathrm{diag}(\mathbf{A})$
$\mathbb{R}_+^{p\times p}$	the set of $p \times p$ positive semidefinite matrices
$\mathbb{R}_{++}^{p\times p}$	the set of $p \times p$ positive-definite matrices
$\mathbb{S}^{p-1}$	the unit sphere in $\mathbb{R}^p$
$\mathcal{U}^{p\times k}$	the set of $p \times k$ matrices with orthonormal rows
$\mathcal{P}^p$	the set of all $p \times p$ permutation matrices
$\mathcal{J}^p$	the set of all $p \times p$ diagonal matrices with diagonal elements ± 1
$\mathcal{D}^p$	the set of all $p \times p$ diagonal matrices with positive diagonal elements
$\mathcal{C}^p$	the set of all $p \times p$ matrices with a single non-zero element in each row and column
$\equiv$	the equivalence relation: $\mathbf{A} \equiv \mathbf{B} \Leftrightarrow \mathbf{A} = \mathbf{JPB}$, $\mathbf{J} \in \mathcal{J}^p$, $\mathbf{P} \in \mathcal{P}^p$
$\overset{*}{\equiv}$	the equivalence relation: $\mathbf{A} \equiv \mathbf{B} \Leftrightarrow \mathbf{A} = \tau\mathbf{JPB}$, $\tau \in \mathbb{R}$, $\mathbf{J} \in \mathcal{J}^p$, $\mathbf{P} \in \mathcal{P}^p$
$\otimes$	the Kronecker (for matrices) or the tensor product (for functions)
vec	the vectorization operator
$\times_m$	the m-mode linear transformation
$\times_{m=1}^r$	the simultaneous m-mode linear transformation from all modes
$\mathcal{H}$	a Hilbert space
$\mathcal{L}(\mathcal{H})$	the set of bounded linear operators from $\mathcal{H}$ to $\mathcal{H}$

List of original publications

The thesis consists of the introductory part and the following five original publications which are reprinted with permission from the copyright holders:

I Virta J., Nordhausen K. and Oja H. (2016), Projection pursuit for non-Gaussian independent components. Submitted, preprint at arXiv:1612.05445.

II Virta J., Li B., Nordhausen K. and Oja H. (2017), Independent component analysis for tensor-valued data. *Journal of Multivariate Analysis*, 162, 172–192.

III Virta J., Li B., Nordhausen K. and Oja H. (2017), JADE for tensor-valued observations. Accepted to Journal of Computational and Graphical Statistics, preprint at arXiv:1603.05406.

IV Virta J., Taskinen S. and Nordhausen K. (2016), Applying fully tensorial ICA to fMRI data. In the proceedings of *2016 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*.

V Virta J., Li B., Nordhausen K. and Oja H. (2017), Independent component analysis for multivariate functional data. Submitted.

Part I

Summary

1 Introduction

The continuous technological advancement has brought with it an ever-increasing output of data and the researchers today are faced with data sets massive in size and complex in structure.

The increased size manifests in huge numbers of observations and variables and to draw any conclusions from the data a reliable way of separating information from the noise is necessary. An attractive option is provided by independent component analysis (ICA) which aims to linearly separate the data into mutually independent source components. To this end ICA has two appealing properties: first, its linearity often translates into low computational complexity, and secondly, it generally yields highly interpretable components.

Similarly, the increasing complexity of data has lead into datasets exhibiting more intricate structures than the basic multivariate methodology taught in universities can account for. While generally there is nothing stopping one from resorting to the classical methods, such behavior tends to ignore the wealth of information available in the special structure. The aim of this thesis is to extend the main ideas behind the classical independent component analysis into the realms of these non-standard data structures. More specifically, we formulate methods for dealing with noisy multivariate data, tensor-valued data and multivariate functional data, all common forms of data encountered in present-day applications, and explore their theoretical properties.

The proposed methods are built by naturally extending classical multivariate methodology to consider the special characteristics of the different structures. The resulting procedures prove powerful tools for the respective forms of data and at the same time still retain the essential properties of the classical methods used as their building blocks. In discussing the methods, special attention is paid to two classical statistical concepts, consistency (showing that the method works) and limiting distributions (showing how well the method works). These two in conjunction with the corresponding algorithms and equivariance properties provide for a statistically comprehensive treatise of the subject.

The summary part is divided into the introduction and four additional sections. Section 2 briefly introduces the notational conventions we adhere to for the remainder of the summary. Some technical issues regarding measure theory and the existence of various probabilistic constructs are also recalled. The next three parts discuss the theory and literature of independent component analysis in the cases of vector-valued, tensor-valued and functional data, respectively. The three treatises have been tried to be kept as unified in content and organization as possible, and for the most parts of Sections 3 and 4 this has been successful (in the author's opinion). However, when we move from finite-dimensional spaces to the infinite-dimensional some fundamental statistical concepts no longer exist in the form we are used to, and consequently in Section 5 discussing functional data some compromises have been made. But the key elements of the theory and most of the familiar intuition gained in spaces of finite dimension still hold.

Throughout the summary part of the thesis we will refrain from giving data analysis examples as numerous ones can be found both in the thesis papers and in Virta et al. (2016c); Virta and Nordhausen (2017a,b). Implementations of the main methods of Sections 3 and 4 can be found in the R-packages Nordhausen et al. (2017b); Virta et al. (2016a).

2 Notation and some technicalities

Throughout the thesis we implicitly assume the existence of some suitable, rich enough probability space $(\Omega, \mathcal{F}, \mathbb{P})$ where all our random constructs of interest are defined as measurable random variables. As the basic theory of random variables taking values in $\mathbb{R}^p$ can be found in any measure-theoretic treatise of probability, see e.g. Billingsley (2008), this construction is given no more thought in the first two parts of the thesis discussing respectively random vectors and random tensors (we use the word "tensor" instead of the mathematically more sound "array" as it is the standard practice). Even though the latter case is in practice quite different from the former, measure-wise it can be reduced to the former by considering random tensors as random vectors with the added ordering of the elements into a lattice form. However, in the third part discussing functional data we will take some time to briefly go through concepts such as random functions and measurability in general Hilbert spaces. This is to ensure that a reader unfamiliar with these concepts can still follow the exposition of the final part of the thesis.

Our notation is mainly standard: The Euclidean spaces are denoted by $\mathbb{R}, \mathbb{R}^p, \mathbb{R}^{p_1 \times p_2}, \ldots, \mathbb{R}^{p_1 \times \cdots \times p_r}$, the convex cones of $p \times p$ positive semidefinite and positive-definite matrices by $\mathbb{R}^{p \times p}_{+}$ and $\mathbb{R}^{p \times p}_{++}$, and the unit sphere in $\mathbb{R}^p$ by $\mathbb{S}^{p-1}$. Orthonormal sets of vectors also play a major role throughout the thesis and by $\mathcal{U}^{p \times k}$ we denote the set of all $p \times k$ matrices with orthonormal columns, $k \leq p$. Univariate random variables will be denoted by lower-case letters, $x, y, z \in \mathbb{R}$, random vectors by bold lower-case, $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^p$, random matrices by bold upper-case, $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \in \mathbb{R}^{p_1 \times p_2}$ and random tensors by the Euler script, $\mathfrak{X}, \mathfrak{Y}, \mathfrak{Z} \in \mathbb{R}^{p_1 \times \cdots \times p_r}$. The components of a random vector/matrix/tensor are denoted in lower-case with suitable indexing such as $x_{i_1 i_2}$ for the elements of the random matrix $\mathbf{X}$. We do not distinguish between a random variable and its realization.

The distribution of a random variable x is denoted by F_x and similarly for a random vector $\mathbf{x}$, a random matrix $\mathbf{X}$, etc. As the aim of the thesis is not to do robust statistics (indeed, quite the opposite as most of the methods to come require fourth moments to operate) we implicitly assume the existence of all required moments and related quantities for all considered distributions F_x, touching the issues related to robustness only briefly in passing. Let then $\mathcal{D}$ be some suitable, rich enough collection of distributions. Many of our constructs and estimators are best defined as statistical functionals $S : \mathcal{D} \to \mathcal{S}$ from $\mathcal{D}$ to some appropriate space, often $\mathbb{R}^{p \times p}_{++}$ and throughout the thesis we abuse the notation by writing $S(x)$ instead of the more proper $S(F_x)$ when $x \sim F_x$. Whenever we discuss the finite sample aspects of an estimator S, we use the notation $\hat{F}_x$ to denote the empirical distribution of the sample and consequently the finite sample estimate of S can be written as $\hat{S} = S(\hat{F}_x)$.

Some important vectors and matrices we repeatedly use include the standard basis vectors $\mathbf{e}_j \in \mathbb{R}^p$, $j \in \{1, \ldots, p\}$ and the matrices $\mathbf{E}^{ij} = \mathbf{e}_i \mathbf{e}_j^\top$ with a single one as the element (i, j) and other elements zero. The set of all $p \times p$ diagonal

matrices with positive diagonal elements is denoted by $\mathcal{D}^p$, the set of all $p \times p$ permutation matrices by $\mathcal{P}^p$ and the set of all $p \times p$ diagonal matrices with diagonal elements equal to ± 1 by $\mathcal{J}^p$. These three classes of matrices can be used respectively to scale, reorder and change the signs of the components of a random vector $\mathbf{x} \in \mathbb{R}^p$. They also naturally combine into the class $\mathcal{C}^p$ of all matrices representable as $\mathbf{DPJ}$ for some $\mathbf{D} \in \mathcal{D}^p$, $\mathbf{P} \in \mathcal{P}^p$ and $\mathbf{J} \in \mathcal{J}^p$. The class $\mathcal{C}^p$ consists then of all $p \times p$ matrices with a single non-zero element on each row and column.

When we move to discuss tensor-valued random variables the following notions prove useful. The vectorization operator $\mathrm{vec} : \mathbb{R}^{p_1 \times \cdots \times p_r} \to \mathbb{R}^{p_1 \cdots p_r}$ takes the tensor $\boldsymbol{\mathcal{X}}$ and stacks its elements into a long vector $\mathrm{vec}(\boldsymbol{\mathcal{X}})$ of length $p_1 \cdots p_r$ in such a way that the first index goes through its cycle the fastest and the last index the slowest. For example, the vectorization of a matrix is obtained by stacking its columns from left to right into a vector. The symbol $\otimes$ denotes the Kronecker product $\mathbf{A} \otimes \mathbf{B}$ between two matrices $\mathbf{A} \in \mathbb{R}^{p_1 \times p_2}$ and $\mathbf{B} \in \mathbb{R}^{q_1 \times q_2}$, defined as the $p_1 q_1 \times p_2 q_2$ block matrix with the (k, l) block equal to $a_{kl}\mathbf{B}$, for $k \in \{1, \ldots, p_1\}$, $l \in \{1, \ldots, p_2\}$. As a binary operation the Kronecker product has numerous useful properties such as associativity and distributivity, see Van Loan (2000). If further $\mathbf{A}, \mathbf{X}$ and $\mathbf{B}$ are matrices of conforming sizes we have the identity $\mathrm{vec}(\mathbf{AXB}^\top) = (\mathbf{B} \otimes \mathbf{A})\mathrm{vec}(\mathbf{X})$, providing a useful connection between the two previous concepts. Similar identity holds also for the vectorization of a linearly transformed tensor, see Section 4.

For both vectors and matrices the notation $\|\cdot\|$ denotes the standard Euclidean (Frobenius) norm unless otherwise stated. For a square matrix $\mathbf{A} \in \mathbb{R}^{p \times p}$ the diagonal $p \times p$ matrix with the same diagonal elements as $\mathbf{A}$ is denoted by $\mathrm{diag}(\mathbf{A})$ and the $p \times p$ matrix obtained by replacing the diagonal of $\mathbf{A}$ with zeroes by $\mathrm{off}(\mathbf{A}) = \mathbf{A} - \mathrm{diag}(\mathbf{A})$. For a $p^2 \times p^2$ matrix $\boldsymbol{\Psi}$ we also introduce the "partial trace" notation $\mathrm{tr}^\circ(\boldsymbol{\Psi})$ to refer to the trace of the matrix obtained by replacing the $k + (k-1)p$, $k \in \{1, \ldots, p\}$, diagonal elements of $\boldsymbol{\Psi}$ by zeroes. If $\boldsymbol{\Psi}$ is a covariance matrix of a vectorized random matrix then $\mathrm{tr}^\circ(\boldsymbol{\Psi})$ is the sum of the variances of its off-diagonal elements.

Throughout the thesis we are particularly concerned with asymptotic results and for that let $\{x_n\}_{n=1}^\infty$ be an infinite sequence of random variables. We say that the sequence x_n belongs to the class $o_p(a_n)$ for some deterministic sequence a_n if the sequence x_n/a_n converges in probability to 0. The class $O_p(a_n)$ in turn contains all sequences x_n for which x_n/a_n is bounded in probability, i.e., for every $\epsilon > 0$ there exists M_ϵ such that $\mathbb{P}(|x_n/a_n| > M_\epsilon) < \epsilon$ for all n. As is standard, we denote the previous two by the abuse of notation, $x_n = o_p(a_n)$ and $x_n = O_p(a_n)$. Our main tools for showing inclusions to the previous two classes are the law of large numbers and the central limit theorem. Namely, if $x_n \to_p x$ then $x_n - x = o_p(1)$ and if $\sqrt{n}(x_n - \mu) \rightsquigarrow \mathcal{N}(0, \psi)$ then $\sqrt{n}(x_n - \mu) = O_p(1)$. The algebra of convergent and bounded sequences is particularly straightforward: we have $o_p(1) + o_p(1) = o_p(1)$, $o_p(1) + O_p(1) = O_p(1)$, $o_p(1)o_p(1) = o_p(1)$ and $o_p(1)O_p(1) = o_p(1)$. Despite their seeming simplicity the previous four rules are sufficient to give us almost all of our asymptotical results. Detailed discussions of the previous concepts can be found in, e.g., Serfling (2009); Van der Vaart (1998).

3 Independent component analysis for vector-valued data

3.1 Location-scatter model and its extensions

Location-scatter model and multivariate normal distribution

Before delving into our main theme, independent component analysis (ICA), we first briefly consider the location-scatter model, see e.g. Oja (2010), discussing some classical methodology associated with it and also deriving our future model of choice in the process. Now, let $\mathbf{x} \in \mathbb{R}^p$ be a random vector coming from the location-scatter model,

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Omega}\mathbf{z}, \tag{3.1}$$

where the location vector $\boldsymbol{\mu} \in \mathbb{R}^p$ and the invertible mixing matrix $\boldsymbol{\Omega} \in \mathbb{R}^{p \times p}$ are the model parameters and $\mathbf{z} \in \mathbb{R}^p$ is an unobserved random vector. Clearly, the parameters are not identifiable without further assumptions on $\mathbf{z}$ and the very least we can do is to fix the location $\boldsymbol{\mu}$ and the scales of the columns of $\boldsymbol{\Omega}$. Assuming finite second moments we achieve this by requiring

$$\mathrm{E}(\mathbf{z}) = \mathbf{0}_p \quad \text{and} \quad \mathrm{Cov}(\mathbf{z}) = \mathrm{E}\left(\mathbf{z}\mathbf{z}^\top\right) = \mathbf{I}_p. \tag{3.2}$$

As a consequence, $\boldsymbol{\Omega}$ is now identifiable up to post-multiplication by an orthogonal matrix as we can write

$$\boldsymbol{\Omega}\mathbf{z} = (\boldsymbol{\Omega}\mathbf{U})\left(\mathbf{U}^\top\mathbf{z}\right) = \boldsymbol{\Omega}^*\mathbf{z}^*, \tag{3.3}$$

where $\mathbf{z}^*$ still satisfies the moment conditions in (3.2). However, despite this unidentifiability the first two moments of $\mathbf{x}$ are still fully identifiable, $\mathrm{E}(\mathbf{x}) = \boldsymbol{\mu}$, $\mathrm{Cov}(\mathbf{x}) = \boldsymbol{\Omega}\boldsymbol{\Omega}^\top$. If moment-based assumptions are to be avoided, some robust alternatives for E and Cov can be used to fix the parameters instead, see Maronna and Yohai (1976).

Imposing additional assumptions on $\mathbf{z}$ in model (3.1), we obtain a variety of classical multivariate models. The traditional choice is to assume multivariate normality, $\mathbf{z} \sim \mathcal{N}_p(\mathbf{0}_p, \mathbf{I}_p)$, yielding a general multivariate normal (Gaussian) distribution for $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with the density function

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

and the covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{\Omega}\boldsymbol{\Omega}^\top$. Two key properties of a random vector $\mathbf{z} \sim \mathcal{N}_p(\mathbf{0}_p, \mathbf{I}_p)$ having the standard multivariate normal distribution are:

i) Spherical symmetry: $\mathbf{z} \sim \mathbf{U}\mathbf{z}$ for all $\mathbf{U} \in \mathcal{U}^{p \times p}$.

ii) Independence: the components of $\mathbf{z}$ are mutually independent.

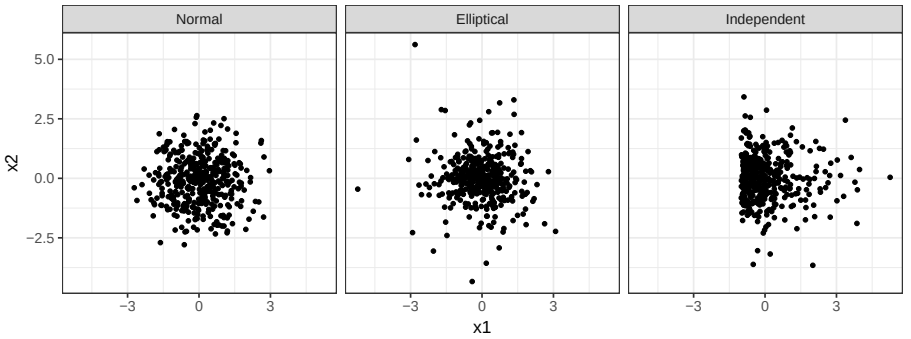

Figure 3.1: From left to right, random samples from the multivariate normal distribution, multivariate t-distribution with 5 degrees of freedom and an independent component model with exponential and logistic components. Each distribution has been standardized to have $\mathrm{E}(\boldsymbol{z}) = \boldsymbol{0}$ *and* $\mathrm{Cov}(\boldsymbol{z}) = \boldsymbol{I}_p$.

Both of the previous properties provide a starting point for generalizing the normal model $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. If we decide to hold onto the spherical symmetry of the normal model we obtain the class of elliptical distributions and if we instead require that the components of $\mathbf{z}$ be kept independent we arrive at the independent component model, the main topic of this thesis. The multivariate normal distribution is the unique family of distributions lying in the intersection of these two models, see Kollo and von Rosen (2006). Figure 3.1 shows bivariate scatter plots of random samples from distributions coming from the three models, showcasing how different their forms can be. The next two sections discuss respectively the elliptical and independent component models.

Elliptical model and principal component analysis

Before we formally define the class of elliptical distributions we first consider its "standardized" counterpart, the class of spherical distributions. See Fang et al. (1990); Kollo and von Rosen (2006); Paindaveine (2012) for detailed treatises of both models. We say that a random vector $\mathbf{z} \in \mathbb{R}^p$ has spherical distribution if it satisfies the first property of the standard multivariate normal distribution highlighted previously,

$$\mathbf{z} \sim \mathbf{U}\mathbf{z}, \quad \text{for all} \quad \mathbf{U} \in \mathcal{U}^{p \times p}.$$

The distribution of a spherical random vector remains unchanged under rotations and reflections implying in particular that its equidensity contours are spheres centered at the origin. Alternative characterizations for spherical distributions based on density functions and characteristic functions exist but for our purposes the above definition is sufficient. All odd moments of a spherically distributed $\mathbf{z}$ are zero and its second moments satisfy $\mathrm{E}(\mathbf{z}\mathbf{z}^\top) = \rho \mathbf{I}_p$ for some constant $\rho > 0$, assuming the moments exist in the first place (Anderson, 1992).

Assume now that $\mathbf{z} \in \mathbb{R}^p$ has a spherical distribution and let

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

where $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$ are as in (3.1) and again the latter is identifiable only up to post-multiplication by an orthogonal matrix. Now $\mathbf{x}$ obeys an elliptical distribution with the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma} = \boldsymbol{\Omega}\boldsymbol{\Omega}^\top$ and, assuming the required moments exist, has the mean and covariance, $\mathrm{E}(\mathbf{x}) = \boldsymbol{\mu}$, $\mathrm{Cov}(\mathbf{x}) = \rho\boldsymbol{\Sigma}$. The reason we stress the existence of moments again is that the most common use of the elliptical model is to craft distributions sharing some key properties of the normal distribution while at the same time having heavier tails and to use the resulting distributions to test the efficiencies of various estimators. That having been said, these concepts are largely irrelevant to the main body of our work and the interested reader is directed to Kariya and Sinha (2014) for more information.

A method closely related to the elliptical family is the classical principal component analysis (PCA), see Pearson (1901); Hotelling (1933); Jolliffe (2002) and the references therein. To provide a foundation for a future comparison between PCA and ICA we next briefly describe PCA in the context of elliptical distributions. Let $\mathbf{x} \in \mathbb{R}^p$ have a centered elliptical distribution, $\mathbf{x} = \boldsymbol{\Omega}\mathbf{z}$, where $\mathbf{z} \in \mathbb{R}^p$ has a spherical distribution. Writing $\boldsymbol{\Omega} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$ for the singular value decomposition of the mixing matrix, we can further assume that $\mathbf{V}^\top = \mathbf{I}_p$, as the transformation by the orthogonal $\mathbf{V}^\top$ leaves the distribution of $\mathbf{z}$ unchanged. Consequently, the covariance matrix of $\mathbf{x}$ has the form $\mathrm{Cov}(\mathbf{x}) = \rho\mathbf{U}\mathbf{D}^2\mathbf{U}^\top$ and projecting the observations on the eigenvectors $\mathbf{U}$ we obtain the principal component scores

$$\mathbf{U}^\top\mathbf{x} = \mathbf{U}^\top\mathbf{U}\mathbf{D}\mathbf{z} = \mathbf{D}\mathbf{z}$$

The covariance matrix of the principal component scores is $\rho\mathbf{D}^2$, implying that the scores are uncorrelated. Of course, this aspect of the result is not a consequence of the elliptical model and some simple linear algebra shows that the above procedure can be used to obtain uncorrelated components regardless of the distribution of the random vector $\mathbf{x}$. A standard practitioner of PCA would next discard a suitable amount of components with the lowest variances and carry out any further analyses with the smaller set of variables. However, we formulated PCA in the context of elliptical distributions for the precise reason of showing that PCA can be used to "solve" the elliptical model (although the original components get lost in the absorbing of $\mathbf{V}^\top$). So although generally ICA is seen as superior to PCA (it finds independent components while PCA finds only uncorrelated), we may also view them in parallel (both solve one generalization of the multivariate normal model). As uncorrelatedness implies independence for Gaussian variables, under the multivariate normal model PCA actually recovers independent components, further making PCA and ICA, the signature methods of the elliptical and independent component model, equivalent in the intersection of the two models.

The previous derivation on elliptical distributions and PCA actually hold not just for the covariance matrix but for any orthogonally equivariant scatter matrix. For example, see Marden (1999); Visuri et al. (2000) for the use of spatial sign covariance matrix in extracting the principal components. A scatter

matrix is any functional $\mathbf{S} : \mathcal{D} \to \mathbb{R}_+^{p \times p}$ which is affine equivariant in the sense that

$$\mathbf{S}(\mathbf{A}\mathbf{x}) = \mathbf{A}\mathbf{S}(\mathbf{x})\mathbf{A}^\top,$$

for all invertible $\mathbf{A} \in \mathbb{R}^{p \times p}$. For orthogonally equivariant scatter matrices we naturally require that the above holds merely for all orthogonal $\mathbf{A} \in \mathcal{U}^{p \times p}$. See Dümbgen et al. (2015) for more information on scatter matrices.

Independent component model

The second extension of the multivariate normal model mimics $\mathcal{N}_p(\mathbf{0}_p, \mathbf{I}_p)$ by equipping the location-scatter model (3.1) with the additional assumption that the components of $\mathbf{z}$ are mutually independent. This move from uncorrelatedness to independence turns out to be strong enough to almost guarantee the identifiability of the model parameters.

Too see how the assumption of independence affects the confounding in (3.3) we invoke the classical Skitovich-Darmois theorem (Ghurye and Olkin, 1962): if we can form independent non-trivial (consisting of more than one summand) linear combinations from a collection of random variables that are themselves mutually independent, then all the random variables must be normally distributed. This means that if at most one component of $\mathbf{z}$ has normal distribution then the confounding matrix $\mathbf{U}$ in (3.3) can have only a single non-zero element in each row (to prevent the formation of non-trivial linear combinations). Thus $\mathbf{U} = \mathbf{J}\mathbf{P}$ for some $\mathbf{J} \in \mathcal{J}^p$ and $\mathbf{P} \in \mathcal{P}^p$ and we can identify the independent components up to order and marginal signs. With this we are now ready to formulate the independent component model.

Definition 1. *We say that the random vector $\boldsymbol{x} \in \mathbb{R}^p$ obeys the independent component (IC) model if*

$$\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{\Omega}\boldsymbol{z}, \tag{3.4}$$

where $\boldsymbol{\mu} \in \mathbb{R}^p$ and the invertible $\boldsymbol{\Omega} \in \mathbb{R}^{p \times p}$ are unknown parameters and the latent random vector $\boldsymbol{z} \in \mathbb{R}^p$ satisfies Assumptions V1, V2 and V3 below.

Assumption V1. *The components of $\boldsymbol{z}$ are mutually independent,*

Assumption V2. *The vector $\boldsymbol{z}$ is standardized,* $\mathrm{E}(\boldsymbol{z}) = \boldsymbol{0}_p$ *and* $\mathrm{Cov}(\boldsymbol{z}) = \boldsymbol{I}_p$.

Assumption V3. *At most one of the components of $\boldsymbol{z}$ is normally distributed.*

Assumption V3 is in a sense backward to the classical multivariate statistics where all methods generally assume full multivariate normality. Also, the assumption is in practice not as strict as it sounds; if $\mathbf{z}$ happens to contain more than one normal component we simply lose our ability to consistently estimate the corresponding columns of $\boldsymbol{\Omega}$, but we can still estimate the remaining columns (and independent components). This approach was taken in the context of ICA in Virta et al. (2016b) and in a wider model in Blanchard et al. (2005) where also dependence between the non-Gaussian components was allowed. In this introduction we however restrict for simplicity to the fully identifiable case, the more general extensions following easily. Finally, Assumptions V2 are V3 are

more formally identifiability constraints but we will nevertheless continue to refer to them as assumptions.

ICA has a long history and was before its modern formulation as a statistical problem considered mainly as a signal separation problem. Two early contributions to ICA include the Fisher consistent fourth cumulant-based methods, *fourth order blind identification* (FOBI) (Cardoso, 1989) and *joint approximate diagonalization of eigen-matrices* (JADE) (Cardoso and Souloumiac, 1993), which will later serve as the primary examples for our extensions of ICA to non-standard data structures. An early version of the IC model was also considered already in Cardoso (1989) for stationary time series. Later approaches based on the same idea of decomposing cumulant matrices or tensors are found, for example, in Moreau (2001); Kollo (2008); Comon et al. (2015).

Comon (1994) defined contrasts as functionals of random vectors that are maximized when the vector has independent components and showed that all marginal cumulants can be used as contrasts to solve the IC problem. The same ideas were later used in conjunction with projection pursuit (Friedman and Tukey, 1974; Huber, 1985) to develop the FastICA-estimator (Hyvärinen, 1999) of which several variants have been proposed over the years, see Hyvärinen and Köster (2006); Koldovský et al. (2006); Nordhausen et al. (2011a); Miettinen et al. (2014a); Virta et al. (2016b); Miettinen et al. (2017).

While the diagonalization of cumulant matrices and tensors along with projection pursuit constitute the two main approaches to ICA in the literature, a diverse array of other perspectives have also been considered: Hastie and Tibshirani (2003); Chen and Bickel (2006); Samworth and Yuan (2012) used non-parametric and semi-parametric marginal density estimation to solve the problem; Ilmonen and Paindaveine (2011); Hallin and Mehta (2015) developed efficient estimators based on marginal ranks and signed ranks; Oja et al. (2006); Taskinen et al. (2007); Nordhausen et al. (2008) based their estimators on pairs of scatter matrices with the independence property; Karvanen et al. (2002); Karvanen and Koivunen (2002) chose the latent score functions from a set of distributions using the method of moments and Matteson and Tsay (2017) minimize a measure of dependency based on the distance covariance.

A different problem is encountered if we allow an arbitrary number of Gaussian components in the IC model and do not fix their number a priori. Now solving the independent component problem amounts to estimating both the signal components and also their true number. Despite the practical implications of this problem inferential treatments of it are rather scarce in the literature. For hypothesis testing of the true dimension using limiting distributions and bootstrapping in the IC model and a related wider model, see Nordhausen et al. (2016, 2017a). A closely related model is the non-Gaussian component analysis (NGCA) model where arbitrary dependencies between the signal components are allowed, see Blanchard et al. (2005); Kawanabe (2005), but where the true dimension of the signal space is usually assumed to be known.

ICA can be seen as a special case of the classical *blind source separation* (BSS) problem where no assumption on the independence of the observations needs to be made. The resulting body of methods covers, for example, time series of varying dimensions (Tong et al., 1990; Belouchrani et al., 1997; Miettinen et al., 2014b; Matilainen et al., 2015; Virta and Nordhausen, 2017a) and spatial data (Nordhausen et al., 2015). While allowing dependent data would

open up the door for a wide range of modern applications, in this thesis we still stick strictly to the classical case of independent and identically distributed observations (and in any case, the methods we discuss have a long history of being successfully applied to data mildly violating some of the key assumptions).

3.2 IC functionals

We next formally define our main tool for estimating the parameters of the independent component model in Definition 1 in the form of statistical functionals. To "solve" the model is equivalent to estimating both $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$. The first task is trivial as Assumption V2 guarantees that $\mathrm{E}(\mathbf{x}) = \boldsymbol{\mu}$ and we may without loss of generality assume throughout the following that $\boldsymbol{\mu} = \mathbf{0}_p$. Instead of $\boldsymbol{\Omega}$ it is customary to estimate its inverse, which is the purpose of IC functionals defined next.

Definition 2. *The functional $\boldsymbol{\Gamma} : \mathcal{D} \to \mathbb{R}^{p \times p}$ is an independent component (IC) functional if we have*

i) $\boldsymbol{\Gamma}(\boldsymbol{z}) \equiv \boldsymbol{I}_p$ *for all standardized* $\boldsymbol{z} \in \mathbb{R}^p$ *with independent components and*

ii) $\boldsymbol{\Gamma}(\boldsymbol{A}\boldsymbol{x}) \equiv \boldsymbol{\Gamma}(\boldsymbol{x})\boldsymbol{A}^{-1}$ *for all* $\boldsymbol{x} \in \mathbb{R}^p$ *and all invertible* $\boldsymbol{A} \in \mathbb{R}^{p \times p}$,

where two square matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{p \times p}$, satisfy $\boldsymbol{A} \equiv \boldsymbol{B}$ if and only if $\boldsymbol{A} = \boldsymbol{J}\boldsymbol{P}\boldsymbol{B}$ for some $\boldsymbol{J} \in \mathcal{J}^p$ and $\boldsymbol{P} \in \mathcal{P}^p$.

To be considered an IC functional a statistical functional $\boldsymbol{\Gamma}$ must thus satisfy two conditions. The first condition requires that in the case of trivial mixing, $\boldsymbol{\Omega} = \mathbf{I}_p$, the transformation by $\boldsymbol{\Gamma}(\mathbf{x}) = \boldsymbol{\Gamma}(\mathbf{z})$ does not mix up the already-found independent components. The second condition is a form of affine equivariance and its implications can be seen by observing that for any change of coordinate system, $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$, the transformation by $\boldsymbol{\Gamma}$ satisfies $\boldsymbol{\Gamma}(\mathbf{x})\mathbf{x} \mapsto \boldsymbol{\Gamma}(\mathbf{A}\mathbf{x})\mathbf{A}\mathbf{x} \equiv \boldsymbol{\Gamma}(\mathbf{x})\mathbf{x}$. That is, the resulting components are not dependent on the used coordinate system making IC functionals share that property with *invariant coordinate system* (ICS) functionals, see Tyler et al. (2009). Notice also that since the second condition of Definition 2 is required to hold for all possible random variables $\mathbf{x}$, it holds particularly for all realized samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$. To summarize, *ii)* is more a general, desirable statistical property of an estimator whereas *i)* is something specific to the current model, a "stopping condition" which halts the estimation procedure when something with independent components is found.

Combining the two conditions shows that for any $\mathbf{x}$ coming from an IC model we have $\boldsymbol{\Gamma}(\mathbf{x}) \equiv \boldsymbol{\Omega}^{-1}$, meaning that $\boldsymbol{\Gamma}$ is Fisher consistent to the inverse of the mixing matrix $\boldsymbol{\Omega}$ up to the order and signs of its rows and can be used to solve the IC model via the linear transformation $\mathbf{x} \mapsto \boldsymbol{\Gamma}(\mathbf{x})\mathbf{x}$. The need for the equivalence $\equiv$ instead of full equality in both of the previous conditions is naturally a consequence of our identifiability constraints leaving both the signs and the order of the components of $\mathbf{z}$ unfixed. However, as this has little bearing on practice we later abuse the language by saying, e.g., that a solution is unique when it is actually unique up to this equivalence class.

After we have later obtained a collection of candidate IC functionals a natural next question is which one should we use. All IC functionals by definition

solve the IC problem and we need some further criteria to differentiate between them. Various choices for such a measure include, e.g., robustness properties of the estimators or their convergence rates. We choose, however, as our criterion the limiting efficiencies of the functionals as the sample estimates of all IC functionals we later encounter can be shown to be root-n consistent with the limiting normal distributions, i.e. $\sqrt{n}\{\mathrm{vec}(\hat{\mathbf{\Gamma}} - \mathbf{\Gamma})\} \rightsquigarrow \mathcal{N}_{p^2}(\mathbf{0}, \mathbf{\Psi})$ for some limiting covariance matrix $\Psi \in \mathbb{R}_+^{p^2 \times p^2}$. The comparison between the IC functionals can then be reduced to the comparison between their limiting covariance matrices and numerous tools for computing the "size" of a square matrix exist, such as trace, $\mathrm{tr}(\cdot)$, determinant, $\det(\cdot)$, or any matrix norm, $\|\cdot\|$. Our preferred measure is closely related to the first of these but motivated in a slightly different way, namely via a connection to a measure of finite sample accuracy of an IC functional, the minimum distance index. But before we delve further into these concepts we first explore two important implications of Definition 2 into the limiting efficiencies.

Using the second property from Definition 2 we have for any invertible matrix $\mathbf{A} \in \mathbb{R}^{p \times p}$ that

$$\begin{aligned}\sqrt{n}[\mathrm{vec}\{\mathbf{\Gamma}(\hat{F}_{\mathbf{Ax}}) - \mathbf{\Gamma}(F_{\mathbf{Ax}})\}] &= \sqrt{n}[\mathrm{vec}\{\mathbf{\Gamma}(\hat{F}_{\mathbf{x}})\mathbf{A}^{-1} - \mathbf{\Gamma}(F_{\mathbf{x}})\mathbf{A}^{-1}\}] \\ &= (\mathbf{A}^{-\top} \otimes \mathbf{I}_p)\sqrt{n}[\mathrm{vec}\{\mathbf{\Gamma}(\hat{F}_{\mathbf{x}}) - \mathbf{\Gamma}(F_{\mathbf{x}})\}],\end{aligned}$$

showing that the limiting distribution of $\mathbf{\Gamma}(\hat{F}_{\mathbf{Ax}})$ is for all $\mathbf{A}$ reduced to that of $\mathbf{\Gamma}(\hat{F}_{\mathbf{x}})$. Consequently, whenever we derive the limiting distributions of the sample estimates we may without loss of generality assume that the IC model is equipped with the trivial mixing $\mathbf{\Omega} = \mathbf{I}_p$, simplifying the calculations greatly. The second effect of Definition 2 on the limiting distributions is caused by the presence of the equivalence $\equiv$ instead of an equality. Namely, as the order and the signs of the rows of $\mathbf{\Gamma}$ are not fixed, an arbitrary sequence of estimators $\hat{\mathbf{\Gamma}}_n$ does not in general converge in probability to any fixed matrix. To obtain the limiting result we thus have to choose our sequence of estimates carefully, by implicitly changing the signs and reordering the rows of $\hat{\mathbf{\Gamma}}_n$, to result into a sequence which does converge in probability to $\mathbf{I}_p$. This fact is addressed in the asymptotic results later on by explicitly saying that we can "choose a sequence of estimates" with the desired properties.

The matrix $\mathbf{\Gamma}(F_{\mathbf{x}})\mathbf{\Omega}$, where $\mathbf{\Omega}$ is the true mixing matrix, is often called the gain matrix and Definition 2 implies that it is invariant under transformations $\mathbf{x} \mapsto \mathbf{Ax}$ for any invertible $\mathbf{A} \in \mathbb{R}^{p \times p}$. In the case of a perfect separation each row of $\mathbf{\Gamma}(F_{\mathbf{x}})\mathbf{\Omega}$ must pick a single, unique element of $\mathbf{z}$ and a natural measure of the success of an IC functional is thus the distance of the gain matrix from the class of matrices $\mathcal{C}^p$.

Definition 3. *Let $\boldsymbol{x} \in \mathbb{R}^p$ come from an IC model with the mixing matrix $\mathbf{\Omega}$ and let $\mathbf{\Gamma} = \mathbf{\Gamma}(F_{\boldsymbol{x}})$ be an IC functional. The minimum distance (MD) index related to $\mathbf{\Gamma}$ is*

$$D(\mathbf{\Gamma}) = D(\mathbf{\Gamma}, \mathbf{\Omega}) = \frac{1}{\sqrt{p-1}} \inf_{\boldsymbol{C} \in \mathcal{C}^p} \|\boldsymbol{C}\mathbf{\Gamma}\mathbf{\Omega} - \boldsymbol{I}_p\|_F.$$

The MD index was introduced in Ilmonen et al. (2010b) where it was shown that $0 \leq D(\mathbf{\Gamma}) \leq 1$ with $D(\mathbf{\Gamma}) = 0$ if and only if $\mathbf{\Gamma\Omega} \in \mathcal{C}^p$. The value zero thus indicates a perfect identification of the independent components. Assuming

identity mixing, $\mathbf{\Omega} = \mathbf{I}_p$, and an IC functional with a limiting normal distribution, $\sqrt{n}\{\mathrm{vec}(\hat{\mathbf{\Gamma}} - \mathbf{I}_p)\} \rightsquigarrow \mathcal{N}_{p^2}(\mathbf{0}, \Psi)$, Ilmonen et al. (2010b) further showed that the sample MD index $D(\hat{\mathbf{\Gamma}})$ satisfies

$$n(p-1)D(\hat{\mathbf{\Gamma}})^2 = n\|\mathrm{off}(\hat{\mathbf{\Gamma}})\|_F^2 + o_p(1).$$

One consequence of the above result that is of particular interest to us is that the transformed index $n(p-1)D(\hat{\mathbf{\Gamma}})^2$ converges to a limiting distribution with the expected value $\mathrm{tr}^\circ(\mathbf{\Psi})$, the sum of the limiting variances of the off-diagonal elements of $\hat{\mathbf{\Gamma}}$. Furthermore, the limiting variances of the diagonal elements of $\hat{\mathbf{\Gamma}}$ do not depend on the choice of the IC functional but only on the distribution of $\mathbf{z}$ (and the standardization method), see Section 3.3. Thus comparing the partial traces of the limiting covariance matrices of different IC functionals within the same IC model is equivalent to comparing their traces. Because of these useful properties we base all our comparisons between different IC functionals on the partial trace $\mathrm{tr}^\circ(\mathbf{\Psi})$. A further advantage of condensing the asymptotic accuracy into a single number is that in simulations we can estimate this quantity as a mean of $n(p-1)D(\hat{\mathbf{\Gamma}})^2$ over several replications and the results can be used in checking whether our computations are correct.

Several other performance measures for IC functionals are also based on the gain matrix, see for example the Amari index (Amari et al., 1996), the interference to signal ratio (ISR) (Ollila, 2010), the inter-channel interference (ICI) (Douglas, 2007) and the review of different indices in Nordhausen et al. (2011b). These however lack the useful limiting properties possessed by the MD index.

3.3 Standardization

All the IC functionals discussed in this thesis use the same preprocessing step, multivariate standardization, known also as whitening. Define the inverse square root of a square matrix $\mathbf{S} \in \mathbb{R}^{p\times p}$ as any matrix $\mathbf{G} \in \mathbb{R}^{p\times p}$ satisfying $\mathbf{GSG}^\top = \mathbf{I}$. If the matrix $\mathbf{S}$ is symmetric positive-definite and has the eigendecomposition $\mathbf{S} = \mathbf{UDU}^\top$, the set of all inverse square roots of $\mathbf{S}$ consists precisely of the matrices $\mathbf{VD}^{-1/2}\mathbf{U}^\top$ where $\mathbf{V} \in \mathcal{U}^{p\times p}$, see Ilmonen et al. (2012). If the diagonal elements of $\mathbf{D}$ are distinct we have the unique symmetric choice $\mathbf{UD}^{-1/2}\mathbf{U}^\top$. Now, given a zero-mean random vector $\mathbf{x} \in \mathbb{R}^p$, its standardized version is defined as $\mathbf{x}_{st} = \mathbf{\Sigma}(\mathbf{x})^{-1/2}\mathbf{x}$ where $\mathbf{\Sigma}(\mathbf{x})^{-1/2}$ is a symmetric inverse square root of $\mathbf{\Sigma}(\mathbf{x})$, the covariance matrix of $\mathbf{x}$. We require, without loss of generality, the symmetry as it makes some asymptotic calculations simpler later on. Naturally $\mathbf{\Sigma}(\mathbf{x}_{st}) = \mathbf{I}_p$.

The multivariate standardization acts particularly nicely under affine transformations: for any invertible $\mathbf{A} \in \mathbb{R}^{p\times p}$ the map $\mathbf{x} \mapsto \mathbf{Ax}$ causes the transformation $\mathbf{\Sigma}_\mathbf{x}^{-1/2} \mapsto \mathbf{U}\mathbf{\Sigma}_\mathbf{x}^{-1/2}\mathbf{A}^{-1}$ where $\mathbf{U} \in \mathcal{U}^{p\times p}$ is some orthogonal matrix depending on both $\mathbf{x}$ and $\mathbf{A}$ (Ilmonen et al., 2012). We next investigate the implications of standardization to the problem of estimating the independent components.

A basic result in ICA says that if we can assume the existence of second moments then under the IC model we have the identity

$$\mathbf{x}_{st} = \mathbf{Uz}, \tag{3.5}$$

for some $\mathbf{U} \in \mathcal{U}^{p\times p}$, see Cardoso and Souloumiac (1993). Consequently, all ICA methods in the following will concentrate solely on estimating the unknown orthogonal matrix $\mathbf{U}$ and all IC functionals we consider will accordingly be of the form $\mathbf{V}(\mathbf{x}_{st})\boldsymbol{\Sigma}(\mathbf{x})^{-1/2}$, where the rotation functional $\mathbf{V}$ taking values in $\mathcal{U}^{p\times p}$ is defined only for standardized random vector. Thus choosing the functional $\mathbf{V}$ is equivalent to choosing the estimation method. While limiting our choice of IC functionals to a smaller class of functionals of a specific form is somewhat restrictive, the members of this class are easier to construct and rather well-behaved as is exemplified by the following two lemmas.

Lemma 1. *Let the functional* $\boldsymbol{\Gamma}$ *be of the form* $\boldsymbol{\Gamma}(\boldsymbol{x}) = \boldsymbol{V}(\boldsymbol{x}_{st})\boldsymbol{\Sigma}(\boldsymbol{x})^{-1/2}$ *with* $\boldsymbol{V} \in \mathcal{U}^{p\times p}$. *Then* $\boldsymbol{\Gamma}$ *is an IC functional if and only if*

i) $\boldsymbol{V}(\boldsymbol{z}_{st}) \equiv \boldsymbol{I}_p$ *for all* $\boldsymbol{z} \in \mathbb{R}^p$ *with independent components and*

ii) $\boldsymbol{V}(\boldsymbol{U}\boldsymbol{x}_{st}) \equiv \boldsymbol{V}(\boldsymbol{x}_{st})\,\boldsymbol{U}^\top$ *for all* $\boldsymbol{x} \in \mathbb{R}^p$ *and all* $\boldsymbol{U} \in \mathcal{U}^{p\times p}$.

The proof of Lemma 1 is given in the Appendix and in order to show that a functional $\boldsymbol{\Gamma}$ is an IC functional it is thus sufficient to show that the corresponding rotation functional $\mathbf{V}$ satisfies the conditions of Lemma 1. The next result shows that the standardization also fixes the asymptotic variances of the diagonal elements of $\hat{\boldsymbol{\Gamma}}$, regardless of $\mathbf{V}$. For its proof, see for example Virta et al. (2016b).

Lemma 2. *Let* $\boldsymbol{x} \in \mathbb{R}^p$ *come from an IC model with* $\boldsymbol{\Omega} = \boldsymbol{I}_p$ *and let* $\boldsymbol{\Gamma}(\boldsymbol{x}) = \boldsymbol{V}(\boldsymbol{x}_{st})\boldsymbol{\Sigma}(\boldsymbol{x})^{-1/2}$ *be an IC functional with the limiting distribution* $\sqrt{n}\{\mathrm{vec}(\hat{\boldsymbol{\Gamma}} - \boldsymbol{I}_p)\} \rightsquigarrow \mathcal{N}_{p^2}(\boldsymbol{0}, \boldsymbol{\Psi})$. *Then the diagonal elements* $\mathrm{ASV}(\gamma_{kk})$ *of* $\boldsymbol{\Psi}$ *are*

$$\mathrm{ASV}(\gamma_{kk}) = \frac{\kappa_k + 2}{4},$$

where $\kappa_k = \mathrm{E}(z_k)^4 - 3$.

The asymptotic variances of the off-diagonal elements of $\hat{\boldsymbol{\Gamma}}$ still depend on the choice of $\mathbf{V}$ and they have to be derived separately in each case. Before we get to the main topic of constructing specific IC functionals, the next section still briefly describes our basic building blocks for formulating the rotation functionals $\mathbf{V}$.

3.4 Cumulants

Our main tools for constructing IC functionals are univariate and multivariate moments, containing information both on the shapes of the marginal distributions and the dependency structure between them. For any $m \in \mathbb{N}$ the set of mth moments of a random vector $\mathbf{x} \in \mathbb{R}^p$ is the set

$$\left\{\mathrm{E}\left(x_{i_1}\cdots x_{i_m}\right) \mid i_1, \ldots i_m \in \{1, \ldots, p\}\right\}.$$

The sets of first two moments are captured conveniently by the mean vector $\mathrm{E}(\mathbf{x})$ and the shifted covariance matrix $\mathrm{Cov}(\mathbf{x}) + \mathrm{E}(\mathbf{x})\,\mathrm{E}(\mathbf{x})^\top$.

However, despite the simple form and the familiarity of moments we prefer to work with an alternative but analogous concept, cumulants, see e.g.

McCullagh (1987). In the same way as moments are defined as the coefficients of the Maclaurin series of the moment-generating function $M_{\mathbf{x}}(\mathbf{t}) = \mathrm{E}\left\{\exp\left(\mathbf{t}^{\top}\mathbf{x}\right)\right\}$, cumulants are obtained from the coefficients of the Maclaurin series of the cumulant-generating function $\log\left\{M_{\mathbf{x}}(\mathbf{t})\right\}$. We denote the mth order joint cumulant between the components $x_{i_1}, \ldots, x_{i_m}$, $i_1, \ldots i_m \in \{1, \ldots, p\}$, by $c(x_{i_1}, \ldots, x_{i_m})$, reserving the standard notation κ for the more frequently appearing excess kurtosis. In case all the indices coincide, we obtain a marginal cumulant of order m and use the shorter notation $c_m(x_i)$. Assuming that the random vector $\mathbf{x}$ has zero mean, a comparison of the two series expansions yields the following relationships between marginal cumulants and moments of low order,

$$c_2(x) = \mathrm{E}\left(x^2\right), \quad c_3(x) = \mathrm{E}\left(x^3\right), \quad c_4(x) = \mathrm{E}\left(x^4\right) - 3\left\{\mathrm{E}\left(x^2\right)\right\}^2,$$

and similarly for joint cumulants and moments,

$$\begin{aligned} c(x_{i_1}, x_{i_2}) &= \mathrm{E}(x_{i_1}x_{i_2}), \\ c(x_{i_1}, x_{i_2}, x_{i_3}) &= \mathrm{E}(x_{i_1}x_{i_2}x_{i_3}), \\ c(x_{i_1}, x_{i_2}, x_{i_3}, x_{i_4}) &= \mathrm{E}(x_{i_1}x_{i_2}x_{i_3}x_{i_4}) - \mathrm{E}(x_{i_1}x_{i_3})\,\mathrm{E}(x_{i_2}x_{i_4}) \\ &\quad - \mathrm{E}(x_{i_1}x_{i_4})\,\mathrm{E}(x_{i_2}x_{i_3}) - \mathrm{E}(x_{i_1}x_{i_2})\,\mathrm{E}(x_{i_3}x_{i_4}). \end{aligned} \tag{3.6}$$

The main reason for preferring cumulants over moments is the number of useful properties they have when working with independent random variables and affine transformations. We next list the most important of these in the form of a lemma, see McCullagh (1987).

Lemma 3. *1. If $x_{i_1}, \ldots, x_{i_s}$ and $x_{j_1}, \ldots, x_{j_t}$ are two independent collections of random variables (the random variables within the same collection may still be dependent) then any cumulant $c(\cdot)$ involving random variables from both collections is zero.*

2. Cumulants are homogeneous of first degree in every component, i.e. for $a_1, \ldots, a_m \in \mathbb{R}$ we have $c(a_1x_{i_1}, \ldots, a_mx_{i_m}) = (\prod_{k=1}^{m} a_k)\, c(x_{i_1}, \ldots, x_{i_m})$. As a special case, marginal cumulants of order m are homogeneous of order m.

3. The marginal cumulants are additive under independence, i.e. if $x_{i_1}, \ldots, x_{i_s}$ are mutually independent then we have for marginal cumulants of any order, $c_m\left(\sum_{k=1}^{s} x_{i_k}\right) = \sum_{k=1}^{s} c_m(x_{i_k})$.

4. The marginal first cumulant is shift-equivariant, $c_1(x_i + b) = c_1(x_i) + b$, and marginal cumulants of higher order are shift-invariant, $c_m(x_i + b) = c_m(x_i)$, for all $b \in \mathbb{R}$.

5. Marginal cumulants of order three and higher vanish for the normal distribution.

Apart from being computationally useful, the properties listed in Lemma 3 lead to some useful interpretations for joint and marginal cumulants. For example, the first property implies that joint cumulants measure the level of dependence between its argument random variables and the final property suggests that marginal cumulants in some sense measure departure from normality.

The number of marginal cumulants of order m is simply m, but the number of joint cumulants of order m quickly grows with m. Therefore it is important to have some more convenient means of handling the collections of cumulants. The simplest one is to arrange the joint cumulants of order m into a tensor $\mathcal{K}^m = \mathcal{K}^m(\mathbf{x}) \in \mathbb{R}^{p \times \cdots \times p}$ of order m so that $(\mathcal{K}^m)_{i_1,\ldots,i_m} = c(x_{1_1}, \ldots, x_{i_m})$. Note that $\mathcal{K}^m$ is symmetric with respect to permutation of its indices meaning that the majority of its elements repeat and the only unique elements are the marginal cumulants on the super-diagonal $(\mathcal{K}^m)_{i,\ldots,i}$. For example, $\mathcal{K}^2$ is equal to the ordinary covariance matrix which is naturally symmetric. The tensorial way of thinking has some useful properties, e.g. under the linear transformation $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$ the mth order cumulant tensor transforms as

$$\mathcal{K}^m \mapsto \mathcal{K}^m \times_{k=1}^m \mathbf{A},$$

where $\times_{k=1}^m$ is the tensor-by-matrix multiplication, see the introduction to tensors in Section 4. Also, if $\mathbf{x}$ has independent components the cumulant tensors $\mathcal{K}^m$, $m \geq 2$ are all super-diagonal. However, the methods we later consider never go beyond fourth cumulants and for our purposes it is actually more beneficial to collect the joint cumulants into matrices.

We start by observing that the last row of (3.6) can be written as

$$\begin{aligned} c(x_{i_1}, x_{i_2}, x_{i_3}, x_{i_4}) = &\operatorname{E}(x_{i_1} x_{i_2} x_{i_3} x_{i_4}) - \operatorname{E}\left(x_{i_1} x_{i_2}^* x_{i_3} x_{i_4}^*\right) \\ &- \operatorname{E}\left(x_{i_1} x_{i_2}^* x_{i_3}^* x_{i_4}\right) - \operatorname{E}\left(x_{i_1} x_{i_2} x_{i_3}^* x_{i_4}^*\right), \end{aligned}$$

where $\mathbf{x}^*$ is an independent copy of the random vector $\mathbf{x} \in \mathbb{R}^p$. Fixing now the first two indices, $i_1 = i$, $i_2 = j$, the obtained p^2 fourth cumulants are captured by the $p \times p$ matrix

$$\begin{aligned} \mathbf{C}^{ij}(\mathbf{x}) = &\operatorname{E}\left(x_i x_j \cdot \mathbf{x}\mathbf{x}^\top\right) - \operatorname{E}\left(x_i x_j^* \cdot \mathbf{x}\mathbf{x}^{*\top}\right) \\ &- \operatorname{E}\left(x_i x_j^* \cdot \mathbf{x}^*\mathbf{x}^\top\right) - \operatorname{E}\left(x_i x_j \cdot \mathbf{x}^*\mathbf{x}^{*\top}\right), \end{aligned} \tag{3.7}$$

and the family of matrices $\{\mathbf{C}^{ij}(\mathbf{x}) \mid i, j \in \{1, \ldots, p\}\}$ collects (with some repetition) all fourth joint cumulants of the random vector $\mathbf{x} \in \mathbb{R}^p$. Assuming that $\mathbf{x}$ is standardized, $\operatorname{E}(\mathbf{x}) = \mathbf{0}_p, \operatorname{E}\left(\mathbf{x}\mathbf{x}^\top\right) = \mathbf{I}_p$, we still have the simpler form

$$\mathbf{C}^{ij}(\mathbf{x}) = \operatorname{E}\left(x_i x_j \cdot \mathbf{x}\mathbf{x}^\top\right) - \mathbf{E}^{ij} - \mathbf{E}^{ji} - \delta_{ij}\mathbf{I}_p. \tag{3.8}$$

Property 1 in Lemma 3 further guarantees that if the components of the standardized $\mathbf{x}$ are also independent then the only non-zero elements in the whole collection of matrices $\mathbf{C}^{ij}(\mathbf{x})$ are the ith diagonal elements of $\mathbf{C}^{ii}(\mathbf{x})$, $i \in \{1, \ldots, p\}$, which equal respectively the marginal kurtoses of the p components, see e.g. Miettinen et al. (2015). This fact will be exploited later in the thesis to craft two classical IC functionals.

However, before using the set of fourth cumulants in its entirety we first investigate two IC functionals that can be constructed by considering only the marginal cumulants in the super-diagonals of the tensors $\mathcal{K}^m$. We are again especially interested in fourth marginal cumulants and introduce the following short-hand notation for some specific cumulants and moments of a standardized random vector $\mathbf{x} = (x_1, \ldots x_p)^\top$,

$$\beta_k(\mathbf{x}) = \operatorname{E}\left(x_k^4\right), \quad \kappa_k(\mathbf{x}) = \beta_k(\mathbf{x}) - 3, \quad \omega_k(\mathbf{x}) = \operatorname{E}\left(x_k^6\right) - \left\{\operatorname{E}\left(x_k^3\right)\right\}^2.$$

If the random vector $\mathbf{x}$ is clear from the context we omit the parenthesis and simply use β_k, κ_k etc. The fourth moment β_k and the fourth cumulant κ_k are conventionally called the kurtosis and excess kurtosis of the random variable x_k, respectively.

3.5 IC functionals based on marginal cumulants

We begin by introducing two lemmas which suggest that specific functions of marginal cumulants are maximized if and only if a random vector has independent components, i.e., in the "solution" of the IC model. Their proofs are given in the Appendix.

Lemma 4. *Let $\boldsymbol{z} \in \mathbb{R}^p$ have independent components and fix $m \in \mathbb{N}, m \geq 3$. Assume further that $\{c_m(z_1)\}^2 \geq \cdots \geq \{c_m(z_p)\}^2$ and that $c_m(z_k) = 0$ for at most one value of $k \in \{1, \ldots, p\}$. Then, for a fixed $k \in \{1, \ldots, p-1\}$, we have for all $\boldsymbol{v}_k \in \mathbb{S}^{p-1}$ with $\boldsymbol{v}_k^\top \boldsymbol{e}_l = 0$, $l \in \{1, \ldots, k-1\}$,*

$$\left\{c_m(\boldsymbol{v}_k^\top \boldsymbol{z})\right\}^2 \leq \{c_m(z_k)\}^2,$$

with equality if and only if $\boldsymbol{v}_k = \pm \boldsymbol{e}_l$ for some $l \in \{k, \ldots, p\}$ with $\{c_m(z_l)\}^2 = \{c_m(z_k)\}^2$.

Lemma 5. *Let $\boldsymbol{z} \in \mathbb{R}^p$ have independent components and let $m \in \mathbb{N}, m \geq 3$. Assume further that $c_m(z_k) = 0$ for at most one value of $k \in \{1, \ldots, p\}$. Then we have for all orthogonal matrices $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p) \in \mathcal{U}^p$,*

$$\sum_{k=1}^{p} \left\{c_m\left(\boldsymbol{v}_k^\top \boldsymbol{z}\right)\right\}^2 \leq \sum_{k=1}^{p} \{c_m(z_k)\}^2,$$

with equality if and only if $\boldsymbol{V}^\top \equiv \boldsymbol{I}_p$.

Lemmas 4 and 5 imply two specific optimization problems for constructing a rotation functional $\mathbf{V}$. In the first one we sequentially search for mutually orthogonal projection directions, maximizing the value of the squared mth marginal cumulant on each step. The components are then found one-by-one in decreasing order corresponding to the values of $\{c_m(z_k)\}^2$. The final conditions for reaching an equality in Lemma 4 are required to accommodate for a case where two or more components share the same non-zero value of the squared cumulant and we can find either of them in a particular step. In the second optimization problem we find all p mutually orthogonal projections at once, maximizing the sum of the squared mth marginal cumulants of the projections. These ideas are formalized in Definitions 4 and 5 below. The inequalities guarantee the consistencies of both these approaches in the sense of condition *i)* in Lemma 1: the values of the objective functions decrease for any non-trivial rotation of a vector of independent components $\mathbf{z}$, assuming that at most one of the independent components has value zero for the chosen cumulant c_m. Based on the final part of Lemma 3 this assumption now replaces the weaker Assumption V3 in the IC model.

In Virta et al. (2016b) equivalent results to those in Lemmas 4 and 5 are given for convex combinations of third and fourth cumulants in the case where we

allow multiple normally distributed independent components. However, as the standard case with only a single cumulant and the basic IC model in Definition 1 better serves instructive purposes we choose to formulate everything under it. The extensions in Virta et al. (2016b) then follow rather straightforwardly with some minor tweaks. Similar results to Lemma 5 can be given also when the second powers are replaced with any qth absolute power $|\cdot|^q$, $q \leq 1$ but as far as we know only the cases $q = 1, 2$ have been considered in the literature, see Miettinen et al. (2015) for the former.

Definition 4. *Let $\boldsymbol{x} \in \mathbb{R}^p$ and fix $m \geq 3$. Then the deflation-based projection pursuit (DPP) functional is $\boldsymbol{\Gamma}^D = \boldsymbol{\Gamma}^D(\boldsymbol{x}) = \boldsymbol{V}\boldsymbol{\Sigma}(\boldsymbol{x})^{-1/2} \in \mathbb{R}^{p\times p}$ where the kth row of the rotation functional $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p)^\top \in \mathcal{U}^{p\times p}$ is found as*

$$\boldsymbol{v}_k = \operatorname{argmax} \left\{ c_m \left(\boldsymbol{v}_k^\top \boldsymbol{x}_{st} \right) \right\}^2,$$

subject to $\boldsymbol{v}_k^\top \boldsymbol{v}_l = \delta_{kl}$ for all $l \in \{1, \ldots, k\}$.

Definition 5. *Let $\boldsymbol{x} \in \mathbb{R}^p$ and fix $m \geq 3$. Then the symmetric projection pursuit (SPP) functional is $\boldsymbol{\Gamma}^S = \boldsymbol{\Gamma}^S(\boldsymbol{x}) = \boldsymbol{V}\boldsymbol{\Sigma}(\boldsymbol{x})^{-1/2} \in \mathbb{R}^{p\times p}$ where the rotation functional $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p)^\top \in \mathcal{U}^{p\times p}$ is found as*

$$\boldsymbol{V} = \operatorname{argmax} \sum_{k=1}^{p} \left\{ c_m \left(\boldsymbol{v}_k^\top \boldsymbol{x}_{st} \right) \right\}^2,$$

subject to $\boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{I}_p$.

The functionals in Definitions 4 and 5 are easily seen to be IC functionals: The first condition of Lemma 1 is satisfied by our earlier discussion and the second condition follows by noting that the optimization problems in Definitions 4 and 5 are equivariant under transformations $\mathbf{x}_{st} \mapsto \mathbf{U}\mathbf{x}_{st}$ where $\mathbf{U} \in \mathcal{U}^{p\times p}$.

The names deflation-based and symmetric come from the signal processing literature where the previous algorithms have the squares replaced with absolute values and go under the names of deflation-based FastICA and symmetric FastICA, see Hyvärinen (1999). In FastICA it is common to use also non-cumulant-based objective functions G, such as the logarithmic hyperbolic cosine, $G(x) = \log\{cosh(x)\}$, or the Gaussian function, $G(x) = \exp(-x^2/x)$. However, neither of these functions satisfies inequalities such as those in Lemmas 4 and 5 and consequently the obtained methods do not provide consistent solutions to the IC model, see Wei (2014). However, they have other useful properties that make them still valuable in practice, see Virta and Nordhausen (2017c).

For solving the two optimization problems in Definitions 4 and 5 the technique of Lagrangian multipliers can be used to obtain sets of fixed-point equations, which in turn lead to corresponding fixed-point algorithms. We refrain from listing these results here as equivalent ones can be found in Virta et al. (2016b) and in numerous papers discussing FastICA (Hyvärinen, 1999; Miettinen et al., 2017). What is more, the fixed-point equations also give us a way of finding the limiting distributions and consequently the asymptotic variances of the two IC functionals. For brevity, we have in the following presented these only in the case $m = 4$, the fourth cumulants. This both simplifies the notation and provides an even ground for comparing the projection pursuit methods to

the IC functionals obtained with fourth joint cumulants in the next section. See Virta et al. (2016b); Miettinen et al. (2017) for more general results and Koldovský et al. (2006); Miettinen et al. (2014a) for using different objective functions for extracting different independent components.

Before the results we still state the assumption on the maximally one vanishing fourth cumulant for easy future reference.

Assumption V4. *At most one of the fourth cumulants κ_k, $k \in \{1, \dots, p\}$, of the independent components is zero.*

Theorem 1. *Let $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ be a random sample having finite eighth moments from an IC model satisfying Assumption V4 and let $\boldsymbol{\Omega} = \boldsymbol{I}_p$. Then for $m = 4$ there exists a sequence of DPP functionals $\hat{\boldsymbol{\Gamma}}^D$ such that $\sqrt{n}\{\mathrm{vec}(\hat{\boldsymbol{\Gamma}}^D - \boldsymbol{I}_p)\} \rightsquigarrow \mathcal{N}_{p^2}(\boldsymbol{0}, \boldsymbol{\Psi})$ where the diagonal elements $\mathrm{ASV}(\gamma_{kl})$, $k \neq l$, of $\boldsymbol{\Psi}$ are*

$$\mathrm{ASV}(\gamma_{kl}) = \frac{\omega_k - \beta_k^2}{\kappa_k^2}, \qquad k < l,$$

$$\mathrm{ASV}(\gamma_{kl}) = \frac{\omega_l - \beta_l^2}{\kappa_l^2} + 1, \qquad k > l.$$

Interestingly the asymptotic variances of the elements of the DPP functional depend on the estimation order of the components and although we have fixed this order in Lemmas 4 and 5 it can happen that this order is compromised in practice (due to local maxima or bad initial values etc.). This behavior is exploited in Nordhausen et al. (2011a) where the extraction order of the independent components is forced to be the one which minimizes the asymptotic variances.

Theorem 2. *Let $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ be a random sample having finite eighth moments from an IC model satisfying Assumption V4 and let $\boldsymbol{\Omega} = \boldsymbol{I}_p$. Then for $m = 4$ there exists a sequence of SPP functionals $\hat{\boldsymbol{\Gamma}}^S$ such that $\sqrt{n}\{\mathrm{vec}(\hat{\boldsymbol{\Gamma}}^S - \boldsymbol{I}_p)\} \rightsquigarrow \mathcal{N}_{p^2}(\boldsymbol{0}, \boldsymbol{\Psi})$ where the diagonal elements $\mathrm{ASV}(\gamma_{kl})$, $k \neq l$, of $\boldsymbol{\Psi}$ are*

$$\mathrm{ASV}(\gamma_{kl}) = \frac{\kappa_k^2(\omega_k - \beta_k^2) + \kappa_l^2(\omega_l - \beta_l^2) + \kappa_l^4}{\left(\kappa_k^2 + \kappa_l^2\right)^2}.$$

The results of Theorems 1 and 2 can now be used to conduct asymptotic comparisons between the DPP and SPP functionals. The asymptotic variances have such differing forms that no simple analytic results can be given but the values of the partial trace $\mathrm{tr}^\circ(\boldsymbol{\Psi})$ can still be computed for specific collections of distributions for the independent components. This has been done in Miettinen et al. (2015); Virta et al. (2016b); Miettinen et al. (2017) with the observation that the SPP functional generally outperforms the DPP functional.

3.6 IC functionals based on joint cumulants

We next turn our attention to joint fourth cumulants and two classical IC functionals based on them. Recall from Section 3 the matrices $\mathbf{C}^{ij}(\mathbf{x})$ collecting all fourth joint cumulants of the standardized random vector $\mathbf{x}$. As p^2, the number of the matrices, grows quite fast with the dimension p it seems reasonable to

consider only a subset of them, ranked using some measure of importance. The matrices $\mathbf{C}^{ii}(\mathbf{x})$, $i = 1, \dots p$, with the repeated index stand out as the foremost ones, containing also the marginal fourth cumulants. To even further condense the information content in these p matrices we take their sum to obtain the matrix used in fourth order blind identification (FOBI) (Cardoso, 1989),

$$\mathbf{C}(\mathbf{x}) = \sum_{i=1}^{p} \mathbf{C}^{ii}(\mathbf{x}) = \mathrm{E}\left(\|\mathbf{x}\|_F^2 \mathbf{x}\mathbf{x}^\top\right) - 3\mathbf{I}_p = \mathrm{E}\left(\mathbf{x}\mathbf{x}^\top \mathbf{x}\mathbf{x}^\top\right) - 3\mathbf{I}_p.$$

FOBI is one of the first methods that can solve the IC problem and using the above FOBI-matrix we define in a minute the corresponding FOBI functional. The main property of the FOBI-matrix that makes it useful to us is that it is diagonal for any random vector with independent components, $\mathbf{C}(\mathbf{z}) = \sum_{k=1}^{p}(\kappa_k + p + 2)\mathbf{E}^{kk}$. This behavior is in the context of scatter matrices called the independence property, see Nordhausen and Tyler (2015); Virta (2016). Replacing $\mathbf{C}$ in the following with any other scatter matrix with the same property would not compromise the method in any way and would lead into collections of new IC functionals, see Nordhausen et al. (2008).

Definition 6. *Let $\boldsymbol{x} \in \mathbb{R}^p$. Then the FOBI-functional is $\boldsymbol{\Gamma}^F = \boldsymbol{\Gamma}^F(\boldsymbol{x}) = \boldsymbol{V}\boldsymbol{\Sigma}(\boldsymbol{x})^{-1/2} \in \mathbb{R}^{p\times p}$ where the rotation functional $\boldsymbol{V} = (\boldsymbol{v}_1, \dots, \boldsymbol{v}_p)^\top \in \mathcal{U}^{p\times p}$ contains the eigenvectors of the matrix*

$$\boldsymbol{C}(\boldsymbol{x}_{st}) = \mathrm{E}\left(\boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{x}\boldsymbol{x}^\top\right) - 3\boldsymbol{I}_p.$$

as its rows in decreasing order according to the corresponding eigenvalues.

To prove that the FOBI-functional is an actual IC functional we again go through the two conditions in Lemma 1. The first one clearly holds if and only if we can identify the eigenbasis of the diagonal matrix $\sum_{k=1}^{p}(\kappa_k + p + 2)\mathbf{E}^{kk}$ uniquely up to the equivalence class $\mathcal{C}^p$. This is guaranteed under the following assumption.

Assumption V5. *The fourth cumulants κ_k, $k \in \{1, \dots, p\}$, of the independent components are distinct.*

Assumption V5 is stronger than Assumption V4 and thus FOBI makes the strictest assumptions of all ICA methods we have seen thus far. Curiously, this does not mean that FOBI would perform better than the projection pursuits when the assumption is satisfied, and the situation is actually quite the opposite, see the discussion after the limiting variances below. The second condition of Lemma 1 is satisfied by the basic properties of the eigendecomposition after noticing the orthogonal equivariance, $\mathbf{C}(\mathbf{U}\mathbf{x}) = \mathbf{U}\mathbf{C}(\mathbf{x})\mathbf{U}^\top$, making $\boldsymbol{\Gamma}^F$ an IC functional.

The above discussion already hints that compressing the cumulant matrices $\mathbf{C}^{ij}$ into the single FOBI-matrix might not have been the best idea and we next seek ways to incorporate all p^2 matrices into the estimation of the rotation functional. Evaluating a single cumulant matrix on the standardized observation $\mathbf{x}_{st} = \mathbf{U}\mathbf{z}$ yields

$$\mathbf{C}^{ij}(\mathbf{U}\mathbf{z}) = \mathbf{U}\left(\sum_{k=1}^{p} u_{ki}u_{kj}\kappa_k \mathbf{E}^{kk}\right)\mathbf{U}^\top,$$

showing that each $\mathbf{C}^{ij}(\mathbf{x}_{st})$ has $\mathbf{U}$ as (one possible set of) its eigenvectors. The ranks of the matrices depend on the unknown $\mathbf{U}$ and thus it is difficult to say whether $\mathbf{U}$ can be identified using a single matrix $\mathbf{C}^{ij}$ only. Our second classical ICA method, joint diagonalization of eigen-matrices (JADE) (Cardoso and Souloumiac, 1993), bypasses this obstacle by as per its name jointly diagonalizing the whole set of cumulant matrices $\mathbf{C}^{ij}(\mathbf{x}_{st})$, $i, j \in \{1, \ldots, p\}$.

To formulate the corresponding IC functional we define the joint diagonalizer of a set of matrices $\mathcal{S} = \{\mathbf{S}_j \mid j \in \{1, \ldots, m\}\}$ as the orthogonal matrix $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_p)^\top \in \mathcal{U}^{p \times p}$ satisfying

$$\mathbf{V} = \operatorname{argmax} \sum_{j=1}^{m} \| \operatorname{diag} \left(\mathbf{V} \mathbf{S}_j \mathbf{V}^\top \right) \|_F^2, \tag{3.9}$$

and with the rows $\mathbf{v}_k$ ordered in decreasing order with respect to the "eigenvalues" $\sum_{j=1}^{m} \mathbf{v}_k^\top \mathbf{S}_j \mathbf{v}_k$, $k \in \{1, \ldots, p\}$. The Frobenius norm of a matrix is invariant under orthogonal transformations from both sides and some simple matrix algebra shows that the above maximization problem is equivalent to minimizing the sum of the squared off-diagonal elements of the matrices, justifying calling the procedure "diagonalization". If the matrices $\mathbf{S}_j$ are diagonal, as is the case with our population level $\mathbf{C}^{ij}(\mathbf{z})$, the joint diagonalizer can be shown to be uniquely determined up to signs and order if, for all pairs (k, l), there exists $j \in \{1, \ldots, m\}$ such that the eigenvalues $\mathbf{v}_k^\top \mathbf{S}_j \mathbf{v}_k$ and $\mathbf{v}_l^\top \mathbf{S}_j \mathbf{v}_l$ are distinct, see Belouchrani et al. (1997).

Definition 7. *Let $\boldsymbol{x} \in \mathbb{R}^p$. Then the JADE functional is $\boldsymbol{\Gamma}^J = \boldsymbol{\Gamma}^J(\boldsymbol{x}) = \boldsymbol{V}\boldsymbol{\Sigma}(\boldsymbol{x})^{-1/2} \in \mathbb{R}^{p \times p}$ where the rotation functional $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p)^\top \in \mathcal{U}^{p \times p}$ is the joint diagonalizer of the set of matrices,*

$$\left\{ \boldsymbol{C}^{ij}(\boldsymbol{x}_{st}) \mid i, j \in \{1, \ldots, p\} \right\}.$$

That the functional $\boldsymbol{\Gamma}^J$ is a true IC functional is somewhat tricky to show, see Miettinen et al. (2015) for a proof, and a sufficient condition for this is Assumption V4. Thus from an assumption point of view, DPP, SPP and JADE start from an even ground while FOBI alone requires more strict conditions for its Fisher consistency. To estimate the JADE-functional in practice the technique of Lagrangian multipliers can be used on the optimization problem (3.9) to obtain a set of fixed-point equations, from which the limiting distributions below are also derived. However, the resulting algorithm is computationally slow and a faster choice is, for example, the Jacobi rotation algorithm (Clarkson, 1988; Belouchrani et al., 1997). Miettinen et al. (2015); Illner et al. (2015) remarked that based on empirical testing both algorithms always yield the same results, justifying the replacement. However, without formal proof the limiting distributions of $\hat{\boldsymbol{\Gamma}}^J$ given next in Theorem 4 still apply only to the estimate computed with the fixed-point algorithm. The proofs of the following asymptotic results can be found in Ilmonen et al. (2010a); Miettinen et al. (2015), see also Bonhomme and Robin (2009); Virta et al. (2015) for similar results.

Theorem 3. *Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be a random sample having finite eighth moments from an IC model satisfying Assumption V5 and let $\boldsymbol{\Omega} = \boldsymbol{I}_p$. Then there exists a sequence of FOBI functionals $\hat{\boldsymbol{\Gamma}}^F$ such that $\sqrt{n}\{\operatorname{vec}(\hat{\boldsymbol{\Gamma}}^F - \boldsymbol{I}_p)\} \rightsquigarrow \mathcal{N}_{p^2}(\boldsymbol{0}, \boldsymbol{\Psi})$*

where the diagonal elements $\mathrm{ASV}(\gamma_{kl})$, $k \neq l$, *of* $\boldsymbol{\Psi}$ *are*

$$\mathrm{ASV}(\gamma_{kl}) = \frac{\omega_k + \omega_l - \beta_k^2 - 6\kappa_l - 9 + \sum_{t \neq k,l}^{p}(\kappa_t + 2)}{(\kappa_k - \kappa_l)^2}.$$

Theorem 4. *Let* $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ *be a random sample having finite eighth moments from an IC model satisfying Assumption V4 and let* $\boldsymbol{\Omega} = \boldsymbol{I}_p$. *Then there exists a sequence of JADE functionals* $\hat{\boldsymbol{\Gamma}}^J$ *such that* $\sqrt{n}\{\mathrm{vec}(\hat{\boldsymbol{\Gamma}}^J - \boldsymbol{I}_p)\} \rightsquigarrow \mathcal{N}_{p^2}(\boldsymbol{0}, \boldsymbol{\Psi})$ *where* $\boldsymbol{\Psi}$ *is as in Theorem 2.*

Again, the results of Theorems 3 and 4 can be used for the asymptotic comparison of FOBI and JADE (and the two projection pursuit functionals), see Miettinen et al. (2015); Virta et al. (2015). The main implication of the comparisons was that regardless of the distributions of the independent components FOBI generally underperforms both the projection pursuit methods and JADE, which is based on Theorem 4 actually asymptotically equivalent to SPP. However, JADE is computationally much more intensive than FOBI and apart from replacing JADE with SPP another way to speed it up would be to diagonalize only a subset of the fourth joint cumulant matrices (Miettinen et al., 2013).

We close the section by giving a heuristic argument for why the limiting distribution of the SPP functional (with fourth cumulants) and JADE-functional should be the same. Recall that the fourth kurtosis tensor $\boldsymbol{\mathcal{K}}^4(\mathbf{x}) \in \mathbb{R}^{p \times p \times p \times p}$ collects all p^4 fourth joint cumulants of the random vectors $\mathbf{x}$. Both SPP and JADE can now be considered as forms of "diagonalization" of $\boldsymbol{\mathcal{K}}^4(\mathbf{x}_{st})$: SPP tries to rotate the data so that the marginal cumulants are maximized which is equivalent to maximizing the diagonal of $\boldsymbol{\mathcal{K}}^4(\mathbf{x}_{st})$. JADE wants to diagonalize the p^2 cumulant matrices which is equivalent to maximizing the diagonal of $\boldsymbol{\mathcal{K}}^4(\mathbf{x}_{st})$ AND some additional off-diagonal elements. However, as in the solution $\mathbf{z}$ the kurtosis tensor $\boldsymbol{\mathcal{K}}^4(\mathbf{z})$ is diagonal these extra elements contribute asymptotically nothing to the estimation, making the result of Theorem 4 somewhat expected.

4 Independent component analysis for tensor-valued data

4.1 Tensor notation

We next move to our first generalization of independent component analysis where both the observations and the latent random variables are assumed to be tensor-valued. To prepare for the theoretical derivations and to gain some visual intuition, this section briefly goes through the main aspects of multilinear algebra essential to our treatise. The used notation follows the standard one found, for example, in De Lathauwer et al. (2000); Kolda and Bader (2009), where more detailed discussions of the following concepts can be found. References to the theory of tensor-valued random variables will be given in the next section when we discuss different models for tensor-valued data.

Let $\mathcal{X} = (x_{i_1 \dots i_r}) \in \mathbb{R}^{p_1 \times \cdots \times p_r}$ be a tensor of rth order. The r "directions" from which we can look at $\mathcal{X}$ are called the *modes* or *ways* of $\mathcal{X}$. For example, a matrix $\mathbf{X} \in \mathbb{R}^{p_1 \times p_2}$ has two modes, 1-mode and 2-mode, corresponding respectively to its columns and rows and when the order $r > 3$ the visualization gets more difficult. As the number of elements in a tensor of high order is generally quite large and tracking all the elements via their positions in the tensor gets quickly quite awkward, two alternative ways of splitting a tensor into smaller pieces are commonly used.

The first is equivalent to dividing a matrix $\mathbf{X}$ either into a collection of columns or a collection of rows, called in the tensor context the 1-mode vectors and 2-mode vectors of $\mathbf{X}$, respectively. In Figure 4.1 we have visualized this decomposition in the case of a tensors of third order. More generally, the collection of m-mode vectors of a tensor $\mathcal{X}$ is obtained when we fix the $r-1$ other indices and vary the value of the mth index to produce a total of $\rho_m = p_1 \cdots p_r / p_m$ vectors of length p_m. If we further collect these ρ_m vectors in some pre-defined order into a matrix $\mathbf{X}_{(m)}$ we obtain what is called the m-mode matricization or m-mode unfolding of $\mathcal{X}$. The actual ordering of the vectors is irrelevant to our causes as long as it stays consistent and one simple option is the cyclical ordering suggested in De Lathauwer et al. (2000). The matricization is a particularly useful concept as it allows us to reduce the derivations of the Fisher consistencies and limiting distributions of the methods into the simplest, non-vector case of tensor-valued observations, matrices. The second division of a tensor $\mathcal{X}$ into several smaller components does the opposite to the previous and fixes the value of the mth index and lets the other $r-1$ indices vary, producing a total of p_m m-mode faces, tensors of size $p_1 \times \cdots \times p_{m-1} \times p_{m+1} \times \cdots \times p_r$. The 1-mode faces and 2-mode faces of a matrix are again its rows and columns but for tensor or order larger than two this no longer holds true. Figure 4.2 illustrates the situation for a tensor of third order. For simplicity we refrain from introducing a notation for the m-mode faces as they will only be used briefly in the context of assumptions.

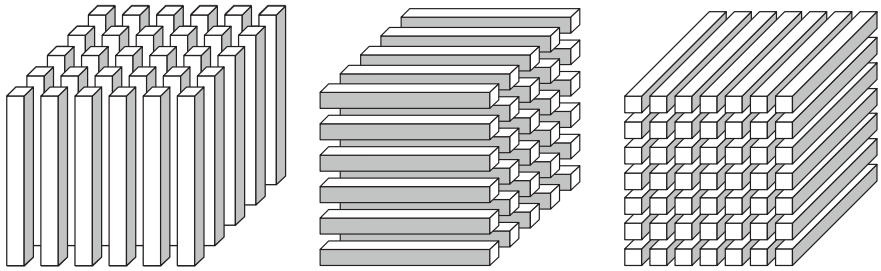

Figure 4.1: The 1-mode, 2-mode and 3-mode vectors of a tensor of third order.

To operate our tensor observations we introduce a specific group of linear transformations of a tensor by matrix. For a tensor $\mathcal{X} \in \mathbb{R}^{p_1 \times \cdots \times p_r}$ and a matrix $\mathbf{A}_m \in \mathbb{R}^{q_m \times p_m}$ we define the m-mode multiplication of $\mathcal{X}$ by $\mathbf{A}$ to be the tensor $\mathcal{X} \times_m \mathbf{A}_m \in \mathbb{R}^{p_1 \times \cdots \times p_{m-1} \times q_m \times p_{m+1} \times \cdots \times p_r}$ with the elements,

$$(\mathcal{X} \times_m \mathbf{A}_m)_{i_1 \ldots i_r} = \sum_{j_m=1}^{p_m} a_{i_m j_m} x_{i_1 \ldots i_{m-1} j_m i_{m+1} \ldots i_r}.$$

An equivalent and more accessible definition can be stated using m-mode vectors: $\mathcal{X} \times_m \mathbf{A}_m$ is the tensor obtained by pre-multiplying each m-mode vector of $\mathcal{X}$ by $\mathbf{A}_m$. The operation can be understood as a linear transformation from the mth direction and is a higher order analogue for the ordinary linear transformation of a vector by a matrix, $\mathbf{x} \mapsto \mathbf{A}_1 \mathbf{x}$. Also the second order cases can be written using basic linear algebra, $\mathbf{X} \mapsto \mathbf{A}_1 \mathbf{X}$ and $\mathbf{X} \mapsto \mathbf{X} \mathbf{A}_2^\top$.

The transformation operation $\times_m$ is associative regardless of the mode m and for two distinct values of m it is also commutative, $\mathbf{X} \times_m \mathbf{A}_m \times_{m'} \mathbf{A}_{m'} = \mathbf{X} \times_{m'} \mathbf{A}_{m'} \times_m \mathbf{A}_m$, for all $m \neq m'$. For two multiplications from the same mode we instead have $\mathbf{X} \times_m \mathbf{A}_m \times_m \mathbf{B}_m = \mathbf{X} \times_m (\mathbf{B}_m \mathbf{A}_m)$. We regularly apply transformations simultaneously from all r modes and introduce the shorthand notation $\mathcal{X} \times_{m=1}^{r} \mathbf{A}_m = \mathcal{X} \times_1 \mathbf{A}_1 \cdots \times_r \mathbf{A}_r$. The simultaneous linear transformation acts particularly nicely under vectorization and matricization. The map $\mathcal{X} \mapsto \mathcal{X} \times_{m=1}^{r} \mathbf{A}_m$ induces the maps $\mathbf{X}_{(m)} \mapsto \mathbf{A}_m \mathbf{X}_{(m)} (\mathbf{A}_{m+1} \otimes \cdots \otimes \mathbf{A}_r \otimes \mathbf{A}_1 \otimes \cdots \otimes \mathbf{A}_{m-1})^\top$ and $\operatorname{vec}(\mathcal{X}) \mapsto (\mathbf{A}_r \otimes \cdots \otimes \mathbf{A}_1) \operatorname{vec}(\mathcal{X})$.

We also introduce a second type of tensor multiplication which takes two tensors, $\mathcal{X} \in \mathbb{R}^{p_1 \times \cdots \times p_{m-1} \times p_m \times p_{m+1} \times \cdots \times p_r}$ and $\mathcal{Y} \in \mathbb{R}^{p_1 \times \cdots \times p_{m-1} \times q_m \times p_{m+1} \times \cdots \times p_r}$, with all modes of equal length except possibly the mth one. The m-mode product of $\mathcal{X}$ and $\mathcal{Y}$ is defined as the matrix $\mathcal{X} \times_{-m} \mathcal{Y} \in \mathbb{R}^{p_m \times q_m}$ with the elements

$$(\mathcal{X} \times_{-m} \mathcal{Y})_{kl} = \sum_{i_1, \ldots, i_{m-1}, i_{m+1}, \ldots, i_r} x_{i_1 \ldots i_{m-1} k i_{m+1} \ldots i_r} y_{i_1 \ldots i_{m-1} l i_{m+1} \ldots i_r}.$$

This operation too has a less intimidating representation using the m-mode vectors: $\mathcal{X} \times_{-m} \mathcal{Y}$ is the sum of all outer products between the corresponding m-mode vectors of $\mathcal{X}$ and $\mathcal{Y}$. We again turn to examples of low order to gain more intuition. For vectors $\mathbf{x} \in \mathbb{R}^{p_1}$, $\mathbf{y} \in \mathbb{R}^{q_1}$ the product is simply $\mathbf{x} \otimes_1 \mathbf{y} = \mathbf{x} \mathbf{y}^\top$ and for matrices $\mathbf{X} \in \mathbb{R}^{p_1 \times p_2}$, $\mathbf{Y} \in \mathbb{R}^{p_1 \times p_2}$ we have $\mathbf{X} \otimes_1 \mathbf{Y} = \mathbf{X} \mathbf{Y}^\top$ and $\mathbf{X} \otimes_2 \mathbf{Y} = \mathbf{X}^\top \mathbf{Y}$. Using the matricization we also have yet another representation for $\times_{-m}$, namely, $\mathcal{X} \times_{-m} \mathcal{Y} = \mathbf{X}_{(m)} \mathbf{Y}_{(m)}$.

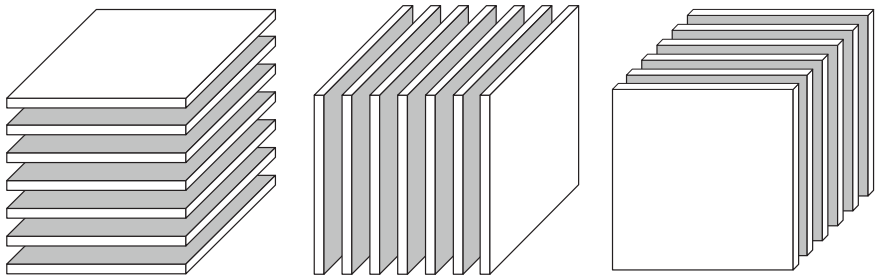

Figure 4.2: The three collections of m-mode faces of a tensor of third order.

The Kronecker product $\otimes$ plays an important part in the following sections. Apart from its previous uses in matricization and vectorization it is also naturally encountered when one considers the covariance matrix of the vectorized transformation $\mathrm{vec}(\mathcal{X} \times_{m=1}^{r} \mathbf{A}_m)$, where the tensor $\mathcal{X}$ has standardized components. Namely, using the formulas in the previous paragraphs we have $\mathrm{Cov}\{\mathrm{vec}(\mathcal{X} \times_{m=1}^{r} \mathbf{A}_m)\} = \mathbf{A}_r \mathbf{A}_r^\top \otimes \cdots \otimes \mathbf{A}_1 \mathbf{A}_1^\top$.

Finally, we define a couple of moment- and cumulant-based quantities for a random tensor $\mathcal{X} \in \mathbb{R}^{p_1 \times \cdots \times p_r}$. These act as tensorial counterparts to the kurtosis and other related quantities of real-valued random variables and play key role in the limiting distributions of the tensorial estimators later on. Let $\mathcal{B}(\mathcal{X}) \in \mathbb{R}^{p_1 \times \cdots \times p_r}$ contain the element-wise fourth moments of the elements of $\mathcal{X}$. For all $m \in \{1, \ldots, r\}$, we define $\bar{\boldsymbol{\beta}}_{(m)}(\mathcal{X}) = (\bar{\beta}_{(m)1}, \ldots, \bar{\beta}_{(m)p_m})^\top \in \mathbb{R}^{p_m}$ as the vector of row means of $\mathbf{B}_{(m)}(\mathcal{X})$, the m-mode flattening of $\mathcal{B}(\mathcal{X})$. Thus, $\bar{\boldsymbol{\beta}}_{(m)}(\mathcal{X})$ contains the average fourth moments of the m-mode faces of $\mathcal{X}$, that is, e.g. for a matrix $\mathbf{X} \in \mathbb{R}^{p_1 \times p_2}$ the vector $\boldsymbol{\beta}_{(1)}(\mathbf{X}) \in \mathbb{R}^{p_1}$ contains the row means of the fourth moments of $\mathbf{X}$ and $\boldsymbol{\beta}_{(2)}(\mathbf{X}) \in \mathbb{R}^{p_2}$ the column means of the fourth moments of $\mathbf{X}$. The vectors $\bar{\boldsymbol{\kappa}}_{(m)}(\mathcal{X}) = \bar{\boldsymbol{\beta}}_{(m)}(\mathcal{X}) - 3, \bar{\boldsymbol{\omega}}_{(m)}(\mathcal{X}) \in \mathbb{R}^{p_m}$ are defined analogously but only with the fourth moments replaced by the element-wise quantities $\mathrm{E}(x^4) - 3$ and $\mathrm{E}(x^6) - \mathrm{E}(x^3)^2$, respectively. Lastly, we define $\bar{\rho}_{(m)kl}(\mathcal{X})$ as the sample covariance of the kth and lth rows of $\mathbf{B}_{(m)}(\mathcal{X})$. If the random tensor $\mathcal{X}$ is clear from the context we again omit it and use simply $\bar{\boldsymbol{\beta}}_{(m)}, \bar{\boldsymbol{\kappa}}_{(m)}$ etc.

The previous list of concepts and their properties shows that it is almost possible to manipulate tensor-valued data with pure linear algebra alone, and in the following we plan to do so whenever possible. But sometimes we still need to resort to the actual tensorial forms, for example, in formulating our models in the next sections.

4.2 On tensorial methodology

Before introducing the tensor independent component model we first briefly discuss canonical examples of tensorial data and the accompanying models and methods commonly used for tensor-valued data in the literature.

To get an idea where tensorial data can be encountered, consider the following list of situations: A multivariate time series of the prices of p assets is a sample of first order tensors (vectors). A collection of grayscale images of size

$w \times h$ constitutes a sample of second order tensors (matrices). If the images are instead coloured, the colour information (RGB) can be thought of as an additional dimension of size three, making the set a sample of tensors of order three. A collection of instantaneous 3D fMRI-measurements of a group of people makes for a sample of tensors of order three, or four if each person has several measurement taken over time, and five if the experiment is repeated in different experimental situations. The list could be continued to even higher dimensions by adding replications under different conditions to the previous examples.

The majority of the approaches to tensor-valued data can be divided roughly into three groups: vectorization-based methods, tensor decompositions and methods based on tensorial samples. The vectorization-based methods constitute the simplest approach to tensor modeling. At their most basic level we simply vectorize the observed sample of tensors $\mathcal{X}_i \in \mathbb{R}^{p_1 \times \cdots \times p_r}$ to obtain a sample of vectors $\mathbf{x}_i \in \mathbb{R}^{p_1 \cdots p_r}$, which can then be subjected to the desired multivariate methods. These kind of procedures are especially common with ICA and functional magnetic resonance imaging (fMRI) data. Temporal ICA and spatial ICA, two staple methods in the previous context, are obtained when one vectorizes the obtained 3D-images and, after dimension reduction with SVD, treats either the rows or columns of the data matrix as observations, respectively, see Stone et al. (2002); Calhoun et al. (2003); Calhoun and Adali (2006). Although a seemingly natural way to approach the problem, the simplicity offered by vectorization comes at a price. As standard multivariate methods are generally invariant to the ordering of the variables in the random vector, the vectorization causes us to lose all spatial structure we had in the original tensors. This is well illustrated with covariance estimation: the covariance matrix $\mathbf{\Sigma}$ of a random vector $\mathbf{x} \in \mathbb{R}^{p_1 p_2}$ contains in general $p_1 p_2 (p_1 p_2 + 1)/2$ parameters. But if $\mathbf{x}$ is known to be a vectorized matrix originating from the matrix location-scatter model introduced later its covariance matrix has the Kronecker structure, $\mathbf{\Sigma} = \mathbf{\Sigma}_2 \otimes \mathbf{\Sigma}_1$, and the number of parameters is instead the considerably smaller $p_1(p_1 + 1)/2 + p_2(p_2 + 1)/2 - 1$. This approach to vectorized data is called structured covariance estimation and it shows that not all methods starting with vectorization are guilty of losing the spatial structure, but it instead depends largely on what we do after the vectorization. See Srivastava et al. (2008); Werner et al. (2008) for different approaches to structured covariance estimation. Several studies for comparing vectorization and truly tensorial methods have been made in recent years, see e.g. Li et al. (2010); Virta et al. (2016c); Virta and Nordhausen (2017a); Virta et al. (2017b,c); Virta and Nordhausen (2017b).

The use of the second category, tensor decompositions, is especially popular among signal processors. The decompositions can be seen as a higher-order analogues of eigendecomposition and singular value decomposition with the two primary methods being the CP-decomposition and the Tucker decomposition, see Kolda and Bader (2009). Vast literature has since been developed around these two and other related decompositions, see Kolda and Bader (2009); Cichocki et al. (2009, 2015); Sidiropoulos et al. (2017) for comprehensive reviews. The biggest difference between the methods of the second category and the other two is that tensorial decompositions rarely incorporate the concept of a random sample that is so central to all statistics. For example, for a sample of matrix-valued data the CP-decomposition takes the observation tensor

$\mathcal{X} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$, where one mode represents the individual samples, and decomposes it as $\mathcal{X} \approx \mathcal{D} \times_{m=1}^{3} \mathbf{A}_m$ where $\mathcal{D}$ is some small diagonal tensor and $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3$ are suitably sized matrices with unit length columns. The decomposition thus acts also on the observation space causing the individual observed matrices being mixed. This is critically in odds with the standard statistical practice where the observations space is left untouched and we instead focus on understanding the structure of the variable space, the two other modes of the data tensor $\mathcal{X}$ in our example. A simple "fix" would be to leave the observation mode untransformed in the decomposition. This approach, which is almost non-existent in the literature, provides fruitful grounds for studying the statistical properties of the wide collection of decomposition methods proposed over the years.

The third group of methods can be seen as a combining the best features of the previous two in a statistical sense. That is, we both retain the concept of a random sample and treat the tensors naturally as tensors. Most often this means generalizing some standard multivariate methodology to tensor-valued data using the multiplication $\times_m$ to operate the observations separately from each mode, and this is the approach we also take in the thesis. Previous works in literature include tensor or matrix versions of, for example, PCA (Ding and Cook, 2014), sufficient dimension reduction (Li et al., 2010; Pfeiffer et al., 2012; Ding and Cook, 2015a,b; Zhong et al., 2015) and (generalized) linear models where either the response or the predictor can be tensor-valued (Hung and Wang, 2012; Zhou et al., 2013; Zhao and Leng, 2014; Zhou and Li, 2014; Li and Zhang, 2017). Tensorial extensions of ICA have also been proposed (Vasilescu and Terzopoulos, 2005; Zhang et al., 2008) but the first model-based treatise appeared only in Virta et al. (2017b). The next sections now go to introduce this approach on the problem.

4.3 Tensorial location-scatter model and its extensions

Tensorial location-scatter model

We motivate the tensorial IC model in the same manner as its vectorial counterpart in Section 3, through the related general location-scatter model. We say that the random tensor $\mathcal{X} \in \mathbb{R}^{p_1 \times \cdots \times p_r}$ obeys the tensorial location-scatter model if

$$\mathcal{X} = \mathcal{M} + \mathcal{Z} \times_{m=1}^{r} \mathbf{\Omega}_m, \tag{4.1}$$

where the location tensor $\mathcal{M} \in \mathbb{R}^{p_1 \times \cdots \times p_r}$ and the invertible mixing matrices $\mathbf{\Omega}_1 \in \mathbb{R}^{p_1 \times p_1}, \ldots, \mathbf{\Omega}_r \in \mathbb{R}^{p_r \times p_r}$ are unknown parameters and the latent tensor $\mathcal{Z}$ satisfies $\mathrm{E}\{\mathrm{vec}(\mathcal{Z})\} = \mathbf{0}_{p_1 \cdots p_r}$ and $\mathrm{Cov}\{\mathrm{vec}(\mathcal{Z})\} = \mathbf{I}_{p_1 \cdots p_r}$. Again all r mixing matrices are identifiable only up to post-multiplications by orthogonal matrices. Vectorizing the model (4.1) reveals it as a structured submodel of the vector-valued location scatter model (3.1) where the mixing matrix has the Kronecker structure $\mathbf{\Omega} = \mathbf{\Omega}_r \otimes \cdots \otimes \mathbf{\Omega}_1$. If one was to proceed with vectorization, this special form of mixing would need to be addressed in the subsequent analyses with the risk of otherwise losing structural information, as discussed in the previous section. The model (4.1) also nicely exhibits the basic paradigm underlying all our tensor constructs: in extending vector-valued methodology to tensor-valued

random variables the focus shifts from individual elements to the modes. That is, the elements of $\mathcal{X}$ are no longer regarded in isolation but in an aggregate way through the corresponding rows, columns and other modes, as exemplified the r simultaneous transformations acting on the individual modes in (4.1).

Additional assumptions on the latent $\mathcal{Z}$ again lead into various useful families of models. Beginning with the simplest choice, fixing the distribution of the latent tensor as $\text{vec}(\mathcal{Z}) \sim \mathcal{N}_{p_1 \cdots p_r}(\mathbf{0}_{p_1 \cdots p_r}, \mathbf{I}_{p_1 \cdots p_r})$ gives $\mathcal{X}$ the tensor normal distribution discussed in Manceur and Dutilleul (2013); Ohlson et al. (2013), see also Gupta and Nagar (1999); Kollo and von Rosen (2006). While the normality assumption simplifies the algebra of linear methods considerably it can be regarded highly unrealistic in practice especially when the tensors are of large size. To obtain more versatile models the tensorial location-scatter model can analogously to Section 3.1 be extended to obtain the tensorial elliptical model and the tensorial IC model. We next discuss these in order.

Tensorial elliptical model

Beginning again from spherical distributions, two natural, alternative definitions for tensorial spherical distributions can be derived. In the first we require that the vectorized latent tensor $\text{vec}(\mathcal{Z})$ has a spherical distribution, see Arashi (2017), and in the second we require that the distribution of the latent tensor is tensorially spherical, that is

$$\mathcal{Z} \sim \mathcal{Z} \times_{m=1}^{r} \mathbf{U}_m, \quad \text{for all} \quad \mathbf{U}_1 \in \mathcal{U}^{p_1 \times p_1}, \ldots, \mathbf{U}_r \in \mathcal{U}^{p_r \times p_r}.$$

Vectorization reveals that the second definition gives a broader class of distributions than the first one with the cost of less structure. However, if one sticks strictly to the natural tensor operations such as $\times_m$ this structure is most likely enough for the majority of methods. The class of tensor elliptical distributions is now readily defined as the collection of all distributions of the form

$$\mathbf{X} = \mathcal{M} + \mathcal{Z} \times_{m=1}^{r} \mathbf{A}_m$$

where $\mathcal{Z}$ has a tensor spherical distribution in one of the previous senses and $\mathbf{A}_m$, $m \in \{1, \ldots, r\}$, are square matrices of suitable sizes. Although the matrix case $r = 2$ has been discussed in the literature (Gupta and Varga, 1995; Gupta and Nagar, 1999; Caro-Lopera et al., 2016), the general tensor case has received very little attention and many of its properties remain still undiscovered. For example, it has been conjectured that the tensorial elliptical model and tensorial PCA do not share the same relationship as their vectorial counterparts, see Section 4.5. Consequently we postpone the discussion of tensorial PCA to Section 4.5 where we generalize the covariance matrix to tensors.

Tensorial IC model

With the previous important but beyond-our-scope models out of the way, we are now ready to define the second natural generalization of the tensor normal distribution and our model of choice, the tensor independent component model.

Definition 8. *We say that the random tensor $\mathcal{X} \in \mathbb{R}^{p_1 \times \cdots \times p_r}$ obeys the tensor independent component (IC) model if*

$$\mathcal{X} = \mathcal{M} + \mathcal{Z} \times_{m=1}^{r} \mathbf{\Omega}_m, \tag{4.2}$$

where $\boldsymbol{\mathcal{M}} \in \mathbb{R}^{p_1 \times \cdots \times p_r}$ *and the invertible* $\boldsymbol{\Omega}_1 \in \mathbb{R}^{p_1 \times p_1}, \ldots, \boldsymbol{\Omega}_r \in \mathbb{R}^{p_r \times p_r}$ *are unknown parameters and the latent random tensor* $\boldsymbol{\mathcal{Z}} \in \mathbb{R}^{p_1 \times \cdots \times p_r}$ *satisfies Assumptions T1, T2, T3 below.*

Assumption T1. *The components of* $\boldsymbol{\mathcal{Z}}$ *are mutually independent,*

Assumption T2. *The vector* $\operatorname{vec}(\boldsymbol{\mathcal{Z}})$ *is standardized,* $\operatorname{E}\{\operatorname{vec}(\boldsymbol{\mathcal{Z}})\} = \boldsymbol{0}_{p_1 \cdots p_r}$ *and* $\operatorname{Cov}\{\operatorname{vec}(\boldsymbol{\mathcal{Z}})\} = \boldsymbol{I}_{p_1 \cdots p_r}$.

Assumption T3. *For each* $m \in \{1, \ldots, r\}$*, at most one of the* m*-mode faces of* $\boldsymbol{\mathcal{Z}}$ *consists entirely of normally distributed components.*

Assumption T1 is again self-evident and Assumptions T2 and T3 serve as identifiability constraints after which the mixing matrices $\boldsymbol{\Omega}_m$ are identifiable up to post-multiplications by matrices $\mathbf{J}_m\mathbf{P}_m$, $\mathbf{J}_m \in \mathcal{J}^{p_m}$, $\mathbf{P}_m \in \mathcal{P}^{p_m}$, for $m \in \{1, \ldots, r\}$. To see how Assumption T3 rids the tensor IC model of the confounding by orthogonal matrices found in the general tensor location-scatter model let, for example, the last two 1-mode faces of $\boldsymbol{\mathcal{Z}}$ consist entirely of standard normal components. The 1-mode matricization of the model now reads

$$\mathbf{X}_{(1)} = \boldsymbol{\Omega}_1 \mathbf{Z}_{(m)} (\boldsymbol{\Omega}_2 \otimes \cdots \otimes \boldsymbol{\Omega}_r)^\top,$$

where the last two rows of $\mathbf{Z}_{(1)}$ now contain only normally distributed components. If we let go of Assumption T3 the block matrix

$$\begin{pmatrix} \mathbf{I}_{p_1-2} & \mathbf{0} \\ \mathbf{0} & \mathbf{U} \end{pmatrix} \begin{pmatrix} \mathbf{I}_{p_1-2} & \mathbf{0} \\ \mathbf{0} & \mathbf{U}^\top \end{pmatrix},$$

where $\mathbf{U} \in \mathcal{U}^{2\times 2}$ is any orthogonal matrix, can be fitted between $\boldsymbol{\Omega}_1$ and $\mathbf{Z}_{(m)}$ all the while preserving the model assumptions, and Assumption T3 is therefore included to prevent situations such as this from happening. Again the Skitovich-Darmois theorem can be used to show that limiting the amount of normal components is indeed a sufficient condition for identifiability as well. Notice that Assumption T3 allows much more freedom on the distirbutions of the individual elements of $\boldsymbol{\mathcal{X}}$ than we would have by vectorizing (4.2) and resorting to the vector-valued IC model. For example, by having $p-1$ non-normal elements on the super-diagonal of the rth order tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{p\times\cdots\times p}$, the tensorial IC model for $\boldsymbol{\mathcal{X}}$ allows $p^r - p + 1$ normal components whereas the corresponding vectorized IC model permits only a single one. This again reflects the importance of the modes over the individual elements of the tensors in tensorial models. However, even if Assumption T3 is violated it again only renders us unable to estimate the corresponding columns of the mixing matrices, meaning that we can still estimate the non-normal parts of the modes. Strategies similar to Nordhausen et al. (2016, 2017a) could also be developed to estimate the number of normal modes in such a case.

The objective in tensor ICA is naturally the estimation of the unknown parameters in the tensor IC model and we can again focus solely on the mixing matrices $\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_r$, the location parameter being found simply as $\operatorname{E}(\boldsymbol{\mathcal{X}}) = \boldsymbol{\mathcal{M}}$. Accordingly, we assume in the following that $\operatorname{E}\{\operatorname{vec}(\boldsymbol{\mathcal{X}})\} = \mathbf{0}_{p_1 \cdots p_r}$.

4.4 Tensorial IC functionals

We follow in our exposition the same structure as we did with the vectorial ICA, making our next topic tensorial IC functionals. As we are now faced with the task of estimating inverses for all r mixing matrices, our definition of an IC functional also consists of the r corresponding parts.

Definition 9. *Fix $m \in \{1, \ldots, r\}$. The functional $\boldsymbol{\Gamma}_m : \mathcal{D} \to \mathbb{R}^{p_m \times p_m}$ is an m-mode independent component functional if we have*

i) $\boldsymbol{\Gamma}_m(\boldsymbol{\mathcal{X}}) \overset{*}{\equiv} \boldsymbol{\Omega}_m^{-1}$ *for all* $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{p_1 \times \cdots \times p_r}$ *coming from a tensorial IC model and*

ii) $\boldsymbol{\Gamma}_m(\boldsymbol{\mathcal{X}} \times_{m=1}^{r} \boldsymbol{U}_m) \equiv \boldsymbol{\Gamma}_m(\boldsymbol{\mathcal{X}}) \boldsymbol{U}_m^\top$ *for all* $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{p_1 \times \cdots \times p_r}$ *and all orthogonal* $\boldsymbol{U}_1 \in \mathcal{U}^{p_1 \times p_1}, \ldots, \boldsymbol{U}_r \in \mathcal{U}^{p_r \times p_r}$,

where $\equiv$ is as in Definition 2 and two square matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{p \times p}$, satisfy $\boldsymbol{A} \overset{*}{\equiv} \boldsymbol{B}$ if and only if $\boldsymbol{A} = \tau \boldsymbol{J}\boldsymbol{P}\boldsymbol{B}$ for some $\tau \in \mathbb{R}$, $\boldsymbol{J} \in \mathcal{J}^p$ and $\boldsymbol{P} \in \mathcal{P}^p$.

The definition of an m-mode IC functional is markedly different from that of an IC functional in Definition 2. First, for condition *ii)* in Definition 9 we require only orthogonal equivariance instead of the affine equivariance we had in the vector case as practice has shown that the latter is simply too strict a requirement; the m-mode functional must, in addition to being able to unmix the mth mode, also be invariant to transformations from other modes. Moreover, all currently existing m-mode IC functionals fail to be affine equivariant, causing us to "settle" for the orthogonal equivariance. Second, the weaker condition *ii)* in conjunction with any possible diagonality criteria we had in Definition 2 would not be strong enough to guarantee the Fisher consistency of the $\boldsymbol{\Gamma}_m$ and we instead have to directly require it in condition *i)*. As in the vector case, we again have an equivalence instead of full equalities in both conditions but this time we also allow arbitrary scale for the quantities. This is necessitated by us not fixing the mutual scaling of the r mixing matrices, i.e., the model remains unchanged if we respectively multiply and divide any two mixing matrices by the same non-zero scalar.

To solve the tensor IC problem it is sufficient to find an m-mode IC functional for all modes $m \in \{1, \ldots, r\}$. These functionals need not necessarily be obtained by the same "formula" and it is instead possible to use different methods to unmix the different modes. This naturally begs the question which combination of functionals is the best one and to answer that we require a measure of the success of an m-mode IC functional. We again resort to the MD index in Definition 3 and its asymptotics for the comparisons but with one difference; the lack of affine equivariance means that our asymptotical results hold only when the mixing matrices $\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_r$ in the tensorial IC model are all orthogonal.

Assume next that we have a full collection of m-mode IC functionals $\boldsymbol{\Gamma}_m$ with limiting normal distributions under the trivial mixing, $\sqrt{n}\{\operatorname{vec}(\hat{\boldsymbol{\Gamma}}_m - \mathbf{I}_{p_m})\} \rightsquigarrow \mathcal{N}_{p^2}(\mathbf{0}, \boldsymbol{\Psi})$, $m \in \{1, \ldots r\}$. We can show that all the asymptotical results derived after Definition 3 still hold within the individual modes, and consequently the partial traces $\operatorname{tr}^\circ(\boldsymbol{\Psi}_m)$ can be used to compare the efficiencies of different m-mode IC functionals within the same mode. However, our goal is still the unmixing of all r modes and thus we are usually more interested in the

joint efficiency of the whole set $\mathbf{\Gamma}_1, \ldots, \mathbf{\Gamma}_r$. Moreover, we would also like to make efficiency comparisons between the tensorial IC methods and the naïve vectorization-based methods, again necessitating the combining of the r functionals into one to allow comparisons with the single functional used by the vectorial methods.

The first step towards the solution is given by the Kronecker product. If we vectorize the tensorial IC model we see that the unmixing of each mode by the respective $\mathbf{\Gamma}_m$ is equivalent to unmixing the vectorized model by the product $\mathbf{\Gamma}_\otimes = \mathbf{\Gamma}_r \otimes \cdots \otimes \mathbf{\Gamma}_1$. Thus to compare tensorial and vectorial ICA methods on an equal level the product $\mathbf{\Gamma}_\otimes$ should directly be compared to the vectorial IC functional $\mathbf{\Gamma}_{\text{vec}}$. The second step is provided by Theorem 6 in Virta et al. (2017b) which connects the limiting behavior of the MD index of $\mathbf{\Gamma}_\otimes$ to those of its factors,

$$n(p-1)D(\hat{\mathbf{\Gamma}}_\otimes)^2 = \sum_{m=1}^{r} \rho_m n(p_m - 1)D(\hat{\mathbf{\Gamma}}_m)^2 + o_p(1).$$

Based on the results in Section 3.2 the average of the left-hand side converges to $\psi_\otimes = \sum_{m=1}^{r} \rho_m \text{tr}^\circ(\mathbf{\Psi}_m)$ and if the compared vector-valued IC method has the limiting covariance matrix $\mathbf{\Psi}_{\text{vec}}$ the efficiency comparisons should be done between $\psi_\otimes$ and $\text{tr}^\circ(\mathbf{\Psi}_{\text{vec}})$. This approach was taken in Virta et al. (2017b,c) where the soon-to-be-introduced tensorial FOBI and tensorial JADE were shown to be highly superior to their vectorial counterparts.

4.5 Tensorial standardization

The first step towards a solution in vectorial ICA is standardization and the same holds for its tensorial counterpart. But before we can extend the Mahalanobis standardization itself we first have to develop a tensorial version of its basic building block, the covariance matrix.

Our definition of m-mode IC functionals acting mode-wise makes clear that also for the standardization we need a collection of r covariance matrices, one for each mode. Staying true to the "mode-paradigm" discussed in Section 4.3, the m-mode covariance matrix $\mathbf{\Sigma}_m(\mathcal{X})$ should then somehow characterize the variation of the tensor $\mathcal{X}$ in the mth direction. As the basic building blocks of the mth mode of $\mathcal{X}$, the individual m-mode vectors and their covariance matrices clearly each capture some aspects of this variability, and to measure the average variation in the mth direction we average over them to obtain,

$$\mathbf{\Sigma}_m(\mathcal{X}) = \frac{1}{\rho_m} \text{E}\left(\mathbf{X}_{(m)}\mathbf{X}_{(m)}^\top\right) \in \mathbb{R}^{p_m \times p_m}, \quad m \in \{1, \ldots, r\},$$

the m-mode covariance matrices of $\mathcal{X}$, see Srivastava et al. (2008). These matrices share various properties of their multivariate counterparts: they are symmetric positive semidefinite, diagonal under independence and, furthermore, they also serve the analogous purpose in tensor independent component analysis. Virta et al. (2017b) showed that if $\mathcal{X}$ comes from a tensorial IC model we have,

$$\mathcal{X}_{st} = \mathcal{X} \times_{m=1}^{r} \mathbf{\Sigma}_m(\mathcal{X})^{-1/2} = \tau \cdot \mathcal{Z} \times_{m=1}^{r} \mathbf{U}_m, \tag{4.3}$$

for some constant $\tau > 0$ and orthogonal $\mathbf{U}_1 \in \mathcal{U}^{p_1 \times p_1}, \ldots, \mathbf{U}_r \in \mathcal{U}^{p_r \times p_r}$. Thus again standardization reduces the problem of estimating inverses of full-rank matrices to that of estimating orthogonal matrices. The unknown factor τ means that in practice we can estimate $\mathfrak{Z}$ only up to scaling, which is completely satisfactory as the scale of our IC model was in any case arbitrarily fixed.

Based on the previous result all our efforts to solve the tensor IC model will start with standardization and consequently the obtained m-mode IC functionals will again be of the form $\mathbf{V}_m(\mathcal{X}_{st})\mathbf{\Sigma}_m(\mathcal{X})^{-1/2}$, where the m-mode rotation functional $\mathbf{V}_m$ taking values in $\mathcal{U}^{p \times p}$ is again defined only for standardized random tensors. This form again both implies simplified criteria for a functional to be an m-mode IC functional and fixes the limiting variances (and distributions) of the diagonal elements of $\mathbf{\Gamma}_m$, as stated in the next two lemmas. The first one is proven in the Appendix and for the proof of the second one, see Virta et al. (2017b).

Lemma 6. *Let the functional $\mathbf{\Gamma}_m$ be of the form $\mathbf{\Gamma}_m(\mathcal{X}) = \boldsymbol{V}_m(\mathcal{X}_{st})\mathbf{\Sigma}_m(\mathcal{X})^{-1/2}$ with $\boldsymbol{V} \colon \mathcal{D} \to \mathcal{U}^{p_m \times p_m}$. Then $\mathbf{\Gamma}_m$ is an IC functional if and only if*

i) $\boldsymbol{V}_m(\mathcal{X}_{st}) \equiv \boldsymbol{I}_{p_m}$ for all $\mathcal{X}$ coming from the tensorial IC model with diagonal mixing, $\mathbf{\Omega}_m = \boldsymbol{D}_m$, for some diagonal matrix $\boldsymbol{D}_m$, $m \in \{1, \ldots, r\}$.

ii) $\boldsymbol{V}_m(\mathcal{X}_{st} \times_{m=1}^{r} \boldsymbol{U}_m) \equiv \boldsymbol{V}_m(\mathcal{X}_{st})\boldsymbol{U}_m^\top$ for all $\mathcal{X} \in \mathbb{R}^{p_1 \times \cdots \times p_r}$ and all orthogonal $\boldsymbol{U}_1 \in \mathcal{U}^{p_1 \times p_1}, \ldots, \boldsymbol{U}_r \in \mathcal{U}^{p_r \times p_r}$.

Lemma 7. *Fix $m \in \{1, \ldots, r\}$. Let $\mathcal{X} \in \mathbb{R}^{p_1 \times \cdots \times p_r}$ come from a tensorial IC model with $\mathbf{\Omega}_1 = \boldsymbol{I}_{p_1}, \ldots, \mathbf{\Omega}_r = \boldsymbol{I}_{p_r}$ and let $\mathbf{\Gamma}_m(\boldsymbol{x}) = \boldsymbol{V}_m(\mathcal{X}_{st})\mathbf{\Sigma}_m(\mathcal{X})^{-1/2}$ be a m-mode IC functional with the limiting distribution $\sqrt{n}\{\mathrm{vec}(\hat{\mathbf{\Gamma}}_m - \boldsymbol{I}_{p_m})\} \rightsquigarrow \mathcal{N}_{p_m^2}(\boldsymbol{0}_{p_m^2}, \mathbf{\Psi}_m)$. Then the diagonal elements $\mathrm{ASV}(\gamma_{m,kk})$ of $\mathbf{\Psi}_m$ are*

$$\mathrm{ASV}(\gamma_{m,kk}) = \frac{\bar{\kappa}_{(m)k} + 2}{4\rho_m}.$$

Again the limiting behaviors of the off-diagonal elements of $\mathbf{\Gamma}_m$ have to be derived separately for each m-mode IC functional and that is one of our main focuses with the proposed methods in the next section.

We close this section by discussing a natural derivative of the m-mode covariance matrices, the tensorial PCA. The main difference between vectorial and tensorial PCA is that in the latter the connection between PCA and elliptical models is most likely lost. Namely, for it to hold we would need an affine equivariant m-scatter functional and Virta et al. (2017b) conjecture that no such functionals exist for general tensors. An affine equivariant m-scatter functional is a functional $\mathbf{S}_m : \mathcal{D} \to \mathbb{R}_+^{p_m \times p_m}$ with the property,

$$\mathbf{S}_m(\mathcal{X} \times_{m=1}^{r} \mathbf{A}_m) = \mathbf{A}_m \mathbf{S}_m(\mathcal{X}) \mathbf{A}_m^\top,$$

for all invertible $\mathbf{A}_1 \in \mathbb{R}^{p_1 \times p_1}, \ldots, \mathbf{A}_r \in \mathbb{R}^{p_r \times p_r}$. Again the requirement to annihilate linear transformations in $r - 1$ modes seems to be too strict for such constructs to exist. Orthogonally equivariant m-scatter functionals still exist, the prime examples being the m-mode covariance matrices above. And although orthogonal equivariance does not render them capable of solving the tensorial elliptical model, their properties are certainly sufficient for the tensorial IC model.

However, despite the above hardships we can still do the analogous to regular PCA and project each mode of $\mathcal{X}$ onto the first $d_m \leq p_m$ eigenvectors of the corresponding $\mathbf{\Sigma}_m(\mathcal{X})$, $m \in \{1, \ldots, r\}$, in order to reduce the dimension of $\mathcal{X}$. This approach has been used in Li et al. (2010) and later in Virta et al. (2016c) under the name TPCA. The procedure is also equivalent to applying the higher order singular value decomposition (HOSVD) (De Lathauwer et al., 2000) to all non-observation modes of the data tensor. A matrix-valued PCA was also proposed in Ding and Cook (2014), with an iterative approach to the estimation, and further research is needed to reveal the relationship between the two methods.

4.6 Tensorial FOBI and JADE

We begin by providing tensorial counterparts for the cumulant matrices used by FOBI and JADE. Our extensions are derived in the same manner as the m-mode covariance matrix in the last section, using the plug-in method for the outer products $\mathbf{x}\mathbf{x}^T$. As the first matrix in (3.7) can be written as $\mathrm{E}(\mathbf{x}\mathbf{x}^\top \mathbf{E}^{ij} \mathbf{x}\mathbf{x}^\top)$, and equivalently for the remaining three matrices, our m-mode tensor extension of the cumulant matrix $\mathbf{C}^{ij}$ is then simply

$$\begin{aligned}\mathbf{C}_m^{ij}(\mathcal{X}) = \rho_m^{-1}\mathrm{E}\left(\mathbf{X}_{(m)}\mathbf{X}_{(m)}^\top \mathbf{E}^{ij} \mathbf{X}_{(m)}\mathbf{X}_{(m)}^\top\right) - \rho_m^{-1}\mathrm{E}\left(\mathbf{X}_{(m)}\mathbf{X}_{(m)}^{*\top} \mathbf{E}^{ij} \mathbf{X}_{(m)}\mathbf{X}_{(m)}^{*\top}\right) \\ -\rho_m^{-1}\mathrm{E}\left(\mathbf{X}_{(m)}\mathbf{X}_{(m)}^{*\top} \mathbf{E}^{ij} \mathbf{X}_{(m)}^*\mathbf{X}_{(m)}^\top\right) - \rho_m^{-1}\mathrm{E}\left(\mathbf{X}_{(m)}\mathbf{X}_{(m)}^\top \mathbf{E}^{ij} \mathbf{X}_{(m)}^*\mathbf{X}_{(m)}^{*\top}\right),\end{aligned}$$

where the division by ρ_m again represents taking average over the m-mode vectors of $\mathcal{X}$. As the above method of constructing $\mathbf{C}_m^{ij}$ is heuristic at best, the matrices have both lost their interpretation as cumulants in the traditional sense and lack any guarantee that they will be useful in tensorial ICA. However, some hope is given when we evaluate the matrix $\mathbf{C}_m^{ij}$ for a tensor $\mathcal{X}$ which has been standardized in the sense of the last section, $\mathbf{\Sigma}_m(\mathcal{X}) = \tau \mathbf{I}_{p_m}$, for all $m \in \{1, \ldots r\}$, and obtain

$$\begin{aligned}\mathbf{C}_m^{ij}(\mathcal{X}) = \frac{1}{\rho_m}\mathrm{E}\left(\mathbf{X}_{(m)}\mathbf{X}_{(m)}^\top \mathbf{E}^{ij} \mathbf{X}_{(m)}\mathbf{X}_{(m)}^\top\right) \\ - \tau^4\left(\rho_m \mathbf{E}^{ij} + \mathbf{E}^{ji} + \delta_{ij}\mathbf{I}_{p_m}\right),\end{aligned} \tag{4.4}$$

having a striking resemblance to the corresponding form in (3.8). The related FOBI-matrix for a standardized tensor is now simply defined as in the vector-valued case,

$$\mathbf{C}_m(\mathcal{X}) = \sum_{i=1}^{p} \mathbf{C}_m^{ii}(\mathcal{X}) = \frac{1}{\rho_m}\mathrm{E}\left(\mathbf{X}_{(m)}\mathbf{X}_{(m)}^\top \mathbf{X}_{(m)}\mathbf{X}_{(m)}^\top\right) - \tau^4(\rho_m + 2)\mathbf{I}_{p_m}.$$

Some mildly tedious algebra along with the ruleset for manipulating tensors in Section 4.1 were now used in Virta et al. (2017b) to show that, under Assumption T5 below, the eigenvectors of $\mathbf{C}_m$ fulfill the conditions of Lemma 6 and the resulting method of solving the tensorial IC model is called TFOBI. The corresponding m-mode IC functionals are given in the next definition.

Assumption T5. *For each $m \in \{1, \ldots, r\}$, the elements of $\bar{\boldsymbol{\kappa}}_{(m)}(\mathfrak{Z})$ are distinct.*

Definition 10. *Fix $m \in \{1, \ldots, r\}$ and let $\mathcal{X} \in \mathbb{R}^{p_1 \times \cdots \times p_r}$. Then the m-mode FOBI-functional is $\boldsymbol{\Gamma}_m^F = \boldsymbol{\Gamma}_m^F(\mathcal{X}) = \boldsymbol{V}_m \boldsymbol{\Sigma}_m^{-1/2}(\mathcal{X}) \in \mathbb{R}^{p_m \times p_m}$ where the rotation functional $\boldsymbol{V}_m = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{p_m})^\top \in \mathcal{U}^{p_m \times p_m}$ contains the eigenvectors of the matrix*

$$\boldsymbol{C}_m(\mathcal{X}_{st}) = \frac{1}{\rho_m} \mathrm{E}\left(\boldsymbol{X}_{st(m)} \boldsymbol{X}_{st(m)}^\top \boldsymbol{X}_{st(m)} \boldsymbol{X}_{st(m)}^\top \right) - \tau^4 (\rho_m + 2) \boldsymbol{I}_{p_m},$$

as its rows in decreasing order according to the corresponding eigenvalues, where τ is as in (4.3).

Even though τ^4 in Definition 10 is an unknown constant there is no need to estimate it as its value does not affect the obtained eigenvectors or their order of extraction in any way.

Comparison between the tensorial and vectorial versions of FOBI now instantly reveals benefits of the former over the latter. First, assume an rth order tensor $\mathcal{X} \in \mathbb{R}^{p \times \cdots \times p}$ coming from the tensorial IC model. Then, computing the r m-mode IC functionals requires a total of $2r$ eigendecompositions of $p \times p$ matrices whereas vectorizing and computing the FOBI IC functional requires 2 eigendecompositions of $p^r \times p^r$ matrices, making the effect of the order r on the computational complexity additive for the tensorial FOBI and multiplicative for the vectorial FOBI. Second, Assumption T5 is a direct tensorial extension of Assumption V5 and the two share the same relationship as Assumptions T3 and V3: again the tensorial version of FOBI gives more freedom for the distributions of the individual components of $\mathcal{Z}$ than the regular FOBI.

However, despite its success over vectorial FOBI, we can still do better than tensorial FOBI by again considering all p_m^2 m-mode cumulant matrices (4.4) in all modes. Plugging in the standardized tensors $\mathcal{X}_{st}$, Virta et al. (2017c) showed that the unknown orthogonal matrices $\mathbf{U}_1, \cdots, \mathbf{U}_r$ diagonalize all cumulant matrices of the respective modes,

$$\mathbf{C}^{ij}(\mathcal{X}_{st}) = \tau^4 \mathbf{U}_m \left(\sum_{k=1}^{p_m} u_{(m)ik} u_{(m)jk} \bar{\kappa}_{(m)k} \mathbf{E}^{kk} \right) \mathbf{U}_m^\top,$$

where $(u_{(m)kl})$ are the elements of $\mathbf{U}_m$, $m \in \{1, \ldots, r\}$. This instantly prompts us to define the tensorial JADE as the joint diagonalization of the cumulant matrices (4.4) in each mode.

Definition 11. *Fix $m \in \{1, \ldots, r\}$ and let $\mathcal{X} \in \mathbb{R}^{p_1 \times \cdots \times p_r}$. Then the m-mode JADE-functional is $\boldsymbol{\Gamma}_m^J = \boldsymbol{\Gamma}_m^J(\mathcal{X}) = \boldsymbol{V}_m \boldsymbol{\Sigma}_m^{-1/2}(\mathcal{X}) \in \mathbb{R}^{p_m \times p_m}$ where the rotation functional $\boldsymbol{V}_m = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{p_m})^\top \in \mathcal{U}^{p_m \times p_m}$ is the joint diagonalizer of the set of matrices,*

$$\left\{ \boldsymbol{C}_m^{ij}(\mathcal{X}_{st}) \mid i, j \in \{1, \ldots, p_m\} \right\},$$

where

$$\begin{aligned} \boldsymbol{C}_m^{ij}(\mathcal{X}_{st}) = {} & \frac{1}{\rho_m} \mathrm{E}\left(\boldsymbol{X}_{st(m)} \boldsymbol{X}_{st(m)}^\top \boldsymbol{E}^{ij} \boldsymbol{X}_{st(m)} \boldsymbol{X}_{st(m)}^\top \right) \\ & - \boldsymbol{\Sigma}_m(\mathcal{X}_{st}) \left(\rho_m \boldsymbol{E}^{ij} + \boldsymbol{E}^{ji} + \delta_{ij} \boldsymbol{I}_{p_m} \right) \boldsymbol{\Sigma}_m(\mathcal{X}_{st})^\top. \end{aligned}$$

Contrary to tensorial FOBI, in Definition 11 the unknown scalar τ needs to be estimated and natural estimators for its square are the m-mode covariance matrices, $\boldsymbol{\Sigma}_m(\boldsymbol{\mathcal{X}}_{st}) = \tau^2 \mathbf{I}_{p_m}$, $m \in \{1, \ldots, r\}$.

Virta et al. (2017c) showed that under Assumption T4 the joint diagonalization indeed extracts the independent components and the obtained method is called TJADE. To show that $\boldsymbol{\Gamma}_m^J$ is a true m-mode IC functional in the sense of Lemma 6 we need to adopt Assumption T4 below and the proof then exactly mimics the corresponding steps for the JADE functional, see Section 3.6.

Assumption T4. *For each $m \in \{1, \ldots, r\}$, at most one element of $\bar{\boldsymbol{\kappa}}_{(m)}$ is zero.*

The comparisons made earlier between the assumptions of vectorial FOBI and tensorial FOBI hold exactly analogously between the assumptions of vectorial JADE and tensorial JADE.

Finally, Definitions 10 and 11 allow for a straightforward but tedious calculation of the limiting distributions of the corresponding sample estimates via Lemma 2 in Miettinen et al. (2015). This was done in Virta et al. (2017b,c) and the following two theorems again describe one specific aspect of the limiting distributions, the limiting variances of the off-diagonal elements of the functionals. Notice that due to the limitations concerning the equivariance properties of tensorial functionals discussed in Sections 4.4 and 4.5 the next results generalize only to orthogonal mixing matrices.

Theorem 5. *Fix $m \in \{1, \ldots, r\}$, let $\boldsymbol{\mathcal{X}}_1, \ldots, \boldsymbol{\mathcal{X}}_n$ be a random sample having finite eighth moments from a tensorial IC model satisfying Assumption T5 and let $\boldsymbol{\Omega}_1 = \boldsymbol{I}_{p_1}, \ldots, \boldsymbol{\Omega}_r = \boldsymbol{I}_{p_r}$. Then there exists a sequence of m-mode FOBI functionals $\hat{\boldsymbol{\Gamma}}_m^F$ such that $\sqrt{n}\{\mathrm{vec}(\hat{\boldsymbol{\Gamma}}_m^F - \boldsymbol{I}_{p_m})\} \rightsquigarrow \mathcal{N}_{{p_m}^2}(\boldsymbol{0}_{p_m^2}, \boldsymbol{\Psi}_m)$ where the diagonal elements $\mathrm{ASV}(\gamma_{m,kl})$, $k \neq l$, of $\boldsymbol{\Psi}_m$ are*

$$\mathrm{ASV}(\gamma_{m,kl}) = \frac{\bar{\omega}_{(m)k} + \bar{\omega}_{(m)l} - \bar{\beta}_{(m)k}^2 + \sum_{t \neq k,l}^{p_m} (\bar{\kappa}_{(m)t} + 2) + b_0}{\rho_m \left(\bar{\kappa}_{(m)k} - \bar{\kappa}_{(m)l}\right)^2},$$

where $b_0 = (\rho_m - 7)\bar{\kappa}_{(m)l} - (\rho_m + 8) + 2\bar{\rho}_{(m)kl} + (\rho_m - 1)(\bar{\beta}_{(m)k} + p_m)$.

Theorem 6. *Fix $m \in \{1, \ldots, r\}$, let $\boldsymbol{\mathcal{X}}_1, \ldots, \boldsymbol{\mathcal{X}}_n$ be a random sample having finite eighth moments from a tensorial IC model satisfying Assumption T4 and let $\boldsymbol{\Omega}_1 = \boldsymbol{I}_{p_1}, \ldots, \boldsymbol{\Omega}_r = \boldsymbol{I}_{p_r}$. Then there exists a sequence of m-mode JADE functionals $\hat{\boldsymbol{\Gamma}}_m^J$ such that $\sqrt{n}\{\mathrm{vec}(\hat{\boldsymbol{\Gamma}}_m^J - \boldsymbol{I}_{p_m})\} \rightsquigarrow \mathcal{N}_{{p_m}^2}(\boldsymbol{0}_{p_m^2}, \boldsymbol{\Psi}_m)$ where the diagonal elements $\mathrm{ASV}(\gamma_{m,kl})$, $k \neq l$, of $\boldsymbol{\Psi}_m$ are*

$$\mathrm{ASV}(\gamma_{m,kl}) = \frac{\psi_{(m)k} + \psi_{(m)l} + \bar{\kappa}_{(m)l}^4 + \xi_{(m)k} + \xi_{(m)l} - 2\bar{\kappa}_{(m)k}^2 \bar{\kappa}_{(m)l}^2 \bar{\rho}_{(m)kl}}{\rho_m \left(\bar{\kappa}_{(m)k}^2 + \bar{\kappa}_{(m)l}^2\right)^2},$$

where $\psi_{(m)k} = \bar{\kappa}_{(m)k}^2(\bar{\omega}_{(m)k} - \bar{\beta}_{(m)k}^2)$ and $\xi_{(m)k} = \bar{\kappa}_{(m)k}^2(\bar{\kappa}_{(m)k} + 2)(\rho_m - 1)$.

The asymptotic variances are again too messy to allow for any analytical comparisons but we can still compare different estimates using the partial traces and the results in Section 4.4. These comparisons were done between tensorial FOBI, tensorial JADE and their vector-valued versions in Virta et al. (2017b,c),

establishing in particular that the tensor methods' lack of affine equivariance is negligible when compared to the effects of vectorization. Finally, to validate the results, plugging in $r = 1, \rho_1 = 1$ and simplifying reverts the asymptotic variances in Theorems 5 and 6 to those in Theorems 3 and 4.

Remark 1. *Virta et al. (2017b,c) defined also a second set of cumulant matrices* $\boldsymbol{C}^{ij}$*, yielding alternative FOBI and JADE* m*-mode functionals to the ones in Definitions 10 and 11. However, the former is generally asymptotically inferior to* $\boldsymbol{\Gamma}_m^F$ *and the latter is asymptotically equivalent to* $\boldsymbol{\Gamma}_m^J$ *and we have thus withheld from reviewing them here.*

5 Independent component analysis for functional data

5.1 Hilbert space theory

Our last topic treats the independent component analysis of multivariate functional data. While functional data analysis can in a sense be seen as statistics in an infinite-dimensional Euclidean space, the notations between this and the previous sections are still anything but compatible. Hence in this section we present yet another, final set of notation. See Conway (2013); Hsing and Eubank (2015) for a more comprehensive discussion.

Let T be a closed interval on $\mathbb{R}$ and let $(\mathcal{H}_0, \langle \cdot, \cdot \rangle_0)$ be a real, separable Hilbert space of functions from T to $\mathbb{R}$. The geometry of the space is entirely determined by the inner product $\langle \cdot, \cdot \rangle_0$, which induces the corresponding norm, $\|f\|_0 = \langle f, f \rangle_0^{1/2}$, and the metric, $d_0(f, g) = \|f - g\|_0$, for $f, g \in \mathcal{H}_0$. The previous concepts already provide analogies for several central concepts of multivariate statistics but some key ones are still missing, most notably matrices, the generalization of which is given by bounded linear operators. An operator $L : \mathcal{H}_0 \to \mathcal{H}_0$ is said to bounded if the operator norm, $\|L\|_{\text{OP}} = \inf\{M \geq 0 \mid \forall f \in \mathcal{H}_0 : \|Lf\|_0 \leq M\|f\|_0\}$, is finite, and linear if $L(af+g) = aL(f)+L(g)$, for all $a \in \mathbb{R}$, $f, g \in \mathcal{H}_0$, both familiar properties of matrices. The set of all bounded linear operators from $\mathcal{H}_0$ to $\mathcal{H}_0$ is denoted by $\mathcal{L}(\mathcal{H}_0)$ and forms a Banach space when equipped with the operator norm $\|\cdot\|_{\text{OP}}$.

We next introduce analogies of several familiar notions of linear algebra to linear operators. The identity operator $id \in \mathcal{L}(\mathcal{H}_0)$ leaves any function unchanged, $id\ f = f$, for all $f \in \mathcal{H}_0$. Similar to the matrix transpose, every bounded linear operator has the adjoint operator $L^* : \mathcal{H}_0 \to \mathcal{H}_0$ determined uniquely by the relation, $\langle Lf, g \rangle_0 = \langle f, L^* g \rangle_0$, for all $f, g \in \mathcal{H}_0$. If an operator satisfies $L = L^*$ it is called self-adjoint and if further $\langle Lf, f \rangle_0 \geq 0$, for all $f \in \mathcal{H}_0$, we say that L is positive semidefinite. A self-adjoint, positive semidefinite operator L is said to be a trace-class operator if $\sum_{k=1}^{\infty} \langle Le_k, e_k \rangle$ converges for some orthonormal basis e_k, $k \in \mathbb{N}_+$, in which case its trace, $\text{tr}(L)$, is defined as the previous quantity (which is independent of the choice of the basis e_k). If both of the identities, $UU^* = id$ and $U^*U = id$, hold for an operator U the operator is called unitary. Finally, the simplest way of constructing bounded linear operators is akin to using the vectors $\mathbf{a}, \mathbf{b}$ to obtain the rank-1 matrix $\mathbf{a}\mathbf{b}^\top$: for every pair of functions, $g, h \in \mathcal{H}_0$, the tensor product operator $(g \otimes h) : \mathcal{H}_0 \to \mathcal{H}_0$ acting as $(g \otimes h)f = \langle h, f \rangle_0 \cdot g$ is a bounded linear operator.

To be able to do statistics in $\mathcal{H}_0$ we introduce randomness to the space by briefly returning to the measure-theoretic notions discussed back in Section 2. Recalling our probability space, $(\Omega, \mathcal{F}, \mathbb{P})$, a random function is defined to be any mapping $X : \Omega \to \mathcal{H}_0$ which is $\mathcal{F}/\mathcal{B}$-measurable where $\mathcal{B}$ is the Borel-algebra generated by the open sets in $\mathcal{H}_0$ with respect to the induced metric d_0. To aid visualization, a random sample of simulated functional data is depicted

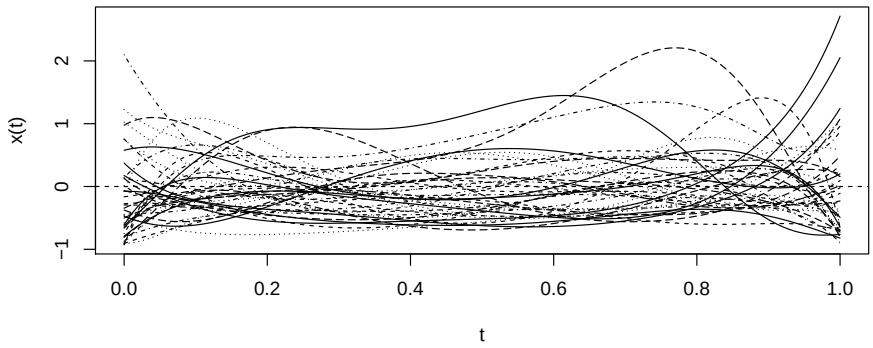

Figure 5.1: A random sample of $n = 40$ functions from a particular distribution.

in Figure 5.1. Similarly, a random operator is defined as any mapping $X : \Omega \to \mathcal{L}(\mathcal{H}_0)$ which is $\mathcal{F}/\mathcal{B}_{\text{OP}}$-measurable where $\mathcal{B}_{\text{OP}}$ is the Borel-algebra generated by the open sets in $\mathcal{L}(\mathcal{H}_0)$ with respect to the metric induced by the operator norm.

The Riesz representation theorem can next be used to define expected values for both random functions and operators. If the random function X is bounded in the sense that $\mathrm{E}(\|X\|_0) < \infty$, then there exists a unique function $\mu \in \mathcal{H}_0$ such that $\mathrm{E}(\langle X, f\rangle_0) = \langle \mu, f\rangle_0$, for all $f \in \mathcal{H}$, and we define $\mathrm{E}(X) = \mu$. Analogously, if the random operator W is bounded in the sense that $\mathrm{E}(\|W\|_{\text{OP}}) < \infty$, then there exists a unique bounded linear operator Υ such that $\mathrm{E}(\langle Wf, g\rangle_0) = \langle \Upsilon f, g\rangle_0$, for all $f, g \in \mathcal{H}_0$ and we again define $E(W) = \Upsilon$.

With the tensor product and the expected values at hand, we are now equipped to quantify variation in functional spaces. Assuming that the random function X is centered, i.e. $\mathrm{E}(X) = 0$, and has finite second moments, $\mathrm{E}(\|X\|_0^2) < \infty$, the covariance operator of X is the self-adjoint, positive semidefinite, trace class operator $\Sigma(X) = \mathrm{E}(X \otimes X)$. Like its multivariate counterpart, also the covariance operator can be shown to admit an eigendecomposition: $\Sigma(X) = \sum_{k=1}^{\infty} \lambda_k(\phi_k \otimes \phi_k)$ where $\lambda_k = \langle \Sigma\phi_k, \phi_k\rangle_0$, $k \in \mathbb{N}_+$, are the nonnegative eigenvalues in decreasing order and the eigenvectors ϕ_k, $k \in \mathbb{N}_+$, form an orthonormal basis of $\mathcal{H}_0$, that is, $\langle \phi_k, \phi_l\rangle_0 = \delta_{kl}$. The concept of functional principal component analysis would now naturally follow but we will postpone its discussion to Section 5.2 where we also review other models for functional data.

From this point onwards, we abandon the familiar course of presentation used in the previous two parts of the summary. Infinite-dimensional data fundamentally differs from the finite-dimensional in various important aspects and this means that generalizations of some everyday objects from multivariate statistics to $\mathcal{H}_0$ are sometimes very difficult to obtain. For example, inverses of operators can be rather ill-behaved in $\mathcal{H}_0$ and as a consequence the standardization of random variables is not feasible in the usual sense, an issue we discuss closer in Section 5.3. However, we will still review the related methodology and models insofar as they exist.

5.2 Functional data models

Whereas in finite-dimensional setting the different classes of methods were unified through their connection to the location-scatter and normal models this no longer holds for functional data and the extensions of the previous models are of a more disconnected variety. We assume in the following, without loss of generality, that all random functions X are centered, $\mathrm{E}(X) = 0$, and begin by describing Gaussian random elements in $\mathcal{H}_0$, the functional version of the normal distribution.

A random function X is said to be Gaussian if $\langle X, f\rangle_0$ is normally distributed for all $f \in \mathcal{H}_0$ (Lifshits, 2012). As an immediate consequence, any finite tuple $(\langle X, f_1\rangle_0, \ldots, \langle X, f_p\rangle_0)$ has a multivariate normal distribution by the Cramér-Wold device. Any Gaussian function X in $\mathcal{H}_0$ is square integrable, $\mathrm{E}(\|X\|^2) < \infty$, and furthermore the distribution of a zero-mean Gaussian random function X is uniquely determined by its covariance operator $\Sigma(X)$. However, contrary to the finite-dimensional Euclidean spaces, not every self-adjoint, positive semidefinite operator $L \in \mathcal{L}(\mathcal{H}_0)$ defines a Gaussian distribution. A counterexample is given by the simplest bounded linear operator, the identity operator id. A heuristic reason for this is that the sequence of eigenvalues of any bounded, self-adjoint, positive semidefinite operator must converge to zero, something id with its infinite sequence of unit eigenvalues fails to do. If $\mathcal{H}_0$ is a reproducing kernel Hilbert space (Paulsen and Raghupathi, 2016), meaning that point evaluation is representable for all $t \in T$ as $f(t) = \langle f, g_t\rangle$, for some $g_t \in \mathcal{H}_0$, any finite collection of time points, $X(t_1), \ldots, X(t_m)$, has a joint Gaussian distribution. The resulting Gaussian functions are in stochastic analysis generally called Gaussian processes and some classes of covariance functions commonly used with Gaussian processes include the set of squared exponential covariance functions and the set of Matern covariance functions (Rasmussen and Williams, 2006).

A classical method of functional data analysis which behaves particularly well for Gaussian functions is functional principal component analysis (FPCA), better known under the name of Karhunen-Loève expansion, see Bosq (2012). Assume that the covariance operator of the random function X has the eigendecomposition $\Sigma(X) = \sum_{k=1}^{\infty} \lambda_k(\phi_k \otimes \phi_k)$. As the eigenvectors form an orthonormal basis of $\mathcal{H}_0$, the identity operator $id : \mathcal{H}_0 \to \mathcal{H}_0$ can be written as $id = \sum_{k=1}^{\infty}(\phi_k \otimes \phi_k)$, and we have the representation $X = \sum_{k=1}^{\infty}\langle \phi_k, X\rangle_0\phi_k$. As in the multivariate case, the "principal components" $m_k = \langle \phi_k, X\rangle_0$ are uncorrelated and have the corresponding eigenvalues as their variances, $\mathrm{E}(m_k m_l) = \delta_{kl}\lambda_k$. If X is now a Gaussian function, our previous derivations reveal that the principal components m_k are independent $\mathcal{N}(0, \lambda_k)$-distributed random variables and dimension reduction into a finite sub-space of $\mathcal{H}_0$ can be carried out by truncating the expansion at a chosen value of k. Numerous extensions of FPCA and the Karhunen-Loève expansion have been proposed over the years, see Yao et al. (2005); Ramsay and Silverman (2006); Bali et al. (2011).

As we saw earlier, id is not a feasible covariance operator in $\mathcal{H}_0$ and consequently pursuing functional elliptical and independent component models via the generalization of the properties of a "standardized" Gaussian function will not work. However some tricks can be used to arrive at extensions of familiar models, such as the elliptical functional models considered in Bali and Boente (2009); Boente et al. (2014). A random function X is said to have an ellip-

tical distribution with parameters $\mu \in \mathcal{H}_0$ and $\Sigma \in \mathcal{L}(\mathcal{H}_0)$, where Σ is self-adjoint, positive semidefinite, compact operator, if for any bounded linear operator $A : \mathcal{H}_0 \to \mathbb{R}^d$ the random vector AX is elliptically distributed with the parameters $A\mu$ and $A\Sigma A^*$. Essentially, a random function is elliptical if all its finite-dimensional reductions are. The resulting class of elliptical random functions shares many of the properties of its finite-dimensional cousin: if the mean and covariance operator of an elliptical function exist they equal respectively μ and $\rho\Sigma$, for some $\rho > 0$; the conditional expectations of finite-dimensional projections of elliptical functions admit simple expressions and the class of elliptical functions is a superset of the set of Gaussian functions.

In addition to PCA, the recent years have seen a wide range of classical multivariate methods being extended to functional data, including non-parametric methods (Yao et al., 2005; Ferraty and Vieu, 2006), linear modelling (Ramsay and Silverman, 2006; Horváth and Kokoszka, 2012), canonical correlation (He et al., 2004; Ramsay and Silverman, 2006; Hsing and Eubank, 2015) and sufficient dimension reduction (Ferré and Yao, 2003, 2005; Hsing and Ren, 2009). In practice, the application of all the previous methods to observed data requires one intermediate step: as we can never observe anything truly infinite-dimensional, the obtained discrete data points need to be *smoothed* into functions by fitting them into an appropriate, (finite-dimensional) functional basis. Excellent account of smoothing can be found, for example, in Ramsay and Silverman (2006) and consequently we will refrain from discussing it here.

Finally, also our main topic has received its functional counterpart. In Li et al. (2015) a functional version of independent component analysis was developed and due to its close connection to the multivarite functional ICA of Virta et al. (2017a) we will devote the next section to it.

5.3 Functional ICA

A first hurdle in defining independent component analysis in spaces of infinite dimension is given by the concept of independence itself. A random function X does not *per se* consist of any distinct components we could consider as independent and the independence must somehow be contained within the function itself. Along these lines Li et al. (2015) defined a random function X to be independent if its coordinates $\langle X, \phi_k \rangle_0$, $k \in \mathbb{N}_+$, with respect to some particular predefined basis ϕ_k, $k \in \mathbb{N}_+$, of $\mathcal{H}_0$ define a set of independent random variables. An infinite collection of random variables is said to be independent if all of its finite subsets consist of mutually independent random variables. The same definition for the independence of X is used also by Gutch and Theis (2012) who in addition require that the orthogonal projection to any subspace spanned by a subset of the basis is independent with the projection to the complement of the subspace. A key question is then the determination of the basis ϕ_k, all following results greatly depending on the choice. In a finite Euclidean space a natural candidate is the corresponding canonical basis but no such concept exists for general Hilbert spaces. The choice of Li et al. (2015) is to use an eigenbasis of the covariance operator of X as this canonical basis and consequently they are able to define the functional IC model.

Definition 12. *Let X be a random element in $\mathcal{H}_0$ with $\mathrm{E}(\|X\|_0^2) < \infty$ and let ϕ_k, $k \in \mathbb{N}_+$, be an eigenbasis of $\Sigma(X)$. Then we say that X obeys the functional*

independent component model if there exists an operator $\Gamma \in \mathcal{L}(\mathcal{H}_0)$ *such that* ΓX *has independent components with respect to the basis* ϕ_k*.*

As the simplest IC method, FOBI seems again like the most reasonable starting point for functional generalization but additional obstacles are encountered already at the first step of the procedure, the standardization: a standardized function would need to have the identity operator as its covariance operator, something we already deemed impossible. As the standardization is an integral part of FOBI and required in the subsequent step, Li et al. (2015) bypass this barrier by assuming that the unmixing operator Γ in Definition 12 is of a very specific form. Namely, they require that for some fixed d, the operator Γ acts only on a subspace spanned by the d first basis vectors ϕ_k, effectively reducing the separation problem from infinite- to finite-dimensional. If we denote $\mathcal{M}_d = \operatorname{span}(\phi_1, \ldots, \phi_d)$ and let P_d and Q_d be respectively the projections onto the span and its orthogonal complement, the assumption can be written as

$$\Gamma = P_d A P_d + Q_d, \tag{5.1}$$

for some bounded linear operator $A \in \mathcal{L}(\mathcal{H}_0)$. In the following this form of the model will be called a d-dimensional functional IC (d-FIC) model and the dimension d is henceforth assumed to be fixed.

One way to motivate the simplification is via Gaussian functions; the mutually independent principal components $m_{k+1}, m_{k+2}, \ldots$ in the Karhunen-Loève expansion of the simplified model behave as if the original function was Gaussian. As normal distribution is a common choice for the distribution of errors in classical models, the previous allows us to consider the tail of the expansion as noise and the subspace $\mathcal{M}_d$ as the "signal subspace". Similar problems involving standardization in functional spaces have been solved using, for example, regularization (Li and Song, 2017). Let next $X^{(d)} = P_d X$ and $Z^{(d)} = P_d Z$ be the projections of the observed and latent function into the signal subspace, to which we may without loss of generality restrict ourselves.

To solve the functional IC model it is thus sufficient to find any operator $\Gamma \in \mathcal{L}(\mathcal{H}_0)$ such that $\Gamma X^{(d)}$ has independent components. However, we again require some additional structure from the estimators, as put forth in the following definition.

Definition 13. *The functional* $\Gamma : \mathcal{D} \to \mathcal{L}(\mathcal{M}_d)$ *is a d-dimensional functional independent component (d-FIC) functional if, for all* $X^{(d)}$ *coming from a d-FIC model, we have,*

i) $\Gamma(X^{(d)})X^{(d)}$ *has independent components and*

ii) $\Gamma(X^{(d)})X^{(d)} = \Gamma(AX^{(d)})AX^{(d)}$ *for all invertible linear operators* $A \in \mathcal{L}(\mathcal{H}_0)$

Definition 13 clearly parallels Definitions 2 and 9 but with one important distinction: the invariance property rests entirely on the assumption that the d-FIC model is accurate. In particular, the invariance is unlikely to hold in any practical situation, a price we have to pay for moving the concepts of ICA to the infinite-dimensional spaces.

Having solved the main issue caused by the infinite dimensionality of the space, Li et al. (2015) next defined a functional counterpart for the FOBI-matrix

using the tensor product,

$$B(X) = \mathrm{E}\left\{(X \otimes X)(X \otimes X)\right\} = \mathrm{E}\left\{\|X\|_0^2 (X \otimes X)\right\},$$

and showed that it exists and has an eigendecomposition under the assumption of finite fourth moments, $\mathrm{E}(\|X\|^4) < \infty$. The pair of operators, $\Sigma(X)$, $B(X)$, now shares many of the properties (equivariance, positive semidefiniteness) of their multivariate progenitors, not the least of which is the ability to solve the functional IC model. Standardization being now possible, Li et al. (2015) proved that,

$$\Sigma(X^{(d)})^{-1/2} X^{(d)} = U \Sigma(Z^{(d)})^{-1/2} Z^{(d)} \tag{5.2}$$

for some unitary operator $U \in \mathcal{L}(\mathcal{H}_0)$, showing that also now standardization solves the first half of the problem. By $\Sigma(X^{(d)})^{-1/2}$ we refer to any inverse square root of the covariance operator $\Sigma(X^{(d)})$ and as we did not make any identifiability constraints on Z also $\Sigma(Z^{(d)})^{-1/2}$ appears in the formula contrary to (3.5) and (4.3). We denote the standardized function in the following by $X_{st}^{(d)} = \Sigma(X^{(d)})^{-1/2} X^{(d)}$ and it naturally satisfies $\Sigma(X_{st}^{(d)}) = P_d$, where the projection operator P_d is equivalent to the identity operator inside $\mathcal{M}_d$. Approaching then the second part ("rotation") via fourth moments as in regular FOBI yields fruitful results and goes to show that the FOBI-methodology seems to know no limits in its areas of application. We define next the corresponding functional IC functional.

Definition 14. *Let X come from a d-FIC model with the covariance operator eigenbasis ϕ_k, $k \in \mathbb{N}_+$. Then the functional FOBI-functional is $\Gamma^{FF} = H\Sigma(X^{(d)})^{-1/2}$ where $H = \sum_{k=1}^{d} (\phi_k \otimes \psi_k)$ and ψ_k, $k \in \{1, \ldots, d\}$, are the eigenvectors of the FOBI-operator,*

$$B(X_{st}^{(d)}) = \sum_{k=1}^{d} \tau_k (\psi_k \otimes \psi_k).$$

Contrary to our earlier IC functionals the current unorthodox relationship of the model with the eigenbasis of the covariance operator forces us to define the IC functional in Definition (17) only for data coming from the corresponding IC model. The functional FOBI-functional can be shown to be a d-FIC functional under the following, familiar-looking assumption.

Assumption F1. *The excess kurtoses $\mathrm{E}(\langle Z, \phi_k \rangle_0^4) - 3$, $k \in \{1, \ldots, d\}$, are distinct.*

To apply the above methods to extract the independent components from a functional data sample assumed to obey the d-FIC model, a *coordinate representation* of the above was developed in Li et al. (2015), essentially translating the whole procedure into the form of linear algebra.

Other excursions into the functional realm from the viewpoint of ICA include: Gutch and Theis (2012) who discuss conditions for the identifability of the infinite-dimensional IC model, Rendón et al. (2014) who defined kurtosis for functional data and used it for clustering and Virta et al. (2017a) who developed a framework for conducting independent component analysis for

multivariate functional data. The multivariate functional setting is in a sense a more logical choice for functional ICA than the univariate one as the concept of independence is in the former naturally defined as a property between the component functions. However, this freedom comes at the price of compromising the identifiability of the corresponding IC model. This and other aspects of the multivariate ICA methodology will be the topics of our final section.

5.4 Multivariate functional ICA

Multivariate functional data

Before moving on to the multivariate functional IC model we first describe the concept of multivariate functional data itself. Let $\mathcal{H} = \mathcal{H}_0 \times \ldots \times \mathcal{H}_0$ be a p-variate functional space with identical component spaces H_0 as described in Section 5.1. The elements of $\mathcal{H}$ are multivariate functions $f = (f_1, \ldots f_p)$ from the interval T to $\mathbb{R}^p$. In practice a sample of multivariate functional data looks very much as the univariate functional sample in Figure 5.1, only that now we have p such curves for each individual. Now, $\mathcal{H}$ is as a direct sum of Hilbert spaces itself a Hilbert space, meaning that everything we wrote in Section 5.1 holds for multivariate functional data as well. However, we next introduce via the choice of inner product and linear operators some additional structure to $\mathcal{H}$ that ties its properties closely to those of its component spaces.

First, the geometry of the space is inherited from the component spaces by defining the inner product as $\langle f, g \rangle = \sum_{j=1}^{p} \langle f_j, g_j \rangle_0$, for all $f = (f_1, \ldots, f_p)$, $g = (g_1, \ldots, g_p) \in \mathcal{H}$. The induced norm then satisfies $\|f\|^2 = \sum_{j=1}^{p} \|f_j\|_0^2$. Our collection of bounded linear operators is likewise defined via the operators of the component spaces: the set $\mathcal{L}(\mathcal{H})$ consists of all operators from $\mathcal{H}$ to $\mathcal{H}$ which act as $f \mapsto (\sum_{j=1}^{p} L_{1j} f_j, \ldots, \sum_{j=1}^{p} L_{pj} f_j)$ for some collection of bounded linear operators $\{L_{ij} \in \mathcal{L}(\mathcal{H}_0) \mid i, j \in \{1, \ldots, p\}\}$. The action resembles closely ordinary matrix multiplication and, indeed, if we represent the operator as

$$L = \begin{pmatrix} L_{11} & \cdots & L_{1p} \\ \vdots & \ddots & \vdots \\ L_{p1} & \cdots & L_{pp} \end{pmatrix}$$

the mapping $f \mapsto Lf$ can be carried out using the rules for matrix multiplication. Basic functional analysis can be used to show that any operator in $\mathcal{L}(\mathcal{H})$ is bounded and linear. Two natural subclasses of $\mathcal{L}(\mathcal{H})$ stand out: unitary operators U which satisfy $\sum_{k=1}^{p} U_{ik} U_{jk} = \delta_{ij} \cdot id$, for all $i, j \in \{1, \ldots, p\}$, and diagonal operators L for which $L_{ij} = 0$ for all $i \neq j$.

Despite the current surge of interest in functional data, multivariate functional data has received considerably less attention in the literature. Multivariate functional PCA has been considered in Ramsay and Silverman (2006); Berrendero et al. (2011); Sato (2013); Chiou et al. (2014); Jacques and Preda (2014); Happ and Greven (2016), multivariate functional clustering in Tokushige et al. (2007); Kayano et al. (2010); Ieva et al. (2011); Jacques and Preda (2014), multivariate functional sufficient dimension reduction in Li and Song (2017) and, finally, multivariate independent component analysis in Virta et al. (2017a), the methodology of which we review next.

Multivariate functional IC model

Call the component functions of $X = (X_1, \ldots, X_p)$ independent if for all measurable mappings $g_1, \ldots, g_p : \mathcal{H}_0 \to \mathbb{R}$ the vector $(g_1(X_1), \ldots, g_p(X_p)) \in \mathbb{R}^p$ of real-valued random variables has independent components. The multivariate functional IC model can then be defined without reference to any basis.

Definition 15. *Let X be a random function in $\mathcal{H}$. Then we say that X obeys the multivariate functional independent component model if there exists an operator $\Gamma \in \mathcal{L}(\mathcal{H})$ such that the component functions of ΓX are independent.*

The lack of the eigenbasis of the covariance operator in Definition 15 means that we also avoid the need for any moment assumptions on X. Yet, such freedom is only temporary as again attempting to fit any analogue of standardization into the model leads into problems. Before that, however, we discuss a more pressing matter. The independence of the component functions of Z is naturally preserved under any map $Z \mapsto DZ$ where $D \in \mathcal{L}(\mathcal{H})$ is a diagonal operator. Thus, we can only hope to estimate the independent components up to transformations by some arbitrary linear operators and the recovering of the "original" components in any form cannot be guaranteed without strong additional assumptions on the model. However, this non-uniqueness sounds more adverse than it actually is in practice and Virta et al. (2017a) showed that various practical uses and interpretations can still be given to the extracted components.

The same restriction as in Li et al. (2015) was used also in Virta et al. (2017a) to solve issue with the standardization. That is, let the covariance operator of the observed multivariate random function $X = (X_1, \ldots, X_p)$ be the matrix of operators,

$$\Sigma(X) = \mathrm{E}\,(X \otimes X) = \begin{pmatrix} \mathrm{E}\,(X_1 \otimes X_1) & \cdots & \mathrm{E}\,(X_1 \otimes X_p) \\ \vdots & \ddots & \vdots \\ \mathrm{E}\,(X_p \otimes X_1) & \cdots & \mathrm{E}\,(X_p \otimes X_p) \end{pmatrix}.$$

The same arguments as in the univariate case show that $\Sigma(X)$ is adjoint, positive semidefinite, trace-class operator and admits a spectral decomposition, $\sum_{k=1}^{\infty} \lambda_k (\phi_k \otimes \phi_k)$ where the eigenvalues λ_k are in a decreasing order. Letting again $\mathcal{M}_d = \mathrm{span}(\phi_1, \ldots, \phi_d)$ for some fixed d, we assume that Γ in Definition 15 is of the form (5.1) for some $A \in \mathcal{L}(\mathcal{M}_d)$, where P_d and Q_d are again respectively the projections onto $\mathcal{M}_d$ and its orthogonal complement. If a random function $X \in \mathcal{H}$ obeys such a model for some fixed d, we say that X follows a d-dimensional multivariate functional independent component (d-MFIC) model.

The definition of a solution of the simplified IC model is given next in the form of a functional, where once again we have been obliged to modify our expectations of an IC functional to best suit the particular model and form of the data.

Definition 16. *The functional $\Gamma : \mathcal{D} \to \mathcal{L}(\mathcal{M}_d)$ is a d-dimensional multivariate functional independent component (d-MFIC) functional if, for all $X^{(d)}$ coming from a d-FIC model, we have,*

i) $\Gamma(X^{(d)})X^{(d)}$ can be divided into subvectors of functions so that each subvector corresponds to exactly one of the components of Z and is independent with the remaining $p - 1$.

The change in the first condition over the univariate functional IC functional in Definition 13 is caused by the interplay of the two conflicting forms of "dimensionality", p and d. Namely, if $d < p$ then we clearly cannot estimate all p component functions and if $d > p$ some of the estimated components must necessarily correspond to the same component of Z. Based on this a rule of thumb $d = p$ was advocated in Virta et al. (2017a). Furthermore, we have completely dispensed with any invariance or equivariance of the functionals as the second of our two primary estimators for the model does not admit such properties. However, this is only an aesthetic setback as the property *ii)* was in Definition 13 in any case completely theoretical and said nothing about the application of the methodology to real data.

Finally, it is not too much of a surprise to see that after the restriction to the finite-dimensional model, standardization by $\Sigma(X^{(d)})$ leaves us a unitary transformation away from our objective. Namely, if we again denote by $\Sigma(X^{(d)})^{-1/2}$ an inverse square root of $\Sigma(X^{(d)})$, we have,

$$\Sigma(X^{(d)})^{-1/2} X^{(d)} = U \Sigma(Z^{(d)})^{-1/2} Z^{(d)},$$

for some unitary $U \in \mathcal{L}(\mathcal{H})$. The estimation of the missing U is then our final task, something Virta et al. (2017a) achieved by generalizing both FOBI and JADE to multivariate functional data.

Multivariate functional FOBI and JADE

For our final extensions of FOBI and JADE we still require another set of "cumulant matrices", this time in the form of linear operators in $\mathcal{M}_d$. In some sense the canonical basis of the space is $\{\psi_1, \ldots, \phi_d\}$ and mimicking the formula for (3.7) prompts us to define,

$$\begin{aligned} C^{ij}(X) = &\operatorname{E}\left\{\langle X, \phi_i\rangle\langle X, \phi_j\rangle \, (X \otimes X)\right\} - \operatorname{E}\left\{\langle X, \phi_i\rangle\langle X^*, \phi_j\rangle \, (X \otimes X^*)\right\} \\ &- \operatorname{E}\left\{\langle X, \phi_i\rangle\langle X^*, \phi_j\rangle \, (X^* \otimes X)\right\} - \operatorname{E}\left\{\langle X, \phi_i\rangle\langle X, \phi_j\rangle \, (X^* \otimes X^*)\right\}, \end{aligned} \tag{5.3}$$

where X^* is an independent copy of X and the assumption of finite fourth moments, $\operatorname{E}(\|X\|^4) < \infty$, is sufficient for the existence of the operators. Again the extensions have no other motivation than analogy and once more that turns to be enough. Our interest lies with standardized random functions X, satisfying $\Sigma(X) = P_d$, and for any such function simplifying (5.3) reveals the familiar form,

$$C^{ij}(X) = \operatorname{E}\left\{\langle X, \phi_i\rangle\langle X, \phi_j\rangle \, (X \otimes X)\right\} - \phi_i \otimes \phi_j - \phi_j \otimes \phi_i - \delta_{ij} \cdot P_d.$$

Further substituting $X = UZ$, where U is unitary and $Z = (Z_1, \ldots, Z_p)$ has independent component functions, to the previous yields $C^{ij}(UZ) = UD^{ij}U^*$ for some diagonal operator $D^{ij} = D^{ij}(UZ)$, the exact form of which is given in Virta et al. (2017a). This implies that the missing unitary operator once again "diagonalizes" all operators $C^{ij}(X_{st})$, soon leading us to yet another forms of FOBI and JADE. However, one needs to remember that we are currently playing with two clashing forms of diagonality, the algebraic diagonality in the sense of spectral decompositions and the physical diagonality in the sense of diagonal operators. The operators D^{ij} belong to the latter group and to be able to use spectral decompositions of the cumulant operators to extract U Virta et al.

(2017a) introduced a connection between the two identically named concepts. To avoid overt repetition we refrain from paraphrasing their results here and simply state later the assumptions which are required for the connections to hold. Note that for Euclidean spaces these problems are averted as the two forms of diagonality exactly coincide.

The corresponding multivariate functional FOBI-operator for a standardized function X also has a recognizable structure,

$$C(X) = \mathrm{E}\left\{(X \otimes X)\,(X \otimes X)\right\} - (d+2) \cdot P_d,$$

and is easily shown to satisfy $C(UX) = UC(X)U^*$ for any unitary U and standardized X. Based on our earlier discussion $C(Z)$ is a diagonal operator and the spectral decomposition of the FOBI-operator proves sufficient grounds for defining the next IC functional.

Definition 17. *Let X come from a d-MFIC model with the covariance operator eigenbasis ϕ_k, $k \in N_+$. Then the multivariate functional FOBI-functional is $\Gamma^{MFF} = H\Sigma(X^{(d)})^{-1/2}$ where H is the linear operator with the action,*

$$f \mapsto (E_1(\psi_1 \otimes \psi_1)f, \ldots, E_1(\psi_d \otimes \psi_d)f)\,,$$

for all $f = (f_1, \ldots f_p) \in \mathcal{M}_d$ where $E_1 : \mathcal{M}_d \to \mathcal{H}_0$ is a linear operator (projection) that "picks" the first component of a multivariate function, $E_1 f = f_1$ and ψ_k, $k \in \{1, \ldots, d\}$, are the eigenvectors of the FOBI-operator $C(X_{st}^{(d)})$.

The multivariate functional FOBI-functional was shown to be a d-MFIC functional in Virta et al. (2017a) under Assumption M1 below. Additionally, it could be shown to be affine equivariant under the d-MFIC-model using analogous arguments to Li et al. (2015).

Assumption M1. *The eigenvalues of $C(Z_{st}^{(d)})$ are distinct.*

Sadly, no simple interpretations for the assumption through, e.g, "functional kurtosis" can be given.

The peculiar form of the FOBI functional stems from the identifiability problems discussed after the introduction of the model in Definition 12. Every projection $(\psi_k \otimes \psi_k)X_{st}^{(d)}$, $k \in \{1, \ldots, d\}$, onto an eigenvector defines a full p-variate random function in a 1-dimensional space where the only dependence on the independent components is through the inner products $\langle \psi_k, X_{st} \rangle$, $k \in \{1, \ldots, d\}$, each of which corresponds to exactly one latent function X_j. Thus the actual functional forms of the projections tell us nothing about the independent components and Virta et al. (2017a) actually equated the solution of the model with this d-vector of inner products. For the same reason the choice of the projection E_1 in Definition 15 is completely arbitrary and any other coordinate projection, $E_2, \ldots, E_p$, could have been used as well.

Moving on to our extension of JADE, one thing we are still missing is a way to combine the information in the cumulant operators C^{ij}. We thus need a concept of joint diagonalizer (understood in the spectral sense) for self-adjoint linear operators. Mimicking the standard finite-dimensional definition (3.9) Virta et al. (2017a) defined the joint diagonalizer of the set of operators in

$\mathcal{M}_d$, $\mathcal{S} = \{S_j \in \mathcal{L}(\mathcal{M}_d) \mid j \in \{1, \ldots, m\}\}$ to be the orthonormal basis ψ_k, $k \in \{1, \ldots, d\}$, of $\mathcal{M}_d$ that maximizes the objective function,

$$w(\psi_1, \ldots, \psi_d) = \sum_{j=1}^{m} \sum_{k=1}^{d} \langle S_j \psi_k, \psi_k \rangle^2. \tag{5.4}$$

With that, the definition of the JADE functional now easily follows.

Definition 18. *Let X come from a d-MFIC model with the covariance operator eigenbasis ϕ_k, $k \in \mathbb{N}_+$. Then the multivariate functional JADE-functional is $\Gamma^{MFJ} = H\Sigma(X^{(d)})^{-1/2}$ where H and E_1 are as in Definition 17 but the functions ψ_k, $k \in \{1, \ldots, d\}$, now constitute the joint diagonalizer of the set $\mathcal{C} = \{C^{ij}(X_{st}^{(d)}) \mid i, j \in \{1, \ldots, d\}\}$.*

To show the Fisher consistency of Γ^{MFJ}, we again need some connection between the spectral diagonalization offered by the joint diagonalization in (5.4) and the physical diagonality of the operators D^{ij}, and, indeed, under two technical conditions (Virta et al., 2017a) this is achieved and the multivariate functional JADE-functional is a d-MFIC functional.

After having extracted the independent components in the form of the inner products using either FOBI or JADE one issue still remains: how to determine which inner products correspond to the same independent component. In practice this has little bearing as the value of the components is most likely assessed by their usefulness in subsequent analyses. However, if some grouping is desired, similar devices to those used in Nordhausen and Oja (2011) in the context of independent subspace analysis (ISA) (Cardoso, 1998) could prove useful in our case as well. ISA can be seen as ICA for block vectors and as such is actually a finite-dimensional analogue for our current problem. Again coordinate representation of the methods can be devised, using which smoothed observed functional data can be subjected to the two methods.

Finally, our proven form of presentation naturally begs for asymptotical results for the functional IC methods too, but due to the current nonexistence of such results and the general difficulty of moving limiting results from multivariate to functional data we choose to stop here.

6 Discussion

We end with a short section collecting various directions for future research. As both tensorial and functional ICA are still emerging fields much work remains to be done. The presented extensions of FOBI and JADE serve well as preliminary excursions to the topic but if any of the intuition from vector-valued ICA carries over, better estimators are to be expected. Especially linear ICA methods are likely to have natural extensions in both cases, see for example the various forms of group ICA (Calhoun et al., 2009) which, while already targeted to tensor-valued data, rely typically on vectorization. Of special interest is also the extension of projection pursuit to the two discussed forms of non-standard data, especially in the light of Theorem 4 showing that SPP reaches the limiting efficiency of JADE with a lighter computational load. Some preliminary calculations already show that the same relationship holds with TJADE and tensorial PP when we define the latter as search for directions $\mathbf{u}$ maximizing the "tensorial kurtosis", $\mathrm{E}\{(\mathbf{u}^\top \mathbf{X}_{(m)} \mathbf{X}_{(m)}^\top \mathbf{u})^2\}$. Replacing then the square with any smooth enough function g would give us a general-purpose "tensorial FastICA". Similarly, projection pursuit in the case of univariate functional data is very straightforwardly postulated and will most likely lead into applying standard FastICA to the vector of functional principal component coefficients.

A major shortcoming of all the presented methods and the key issue preventing their large-scale application in practice is the lack of tools for dimension estimation. To reliably reduce the dimension of tensorial or functional observations a systematic procedure for determining the "correct" dimensionality of the latent tensors and functions is called for and the asymptotic and bootstrapping schemes developed in Nordhausen et al. (2016, 2017a) for vector-valued dimension reduction will likely provide a valuable starting point for hypothesis testing in the former case, assuming normally distributed noise. Similarly the general order determination technique of Luo and Li (2016) allowing the estimation of the rank of any scatter matrix estimate could be used for the same purpose. Assuming still Gaussian noise, for univariate functional ICA testing for the dimension is equivalent to testing whether the "tail" of the observed function is a Gaussian process. A naïve example test is obtained by choosing multiple subsets of the tail coordinates with different cardinality and conducting a multivariate test of normality for each.

Third interesting prospect is the studying of the connection between the statistical tensor methods' and the traditional tensor decompositions discussed in Section 4.2. Practically any tensor decomposition can be used to define a dimension reduction method by leaving the mode corresponding to the identically and independently distributed sample uncompressed and if the resulting methods are subjected to careful scrutiny, many interesting connections to statistical methodology are bound to be revealed.

Appendix

Proof of Lemma 1. Without loss of generality we assume that all considered roots of positive-definite matrices are selected to be the unique positive-definite choice. Assume first that the conditions of Definition 2 hold for $\mathbf{\Gamma}$. Then for any $\mathbf{z} \in \mathbb{R}^p$ with independent components we have, $\mathbf{V}(\mathbf{z}_{st}) = \mathbf{\Gamma}(\mathbf{z})\mathbf{\Sigma}(\mathbf{z})^{1/2} \equiv \mathbf{\Gamma}\{\mathbf{\Sigma}(\mathbf{z})^{-1/2}\mathbf{z}\}\mathbf{\Sigma}(\mathbf{z})^{-1/2}\mathbf{\Sigma}(\mathbf{z})^{1/2} \equiv \mathbf{I}_p$, where the first equivalence uses condition *ii)* of Definition 2 and the second equivalence condition *i)*. Similarly, for arbitrary $\mathbf{x} \in \mathbb{R}^p$ and $\mathbf{U} \in \mathcal{U}^{p\times p}$ we have,

$$\mathbf{V}(\mathbf{U}\mathbf{x}_{st}) = \mathbf{V}\{\mathbf{U}\mathbf{\Sigma}(\mathbf{x})^{-1/2}\mathbf{U}^\top\mathbf{U}\mathbf{x}\} = \mathbf{V}\{\mathbf{\Sigma}(\mathbf{U}\mathbf{x})^{-1/2}\mathbf{U}\mathbf{x}\} = \mathbf{V}\{(\mathbf{U}\mathbf{x})_{st}\}.$$

Consequently, by the form of the functional, $\mathbf{V}\{\mathbf{U}\mathbf{x}_{st}\} = \mathbf{\Gamma}(\mathbf{U}\mathbf{x})\mathbf{\Sigma}(\mathbf{U}\mathbf{x})^{1/2} \equiv \mathbf{\Gamma}(\mathbf{x})\mathbf{U}^\top\mathbf{U}\mathbf{\Sigma}(\mathbf{x})^{1/2}\mathbf{U}^\top \equiv \mathbf{\Gamma}(\mathbf{x})\mathbf{\Sigma}(\mathbf{x})^{1/2}\mathbf{U}^\top \equiv \mathbf{V}(\mathbf{x}_{st})\mathbf{U}^\top$, where the first equivalence uses condition *ii)* of Definition 2.

Assume then that the two conditions of Lemma 1 hold. Then for any standardized $\mathbf{z} \in \mathbb{R}^p$ with independent components, $\mathbf{\Gamma}(\mathbf{z}) = \mathbf{V}(\mathbf{z}_{st})\mathbf{\Sigma}(\mathbf{z})^{-1/2} = \mathbf{V}(\mathbf{z}_{st}) \equiv \mathbf{I}_p$ where the second equality holds as $\mathbf{\Sigma}(\mathbf{z}) = \mathbf{I}$ and the equivalence uses condition *i)* of Lemma 1. Finally, for arbitrary $\mathbf{x} \in \mathbb{R}^p$ and $\mathbf{A} \in \mathbb{R}^{p\times p}$ we have, $\mathbf{\Gamma}(\mathbf{A}\mathbf{x}) = \mathbf{V}((\mathbf{A}\mathbf{x})_{st})\mathbf{\Sigma}(\mathbf{A}\mathbf{x})^{-1/2}$. By Ilmonen et al. (2012) $\mathbf{\Sigma}(\mathbf{A}\mathbf{x})^{-1/2} = \mathbf{W}\mathbf{\Sigma}(\mathbf{x})^{-1/2}\mathbf{A}^{-1}$, where the orthogonal matrix $\mathbf{W} \in \mathcal{U}^{p\times p}$ is chosen so that the inverse root is symmetric. This in turn implies that $(\mathbf{A}\mathbf{x})_{st} = \mathbf{W}\mathbf{x}_{st}$ and we have $\mathbf{\Gamma}(\mathbf{A}\mathbf{x}) = \mathbf{V}(\mathbf{W}\mathbf{x}_{st})\mathbf{W}\mathbf{\Sigma}(\mathbf{x})^{-1/2}\mathbf{A}^{-1} \equiv \mathbf{\Gamma}(\mathbf{x})\mathbf{A}^{-1}$ where the equivalence follows from the condition *ii)* of Lemma 1. □

Proof of Lemma 4. Starting from the left-hand side of the claim we have,

$$c_m\left(\mathbf{v}_k^\top\mathbf{z}\right)^2 = \left\{\sum_{l=1}^p v_{kl}^m c_m(z_l)\right\}^2 \leq \sum_{l=1}^p v_{kl}^{2(m-1)} c_m^2(z_l),$$

where v_{kl} is the lth element of $\mathbf{v}_k$, the equality uses properties 2 and 3 from Lemma 3 and the inequality uses Cauchy-Schwarz to the partition $v_{kl}^m c_m(z_l) = \{v_{kl}\}\{v_{kl}^{m-1}c_m(z_l)\}$. By our assumptions $v_{kl} = 0$, for $1 \leq l \leq k-1$, and $c_m(z_l)^2 \leq c_m(z_k)^2$, for $k \leq l \leq p$. Using these, we get

$$c_m\left(\mathbf{v}_k^\top\mathbf{z}\right)^2 \leq \sum_{l=k}^p v_{kl}^{2(m-1)} c_m^2(z_l) \leq c_m^2(z_k)\sum_{l=k}^p v_{kl}^{2(m-1)} \leq c_m^2(z_k)\sum_{l=k}^p v_{kl}^2 \leq c_m^2(z_k),$$

where the second-to-last inequality uses the bound, $v_{kl}^{2(m-1)} = v_{kl}^2 v_{kl}^{2(m-2)} \leq v_{kl}^2$ and the last inequality the unit length of $\mathbf{v}_k$.

Now, plugging in $\mathbf{v}_k = \pm\mathbf{e}_l$ for some $l \in \{k,\ldots,p\}$ with $\{c_m(z_l)\}^2 = \{c_m(z_k)\}^2$ instantly reveals that equality holds. The converse is seen to be true by checking what is implied if the second-to-last inequality above actually preserves equality. In that case we must have $c_m^2(z_k)\sum_{l=k}^p v_{kl}^2(1-v_{kl}^{2(m-2)}) = 0$, all terms and factors of which are non-negative. Thus either $c_m^2(z_k) = 0$ or, for

all $l \in \{k, \ldots, p\}$, $v_{kl} \in \{-1, 0, 1\}$. As $k \leq p$ we cannot have $c_m^2(z_k) = 0$ and thus by the unit length constraint $\mathbf{v}_k = \pm\mathbf{e}_l$ for some $l \in \{k, \ldots, p\}$. Assume then that the whole chain of inequalities preserves equality. Then both the previous must hold and the second inequality in the chain must preserve equality. This implies that $c_m^2(z_l) = c_m^2(z_k)$ for the same $l \in \{k, \ldots, p\}$ for which $\mathbf{v}_k = \pm\mathbf{e}_l$ holds, showing that the postulated condition is not just sufficient but also necessary for the equality.

□

Proof of Lemma 5. As in the proof of Lemma 4 we have,

$$\sum_{k=1}^{p} c_m \left(\mathbf{v}_k^\top \mathbf{z}\right)^2 = \sum_{k=1}^{p} \left\{ \sum_{l=1}^{p} v_{kl}^m c_m(z_l) \right\}^2 \leq \sum_{k=1}^{p} \sum_{l=1}^{p} v_{kl}^{2(m-1)} c_m^2(z_l).$$

Using again the upper bound, $v_{kl}^{2(m-1)} \leq v_{kl}^2$, now gives the desired inequality.

Straightforward substitution $\mathbf{V}^\top = \mathbf{PJ}$, where $\mathbf{P} \in \mathcal{P}^p$ and $\mathbf{J} \in \mathcal{J}^p$ are arbitrary, next reveals that the equality is preserved in this case. To show the converse we observe exactly when the application of the bound $v_{kl}^{2(m-1)} \leq v_{kl}^2$ preserves equality: for all k we must have $\sum_{l=1}^{p} v_{kl}^{2(m-1)} c_m^2(z_l) = \sum_{l=1}^{p} v_{kl}^2 c_m^2(z_l)$ which is equivalent to requiring $\sum_{l=1}^{p} v_{kl}^2 c_m^2(z_l)(1 - v_{kl}^{2(m-2)})$. All factors in the terms of the previous sum are non-negative and thus if equality is reached then we have for all pairs (k, l) either $c_m^2(z_l) = 0$ or $v_{kl} \in \{-1, 0, 1\}$. By our assumption at most one of the cumulants is zero so at least $p - 1$ columns of $\mathbf{V}^\top$ consist entirely of the elements $-1, 0, 1$. As $\mathbf{V}^\top$ is orthogonal each of these columns must then contain exactly one ± 1 and rest of the entries must be zero, meaning that to achieve orthogonality also the final column must be of the same form. Thus $\mathbf{V}^\top = \mathbf{PJ}$ and we have shown this to be both a sufficient and a necessary condition for equality.

□

Proof of Lemma 6. The proof is similar to that of Lemma 1. Assume first that the conditions of Definition 9 hold. Then, for any $\mathcal{X}$ coming from the tensorial IC model with diagonal mixing $\mathbf{D}_m$, $m \in \{1, \ldots, r\}$, we have,

$$\mathbf{V}_m(\mathcal{X}_{st}) = \mathbf{\Gamma}_m(\mathcal{X})\mathbf{\Sigma}_m(\mathcal{X})^{1/2} \overset{*}{\equiv} \mathbf{D}_m^{-1}(\tau_m \mathbf{D}_m^2)^{1/2} \overset{*}{\equiv} \mathbf{I}_{p_m} \equiv \mathbf{I}_{p_m},$$

where the first equivalence uses condition *i)* from Definition 9 and a result from Virta et al. (2017b) that under arbitrary mixing $\mathbf{\Sigma}_m(\mathcal{X}) \propto \mathbf{\Omega}_m \mathbf{\Omega}_m^\top$. The final equivalence $\equiv$ must hold because both $\mathbf{V}_m(\mathcal{X}_{st})$ and $\mathbf{I}_{p_m}$ are orthogonal matrices and thus have fixed scale. Assuming then an arbitrary tensor $\mathcal{X}$ and orthogonal matrices $\mathbf{U}_1, \ldots, \mathbf{U}_r$, we have,

$$\begin{aligned}
\mathbf{V}_m(\mathcal{X}_{st} \times_{m=1}^{r} \mathbf{U}_m) &= \mathbf{V}_m\{\mathcal{X} \times_{m=1}^{r} (\mathbf{U}_m \mathbf{\Sigma}_m(\mathcal{X})^{-1/2} \mathbf{U}_m^\top \mathbf{U}_m)\} \\
&= \mathbf{V}_m[\mathcal{X} \times_{m=1}^{r} \{\mathbf{\Sigma}_m(\mathcal{X} \times_{m=1}^{r} \mathbf{U}_m)^{-1/2} \mathbf{U}_m\}] \\
&= \mathbf{V}_m\{(\mathcal{X} \times_{m=1}^{r} \mathbf{U}_m)_{st}\}
\end{aligned}$$

where we have used the same trick as in the proof of Lemma 1 along with the orthogonal equivariance of $\mathbf{\Sigma}_m$. Consequently,

$$\begin{aligned}\mathbf{V}_m(\mathcal{X}_{st} \times_{m=1}^r \mathbf{U}_m) &= \mathbf{\Gamma}_m(\mathcal{X} \times_{m=1}^r \mathbf{U}_m)\mathbf{\Sigma}_m(\mathcal{X} \times_{m=1}^r \mathbf{U}_m)^{1/2} \\ &\equiv \mathbf{\Gamma}_m(\mathcal{X})\mathbf{U}_m^\top \mathbf{U}_m \mathbf{\Sigma}_m(\mathcal{X})^{1/2}\mathbf{U}_m^\top \\ &\equiv \mathbf{V}_m(\mathcal{X}_{st})\mathbf{U}_m^\top,\end{aligned}$$

where the first equivalence uses condition *ii)* from Definition 9.

Assume then that $\mathbf{\Gamma}_m$ satisfies the conditions of Lemma 6. Now, for any $\mathcal{X}$ coming from the tensorial IC model,

$$\mathbf{\Gamma}_m(\mathcal{X}) = \mathbf{V}_m(\mathcal{X}_{st})\mathbf{\Sigma}_m(\mathcal{X})^{-1/2} = \mathbf{V}_m(\tau\mathcal{Z} \times_{m=1}^r \mathbf{U}_m\mathbf{V}_m^\top)\tau_m\mathbf{U}_m\mathbf{D}_m^{-1}\mathbf{U}_m^\top,$$

where $\tau, \tau_m \in \mathbb{R}$ and we have used the singular value decomposition $\mathbf{\Omega}_m = \mathbf{U}_m\mathbf{D}_m\mathbf{V}_m^\top$ along with Lemma 2 and Theorem 2 from Virta et al. (2017b). The same theorem also shows that $\tau\mathcal{Z} = (\mathcal{Z} \times_{m=1}^r \mathbf{D}_m)_{st}$ and using conditions *ii)* and *i)* of Lemma 6 we get,

$$\begin{aligned}\mathbf{\Gamma}_m(\mathcal{X}) &\equiv \tau_m\mathbf{V}_m\{(\mathcal{Z} \times_{m=1}^r \mathbf{D}_m)_{st}\}\mathbf{V}_m\mathbf{U}_m^\top\mathbf{U}_m\mathbf{D}_m^{-1}\mathbf{U}_m^\top \\ &\overset{*}{\equiv} \mathbf{I}_{p_m}\mathbf{V}_m\mathbf{D}_m^{-1}\mathbf{U}_m^\top \\ &\overset{*}{\equiv} \mathbf{\Omega}_m^{-1}.\end{aligned}$$

Let next the tensor $\mathcal{X}$ and the orthogonal $\mathbf{U}_1, \ldots, \mathbf{U}_r$ be arbitrary. Then, as in the proof of the second condition of Lemma 6 above,

$$\begin{aligned}\mathbf{\Gamma}_m(\mathcal{X} \times_{m=1}^r \mathbf{U}_m) &= \mathbf{V}_m\{(\mathcal{X} \times_{m=1}^r \mathbf{U}_m)_{st}\}\mathbf{\Sigma}_m(\mathcal{X} \times_{m=1}^r \mathbf{U}_m)^{-1/2} \\ &= \mathbf{V}_m(\mathcal{X}_{st} \times_{m=1}^r \mathbf{U}_m)\mathbf{U}_m\mathbf{\Sigma}_m(\mathcal{X})^{-1/2}\mathbf{U}_m^\top \\ &\equiv \mathbf{\Gamma}_m(\mathcal{X})\mathbf{U}_m^\top,\end{aligned}$$

where the equivalence uses condition *ii)* of Lemma 6. □

Summaries of original publications

I ICA is well-known for its limitation concerning the number of Gaussian variables. In Virta et al. (2016b) this constraint is relaxed and the statistical properties of two types of projection pursuit estimators are derived under the obtained non-Gaussian ICA model. The efficiency of the estimators can be divided into two parts: how well they separate the signal from the noise and how well they separate the individual signals from each other. The two estimators are shown to be asymptotically equivalent with respect to the former and simulations are used to investigate their differences with respect to the latter. A sequential hypothesis test for estimating the dimension of the signal subspace is also proposed.

II Virta et al. (2017b) present the novel concept of tensorial independent component analysis, a dimension reduction framework for higher-order data. An estimator based on the classical fourth-order method FOBI is proposed and its consistency and limiting distribution are derived under the tensorial independent component model. Simulations and a real data example reveal that for both finite samples and in the limit the proposed estimator is superior to the commonly used alternative of vectorization.

III An improvement over FOBI in standard independent component analysis is given by JADE which better utilizes the fourth-order information in the estimation. Motivated by this, Virta et al. (2017c) present tensorial JADE and derive its statistical properties under the tensorial independent component model. Comparisons with both the tensorial FOBI and vectorial ICA methods show that the previously mentioned relationship between FOBI and JADE continues to hold also for tensorial random variables.

IV A popular example of higher order data is given by fMRI-measurements and in article Virta et al. (2016c) the methods developed in Virta et al. (2017b,c) are put to use in the context of simulated noisy fMRI-data. The results imply that the tensorial ICA model can be used to approximate the structure of brain image data and that the methods of tensorial ICA can reliably extract signals from noisy, high-dimensional tensor-valued data.

V The growing popularity of functional data has also increased the need for functional dimension reduction methodology. In Virta et al. (2017a) an independent component analysis framework is developed for multivariate functional data. The methodology is again based on the ideas behind FOBI and JADE and the obtained two estimators are shown to be Fisher consistent under the introduced multivariate functional independent component model. Both simulations and a real data example reveal that the proposed estimators surpass functional principal component analysis in efficiency.

References

S.-I. Amari, A. Cichocki, and H. H. Yang. A new learning algorithm for blind signal separation. In *Advances in Neural Information Processing Systems*, pages 757–763, 1996.

T. W. Anderson. Nonnormal multivariate distributions: Inference based on elliptically contoured distributions. Technical report, Stanford University, Department of Statistics, 1992.

M. Arashi. Some theoretical results on tensor elliptical distribution. *arXiv preprint arXiv:1709.00801*, 2017.

J. L. Bali and G. Boente. Principal points and elliptical distributions from the multivariate setting to the functional case. *Statistics & Probability Letters*, 79:1858–1865, 2009.

J. L. Bali, G. Boente, D. E. Tyler, and J.-L. Wang. Robust functional principal components: A projection-pursuit approach. *The Annals of Statistics*, 39: 2852–2882, 2011.

A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines. A blind source separation technique using second-order statistics. *IEEE Transactions on Signal Processing*, 45:434–444, 1997.

J. R. Berrendero, A. Justel, and M. Svarc. Principal components for multivariate functional data. *Computational Statistics & Data Analysis*, 55:2619–2634, 2011.

P. Billingsley. *Probability and measure*. John Wiley & Sons, 2008.

G. Blanchard, M. Sugiyama, M. Kawanabe, V. Spokoiny, and K.-R. Müller. Non-Gaussian component analysis: A semi-parametric framework for linear dimension reduction. In *Advances in Neural Information Processing Systems*, pages 131–138, 2005.

G. Boente, M. S. Barrera, and D. E. Tyler. A characterization of elliptical distributions and some optimality properties of principal components for functional data. *Journal of Multivariate Analysis*, 131:254–264, 2014.

S. Bonhomme and J.-M. Robin. Consistent noisy independent component analysis. *Journal of Econometrics*, 149:12–25, 2009.

D. Bosq. *Linear processes in function spaces: theory and applications*, volume 149. Springer Science & Business Media, 2012.

V. Calhoun, T. Adali, L. K. Hansen, J. Larsen, and J. Pekar. ICA of functional MRI data: An overview. In *Proceedings of the 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, pages 281–288, 2003.

V. D. Calhoun and T. Adali. Unmixing fMRI with independent component analysis. *IEEE Engineering in Medicine and Biology Magazine*, 25:79–90, 2006.

V. D. Calhoun, J. Liu, and T. Adali. A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. *Neuroimage*, 45: S163–S172, 2009.

J.-F. Cardoso. Source separation using higher order moments. In *Proceedings of the 1989 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2109–2112. IEEE, 1989.

J.-F. Cardoso. Multidimensional independent component analysis. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 4, pages 1941–1944. IEEE, 1998.

J.-F. Cardoso and A. Souloumiac. Blind beamforming for non-Gaussian signals. In *IEE proceedings F (radar and signal processing)*, volume 140, pages 362–370. IET, 1993.

F. J. Caro-Lopera, G. G. Farías, and N. Balakrishnan. Matrix-variate distribution theory under elliptical models-4: Joint distribution of latent roots of covariance matrix and the largest and smallest latent roots. *Journal of Multivariate Analysis*, 145:224–235, 2016.

A. Chen and P. J. Bickel. Efficient independent component analysis. *The Annals of Statistics*, 34:2825–2855, 2006.

J.-M. Chiou, Y.-T. Chen, and Y.-F. Yang. Multivariate functional principal component analysis: A normalization approach. *Statistica Sinica*, 2014:1571–1596, 2014.

A. Cichocki, R. Zdunek, A. H. Phan, and S.-I. Amari. *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009.

A. Cichocki, D. Mandic, L. De Lathauwer, G. Zhou, Q. Zhao, C. Caiafa, and H. A. Phan. Tensor decompositions for signal processing applications: From two-way to multiway component analysis. *IEEE Signal Processing Magazine*, 32:145–163, 2015.

D. B. Clarkson. A least squares version of algorithm AS 211: The FG diagonalization algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 37:317–321, 1988.

P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36:287–314, 1994.

P. Comon, Y. Qi, and K. Usevich. A polynomial formulation for joint decomposition of symmetric tensors of different orders. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 22–30. Springer, 2015.

J. B. Conway. *A course in functional analysis*, volume 96. Springer Science & Business Media, 2013.

L. De Lathauwer, B. De Moor, and J. Vandewalle. A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications*, 21:1253–1278, 2000.

S. Ding and R. D. Cook. Dimension folding PCA and PFC for matrix-valued predictors. *Statistica Sinica*, 24:463–492, 2014.

S. Ding and R. D. Cook. Higher-order sliced inverse regressions. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7:249–257, 2015a.

S. Ding and R. D. Cook. Tensor sliced inverse regression. *Journal of Multivariate Analysis*, 133:216–231, 2015b.

S. C. Douglas. Fixed-point algorithms for the blind separation of arbitrary complex-valued non-Gaussian signal mixtures. *EURASIP Journal on Advances in Signal Processing*, 2007:036525, 2007.

L. Dümbgen, M. Pauly, and T. Schweizer. M-functionals of multivariate scatter. *Statistics Surveys*, 9:32–105, 2015.

K.-T. Fang, S. Kotz, and K. W. Ng. *Symmetric multivariate and related distributions*. Chapman and Hall, 1990.

F. Ferraty and P. Vieu. *Nonparametric functional data analysis: Theory and practice*. Springer Science & Business Media, 2006.

L. Ferré and A.-F. Yao. Functional sliced inverse regression analysis. *Statistics*, 37:475–488, 2003.

L. Ferré and A.-F. Yao. Smoothed functional inverse regression. *Statistica Sinica*, 15:665–683, 2005.

J. H. Friedman and J. W. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, 100:881–890, 1974.

S. Ghurye and I. Olkin. A characterization of the multivariate normal distribution. *The Annals of Mathematical Statistics*, 33:533–541, 1962.

A. Gupta and T. Varga. Normal mixture representations of matrix variate elliptically contoured distributions. *Sankhyā: The Indian Journal of Statistics, Series A*, 57:68–78, 1995.

A. K. Gupta and D. K. Nagar. *Matrix variate distributions*. CRC Press, 1999.

H. W. Gutch and F. J. Theis. To infinity and beyond: On ICA over Hilbert spaces. In *Latent Variable Analysis and Signal Separation*, pages 180–187. Springer, 2012.

M. Hallin and C. Mehta. R-estimation for asymmetric independent component analysis. *Journal of the American Statistical Association*, 110:218–232, 2015.

C. Happ and S. Greven. Multivariate functional principal component analysis for data observed on different (dimensional) domains. *Journal of the American Statistical Association*, 2016. Accepted.

T. Hastie and R. Tibshirani. Independent components analysis through product density estimation. In *Advances in Neural Information Processing Systems*, pages 665–672, 2003.

G. He, H.-G. Müller, and J.-L. Wang. Methods of canonical analysis for functional data. *Journal of Statistical Planning and Inference*, 122:141–159, 2004.

L. Horváth and P. Kokoszka. *Inference for functional data with applications*. Springer Science & Business Media, 2012.

H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417, 1933.

T. Hsing and R. Eubank. *Theoretical foundations of functional data analysis, with an introduction to linear operators*. John Wiley & Sons, 2015.

T. Hsing and H. Ren. An RKHS formulation of the inverse regression dimension-reduction problem. *The Annals of Statistics*, 37:726–755, 2009.

P. J. Huber. Projection pursuit. *The Annals of Statistics*, 13:435–475, 1985.

H. Hung and C.-C. Wang. Matrix variate logistic regression model with application to EEG data. *Biostatistics*, 14:189–202, 2012.

A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10:626–634, 1999.

A. Hyvärinen and U. Köster. FastISA: A fast fixed-point algorithm for independent subspace analysis. In *EsANN*, pages 371–376, 2006.

F. Ieva, A. M. Paganoni, D. Pigoli, and V. Vitelli. ECG signal reconstruction, landmark registration and functional classification. In *7th Conference on Statistical Computation and Complex System*, 2011.

K. Illner, J. Miettinen, C. Fuchs, S. Taskinen, K. Nordhausen, H. Oja, and F. J. Theis. Model selection using limiting distributions of second-order blind source separation algorithms. *Signal Processing*, 113:95–103, 2015.

P. Ilmonen and D. Paindaveine. Semiparametrically efficient inference based on signed ranks in symmetric independent component models. *The Annals of Statistics*, 39:2448–2476, 2011.

P. Ilmonen, J. Nevalainen, and H. Oja. Characteristics of multivariate distributions and the invariant coordinate system. *Statistics & Probability Letters*, 80:1844–1853, 2010a.

P. Ilmonen, K. Nordhausen, H. Oja, and E. Ollila. A new performance index for ICA: Properties, computation and asymptotic analysis. *Latent Variable Analysis and Signal Separation*, pages 229–236, 2010b.

P. Ilmonen, H. Oja, and R. Serfling. On invariant coordinate system (ICS) functionals. *International Statistical Review*, 80:93–110, 2012.

J. Jacques and C. Preda. Model-based clustering for multivariate functional data. *Computational Statistics & Data Analysis*, 71:92–106, 2014.

I. Jolliffe. *Principal component analysis*. Springer Verlag, 2002.

T. Kariya and B. K. Sinha. *Robustness of statistical tests*. Academic Press, 2014.

J. Karvanen and V. Koivunen. Blind separation methods based on Pearson system and its extensions. *Signal Processing*, 82:663–673, 2002.

J. Karvanen, J. Eriksson, and V. Koivunen. Adaptive score functions for maximum likelihood ICA. *Journal of VLSI signal processing systems for signal, image and video technology*, 32:83–92, 2002.

M. Kawanabe. Linear dimension reduction based on the fourth-order cumulant tensor. In *International Conference on Artificial Neural Networks*, pages 151–156. Springer, 2005.

M. Kayano, K. Dozono, and S. Konishi. Functional cluster analysis via orthonormalized Gaussian basis expansions and its application. *Journal of Classification*, 27:211–230, 2010.

T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Review*, 51:455–500, 2009.

Z. Koldovský, P. Tichavský, and E. Oja. Efficient variant of algorithm FastICA for independent component analysis attaining the Cramér-Rao lower bound. *IEEE Transactions on Neural Networks*, 17:1265–1277, 2006.

T. Kollo. Multivariate skewness and kurtosis measures with an application in ICA. *Journal of Multivariate Analysis*, 99:2328–2338, 2008.

T. Kollo and D. von Rosen. *Advanced multivariate statistics with matrices*. Springer Science & Business Media, 2006.

B. Li and J. Song. Nonlinear sufficient dimension reduction for functional data. *The Annals of Statistics*, 45:1059–1095, 2017.

B. Li, M. K. Kim, and N. Altman. On dimension folding of matrix-or array-valued statistical objects. *The Annals of Statistics*, 38:1094–1121, 2010.

B. Li, G. Van Bever, H. Oja, R. Sabolová, and F. Critchley. Functional independent component analysis: an extension of the fourth-order blind identification. 2015. Submitted.

L. Li and X. Zhang. Parsimonious tensor response regression. *Journal of the American Statistical Association*, pages 1–16, 2017.

M. Lifshits. *Lectures on Gaussian processes*. Springer, 2012.

W. Luo and B. Li. Combining eigenvalues and variation of eigenvectors for order determination. *Biometrika*, 103:875–887, 2016.

A. M. Manceur and P. Dutilleul. Maximum likelihood estimation for the tensor normal distribution: Algorithm, minimum sample size, and empirical bias and dispersion. *Journal of Computational and Applied Mathematics*, 239: 37–49, 2013.

J. I. Marden. Some robust estimates of principal components. *Statistics & Probability Letters*, 43:349–359, 1999.

R. A. Maronna and V. J. Yohai. Robust estimation of multivariate location and scatter. *Wiley StatsRef: Statistics Reference Online*, 1976.

M. Matilainen, K. Nordhausen, and H. Oja. New independent component analysis tools for time series. *Statistics & Probability Letters*, 105:80–87, 2015.

D. S. Matteson and R. S. Tsay. Independent component analysis via distance covariance. *Journal of the American Statistical Association*, pages 1–16, 2017.

P. McCullagh. *Tensor methods in statistics*. Chapman and Hall London, 1987.

J. Miettinen, K. Nordhausen, H. Oja, and S. Taskinen. Fast equivariant JADE. In *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6153–6157. IEEE, 2013.

J. Miettinen, K. Nordhausen, H. Oja, and S. Taskinen. Deflation-based FastICA with adaptive choices of nonlinearities. *IEEE Transactions on Signal Processing*, 62:5716–5724, 2014a.

J. Miettinen, K. Nordhausen, H. Oja, and S. Taskinen. Deflation-based separation of uncorrelated stationary time series. *Journal of Multivariate Analysis*, 123:214–227, 2014b.

J. Miettinen, S. Taskinen, K. Nordhausen, and H. Oja. Fourth moments and independent component analysis. *Statistical Science*, 30:372–390, 2015.

J. Miettinen, K. Nordhausen, H. Oja, S. Taskinen, and J. Virta. The squared symmetric FastICA estimator. *Signal Processing*, 131:402–411, 2017.

E. Moreau. A generalization of joint-diagonalization criteria for source separation. *IEEE Transactions on Signal Processing*, 49:530–541, 2001.

K. Nordhausen and H. Oja. Independent subspace analysis using three scatter matrices. *Austrian Journal of Statistics*, 40:93–101, 2011.

K. Nordhausen and D. E. Tyler. A cautionary note on robust covariance plug-in methods. *Biometrika*, 102:573–588, 2015.

K. Nordhausen, H. Oja, and E. Ollila. Robust independent component analysis based on two scatter matrices. *Austrian Journal of Statistics*, 37:91–100, 2008.

K. Nordhausen, P. Ilmonen, A. Mandal, H. Oja, and E. Ollila. Deflation-based FastICA reloaded. In *Signal Processing Conference, 2011 19th European*, pages 1854–1858. IEEE, 2011a.

K. Nordhausen, E. Ollila, and H. Oja. On the performance indices of ICA and blind source separation. In *IEEE 12th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), 2011*, pages 486–490. IEEE, 2011b.

K. Nordhausen, H. Oja, P. Filzmoser, and C. Reimann. Blind source separation for spatial compositional data. *Mathematical Geosciences*, 47:753–770, 2015.

K. Nordhausen, H. Oja, and D. E. Tyler. Asymptotic and bootstrap tests for subspace dimension. *arXiv preprint arXiv:1611.04908*, 2016.

K. Nordhausen, H. Oja, D. E. Tyler, and J. Virta. Asymptotic and bootstrap tests for the dimension of the non-Gaussian subspace. *IEEE Signal Processing Letters*, 24:887–891, 2017a.

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

M. Ohlson, M. R. Ahmad, and D. Von Rosen. The multilinear normal distribution: Introduction and some basic properties. *Journal of Multivariate Analysis*, 113:37–47, 2013.

H. Oja. *Multivariate nonparametric methods with R: An approach based on spatial signs and ranks*. Springer Science & Business Media, 2010.

H. Oja, S. Sirkiä, and J. Eriksson. Scatter matrices and independent component analysis. *Austrian Journal of Statistics*, 35:175–189, 2006.

E. Ollila. The deflation-based FastICA estimator: Statistical analysis revisited. *IEEE Transactions on Signal Processing*, 58:1527–1541, 2010.

D. Paindaveine. Elliptical symmetry. *Encyclopedia of Environmetrics*, 2012.

V. I. Paulsen and M. Raghupathi. *An introduction to the theory of reproducing kernel Hilbert spaces*. Cambridge University Press, 2016.

K. Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2:559–572, 1901.

R. M. Pfeiffer, L. Forzani, and E. Bura. Sufficient dimension reduction for longitudinally measured predictors. *Statistics in Medicine*, 31:2414–2427, 2012.

J. O. Ramsay and B. Silverman. *Functional data analysis*. Wiley Online Library, 2006.

C. E. Rasmussen and C. K. Williams. *Gaussian processes for machine learning*. MIT press Cambridge, 2006.

C. Rendón, F. J. Prieto, and D. Peña. Independent components techniques based on kurtosis for functional data analysis. Technical report, Charles III University of Madrid, Departament of Statistics, 2014.

R. J. Samworth and M. Yuan. Independent component analysis via nonparametric maximum likelihood estimation. *The Annals of Statistics*, 40:2973–3002, 2012.

Y. Sato. Theoretical considerations for multivariate functional data analysis. In *Proceedings 59th ISI World Statistics Congress*, pages 25–30, August 2013.

R. J. Serfling. *Approximation theorems of mathematical statistics*. John Wiley & Sons, 2009.

N. D. Sidiropoulos, L. De Lathauwer, X. Fu, K. Huang, E. E. Papalexakis, and C. Faloutsos. Tensor decomposition for signal processing and machine learning. *IEEE Transactions on Signal Processing*, 65:3551–3582, 2017.

M. S. Srivastava, T. von Rosen, and D. Von Rosen. Models with a Kronecker product covariance structure: estimation and testing. *Mathematical Methods of Statistics*, 17:357–370, 2008.

J. Stone, J. Porrill, N. Porter, and I. Wilkinson. Spatiotemporal independent component analysis of event-related fMRI data using skewed probability density functions. *NeuroImage*, 15:407–421, 2002.

S. Taskinen, S. Sirkiä, and H. Oja. Independent component analysis based on symmetrised scatter matrices. *Computational Statistics & Data Analysis*, 51: 5103–5111, 2007.

S. Tokushige, H. Yadohisa, and K. Inada. Crisp and fuzzy k-means clustering algorithms for multivariate functional data. *Computational Statistics*, 22: 1–16, 2007.

L. Tong, V. Soon, Y. Huang, and R. Liu. AMUSE: a new blind identification algorithm. In *IEEE International Symposium on Circuits and Systems*, pages 1784–1787. IEEE, 1990.

D. E. Tyler, F. Critchley, L. Dümbgen, and H. Oja. Invariant co-ordinate selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71:549–592, 2009.

A. W. Van der Vaart. *Asymptotic statistics*. Cambridge University Press, 1998.

C. F. Van Loan. The ubiquitous Kronecker product. *Journal of Computational and Applied Mathematics*, 123:85–100, 2000.

M. A. O. Vasilescu and D. Terzopoulos. Multilinear independent components analysis. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 547–553. IEEE, 2005.

J. Virta. One-step M-estimates of scatter and the independence property. *Statistics & Probability Letters*, 110:133–136, 2016.

J. Virta and K. Nordhausen. Blind source separation of tensor-valued time series. *Signal Processing*, 141:204–216, 2017a.

J. Virta and K. Nordhausen. Blind source separation for non-stationary tensor-valued time series. In *IEEE International Workshop on Machine Learningfor Signal Processing*, 2017b.

J. Virta and K. Nordhausen. On the optimal non-linearities for Gaussian mixtures in FastICA. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 427–437. Springer, 2017c.

J. Virta, K. Nordhausen, and H. Oja. Joint use of third and fourth cumulants in independent component analysis. *Unpublished manuscript, preprint at arXiv:1505.02613*, 2015.

J. Virta, B. Li, K. Nordhausen, and H. Oja. *tensorBSS: Blind Source Separation Methods for Tensor-Valued Observations*, 2016a. URL `https://CRAN.R-project.org/package=tensorBSS`. R package version 0.3.3.

J. Virta, K. Nordhausen, and H. Oja. Projection pursuit for non-Gaussian independent components. *arXiv preprint arXiv:1612.05445*, 2016b.

J. Virta, S. Taskinen, and K. Nordhausen. Applying fully tensorial ICA to fMRI data. In *Signal Processing in Medicine and Biology Symposium (SPMB), 2016 IEEE*, pages 1–6. IEEE, 2016c.

J. Virta, B. Li, K. Nordhausen, and H. Oja. Independent component analysis for multivariate functional data. *Submitted to the Journal of the American Statistical Association*, 2017a.

J. Virta, B. Li, K. Nordhausen, and H. Oja. Independent component analysis for tensor-valued data. *Journal of Multivariate Analysis*, 162:172–192, 2017b.

J. Virta, B. Li, K. Nordhausen, and H. Oja. JADE for tensor-valued observations. *Accepted to Journal of Computational and Graphical Statistics*, 2017c. arXiv preprint arXiv:1603.05406.

S. Visuri, V. Koivunen, and H. Oja. Sign and rank covariance matrices. *Journal of Statistical Planning and Inference*, 91:557–575, 2000.

T. Wei. On the spurious solutions of the FastICA algorithm. In *IEEE Workshop on Statistical Signal Processing (SSP)*, pages 161–164. IEEE, 2014.

K. Werner, M. Jansson, and P. Stoica. On estimation of covariance matrices with Kronecker product structure. *IEEE Transactions on Signal Processing*, 56:478–491, 2008.

F. Yao, H.-G. Müller, and J.-L. Wang. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100:577–590, 2005.

L. Zhang, Q. Gao, and D. Zhang. Directional independent component analysis with tensor representation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–7. IEEE, 2008.

J. Zhao and C. Leng. Structured lasso for regression with matrix covariates. *Statistica Sinica*, 24:799–814, 2014.

W. Zhong, X. Xing, and K. Suslick. Tensor sufficient dimension reduction. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7:178–184, 2015.

H. Zhou and L. Li. Regularized matrix regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76:463–483, 2014.

H. Zhou, L. Li, and H. Zhu. Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108:540–552, 2013.

Annales Universitatis Turkuensis

Turun yliopisto
University of Turku

Painosalama Oy, Turku, Finland 2018

ISBN 978-951-29-7148-0 (PRINT)
ISBN 978-951-29-7149-7 (PDF)
ISSN 0082-7002 (PRINT) | ISSN 2343-3175 (PDF)