Helsinki University of Technology
Department of Signal Processing and Acoustics
Espoo 2010 Report 14

CONTRIBUTIONS TO INDEPENDENT COMPONENT ANALYSIS, SENSOR ARRAY AND COMPLEX-VALUED SIGNAL PROCESSING

Esa Ollila

Dissertation for the degree of Doctor of Science in Technology to be presented with due permission of the Faculty of Electronics, Communications and Automation for public examination and debate in Auditorium S4 at Helsinki University of Technology (Espoo, Finland) on the 5th of March, 2010, at 12 o'clock noon.

Helsinki University of Technology

Faculty of Electronics, Communications and Automation

Department of Signal Processing and Acoustics

Teknillinen korkeakoulu

Elektroniikan, tietoliikenteen ja automaation tiedekunta

Signaalinkäsittelyn ja akustiikan laitos

Helsinki University of Technology
Department of Signal Processing and Acoustics
P.O. Box 3000
FI-02015 TKK
Tel. +358-9-451 3211
Fax +358-9-452 3641
E-mail Mirja.Lemetyinen@hut.fi
Web page http://signal.hut.fi

ISBN 978-952-60-3030-2 (Printed)
ISBN 978-952-60-3031-9 (PDF)
ISSN 1797-4267

Multiprint Oy
Espoo, Finland 2010

ABSTRACT OF DOCTORAL DISSERTATION	HELSINKI UNIVERSITY OF TECHNOLOGY P. O. BOX 1000, FI-02015 TKK http://www.tkk.fi

Author Esa Ollila

Name of the dissertation

Contributions to independent component analysis, sensor array and complex-valued signal processing

Manuscript submitted November 1, 2009	Manuscript revised February 5, 2010
Date of the defence March 5, 2010	
☐ Monograph	☒ Article dissertation (summary + original articles)

Faculty	Faculty of Electronics, Communications and Automation
Department	Department of Signal Processing and Acoustics
Field of research	Signal Processing for Communications
Opponent(s)	Prof. Abdelhak Zoubir and Prof. Gonzalo Arce
Supervisor	Prof. Visa Koivunen
Instructor	Prof. Visa Koivunen

Abstract

Array and multichannel signal processing techniques are key technologies in wireless communications, radar, sonar and biomedical systems. In array signal processing, signals from multiple sources arrive simultaneously at a sensor array, so that each sensor array output contains a mixture of source signals. The multichannel output is then processed to provide information about the parameters of interest, e.g. the Direction-of-Arrival (DOA) of the source signals or the mixing system in the case of independent component analysis (ICA). Application areas include communications, radar, sonar and biomedicine. An important aspect is that the multichannel output is commonly complex-valued.

In this thesis, new statistical procedures and several analytical results for array and multichannel signal processing are developed and derived. Also theoretical performance bounds of estimators are established. Experimental results showing reliable performance are given on all of the presented methods.

In the area of array signal processing, the work concentrates on beamforming, high-resolution DOA estimation and estimation of the number of sources. The methods developed are robust in the sense that they are insensitive to largely deviating observations called outliers and to non-Gaussian noise environments.

In the area of complex-valued ICA, we propose two new classes of demixing matrix estimators that add a new dimension of flexibility and versatility to complex-valued ICA since distinct estimators within the same class can have largely different statistical (robustness, accuracy) properties. Hence one can choose an estimator from the class that yields the best results to the specific application at hand. A simple closed form expression for the Cramér-Rao bound (CRB) is derived for demixing matrix estimation problem as well. Its usefulness is illustrated with a simulation study.

In this thesis, the mathematical and statistical aspects of complex-valued signal processing are also addressed. Probability models, estimation bounds and novel statistics characterizing complex-valued signals are proposed. Specifically, complex elliptically symmetric (CES) distributions are proposed and studied, CRB for constrained and unconstrained complex-valued parameter estimation are derived, detectors of circularity are proposed and statistics such as circularity quotient and complex cumulants are derived.

Keywords Independent component analysis, sensor array and complex-valued signal processing, robust estimation	
ISBN (printed) 978-952-60-3030-2	ISSN (printed) 1797-4267
ISBN (pdf) 978-952-60-3031-9	ISSN (pdf)
Language English	Number of pages 106 + 106
Publisher Helsinki University of Technology, Department of Signal Processing and Acoustics	
Print distribution Helsinki University of Technology, Department of Signal Processing and Acoustics	
☒ The dissertation can be read at http://lib.tkk.fi/Diss/2010/isbn9789526030319	

VÄITÖSKIRJAN TIIVISTELMÄ	TEKNILLINEN KORKEAKOULU PL 1000, 02015 TKK http://www.tkk.fi

Tekijä Esa Ollila

Väitöskirjan nimi

Kontribuutioita riippumattomien komponenttien analyysiin, anturiryhmien ja kompleksiarvoisten signaalien käsittelyyn

Käsikirjoituksen päivämäärä 1.11.2009	Korjatun käsikirjoituksen päivämäärä 5.2.2010
Väitöstilaisuuden ajankohta 5.3.2010	
☐ Monografia	☒ Yhdistelmäväitöskirja (yhteenveto + erillisartikkelit)

Tiedekunta	Elektroniikan, tietoliikenteen ja automaation tiedekunta
Laitos	Signaalinkäsittelyn ja akustiikan laitos
Tutkimusala	Tietoliikenteen signaalinkäsittely
Vastaväittäjä(t)	Prof. Abdelhak Zoubir and Prof. Gonzalo Arce
Työn valvoja	Prof. Visa Koivunen
Työn ohjaaja	Prof. Visa Koivunen

Tiivistelmä

Anturiryhmät ja niihin liittyvä monikanavaiset signaalinkäsittelytekniikat ovat avainteknologioita langattomissa tiedonsiirtojärjestelmissä ja radiotaajuisissa mittauksissa. Tutkimuksen sovellusalueet löytyvät erityisesti tietoliikennetekniikasta, sensoriverkoista ja biolääketieteistä. Anturiryhmiin perustuvassa signaalinkäsittelyn sovelluksissa havaittu monikanavainen mittaus voidaan esittää lineaarisena sekoitteena alkuperäisistä lähdesignaaleista jotka saapuvat useasta lähteestä samanaikaisesti anturiryhmään. Havaitun moniulotteisen mittausdatan avulla on tarkoitus selvittää kiinnostuksen kohteena olevien parametrien, kuten lähdesignaalien tulosuunnat tai sekoitematriisin arvo riippumattomien komponenttien analyysin ongelmassa.

Tässä tutkimuksessa on kehitetty uusia tilastollisia menetelmiä monikanavaisten signaalien käsittelyyn. Sen lisäksi on analyyttisesti tarkasteltu teoreettisia estimointitarkkuuden alarajoja. Anturiryhmien signaalinkäsittelyssä työssä keskitytään keilanmuodostukseen, signaalien tulosuuntien ja lukumäärän estimointiin. Kehitetyt tekniikat ovat vankkoja siinä mielessä, että ne toimivat hyvin luotettavasti virheellisien tai poikkevien havaintojen sekä ei-Gaussisen kohinan tapauksissa. Kehitettyjen tekniikoiden luotettavuus on todennettu sekä analyyttisesti että simulaatioiden avulla.

Kompleksiarvoisen riippumattomien komponenttien analyysin ongelmaan työssä on kehitetty kaksi uutta estimaattoriperhettä sekoitematriisin estimointiin. Tutkimus parantaa analyysimenetelmän käytettävyyttä, sillä käyttäjä voi valita laajan estimaattoriperheen sisältä sen estimaattorin joka tuottaa parhaita tuloksia kussakin ongelmassa. Lisäksi työssä on johdettu Cramér-Rao estimointitarkkuuden alaraja sekoitematriisin estimoinnissa. Kehitettyjen tekniikoiden luotettavuus ja johdetun alarajan hyödyllisyys on todennettu simulaatioiden avulla.

Tässä työssä on johdettu uutta matemaattista ja tilastollista teoriaa kompleksiarvoisten signaalien käsittelyyn. Työssä on kehitetty uusia todennäköisyysmalleja sekä tilastollisia tunnuslukuja jotka karakterisoivat ja luokittelevat kompleksiarvoisia signaaleja. Sen lisäksi on tarkasteltu Cramér-Rao estimointitarkkuuden alarajoja kompleksiarvoisten parametrien estimoinnissa.

Asiasanat Riippumattomien komponenttien analyysi, anturiryhmän ja kompleksiarvoisten signaalien käsittely

ISBN (painettu) 978-952-60-3030-2	ISSN (painettu) 1797-4267
ISBN (pdf) 978-952-60-3031-9	ISSN (pdf)
Kieli Englanti	Sivumäärä 106 + 106

Julkaisija Teknillinen korkeakoulu, Signaalinkäsittelyn ja akustiikan laitos

Painetun väitöskirjan jakelu Teknillinen korkeakoulu, Signaalinkäsittelyn ja akustiikan laitos

☒ Luettavissa verkossa osoitteessa http://lib.tkk.fi/Diss/2010/isbn9789526030319

Preface

The research work for thesis was carried out at the Department of Signal Processing and Acoustics, Helsinki University of Technology, during the years 2003-2009. The Statistical Signal Processing group led by Prof. Visa Koivunen is part of SMARAD (Smart and Novel Radios Research Unit) Centre of Excellence in research nominated by the Academy of Finland.

First, I wish to express my deepest gratitude to Professor Visa Koivunen who has been the supervisor and the main collaborator during the course of this work. It has has been a great pleasure to work with him and I greatly appreciate all the advices he has given me during these years.

I also wish to thank all of my co-authors, Prof. Visa Koivunen, Dr. Hyon-Jung Kim, Academy Prof. Hannu Oja and Dr. Jan Eriksson for fruitful collaboration. The department secretary Mirja Lemetyinen deserves many thanks for assisting with many practical issues and arrangements. I would like to give my warmest thank to my parents, Elvi and Antti, for all the support. The financial support from the Academy of Finland is gratefully acknowledged.

Finally, the work put on this thesis was worthwhile only due to my family, my wife Hyon, our daughter Sehi and our tummy-baby. This thesis would not have been written without them.

Oulu, February 2010

Esa Ollila

To Hyon and Sehi

Contents

List of original publications

[I] Esa Ollila and Visa Koivunen, "Complex ICA using generalized uncorrelating transform", *Signal Processing*, vol. 89, no. 4, pp. 365–377, 2009.

[II] Esa Ollila, Hannu Oja and Visa Koivunen, "Complex-valued ICA based on a pair of generalized covariance matrices", *Computational Statistics and Data Analysis*, vol. 52, no. 7, pp. 3789–3805, 2008.

[III] Esa Ollila, Hyon-Jung Kim, and Visa Koivunen, "Compact Cramér-Rao bound expression for independent component analysis", *IEEE Transactions on Signal Processing*, vol. 56, no. 4, pp. 1421–1428, 2008.

[IV] Esa Ollila and Visa Koivunen, "Influence function and asymptotic efficiency of scatter matrix based array processors: case MVDR Beamformer", *IEEE Transactions on Signal Processing*, vol. 57, no. 1, pp. 247–259, 2009.

[V] Esa Ollila and Visa Koivunen, "Robust antenna array processing using M-estimators of pseudo-covariance", In Proc. *14th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC'03)*, Beijing, China, Sept. 7-10, 2003, vol. 3, pp. 2659 – 2663.

[VI] Esa Ollila, "On the circularity of a complex random variable", *IEEE Signal Processing Letters*, vol. 15, pp. 841–844, 2008.

[VII] Esa Ollila and Visa Koivunen, "Adjusting the generalized likelihood ratio test of circularity robust to non-normality", In Proc. *10th IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC'09)*, Perugia, Italy, June 21-24, 2009, pp. 558 – 562.

[VIII] Jan Eriksson, Esa Ollila, and Visa Koivunen, "Statistics for complex random variables revisited", In Proc. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'09)*, Taipei, Taiwan, April 19-24, 2009, pp. 3565 – 3568.

[IX] Esa Ollila, Visa Koivunen, and Jan Eriksson,"On the Cramer-Rao bound for the constrained and unconstrained complex parameters", In Proc. *IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM'08)*, Darmstadt, Germany, July 21-23, 2008, pp. 414–418.

[X] Esa Ollila and Visa Koivunen, "Generalized complex elliptical distributions", In Proc. *3rd IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM'04)*, Sitges, Spain, July 18-21, 2004, pp. 460–464.

List of abbreviations and symbols

Abbreviations

BSS	blind signal processing
CES	complex elliptically symmetric
CRB	Cramér-Rao bound
DOA	direction-of-arrival
d.f.	degrees of freedom
DOGMA	diagonalizer of the generalized covariance matrices
EIF	empirical influence function
EVD	eigenvalue decomposition
FOBI	fourth-order blind identification
fMRI	functional magnetic resonance imaging
GUT	generalized uncorrelating transform
GLRT	generalized likelihood ratio test
HOS	higher-order statistics
IC	independent component
ICA	independent component analysis
JADE	joint approximate diagonalization of eigen-matrices
MIMO	multiple-input multiple-output
MDL	minimum description length
ML	maximum likelihood
MLE	maximum likelihood estimator
MVDR	minimum variance distortionless response
PCA	principal component analysis
p.d.f.	probability density function
RES	real elliptically symmetric
r.v.	random vector
r.va.	random variable
SCM	sample covariance matrix
SMB-MVDR	scatter matrix based minimum variance distortionless response
SNR	signal to noise ratio
SOI	signal of interest
SSF	signal subspace fitting
SUT	strong uncorrelating transform
SVD	singular value decomposition
ULA	uniform linear array
w.l.o.g.	without loss of generality

Symbols

$\mathbb{C}$	field of complex numbers
$\mathbb{R}$	field of real numbers
i, j, k	discrete index
$\exp(\cdot)$	real or complex exponential function
$\arg(\cdot)$	argument of a complex number
$\mathrm{Arg}(\cdot)$	principal argument of a complex number
$\mathrm{Re}[\cdot]$	operator extracting the real part of its argument
$\mathrm{Im}[\cdot]$	operator extracting the imaginary part of its argument
$(\cdot)^*$	complex conjugate
$(\cdot)^T$	transpose of a matrix
$(\cdot)^H$	Hermitian transpose of a matrix
$(\cdot)^{-1}$	inverse of a matrix
$\det(\cdot)$	determinant of a matrix
$\mathrm{Tr}(\cdot)$	matrix trace
$\mathrm{Off}(\cdot)$	sum of squares of the off-diagonal elements of the matrix argument
$\lvert\cdot\rvert$	modulus of a complex number
$\lVert\cdot\rVert_2$	Euclidean (L_2-)norm
$\lVert\cdot\rVert$	norm w.r.t. to the covariance matrix
$\mathrm{diag}(c_i)$	diagonal matrix of the set $\{c_i\}$ of scalars
$\langle\cdot,\cdot\rangle$	inner product w.r.t. covariance matrix
$\arg\max$	argument of the maximum
$\mathbb{E}[\cdot]$	expectation operator
$\jmath$	imaginary number
$\triangleq$	defined as
$\mathrm{PDH}(d)$	the set of complex positive definite Hermitian $d \times d$ matrices
$\mathrm{PDS}(d)$	the set of real positive definite symmetric $d \times d$ matrices
$\mathrm{CS}(d)$	the set of complex symmetric $d \times d$ matrices
$\mathrm{cum}_4(\cdot)$	4th-order cumulant of a random variable
$\sigma^2(\cdot)$	variance of a random variable
$\tau(\cdot)$	pseudo-variance of a random variable
$\varrho(\cdot)$	circularity quotient of a random variable
$\mathrm{kurt}(\cdot)$	kurtosis of a random variable
$\gamma(\cdot)$	standardized 4th order moment of a random variable
$\boldsymbol{\mathcal{C}}(\cdot)$	covariance matrix of a random vector
$\boldsymbol{\mathcal{P}}(\cdot)$	pseudo-covariance matrix of a random vector
$\boldsymbol{\mathcal{K}}(\cdot)$	kurtosis matrix of a random vector
$\boldsymbol{\mathcal{K}}_{\mathbf{M}}(\cdot)$	cumulant matrix of a random vector
$\boldsymbol{\varrho}(\cdot)$	circularity matrix of a random vector
$\mathrm{MI}(\cdot)$	mutual information
$\mathbf{A}$	system matrix, e.g. mixing matrix in ICA
$\tilde{\mathbf{A}}$	mixing matrix of the whitened mixture
$\mathbf{a}_i$	ith column vector of $\mathbf{A}$, e.g. mixing vector in ICA

$\mathbf{W}$	demixing matrix of ICA
$\mathbf{w}_i$	ith (transposed) row vector of $\mathbf{W}$
d	number of sources
n	number of samples
$\mathbf{v}$	whitened mixture
$\mathbf{w}$	a weight vector
$\mathbf{n}$	additive random noise vector
$\mathbf{s}$	source vector
s_j	jth source
σ_j	standard deviation of the jth source
σ_j^2	variance of the jth source
κ_j	kurtosis of the jth source
c_j	4th-order cumulant of the jth source
$\boldsymbol{\Delta}$	diagonal matrix of standard deviations of the sources
$\mathcal{J}(\cdot)$	criterion function in ICA
$\boldsymbol{\mathcal{B}}$	whitening matrix of a random vector
$\mathbf{I}$	identity matrix
$\hat{\mathbf{c}}$	mapping forming an augmented complex $2d$-vector from $\mathbf{c} \in \mathbb{C}^d$
$\bar{\mathbf{c}}$	mapping forming a composite real $2d$-vector from $\mathbf{c} \in \mathbb{C}^d$
$\boldsymbol{\Gamma}$	scatter parameter of a RES distribution
$\boldsymbol{\Sigma}, \boldsymbol{\Omega}$	scatter and pseudo-scatter parameters of a CES distribution
$\hat{\boldsymbol{\Gamma}}$	augmented scatter matrix
$\boldsymbol{\mathcal{I}}_{\boldsymbol{\theta}}$	information matrix
$\boldsymbol{\mathcal{P}}_{\boldsymbol{\theta}}$	pseudo-information matrix
$\nabla_{\boldsymbol{\theta}^*}$	complex gradient
$\mathbf{S}$	sample covariance matrix
$\psi(\cdot)$	a weight function
$\psi_{ML}(\cdot)$	weight function of the MLE

Chapter 1

Introduction

1.1 Motivation of the thesis

In signal processing and related fields, multichannel measurements are often encountered. For example, biomedical measurements such as MEG and EEG, radar signals, many communications signals are multivariate. The m-variate received signal $\mathbf{z} = (z_1, \ldots, z_m)^T$ (sensor outputs) may be modelled in terms of the transmitted *source signals* $s_1, \ldots, s_d$ possibly corrupted by additive *noise vector* $\mathbf{n}$, *i.e.*

$$\begin{aligned} \mathbf{z} &= \mathbf{A}\mathbf{s} + \mathbf{n} \\ &= \mathbf{a}_1 s_1 + \ldots + \mathbf{a}_d s_d + \mathbf{n} \end{aligned} \tag{1.1}$$

where $\mathbf{A} = (\mathbf{a}_1 \ \cdots \ \mathbf{a}_d)$ is the unknown $m \times d$ *system matrix* and $\mathbf{s} = (s_1, \ldots, s_d)^T$ contains the source signals. It is commonly assumed that $d \leq m$. In practice, the system matrix is used to describe the array geometry in sensor array applications, MIMO (multiple-input multiple-output) channel in wireless multiantenna communication systems and mixing system in the case of signal separation problems, for example. The vector and matrix quantities in the model can be real-valued or complex-valued depending on the application and problem at hand. Source vector $\mathbf{s}$ can be modelled as random or deterministic, observable or unobservable, and it can be referred to as independent components, source signals, latent variables, common factors, principal components, etc, again depending on the application. Also the system matrix $\mathbf{A}$ is often named differently. In most cases, $\mathbf{s}$ and $\mathbf{n}$ are assumed to be mutually statistically independent with zero mean. An example of a multiantenna sensing system with uniform linear array (ULA) configuration is depicted in Figure 1.1.

The model (1.1) is indeed very general, and covers for example the following important applications that constitute the core topics of this thesis:

In *narrowband array signal processing* [1–5] each vector $\mathbf{a}_i$ represents a point in known array manifold (array transfer function, steering vector) $\mathbf{a}(\theta)$, *i.e.* $\mathbf{a}_i = \mathbf{a}(\theta_i)$, where θ_i is an unknown parameter, typically the direction-of-arrival (DOA) of the ith

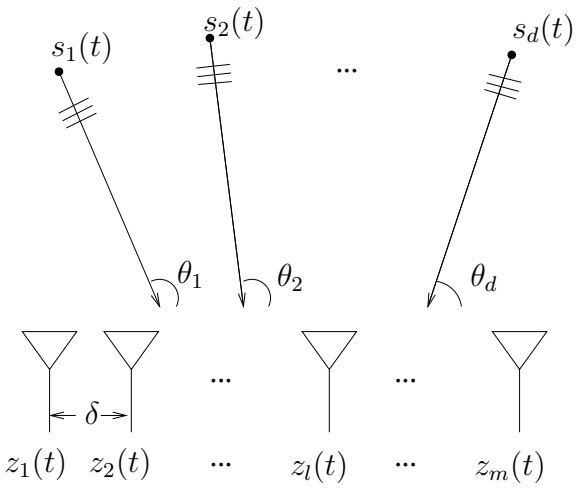


Figure 1.1: A uniform linear array (ULA) of m sensors with sensor displacement δ receiving plane waves from d far-field point sources.

source, $i = 1, \ldots, d$. Identifying $\mathbf{A}$ is then equivalent to the problem of identifying $\theta_1, \ldots, \theta_d$. For example, in case of ULA with identical sensors,

$$\mathbf{a}(\theta) = \begin{pmatrix} 1 & e^{-\jmath\omega} & \cdots & e^{-\jmath(m-1)\omega} \end{pmatrix}^T,$$

where $\omega = 2\pi(\delta/\lambda)\sin(\theta)$ depends on the signal wavelength λ, the DOA θ of the signal w.r.t. broadside, and the sensor spacing δ. The source signal vector $\mathbf{s}$ is modelled as either deterministic or random, depending on the application. The application domain include radar, wireless communications and sensor array processing.

In blind signal separation (BSS) based on *independent component analysis (ICA)* [6–9], both the mixing system $\mathbf{A}$ and the sources $\mathbf{s}$ are unknown. The goal in ICA is to solve the mixing matrix and consequently to recover the original sources from their mixtures exploiting only the assumption that sources are mutually statistically independent. ICA has been successfully applied in several applications including wireless communications, audio and speech signal separation, biomedical signal processing, image processing, feature extraction and data-mining.

A common assumption imposed on the signal model (1.1) is that

Assumption (A1) noise $\mathbf{n}$ and/or source $\mathbf{s}$ possess circularly (or, spherically) symmetric distributions.

In addition, in the process of deriving optimal array processors, the distribution of the noise $\mathbf{n}$ is typically assumed to be known also, the conventional assumption being that

Assumption (A2) noise $\mathbf{n}$ possess (circular complex, or, real) multivariate Gaussian distribution.

Furthermore, commonly in complex-valued case, if $\mathbf{s}$ is modelled as stochastic, then $\mathbf{s}$ and $\mathbf{n}$ are both assumed to be independent with circular complex Gaussian distribution, and consequently, sensor output $\mathbf{z}$ also has m-variate circular complex Gaussian distribution. Note that noise and interference in indoor and outdoor mobile communication channels as well as in sonar and radar signals have been shown to be contaminated by non-Gaussian noise and interference thus violating Assumption (A2). See e.g. Middeleton [10] and Williams and Johnson [11] and references therein.

1.2 Scope of the thesis

The objective of this thesis is to develop the theory and methodology as well as to construct practical and reliable robust statistical inference procedures (estimators, detectors) for multi-channel and sensor array signal processing applications. Another objective is to establish theory and relevant statistical inference procedures for complex-valued signal processing. The applications considered are beamforming, DOA estimation, estimation of the number of source signals, real- and complex-valued ICA, and complex-valued signal processing. Particular attention is paid to violations of the Assumption (A1) or (A2). The design goal of the developed methods is that they work reliably and robustly when the conventional assumptions (e.g. of normality, or, circularity) are not valid, and produce highly reliable estimates otherwise. Rigorous mathematical theory should accompany the derived methods.

1.3 Contributions

This work contributes to the fields of multi-channel and array signal processing, probability theory and communications.

The contributions of the thesis include the following

- Two new classes of demixing matrix estimators for complex-valued ICA are proposed as generalizations of FOBI [12] and SUT [13]. The proposed classes of demixing matrix estimators add flexibility and versatility to complex-valued ICA since distinct estimators within the same class can have largely different statistical (robustness, accuracy) properties. Hence one can choose an estimator from the class that yields the best results to the specific application at hand.
- A simple closed form expression of the Cramér-Rao bound (CRB) for the demixing matrix estimation is derived, thus filling an important gap in the theoretical foundations of real-valued ICA.
- A class of scatter matrix based Minimum Variance Distortionless Response (SMB-MVDR) beamformers are proposed. Theoretical properties of SMB-MVDR beam-

former weight vectors are studied by deriving their influence function and asymptotic covariance matrix under the class of complex elliptically symmetric (CES) distributions.

- Maronna's [14] celebrated M-estimators of scatter are extended to the complex-valued case and their usefulness are illustrated in several practical signal processing applications.

- The concept of circularity is studied and a degree of circularity, called circularity quotient, is proposed and its properties are established. A generalized likelihood ratio test (GLRT) of circularity is derived assuming complex normal sample. It is shown that with a slight adjustment, the GLRT can be made asymptotically robust with respect to departures from Gaussianity within the CES distributions. The asymptotic distributions of the tests under the null hypothesis are established.

- The unconstrained and constrained CRB for complex-valued parameter estimation are derived. An advantage of the complex CRB is that it is often easier to calculate than its real form.

- Complex elliptically symmetric (CES) distributions are proposed and studied. This wide class of distributions include the circular CES distribution [15], the Cauchy distribution and the complex normal distribution [16, 17] as special cases and hence can be used to model wide variety of random phenomena.

- Complex cumulants are derived in a mathematically rigorous manner and a novel complex-valued extension of Taylor series is introduced.

1.4 Structure of the thesis

This thesis consists of an introductory part and 10 original publications. The publications are listed at page viii and appended at the end of the manuscript. They will be referred in the text by Publication [I], Publication [II], or simply [I], [II], etc.

The introductory part of this thesis is organized as follows. Chapter 2 provides an introduction to ICA. The real-valued instantaneous ICA model and the underlying probabilistic concepts and assumptions are discussed and some benchmark ICA methods are reviewed. Also an image analysis example is provided and the concern of robust estimation in the ICA model is illustrated with simulation studies and plots of the empirical influence functions of the estimators.

Chapter 3 reviews the fundamentals of processing of complex-valued signals, e.g. complex-field and functions, Taylor's $\mathbb{R}$-theorem, main probabilistic tools, statistics and concepts needed in uni- and multivariate complex-valued signal processing, such

as covariance, pseudo-covariances, circularity quotient, complex-valued kurtosis. CES distributions and circularity detectors withing this class are discussed with illustrative examples. Also MLE of the scatter parameter of the CES distribution is derived.

Chapter 4 introduces the common signal model and the basic concepts employed in array signal processing. A brief overview of widely used DOA estimation techniques are given and the problem of estimating the number of signals is addressed. The emphasis is on scatter matrix based (SMB) array processing. In particular, M-estimators of scatters are reviewed and their robust performance is shown with several illustrative array processing examples.

The chapters serve as a review of the work in Pubications [I]-[X] but also as a survey of the state-of-the-art of the topics studied in this thesis.

1.5 Summary of publications

In this subsection, a brief overview of the publications are given.

In Publication [I], an extension of the whitening transformation for complex random vectors, called the generalized uncorrelating transformation (GUT), is introduced. GUT is a generalization of the SUT [13] based upon generalized estimators of the covariance and pseudo-covariance matrix, called the scatter matrix and spatial pseudo-scatter matrix, respectively. It is shown that GUT is a demixing matrix estimator for complex-valued ICA when at most one source random variable possess circularly symmetric distribution and sources do not have identical distribution. Special emphasis is put on robust GUT estimators.

In Publication [II], a new class of demixing matrix estimators, called the diagonalizer of generalized covariance matrices (DOGMA), for complex-valued ICA are proposed. The DOGMA class is a a generalization of FOBI [12] based upon two distinct matrix-valued statistics, called the scatter matrix and the spatial scatter matrix. The proposed approach is computationally attractive and an efficient algorithm that avoids decorrelation of the data is proposed. Special emphasis is put on robust DOGMA estimators.

In Publication [III], a simple closed form expression of the CRB for the demixing matrix estimation is derived, thus filling an important gap yet existing in the theoretical foundations of real-valued ICA. A simulation study comparing the performance of some widely used ICA estimators with the CRB is given.

In Publication [IV], a class of scatter matrix based Minimum Variance Distortionless Response (SMB-MVDR) beamformers are proposed. Statistical properties of SMB-MVDR beamformer weight vectors are investigated by deriving their influence function and asymptotic covariance matrix under the wide class of circular CES distributions. The results clearly reveal the lack of robustness and inefficiency of the conventional

MVDR beamformer in the face of non-Gaussianity.

In Publication [V], Maronna's celebrated M-estimators of scatter are extended to complex-valued case. Estimates of the noise and signal subspaces based on M-estimators are then used to robustify the subspace DOA estimation methods. In addition, eigenvalues based on M-estimators of scatter are used to robustify the estimation of number of signals using the minimum description length (MDL) criterion [18, 19].

In Publication [VI], a degree of circularity, called circularity quotient, is proposed and studied. Its connection with the Pearson correlation coefficient ρ is established and bounds on ρ given the circularity quotient (and vice versa) are derived. The GLRT of circularity is shown to be a function of the modulus of the circularity quotient with asymptotic χ^2_2 distribution.

In Publication [VII], it is shown that with a slight adjustment the GLRT of circularity can be made asymptotically robust with respect to departures from Gaussianity within the CES distributions. The asymptotic distribution of the test under the null hypothesis is established and simulations and a communication example are provided to illustrate the usefulness and applicability of the proposed test. Connection between the complex kurtosis and the marginal real kurtosis of a complex random variable with a CES distributions is established.

In Publication [VIII], a concise and rigorous treatment of mathematical and statistical foundations of complex-valued signal processing is presented. Specifically, complex-valued cumulants are derived in a mathematically rigorous manner and a novel complex-valued extension of Taylor series is introduced.

In Publication [IX], the unconstrained and constrained CRB for complex-valued parameter estimation are derived. The advantage of the complex CRB is that it is often easier to calculate than its real form. It is shown that a statistic that attains a bound on the complex covariance matrix alone do not necessarily attain the CRB.

In Publication [X], the CES distributions are proposed and its properties are studied. Also the conditional mean estimator within this class is studied and a likelihood ratio test and the GLRT of circularity is derived assuming complex normality.

The results in Publications [I–VII], [IX] and [X] were derived independently by the author of this thesis. The co-authors have helped in writing and structuring the manuscript, planning the examples and steering/defining the research. In publication [VIII], the idea of $\mathbb{R}$-linearity, $\mathbb{R}$-differential and the proposed circularity measure based on characteristic function are due to Jan Eriksson. The proof of Theorem 2 is also by him. The idea of complex cumulants using the 2nd characteristic function are due to the author of this thesis. The proof of Theorem 1 is due to the author of this thesis. The writing of the publication [VIII] and other concepts were done in close collaboration with the co-authors. All the simulation software for the proposed methods in this dissertation were written by the author of this thesis.

Chapter 2

Independent Component Analysis

Over the past two decades independent component analysis (ICA) has become a widely used data analysis and signal processing technique with applications in many diverse fields such as wireless communications, blind source separation, medical imaging, audio and speech signal processing, image processing, feature extraction and data mining. See text-books [8, 9] and their bibliographies for more details.

In this section we review the real-valued instantaneous ICA model and some commonly used methods of ICA. Also an image analysis example is provided and the concern of robust estimation in the ICA model is illustrated with simulation studies and plots of the empirical influence functions of the estimators. It is demonstrated that the DOGMA method [II], which extends to real-valued case, offers a robust and practical alternative.

2.1 ICA model

We consider the instantaneous noise-free real linear ICA model in which the observed random vector $\mathbf{x} = (x_1, \ldots, x_m)^T$ is modelled as a linear mixture of the unobserved (latent) source r.v. $\mathbf{s} = (s_1, \ldots, s_d)^T$,

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{2.1}$$

where the real $m \times d$ *mixing matrix* $\mathbf{A}$ as well as the distributions of the sources are assumed to be unknown and satisfy the assumptions

IC1 *Sources $s_1, \ldots, s_d$ are statistically independent.*

IC2 *The number of mixtures m equals the number of sources d: $m = d$.*

IC3 *The columns $\mathbf{a}_i$ of $\mathbf{A}$, called the* mixing vectors*, are linearly independent.*

IC4 *At most one source has a Gaussian distribution.*

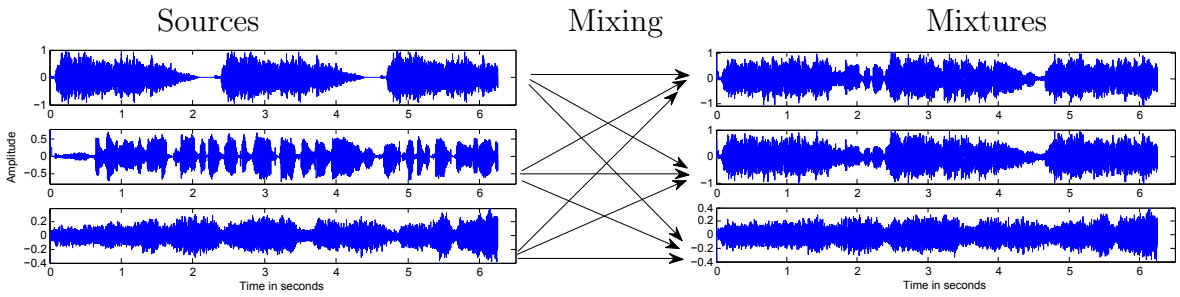


Figure 2.1: An illustration of the mixing system: three sound sources are linearly mixed by a random mixing matrix. The mixing system and the original sound sources are unknown and only the (microphone) recordings of the mixtures of the sound sources are observed.

The goal in ICA is to estimate the demixing matrix $\mathbf{W} = \mathbf{A}^{-1}$ allowing to separate the independent components (IC's) as $\mathbf{s} = \mathbf{W}\mathbf{x}$, where the (transposed) ith row vector $\mathbf{w} \in \mathbb{R}^d$ of $\mathbf{W}$ is called the ith *demixing vector*. Due to the assumptions IC2-IC4 separation is possible up to the fundamental indeterminacy (that allows permutation, sign and scale changes) [6, 20]; We shall return to the ambiguities in the model in Section 2.1.1 and Section 2.1.2. We shall write $\mathbf{x} \sim F_{\mathbf{A}}$ to denote that $\mathbf{x}$ follows ICA model.

We wish to point out that IC2 could be replaced by a more general assumption that the number of mixtures is larger or equal to the number of sources ($m \geq d$) in which case the left inverse of $\mathbf{A}$, $\mathbf{W} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$, is a demixing matrix achieving the separation of the sources. Assumption IC2 is not a limitation since (assuming that the number of sources are known) the dimensionality of $\mathbf{x}$ can be reduced e.g. by Principal Component Analysis (PCA) [21].

■ EXAMPLE 1. Suppose that two mixing vectors $\mathbf{a}_1$ and $\mathbf{a}_2$ are parallel (*i.e.* linearly dependent, $\mathbf{a}_1 = a\mathbf{a}_2$, for some $a \in \mathbb{R}$). Then $\mathbf{x}$ has also statistical representation with only $d-1$ sources by combining the 1st and 2nd source to a single source $as_1 + s_2$. This case and also the simplest pathological cases, for example that $\mathbf{A} = (c)_{ij}$ (*i.e.* $a_{ij} \equiv c$ for all i, j) are excluded by requiring IC3. ■

The utility of ICA is commonly illustrated by the so called *cocktail party problem*. Suppose there are three microphones and three sound sources. In a simplified model (e.g. omitting multipath propagation, time delays), the microphone recordings are unknown mixtures of the sound sources. The mixing depends naturally on the distance, position and angle of the microphones relative to sound sources. Using ICA the original sound sources can then be separated from the mixtures. See Figure 2.1 for an illustration.

In the model and in the example above we have omitted the noise term that is always present in real-world physical measurements. This so called *noisy ICA model*

can be expressed as $\mathbf{x} = \mathbf{A}\mathbf{s} + \mathbf{n}$, where the noise $\mathbf{n}$ is typically assumed to be d-variate Gaussian random vector independent of the sources. With the additive noise, the ICA model is more realistic. If the the signal to noise ratio (SNR) is high, the methods for noiseless ICA typically yield satisfactory results; See e.g. simulation examples in Publications [I,II]. Methods for noisy ICA are considered in [8, 22–24]

2.1.1 Fundamental indeterminacy of the ICA model

Let $i_1, i_2, \ldots, i_d$ denote any permutation of the set $1, 2, \ldots, d$. Vector

$$\mathbf{s}_0 = (s_{i_1}\alpha_1, \ldots, s_{i_d}\alpha_d)^T, \quad \alpha_i \neq 0,$$

whose elements are permuted and scaled version of the source r.v. $\mathbf{s}$, is called the *copy of* $\mathbf{s}$. Note that also $\mathbf{s}_0$ has statistically independent components. Write

$$\mathbf{A}_0 = \begin{pmatrix} \mathbf{a}_{i_1}\alpha_1^{-1} & \mathbf{a}_{i_2}\alpha_2^{-1} & \cdots & \mathbf{a}_{i_d}\alpha_d^{-1} \end{pmatrix}$$

for the matrix $\mathbf{A}$ whose columns are permuted accordingly and scaled by the divisor of the scalar multiplier α_i. Clearly, the mixture $\mathbf{x}$ does not have unique statistical representation since

$$\mathbf{x} = \mathbf{A}_0\mathbf{s}_0 = \mathbf{A}\mathbf{s}.$$

and both $\mathbf{A}_0\mathbf{s}_0$ and $\mathbf{A}\mathbf{s}$ are valid generative ICA models for the observed mixture $\mathbf{x}$ since the couples, $(\mathbf{A}_0, \mathbf{s}_0)$ and $(\mathbf{A}, \mathbf{s})$, both satisfy assumptions IC1–IC3. This lead to the so called *fundamental indeterminacy*: it is possible to identify $\mathbf{A}$ (and hence $\mathbf{W} = \mathbf{A}^{-1}$) only up to scaling, sign and permutation of its column vectors $\mathbf{a}_i$ (resp. row vectors $\mathbf{w}_i$); In addition, the scales, signs and the order of the IC's $s_1, \ldots, s_d$ cannot be determined without additional assumptions.

Hence separation of the IC's from their mixtures should be understood as the determination of a demixing matrix $\mathbf{W}$, such that $\mathbf{W}\mathbf{x} = \mathbf{s}_0$, *i.e.* $\mathbf{W}$ maps the mixtures $\mathbf{x}$ to any copy of $\mathbf{s}$. Thus $\mathbf{A}^{-1}$ is a demixing matrix, but, any matrix $\mathbf{W}$ that is equal to $\mathbf{A}^{-1}$ up to permutation, sign and scale change of its row vectors, is a valid demixing matrix as well.

2.1.2 Non-Gaussianity

Let us now shed some light on the non-Gaussianity requirement IC4. Suppose that sources $s_1, \ldots, s_d$ have zero-mean normal (Gaussian) distribution. Due to scale ambiguity, we can assume that they are of unit variance. Thus $s_i \sim N(0, 1)$, $i = 1, \ldots, d$ and hence $\mathbf{s}$ has d-variate standard normal distribution, denoted $\mathbf{s} \sim N_d(\mathbf{0}, \mathbf{I})$, where $\mathbf{I}$ denotes the identity matrix. Next recall that the standard normal distribution remains invariant under orthogonal linear transformations [25], that is, also $\mathbf{V}\mathbf{s}$ possesses

$N_d(\mathbf{0}, \mathbf{I})$ distribution for any orthogonal matrix $\mathbf{V}$ (*i.e.* $\mathbf{V}^T\mathbf{V} = \mathbf{I}$). Hence a couple $(\mathbf{AV}, \mathbf{Vs})$ yields a valid ICA model since

$$\mathbf{x} = (\mathbf{AV}^T)\mathbf{Vs}$$

and they satisfy IC1-IC3 (since $\mathbf{Vs}$ has independent $N(0,1)$ components and $\mathbf{AV}^T$ has full rank). Thus $\mathbf{A}$ can be at best identified up to a right multiplication by an orthogonal matrix. Hence a necessary condition for $\mathbf{A}$ to be identifiable up to the fundamental indeterminacy is that IC4 holds. In fact, in [20] it was shown that the non-Gaussianity of the sources (except for possibly one) is also sufficient condition for $\mathbf{A}$ to be identifiable up to the fundamental indeterminacy.

Note that IC4 is a necessary condition to estimate *all* the sources from the mixture. If the goal is to extract all (or a subset of the) non-Gaussian sources IC4 is not required. ICA methods that utilize deflation strategy [26] do not need IC4.

One of the reasons, why higher-order statistics (HOS) have attained popularity in ICA, is their ability to measure non-Gaussianity. Commonly used HOS to measure non-Gaussianity is the *kurtosis*, standardized and shifted 4th-order moment defined as

$$\mathrm{kurt}(x) \triangleq \gamma(x) - 3, \quad \gamma(x) \triangleq \frac{\mathbb{E}[(x - \mathbb{E}[x])^4]}{(\sigma^2(x))^2} \tag{2.2}$$

where $\sigma^2(x) \triangleq \mathbb{E}[(x - \mathbb{E}[x])^2]$ denotes the variance of a r.va. x. Kurtosis has the property that it vanishes when x is a Gaussian random variable. Note however that there exists non-Gaussian distributions that have vanishing kurtosis as well. In ICA literature, kurtosis is also used to classify random variables: the term *sub-Gaussian* (resp. *super-Gaussian*) refers to a r.va. whose kurtosis is strictly smaller (resp. strictly larger) than zero [8]. The pitfall of employing HOS such as kurtosis is their non-robustness and inaccuracy for small sample sizes.

2.2 Data pre-processing in ICA

Many ICA algorithms require that the data is centered (has zero mean) and whitened (uncorrelated, or sphered). If the 2nd-order moments of the sources are assumed to exist, then the mean vector $\mathbb{E}[\mathbf{x}]$ and the covariance matrix $\boldsymbol{\mathcal{C}}(\mathbf{x}) = [\mathrm{Cov}(x_i, x_j)]$ of the mixture $\mathbf{x}$ are

$$\mathbb{E}[\mathbf{x}] = \mathbf{A}\mathbb{E}[\mathbf{s}]$$

and

$$\boldsymbol{\mathcal{C}}(\mathbf{x}) \triangleq \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^T] = \mathbf{A}\boldsymbol{\mathcal{C}}(\mathbf{s})\mathbf{A}^T. \tag{2.3}$$

Note that $\boldsymbol{\mathcal{C}}(\mathbf{x}) \in \mathrm{PDS}(d)$, where $\mathrm{PDS}(d)$ denote the set of all $d \times d$ real positive definite symmetric $d \times d$ matrices.

The most common pre-processing step in ICA is to center the data,

$$\mathbf{x} - \mathbb{E}[\mathbf{x}] = \mathbf{A}(\mathbf{s} - \mathbb{E}[\mathbf{s}]),$$

since advantage can be taken from the property that the centered data follows ICA model with an additional feature that sources are of zero mean. In addition, thanks to the scale ambiguity of the sources, one can now tacitly assume without loss of generality (w.l.o.g.) that

IC5 Sources are of zero mean, $\mathbb{E}[s_i] = 0$, and of unit variance, $\sigma_i^2 \equiv \sigma^2(s_i) = 1$, $i = 1, \ldots, d$.

Another common pre-processing step in ICA is the whitening (or sphering, decorrelating) of the data. Let $\boldsymbol{\mathcal{B}} = \boldsymbol{\mathcal{B}}(\mathbf{x})$ denote any *whitening matrix* of a r.v. $\mathbf{x}$, *i.e.* $\boldsymbol{\mathcal{C}}(\mathbf{x})^{-1} = \boldsymbol{\mathcal{B}}^T\boldsymbol{\mathcal{B}}$ holds. For example, the (unique) principal square-root matrix of $\boldsymbol{\mathcal{C}}(\mathbf{x})^{-1}$, $\boldsymbol{\mathcal{B}} = \boldsymbol{\mathcal{C}}(\mathbf{x})^{-1/2}$. Then the whitened data

$$\mathbf{v} = \boldsymbol{\mathcal{B}}\mathbf{x}$$

has uncorrelated components of unit variance, *i.e.* $\boldsymbol{\mathcal{C}}(\mathbf{v}) = \mathbf{I}$.

It is crucial to realise the difference between independence and uncorrelatedness. Namely, if random variables x_i and x_j are independent, then they are are uncorrelated, but the converse is not necessarily true. Indeed there can exists a strong dependency between r.va.'s x_i and x_j, yet being uncorrelated.

■ Example 2. Let x_1 be any r.va. symmetric about zero (*i.e.* $x_1 =_d -x_1$), y any random variable independent of x_1 and $x_2 = ax_1^2 + by$ where a and b can be any non-zero constants. Hence there is a strong dependency between x_1 and x_2, but they are nevertheless uncorrelated. ■

If $\mathbf{x}$ follows the ICA model, then the *whitened mixture*

$$\mathbf{v} = \tilde{\mathbf{A}}\mathbf{s} \quad \text{where} \quad \tilde{\mathbf{A}} = \boldsymbol{\mathcal{B}}\mathbf{A}$$

also follows ICA models as the couple $(\tilde{\mathbf{A}}, \mathbf{s})$ necessarily satisfy assumptions IC1-IC4 as well. Let us denote by $\boldsymbol{\Delta}$ the diagonal matrix with the standard deviations σ_i of the sources as diagonal elements, *i.e.* $\boldsymbol{\mathcal{C}}(\mathbf{s}) = \boldsymbol{\Delta}^2$. The fact that $\boldsymbol{\mathcal{C}}(\mathbf{v}) = \boldsymbol{\mathcal{B}}\boldsymbol{\mathcal{C}}(\mathbf{x})\boldsymbol{\mathcal{B}}^T = \mathbf{I}$ together with (2.3) indicate that

$$\tilde{\mathbf{A}}\boldsymbol{\Delta}^2\tilde{\mathbf{A}}^T = (\tilde{\mathbf{A}}\boldsymbol{\Delta})(\tilde{\mathbf{A}}\boldsymbol{\Delta})^T = \mathbf{I},$$

which, in turn, implies that the scaled mixing matrix $\tilde{\mathbf{A}}\boldsymbol{\Delta}$ is orthogonal. But since the scales of the columns $\tilde{\mathbf{a}}_i$ of $\tilde{\mathbf{A}}$ can not be identified we may contend w.l.o.g. that $\tilde{\mathbf{A}}$ is an orthogonal matrix[1]. Thus the demixing matrix of the whitened mixture must be

[1] Or equivalently, since IC5 can be assumed, we may contend that $\boldsymbol{\Delta} = \mathbf{I}$, which in turn implies that $\tilde{\mathbf{A}}$ is an orthogonal matrix

orthogonal as well. Hence there exists an orthogonal matrix $\tilde{\mathbf{W}}$ such that

$$\mathbf{s}_0 = \tilde{\mathbf{W}}\mathbf{v} = (\tilde{\mathbf{W}}\boldsymbol{\mathcal{B}})\mathbf{A}\mathbf{s}$$

is a copy of $\mathbf{s}$. Thus $\tilde{\mathbf{W}}\boldsymbol{\mathcal{B}}$ is a demixing matrix for the original mixture $\mathbf{x} = \mathbf{A}\mathbf{s}$.

By whitening the data, the ICA problem is reduced to a simpler problem of finding orthogonal demixing matrix of the whitened mixture. Since the orthogonal demixing matrix $\tilde{\mathbf{W}}$ has roughly one half unknown parameters compared to the d^2 unknown coefficients of the demixing matrix $\mathbf{W}$, the ICA problem is now considerably simpler. Most ICA methods employ pre-whitening and thus they differ only in the way they estimate the orthogonal demixing matrix of the whitened mixture.

It is important to realise that the ICA model itself does not require any moment assumptions on the sources. Whitening implicitly assumes that the 2nd-order moments of the sources exist. Many heavy-tailed distribution, however, do not possess finite variance, take Cauchy or t_3-distribution as examples. Thus ICA methods that require whitening often perform poorly for heavy-tailed sources or when outliers (*i.e.* highly deviating observations) are present.

2.3 Review of ICA methods

In this section, we provide a short review of ICA methods. Up to date, thanks to the vast interest in ICA during the past two decades, there exists a broad array of ICA methods; see e.g. [6, 8, 9, 27] for a comprehensive account. We have chosen to represent the FastICA [28–30] and JADE [7, 31] methods in more detail since they have become benchmark methods of ICA. Also reviewed are FOBI [12] and its generalization, DOGMA (publication [II]) as it extends to the real-valued case [32].

2.3.1 Anatomy of ICA algorithms

Roughly speaking, there exists two main branches of ICA methods, the *optimization group* and the *algebraic group* of ICA algorithms.

In the optimization group, the first step is to formulate a *criterion function* which serves as a measure of independence (or of non-Gaussianity). Criterion function $\mathcal{J}$ is a *statistical functional* $\mathcal{J} : \mathcal{F} \to \mathbb{R}^+ = [0, \infty)$, *i.e.* a real-valued function of d-variate probability distributions $F \in \mathcal{F}$, where $\mathcal{F}$ denotes a set of all distributions on $\mathbb{R}^d$ (or a large subset of it) such that $\mathcal{J}(F)$ exists. A sensible criterion function should have at least the following properties:

(i) $\mathcal{J}(\mathbf{x}) = 0$ if r.v. $\mathbf{x}$ with the distribution $F \in \mathcal{F}$ has independent components.

(ii) $\mathcal{J}(\mathbf{x}) > 0$ for all or at least most distributions $F \in \mathcal{F}$ of $\mathbf{x}$ that do not have independent components.

(iii) $\mathcal{J}(\mathbf{Px}) = \mathcal{J}(\mathbf{x})$ for all permutation matrices $\mathbf{P}$.

The idea is to find an invertible $d \times d$ matrix $\mathbf{W}$ that minimizes the criterion function $\mathcal{J}(\mathbf{Wx})$ over all $d \times d$ real invertible matrices. For judicious choice of $\mathcal{J}(\cdot)$, the found minima $\mathbf{W}_{\mathcal{J}}$ is a demixing matrix when $\mathbf{x}$ follows ICA model. If $\mathbf{x}$ is pre-whitened, then the optimization problem is simplified to that of finding an orthogonal $d \times d$ matrix $\mathbf{W}$ minimizing $\mathcal{J}(\mathbf{Wx})$.

If the r.v. $\mathbf{x} \sim F$ possess a p.d.f., then the *mutual information* (MI)

$$\mathrm{MI}(\mathbf{x}) = \mathbb{E}_F\left[\log \frac{f(\mathbf{x})}{\prod_{i=1} f_i(x_i)}\right] = \int_{-\infty}^{\infty} \log \frac{f(\mathbf{x})}{\prod_{i=1} f_i(x_i)} f(\mathbf{x}) d\mathbf{x}$$

where $f(\cdot)$ and $f_i(\cdot)$ denotes the p.d.f. of $\mathbf{x}$ and x_i respectively ($i = 1, \ldots, d$), then satisfies property (iii), and also properties (i) and (ii) in a strict sense: $\mathrm{MI}(\mathbf{x}) \geq 0$ with equality if and only if $\mathbf{x}$ has independent components. Hence MI is a *contrast function* [6]. Comon [6] also proposed of using $\mathcal{J}(\mathbf{x}) = \sum_{i=1}^{d} |\mathrm{cum}_4(x_i)|^2$ where $\mathbf{x}$ is assumed to be whitened and $\mathrm{cum}_4(x) \triangleq \mathbb{E}[(x - \mathbb{E}[x])^4] - 3(\mathbb{E}[(x - \mathbb{E}[x])^2])^2$ denotes the *4th-order cumulant* of a r.va. x.

Alternatively, one can search for a single demixing vector $\mathbf{w} \in \mathbb{R}^d$ such that the projection $s = \mathbf{w}^T\mathbf{x}$ of the r.v. $\mathbf{x} \sim F$ minimizes/maximizes some criterion function $\mathcal{J}(\mathbf{w}^T\mathbf{x})$. Such approaches are strongly related to projection pursuit [33, 34] method. For example, FastICA (subject of Section 2.3.2) formulates a criterion function that measures "non-Gaussianity". To find all demixing vectors one can then use *deflation approach* (one-unit, sequential extraction) [26], where the IC's are extracted sequentially, *i.e.* one after another. Typically, the constraint of uncorrelatedness with the previously found sources is then used to prevent the optimization algorithm from converging to previously found components.

In the algebraic group, the demixing matrix is sought using matrix algebra and matrix-valued statistics. Several ICA methods, e.g. FOBI, DOGMA, CHESS (CHaracteristic function Enabled Source Separation) [35], JADE (and inherently related tensorial methods [8, 23]) are examples of ICA methods that can be classified to this group. Naturally, the distinction in not always so clear cut. For example, JADE could also be classified to the optimization group (see e.g. [7]).

2.3.2 FastICA

Arguably, FastICA [28, 29] is one of the most popular and widespread method. Its popularity can be attributed to its simplicity, ease of implementation, fast computation, a user-friendly public-domain software [30] and flexibility to choose the *nonlinearity* function.

We restrict our attention to the deflation-based FastICA method, referred as deflationary FastICA or FastICA for short. There also exist *symmetric* (or, joint, simulta-

neous extraction) mode, but originally [28] FastICA was put forth in deflation mode. Advantage of the deflation-based FastICA over the symmetric FastICA is the ability to estimate a single or a subset of the original IC's which can be desirable in some applications [36]. Another advantage is the reduced computational load which can be significant if only a small subset of sources needs to be extracted from a high-dimensional data set. The downside is that errors can accumulate in successive deflation stages in which case symmetric approach can provide better overall separation performance.

FastICA needs to assume that the sources have finite 2nd-order moments. Hence we assume w.l.o.g. that IC5 holds. Let us now assume that the data is centered so that $\mathbb{E}[\mathbf{x}] = \mathbb{E}[\mathbf{s}] = \mathbf{0}$ holds and define the *inner product* in the vector space $\mathbb{R}^d$ as

$$\langle \mathbf{u}, \mathbf{y} \rangle \triangleq \mathbb{E}[(\mathbf{u}^T\mathbf{x})(\mathbf{y}^T\mathbf{x})] = \mathbf{u}^T\boldsymbol{\mathcal{C}}\mathbf{y}, \tag{2.4}$$

where $\boldsymbol{\mathcal{C}}$ is the positive definite $d \times d$ covariance matrix of the centered $\mathbf{x}$. This induces the norm $\|\cdot\|$,

$$\|\mathbf{u}\|^2 \triangleq \langle \mathbf{u}, \mathbf{u} \rangle = \mathrm{var}(\mathbf{u}^T\mathbf{x}) = \mathbf{u}^T\boldsymbol{\mathcal{C}}\mathbf{u}.$$

Note that the standard inner product of $\mathbb{R}^d$ is the dot product with the corresponding norm being the Euclidean distance (L_2-norm) $\|\mathbf{w}\|_2 \triangleq \sqrt{\mathbf{w}^T\mathbf{w}}$. Geometrically, a vector $\mathbf{w}$ of $\mathbb{R}^d$ with unit norm $\|\mathbf{w}\| = 1$ then lies on the ellipsoid (centered at the origin) whose axis have endpoints at $\pm(1/\sqrt{\lambda_i})\mathbf{e}_i$, $i = 1, \ldots, d$, where $(\lambda_i, \mathbf{e}_i)$, $i = 1, \ldots, d$ denote the eigenvalue-eigenvector pair of the positive definite covariance matrix $\boldsymbol{\mathcal{C}}$.

Nonlinearity in FastICA

As in [37, 38], we formulate the FastICA method without the unnecessary pre-whitening stage. FastICA method is based on the idea of maximizing a "non-Gaussianity" measure $|\,\mathbb{E}_F[G(\mathbf{w}^T\mathbf{x})]\,|$, where G can be any twice continuously differentiable nonlinear and nonquadratic function with $G(0) = 0$, and write $g = G'$ and $g' = G''$ for the 1st and 2nd derivative of G, respectively. Function g is then called the *nonlinearity*. The standard nonlinearities and their acronyms implemented in FastICA software [30] are

$$\begin{aligned} \textit{pow3}: \; & g(s) = s^3 \\ \textit{tanh}: \; & g(s) = \tanh(s) \\ \textit{gaus}: \; & g(s) = se^{-s^2/2} \\ \textit{skew}: \; & g(s) = s^2 \end{aligned}$$

Figure 2.2 shows their graphs. Nonlinearity *pow3* (that corresponds to the original FastICA estimator [28]) is recommended for sub-Gaussian sources, *gaus* for super-Gaussian sources (as it redescends to zero and thus gives outliers less weight) whereas *tanh* (which is bounded, but less robust than *gaus*) is described as a "good general-purpose contrast function" [29]. The nonlinearity *skew* corresponds to skewness optimization and it can

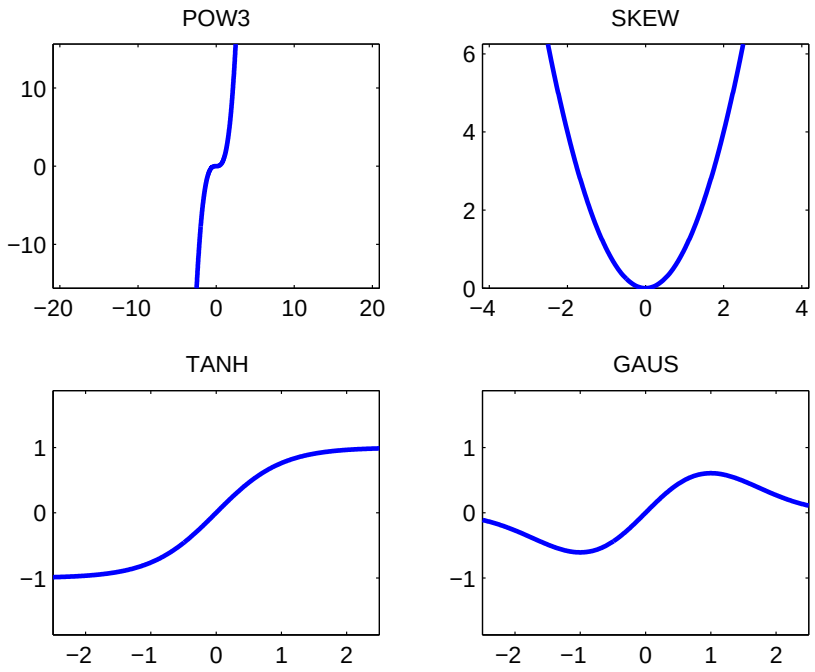


Figure 2.2: The standard nonlinearities implemented in FastICA software [30]

be used only to extract a non-symmetric source; it is commonly used in applications of FastICA to functional magnetic resonance imaging (fMRI) data since the estimated spatial maps often possess skewed distributions [39]. If the distribution of the sources is known, then the optimal nonlinearity is the location score function $\varphi(s) \triangleq -\frac{d}{ds}\log f(s)$, where f denotes the p.d.f. of the extracted source. In [40] symmetric FastICA approach combined with adaptive estimation of the location score function using the Pearson system is utilized.

Deflationary (k-unit) FastICA

The criterion must be optimized under a constraint on the scale of $\mathbf{w}$, e.g. $\|\mathbf{w}\|^2 = \mathbf{w}^T\boldsymbol{\mathcal{C}}\mathbf{w} = 1$. The *1-unit FastICA functional* $\mathbf{w}_{g,1}(F)$ is then defined as

$$\mathbf{w}_{g,1} = \arg\max_{\|\mathbf{w}\|=1} |\,\mathbb{E}[G(\mathbf{w}^T\mathbf{x})]\,|.$$

The dependence of the solution on the choice of G is indicated in the subindex via its derivative (nonlinearity) g. The above constrained optimization problem can be solved by the FastICA algorithm that uses a fixed point scheme for finding a local extrema of $\mathbb{E}[G(\mathbf{w}^T\mathbf{x})]$ subject to $\|\mathbf{w}\| = 1$.

If we wish to find more than one source, then at the kth deflation-stage, an additional constraint of orthogonality of $\mathbf{w}$ with the previously found vectors $\mathbf{w}_{g,1}, \ldots, \mathbf{w}_{g,k-1}$

is required, *i.e.*

$$\langle \mathbf{w}, \mathbf{w}_{g,i} \rangle = \mathbf{w}^T \boldsymbol{\mathcal{C}} \mathbf{w}_{g,i} = 0 \quad \text{for} \;\; i = 1, \ldots, k-1. \tag{2.5}$$

This means that $\mathbf{w}$ yields a projection $s = \mathbf{w}^T\mathbf{x}$ that is uncorrelated with the previously found sources $\mathbf{w}_{g,i}^T\mathbf{x}$, $i = 1, \ldots, k-1$. Such a constraint is natural as independence implies uncorrelatedness. The *kth FastICA functional* $\mathbf{w}_{g,k}(F)$ is thus defined as

$$\mathbf{w}_{g,k} = \arg\max_{\|\mathbf{w}\|=1} \big| \mathbb{E}[G(\mathbf{w}^T\mathbf{x})] \big| \quad \text{subject to (2.5)}.$$

If $\mathbf{x}$ follows ICA model, *i.e.* $\mathbf{x} \sim F_{\mathbf{A}}$, then under general assumptions (given below), $\mathbf{w}_{g,k}(F_{\mathbf{A}})$ is equal to one of the demixing vectors that have not been found at the earlier deflation stages *but it is not known in advance which one*. Therefore, we can assume w.l.o.g. (since it is always possible to shuffle the sources s_i's in such order due to permutation ambiguity of the ICA problem), that the solution (local maxima) is the kth demixing vector, *i.e.* $\mathbf{w}_{g,k}(F_{\mathbf{A}}) = \mathbf{w}_k$.

To find $k \in \{1, \ldots, d\}$ sources the FastICA requires the assumption

$$\begin{cases} \text{If } k < d: \; \mathbb{E}[g(s_j)s_j] - \mathbb{E}[g'(s_j)] \neq 0 \;\; \forall j = 1, \ldots, k \\ \text{If } k = d: \; \mathbb{E}[g(s_j)s_j] - \mathbb{E}[g'(s_j)] \neq 0 \;\; \forall j = 1, \ldots, d-1. \end{cases} \tag{2.6}$$

The last deflation stage requires special attention since the last source is fully determined based on the previous extractions: $\mathbf{w}_{g,d}$ is the unit norm vector of $\mathbb{R}^d$ that is orthogonal (in the inner vector space) to the previously found set of orthonormal demixing vectors $\mathbf{w}_{g,1}, \ldots, \mathbf{w}_{g,d-1}$. This also means that the last extracted source can be Gaussian. For example, in case of *pow3* nonlinearity ($g(s) = s^3$ and $g'(s) = 3s^2$), assumption (2.6) is satisfied if $\mathbb{E}[s_j^4] - 3\mathbb{E}[s_j^2] \neq 0$, implying that the (zero mean and unit variance) sources s_j, $j = 1, \ldots, k$ need to have finite 4th-order moments with non-zero kurtosis ($k < d$). Thus none of the s_j's, $j = 1, \ldots, k$, can be Gaussian. Nonlinearities *tanh* or *gaus*, however, do not require additional higher-order moment assumptions, and in this respect, they are more appropriate choices for super-Gaussian sources.

The FastICA algorithm

The FastICA algorithm finds $\mathbf{w}_{g,k}$ (for $k = 1, \ldots, d$) by iterating the steps

Step 1. $\mathbf{w} \leftarrow \boldsymbol{\mathcal{C}}^{-1}\mathbb{E}[g(\mathbf{w}^T\mathbf{x})\mathbf{x}] - \mathbb{E}[g'(\mathbf{w}^T\mathbf{x})]\mathbf{w}$

Step 2. $\mathbf{w} \leftarrow \text{proj}^{\perp}(\mathbf{w})$

Step 3. $\mathbf{w} \leftarrow \mathbf{w}/\|\mathbf{w}\|$

until convergence, *i.e.* until current and previously found vectors are practically parallel. Above

$$\text{proj}^{\perp}(\mathbf{w}) = \mathbf{w} - \sum_{j=1}^{k-1} \langle \mathbf{w}, \mathbf{w}_{g,j} \rangle \mathbf{w}_{g,j}$$

is the projection of $\mathbf{w}$ onto the orthogonal complement of the subspace (of the inner product space $\mathbb{R}^d$) spanned by the previously found demixing vectors. This is required for $\mathbf{w}$ to satisfy the constraint (2.5) whereas in Step 3 the vector is normalized to satisfy the unit norm constraint. Observe that at the last deflation stage ($k = d$), the algorithm needs only one iteration and step 1 can be omitted. Naturally, in practice the expectations above are replaced by sample means. As highlighted earlier, it is not known in advance which component the algorithm finds. It mainly depends on the initial value of the iteration.

Note that more commonly the FastICA algorithm is represented for whitened data. In the whitened space, the inner product (2.4) reduces to the dot product and the algorithm above reduces to common form represented e.g. in [8, Table 8.3]. In the whitened space, after finding the orthogonal demixing matrix by the FastICA algorithm, the final step is to calculate the original demixing matrix as $\mathbf{W}_g \leftarrow \mathbf{W}_g \boldsymbol{\mathcal{C}}^{-1/2}$, where $\mathbf{W}_g = (\mathbf{w}_{g,1} \cdots \mathbf{w}_{g,d})^T$. A bulk of the research on FastICA so far has concentrated on the convergence speed of the FastICA algorithm (e.g. [8, 28, 29, 41–43]) and only few to statistical properties of FastICA (e.g. [37, 38, 44, 45]).

A straightforward attempt to robustify FastICA algorithm by employing a robust covariance matrix estimator in place of the conventional covariance matrix fails for at least two reasons. First, the derivation that leads to FastICA algorithm essentially depends on the conventional covariance matrix. Hence, when the covariance matrix is replaced by some robust estimator, the algorithm experiences convergence problems (also reported e.g. in [46, Sect. 2.4]). Second, although many robust covariance estimators have been proven to be consistent estimators of the covariance matrix (up to a multiplicative scalar constant) in the elliptical model, a robust estimator may not estimate the covariance matrix in the ICA model.

2.3.3 FOBI

FOBI (Fourth-Order Blind Identification) method [12] was one of the first methods to solve the ICA problem. Recently, it has also been used to discriminate between multivariate models [47]. In FOBI, the demixing matrix is calculated algebraically from the matrix product of the inverse of the covariance matrix $\boldsymbol{\mathcal{C}}(\mathbf{x})$ and the *kurtosis matrix* [II]

$$\boldsymbol{\mathcal{K}}(\mathbf{x}) \triangleq \mathbb{E}[(\mathbf{x}^T \boldsymbol{\mathcal{C}}(\mathbf{x})^{-1}\mathbf{x})\mathbf{x}\mathbf{x}^T]$$

where $\mathbf{x}$ is assumed to be centered so that $\mathbb{E}[\mathbf{x}] = \mathbf{0}$ holds. Kurtosis matrix has the properties:

- *Equivariance under invertible linear transformations*:

$$\boldsymbol{\mathcal{K}}(\mathbf{G}\mathbf{x}) = \mathbf{G}\boldsymbol{\mathcal{K}}(\mathbf{x})\mathbf{G}^T$$

  for any non-singular $d \times d$ matrix $\mathbf{G}$, *i.e.* it is a scatter matrix [II].

- *IC-property* [II]: If $\mathbf{s}$ has independent components of zero mean, then $\boldsymbol{\mathcal{K}}(\mathbf{s})$ reduces to a diagonal matrix,

$$\boldsymbol{\mathcal{K}}(\mathbf{s}) = \boldsymbol{\Delta}^2 \operatorname{diag}(\kappa_i + d + 2\,),$$

  where $\boldsymbol{\Delta} = \operatorname{diag}(\sigma_i)$ and $\kappa_i \triangleq \operatorname{kurt}(s_i)$ denotes the kurtosis of the ith source.

These properties are essential in proving the following result.

Theorem 2.3.1 (Publication [II])**.** *Assume that* $\mathbf{x}$ *follows ICA model, and that the kurtosis of the sources* $s_1, \ldots, s_d$ *exists and are distinct, i.e.* $\kappa_i \neq \kappa_j$*. Then, it holds that*

$$[\boldsymbol{\mathcal{C}}(\mathbf{x})^{-1}\boldsymbol{\mathcal{K}}(\mathbf{x})]\mathbf{W}^T = \operatorname{diag}(\kappa_i + d + 2)\mathbf{W}^T,$$

that is, the demixing vectors $\mathbf{w}_1, \ldots, \mathbf{w}_d$ *are the eigenvectors of the matrix* $\boldsymbol{\mathcal{C}}(\mathbf{x})^{-1}\boldsymbol{\mathcal{K}}(\mathbf{x})$ *and the corresponding eigenvalues are* $\kappa_i + d + 2$*,* $i = 1, \ldots, d$*.*

Recall that the eigenvectors are subject to the same sign and scale ambiguity as the demixing vectors, namely they are uniquely defined only up to a sign and positive constant scalar. Usually, eigenvectors are defined to have a unit Euclidean norm to get rid of the scale ambiguity and most eigenvector-eigenvalue extraction routines in commercial software packages do so. By Theorem 2.3.1, a FOBI demixing matrix estimator

$$\mathbf{W}_{\text{fobi}} = \begin{pmatrix}\mathbf{e}_1 & \cdots & \mathbf{e}_d\end{pmatrix}^T,$$

that contains the eigenvectors $\mathbf{e}_1, \ldots, \mathbf{e}_d$ of $\boldsymbol{\mathcal{C}}(\mathbf{x})^{-1}\boldsymbol{\mathcal{K}}(\mathbf{x})$ as rows (regardless of the normalization of the eigenvectors used) is a valid demixing matrix estimator provided that the sources have distinct kurtosis, that is, $\mathbf{s}_0 = \mathbf{W}_{\text{fobi}}\mathbf{x}$ is a copy of $\mathbf{s}$.

FOBI is arguably among the simplest methods to solve the ICA problem. It can be computed via standard eigenvector decomposition operating on matrix $\boldsymbol{\mathcal{C}}(\mathbf{x})^{-1}\boldsymbol{\mathcal{K}}(\mathbf{x})$. Hence it is also computationally among the most efficient approaches to ICA.

FOBI algorithm (Publication [II], pp. 3798):

Step 1. Calculate the inverse of the covariance matrix $\mathbf{Q}(\mathbf{x}) = \boldsymbol{\mathcal{C}}(\mathbf{x})^{-1}$.
Matlab code: `Q = (X*X'/n)\eye(d); % X is d x n centered data matrix`

Step 2. Calculate the kurtosis matrix $\boldsymbol{\mathcal{K}}(\mathbf{x}) = \mathbb{E}[(\mathbf{x}^T\mathbf{Q}(\mathbf{x})\mathbf{x})\mathbf{x}\mathbf{x}^T]$.
Matlab code: `K = ones(d,1)*sum(X.*(Q*X)).*X*X'/n;`

Step 3. Calculate the eigenvectors $\mathbf{e}_1, \ldots, \mathbf{e}_d$ of $\mathbf{Q}(\mathbf{x})\boldsymbol{\mathcal{K}}(\mathbf{x})$ and set $\mathbf{W}_{\text{fobi}} = \begin{pmatrix}\mathbf{e}_1 & \cdots & \mathbf{e}_d\end{pmatrix}^T$.
Matlab code: `[Wt L] = eig(Q*K); Wfobi = Wt';`

Note that an efficient implementation of FOBI in Matlab requires only three lines of code. This is in deep contrast to most ICA methods proposed thus far as it is not based on high complexity iterative optimization of a non-linear function.

Alternatively, as was put forth in the original publication [12], the FOBI estimator can be calculated via the steps :

1. Calculate the whitened data $\mathbf{v} = \boldsymbol{\mathcal{B}}\mathbf{x}$, where $\boldsymbol{\mathcal{B}}$ is any whitening matrix.

2. Calculate the eigenvalue decomposition (EVD) of the kurtosis matrix $\boldsymbol{\mathcal{K}}(\mathbf{v})$ of the whitened mixture $\mathbf{v}$:

$$\boldsymbol{\mathcal{K}}(\mathbf{v}) = \mathbf{U}\mathbf{L}\mathbf{U}^T, \tag{2.7}$$

   where $\mathbf{U}$ is the $d \times d$ orthogonal matrix of eigenvectors of $\boldsymbol{\mathcal{K}}(\mathbf{v})$ as columns and $\mathbf{L}$ is the $d \times d$ diagonal matrix of respective eigenvalues as diagonal elements.

3. Set $\mathbf{W}_{\text{fobi}} = \mathbf{U}^T\boldsymbol{\mathcal{B}}$.

The original approach involves two steps, whitening of the mixture $\mathbf{x}$ followed by the computation of the EVD of $\boldsymbol{\mathcal{K}}(\mathbf{v})$. The earlier approach, however, is to be preferred as it involves only one eigenvector extraction instead of two.

2.3.4 Extensions of FOBI

The FOBI method explained above has the limitation that it can only separate sources with distinct kurtosis. If any two sources have identical distribution (up to location and scale), then they have necessarily identical kurtosis due to location-scale invariance property of the kurtosis: $\text{kurt}(as_i + b) = \text{kurt}(s_i)$ for all $a \neq 0$ and $b \in \mathbb{R}$. Thus sources can not have identical distribution.

This limitation of FOBI originates from the lack of uniqueness of eigenvectors corresponding to an eigenvalue with multiplicity greater than one. Recall that FOBI identifies demixing vectors as the eigenvectors $\mathbf{e}_1, \ldots, \mathbf{e}_d$ of $\boldsymbol{\mathcal{C}}^{-1}(\mathbf{x})\boldsymbol{\mathcal{K}}(\mathbf{x})$; the respective eigenvalues being $\lambda_i = \kappa_i + d + 2$ (Theorem 2.3.1). Thus two sources, say s_1 and s_2, with identical kurtosis $\kappa = \kappa_1 = \kappa_2$, indicates an eigenvalue with multiplicity two, $\lambda = \lambda_1 = \lambda_2$. In such instances, eigenvectors $\mathbf{e}_1$ and $\mathbf{e}_2$ no longer identify demixing vectors $\mathbf{w}_1$ and $\mathbf{w}_2$ up to their sign and scale. Namely, although $\mathbf{w}_1$ and $\mathbf{w}_2$ are eigenvectors (corresponding to the eigenvalue λ), so is any $\mathbf{w}_0 = a_1\mathbf{w}_1 + a_2\mathbf{w}_2$ $(a_1, a_2 \in \mathbb{R})$, *i.e.* $[\boldsymbol{\mathcal{C}}^{-1}(\mathbf{x})\boldsymbol{\mathcal{K}}(\mathbf{x})]\mathbf{w}_0 = \lambda\mathbf{w}_0$. Hence the computed eigenvectors $\mathbf{e}_1$ or $\mathbf{e}_2$ are some unknown linear combinations of $\mathbf{w}_1$ and $\mathbf{w}_2$, but we are not able to identify $\mathbf{w}_1$ and $\mathbf{w}_2$ from the sole knowledge of $\mathbf{e}_1$ and $\mathbf{e}_2$.

Generalized FOBI

There exists a simple generalization of FOBI that avoids the assumption of distinct kurtosis. See [7]. Let us define a *cumulant matrix* [7, Sect. 3.2.1] as follows

$$\boldsymbol{\mathcal{K}}_{\mathbf{M}}(\mathbf{x}) \triangleq \mathbb{E}[(\mathbf{x}^T\mathbf{M}\mathbf{x})\mathbf{x}\mathbf{x}^T] - \boldsymbol{\mathcal{C}}(\mathbf{x})\mathrm{Tr}[\mathbf{M}\boldsymbol{\mathcal{C}}(\mathbf{x})] - \boldsymbol{\mathcal{C}}(\mathbf{x})\mathbf{M}\boldsymbol{\mathcal{C}}(\mathbf{x}) - \boldsymbol{\mathcal{C}}(\mathbf{x})\mathbf{M}^T\boldsymbol{\mathcal{C}}(\mathbf{x}),$$

where $\mathbf{M}$ is any $d \times d$ matrix and $\mathbf{x}$ is assumed to be centered (so $\mathbb{E}[\mathbf{x}] = \mathbf{0}$). The matrix parameter $\mathbf{M}$ is a *tuning parameter* chosen by the user. How to choose $\mathbf{M}$ is discussed at the end of this subsection. Observe that the cumulant matrix is symmetric, so $\boldsymbol{\mathcal{K}}_{\mathbf{M}}(\mathbf{x})^T = \boldsymbol{\mathcal{K}}_{\mathbf{M}}(\mathbf{x})$, and possesses similar properties as the kurtosis matrix:

(C.1) Let $\mathbf{x}' = \mathbf{G}\mathbf{x}$ denote a linear transformation of d-variate r.v. $\mathbf{x}$ for any $q \times d$ matrix $\mathbf{G}$. Then

$$\boldsymbol{\mathcal{K}}_{\mathbf{M}}(\mathbf{x}') = \mathbf{G}\,\boldsymbol{\mathcal{K}}_{\mathbf{N}}(\mathbf{x})\,\mathbf{G}^T,$$

where $\mathbf{N} = \mathbf{G}^T\mathbf{M}\mathbf{G}$ is a $d \times d$ matrix and $\mathbf{M}$ is any $d \times d$ matrix.

(C.2) if $\mathbf{s} = (s_1, \ldots, s_d)^T$ has independent components of zero mean ($\mathbb{E}[\mathbf{s}] = \mathbf{0}$), then the cumulant matrix $\boldsymbol{\mathcal{K}}_{\mathbf{M}}(\mathbf{s})$ is a diagonal matrix,

$$\boldsymbol{\mathcal{K}}_{\mathbf{M}}(\mathbf{s}) = \mathrm{diag}(c_1 m_{11}, \ldots, c_d m_{dd}),$$

where $c_i \equiv \mathrm{cum}_4(s_i)$ denotes the 4th-order cumulant of the ith source and m_{ii} is the ith diagonal element of $d \times d$ matrix $\mathbf{M}$, $i = 1, \ldots, d$.

Hence if $\mathbf{x}$ follows ICA model and is centered (so $\mathbb{E}[\mathbf{s}] = \mathbf{0}$), then by (C.1) and (C.2), we have that

$$\boldsymbol{\mathcal{K}}_{\mathbf{M}}(\mathbf{x}) = \mathbf{A}\boldsymbol{\mathcal{K}}_{\mathbf{N}}(\mathbf{s})\mathbf{A}^T,$$

where $\mathbf{N} = \mathbf{A}^T\mathbf{M}\mathbf{A}$ and $\boldsymbol{\mathcal{K}}_{\mathbf{N}}(\mathbf{s}) = \mathrm{diag}(c_i\mathbf{a}_i^T\mathbf{M}\mathbf{a}_i)$, For example, the choice $\mathbf{M} = \mathbf{I}$ yields $\boldsymbol{\mathcal{K}}_{\mathbf{N}}(\mathbf{s}) = \mathrm{diag}(c_i\|\mathbf{a}_i\|^2)$.

An obvious generalization of the FOBI method is described next. The proof proceeds similarly as the proof of Theorem 2.3.1.

Theorem 2.3.2. *Assume that centered* $\mathbf{x}$ *follows ICA model, a) kurtosis* κ_i *of the sources exists, and b)* $\lambda_i = (\sigma_i^2\kappa_i)\mathbf{a}_i^T\mathbf{M}\mathbf{a}_i$ *are distinct for* $i = 1, \ldots, d$. *Then, it holds that,*

$$[\boldsymbol{\mathcal{C}}(\mathbf{x})^{-1}\boldsymbol{\mathcal{K}}_{\mathbf{M}}(\mathbf{x})]\mathbf{W}^T = \boldsymbol{\Lambda}\mathbf{W}^T,$$

where $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_i)$, *that is, the demixing vectors* $\mathbf{w}_1, \ldots, \mathbf{w}_d$ *are the eigenvectors of the matrix* $\boldsymbol{\mathcal{C}}(\mathbf{x})^{-1}\boldsymbol{\mathcal{K}}_{\mathbf{M}}(\mathbf{x})$ *and the corresponding eigenvalues are* $\lambda_1, \ldots, \lambda_d$.

Thus the *generalized FOBI demixing matrix estimator* $\mathbf{W}_{\mathbf{M}}$ that contains eigenvectors of $\boldsymbol{\mathcal{C}}(\mathbf{x})^{-1}\boldsymbol{\mathcal{K}}_{\mathbf{M}}(\mathbf{x})$ as rows (regardless of their normalization) is a valid demixing

matrix estimator provided that the assumptions of Theorem 2.3.2 hold. If we choose matrix $\mathbf{M}$ randomly, then the eigenvalues $\lambda_1, \ldots, \lambda_d$ are distinct with probability one. It is easy to show that if $\mathbf{M} = \boldsymbol{\mathcal{C}}^{-1}(\mathbf{x})$, then the corresponding estimator $\mathbf{W}_{\mathbf{M}}$ is the FOBI estimator $\mathbf{W}_{\text{fobi}}$. We note that originally Cardoso [7, Sect. 3.3.1] proposed using another cumulant matrix $\boldsymbol{\mathcal{K}}_{\mathbf{M}}(\mathbf{x})$ with distinct $\mathbf{M}$ in place of the covariance matrix in calculating the demixing matrix. Herein we chose to use $\boldsymbol{\mathcal{C}}(\mathbf{x})$ since then the method can be interpreted as a generalization of FOBI.

Some problems still remains with the generalized FOBI approach. For example,

- *The estimator lacks ICA-equivariance* (in the sense of [48, Sect. II-C], [II,Def. 5]).
- *The problem of selecting* $\mathbf{M}$. Although choosing $\mathbf{M}$ randomly guarantees distinct eigenvalues, different choices can lead to different sample performance. Since the eigenvalues λ_i also depend on the unknown mixing matrix $\mathbf{A}$, it is not possible to choose matrix $\mathbf{M}$ optimally, e.g. to guarantee large dispersion of eigenvalues. One could, for example, compute the sample estimator $\hat{\mathbf{W}}_{\mathbf{M}}$ for several randomly chosen matrices $\mathbf{M}$ and retain the one with the largest eigenvalue spread.
- *Inaccuracy for small sample sizes.* By (C.2), $\boldsymbol{\mathcal{K}}_{\mathbf{M}}(\mathbf{s})$ is a diagonal matrix. However, for small sample lengths (say $n < 1000$), the sample estimator calculated from the source data matrix can be rather far from a diagonal matrix.

JADE

JADE (Joint Approximate Diagonalization of Eigenmatrices) [7, 31] is an elegant approach that solves the above problems associated with generalized FOBI estimator. Let $\mathbf{v} = \boldsymbol{\mathcal{B}}\mathbf{x}$ denote the centered and whitened data. Hence $\mathbb{E}[\mathbf{v}] = \mathbf{0}$ and $\boldsymbol{\mathcal{C}}(\mathbf{v}) = \mathbf{I}$. Recall that the whitened r.v. $\mathbf{v}$ also follows ICA model $\mathbf{v} = \tilde{\mathbf{A}}\mathbf{s}$, where $\tilde{\mathbf{A}} = \boldsymbol{\mathcal{B}}\mathbf{A}$ is an orthogonal mixing matrix and $\mathbb{E}[\mathbf{s}] = \mathbf{0}$ due to centering. In an analogous fashion, the properties (C.1) and (C.2) imply that

$$\boldsymbol{\mathcal{K}}_{\mathbf{M}}(\mathbf{v}) = \tilde{\mathbf{A}} \operatorname{diag}(c_i \tilde{\mathbf{a}}_i^T \mathbf{N} \tilde{\mathbf{a}}_i)\, \tilde{\mathbf{A}}^T,$$

where $\tilde{\mathbf{a}}_i$ denotes the ith column of $\tilde{\mathbf{A}}$. Let $\{\mathbf{M}_1, \ldots, \mathbf{M}_p\}$ be a set of $d \times d$ tuning matrices and $\boldsymbol{\mathcal{K}}_i = \boldsymbol{\mathcal{K}}_{\mathbf{M}_i}(\mathbf{v})$, $i = 1, \ldots, p$ the corresponding p cumulant matrices. Let us define a non-negative measure of non-diagonality of a matrix as $\operatorname{Off}(\mathbf{G}) \triangleq \sum_{i \neq j} (g_{ij})^2$, *i.e.* as the sum of squares of the off-diagonal elements of its matrix argument. Then $\operatorname{Off}(\mathbf{G}) = 0$ iff $\mathbf{G}$ is diagonal, and the larger the measure $\operatorname{Off}(\mathbf{G})$ is the more 'non-diagonal' it looks like. Particularly, $\operatorname{Off}(\tilde{\mathbf{W}} \boldsymbol{\mathcal{K}}_i \tilde{\mathbf{W}}^T) = 0$ when $\tilde{\mathbf{W}} = \tilde{\mathbf{A}}^{-1} = \tilde{\mathbf{A}}^T$. For an orthogonal matrix $\mathbf{U}$ define the *joint-diagonality (JD) criterion* as

$$\operatorname{JD}(\mathbf{U}) \triangleq \sum_{i=1}^{p} \operatorname{Off}(\mathbf{U} \boldsymbol{\mathcal{K}}_i \mathbf{U}^T)$$

which measures how close to diagonality an orthogonal matrix $\mathbf{U}$ can simultaneously bring the set of p cumulant matrices generated by $\{\mathbf{M}_i\}$. The idea in JADE is to find the demixing matrix $\tilde{\mathbf{W}}$ of the whitened data as the minimizer of the JD criterion where the set $\{\mathbf{M}_i\}$ is chosen cleverly as a set of 'eigenmatrices'. At the population level, the set of cumulant matrices can be exactly jointly diagonalized, but for finite samples, the set $\{\hat{\boldsymbol{\mathcal{K}}}_i\}$ calculated from the sample can only be approximately jointly diagonalized. In finding the orthogonal matrix that minimizes the JD criterion, a Jacobi algorithm [7, 31] is utilized. We note that Jacobi type of optimization is used by other ICA methods as well, e.g. [6, 49]. Also the (approximate) joint diagonalization is a commonly used approach that has a long history in ICA; see e.g. [35, 50–53].

Although JADE in essence solves the other two problems of generalized FOBI method, yet the lack of ICA-equivariance remains. The downside is that due to the pairwise processing (Jacobi technique), JADE is not well suited for high-dimensional data sets.

DOGMA

FOBI suffers from its non-robustness (*i.e.* high sensitivity to outliers) and limited versatility (e.g. existence of the 4th-order moments of the sources is required). DOGMA (Diagonalizers Of Generalized covariance MAtrices) is a generalization of FOBI utilizing any distinct pair of scatter matrices with independent components (IC-)property in place of the covariance and kurtosis matrix; see [32] for the real-valued case and Publication [II] for the complex-valued case. Thus DOGMA estimators constitute a large family of estimators that include FOBI as a particular special case.

A scatter matrix is a generalization of the covariance matrix. A positive definite symmetric $d \times d$ matrix $\mathbf{C}(\mathbf{x})$ is called a *scatter matrix* if it is equivariant in the sense that $\mathbf{C}(\mathbf{G}\mathbf{x}) = \mathbf{G}\mathbf{C}(\mathbf{x})\mathbf{G}^T$ for any nonsingular $d \times d$ matrix $\mathbf{G}$. The covariance matrix $\boldsymbol{\mathcal{C}}(\cdot)$ and the kurtosis matrix $\boldsymbol{\mathcal{K}}(\cdot)$ are scatter matrices for distributions with finite 2nd- and 4th-order moments, respectively. Scatter matrix of a real-valued r.v. $\mathbf{x}$ is a key concept in multivariate statistics and up to date there exists a broad array of estimators one can choose from. One of the first proposals were M-estimators of scatter by Maronna [14]. Since this pioneering work several competing robust estimators have been proposed, e.g. minimum volume ellipsoid estimator [54], minimum covariance determinant estimator [54], S-estimators [55, 56], τ-estimators [57], *CM*-estimators [58], *MM*-estimators [59], sign and rank based scatter matrices [60, 61] to cite only a few.

Let $\mathbf{C}_1(\cdot)$ and $\mathbf{C}_2(\cdot)$ denote any pair of distinct scatter matrices. For purposes of ICA we require that the selected scatter matrices possess *IC-property*, namely, $\mathbf{C}_1(\mathbf{s})$ and $\mathbf{C}_2(\mathbf{s})$ are diagonal, *i.e.* $[\mathbf{C}_1(\mathbf{s})]_{ij} = 0$ and $[\mathbf{C}_2(\mathbf{s})]_{ij} = 0$ for all $i \neq j$ when r.v. $\mathbf{s}$ has independent components. Covariance matrix and kurtosis matrix possess IC-property. Robust scatter matrix estimators however do not necessarily possess IC-property. If

sources $s_1, \ldots, s_d$ have symmetric distribution (*i.e.* s_i has the same distribution as $-s_i$), then a scatter matrix automatically has the IC-property.

Theorem 2.3.3 (Publication [II]). *Assume that* $\mathbf{x}$ *follows ICA model and* $\mathbf{x}$ *is centered. Suppose that a pair of scatter matrices* $\mathbf{C}_1(\cdot)$ *and* $\mathbf{C}_2(\cdot)$ *possess IC-property and that* $\lambda_i = [\mathbf{C}_2(\mathbf{s})]_{ii}/[\mathbf{C}_1(\mathbf{s})]_{ii}$, $i = 1, \ldots, d$, *are distinct. Then, it holds that*

$$[\mathbf{C}_1(\mathbf{x})^{-1}\mathbf{C}_2(\mathbf{x})]\mathbf{W}^T = \mathbf{\Lambda}\mathbf{W}^T,$$

where $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_d)$. *This means that the demixing vectors* $\mathbf{w}_1, \ldots, \mathbf{w}_d$ *are the eigenvectors of the matrix* $\mathbf{C}_1(\mathbf{x})^{-1}\mathbf{C}_2(\mathbf{x})$ *and the corresponding eigenvalues are* $\lambda_1, \ldots, \lambda_d$.

Again, due to fundamental indeterminacy of ICA model, a DOGMA demixing matrix estimator containing the eigenvectors of $\mathbf{C}_1(\mathbf{x})^{-1}\mathbf{C}_2(\mathbf{x})$ as rows (regardless of the used normalization) is a valid demixing matrix provided that the eigenvalues are distinct. If $\mathbf{C}_1 = \boldsymbol{\mathcal{C}}$ and $\mathbf{C}_2 = \boldsymbol{\mathcal{K}}$, then FOBI is obtained. The assumption of distinct eigenvalues is required to separate all the sources. Namely, DOGMA demixing matrix is not able to separate the set of sources with identical eigenvalue but the rest of the sources are separated. Hence, the DOGMA method contains a *built-in warning*: detection of two close eigenvalues is an indication that the corresponding sources may not be reliably separated.

2.4 Image analysis example

2.4.1 PC, whitening and IC-transform illustrated

Let $\mathbf{x}$ be a d-dimensional random vector and denote by $\mathbf{E}\mathbf{\Lambda}\mathbf{E}^T$ the EVD of its covariance matrix $\boldsymbol{\mathcal{C}}(\mathbf{x})$. Consider the following linear transformations of a r.v. $\mathbf{x}$:

- *Principal Components (PC-)transformation*:

$$\mathbf{z} = \mathbf{E}^T\mathbf{x}.$$

- *Whitening transformation*:

$$\mathbf{v} = \boldsymbol{\mathcal{B}}\mathbf{x} = \mathbf{\Lambda}^{-1/2}\mathbf{z},$$

  where $\boldsymbol{\mathcal{B}} = \mathbf{\Lambda}^{-1/2}\mathbf{E}^T$ is a whitening matrix.

- *IC-transformation* using FOBI:

$$\mathbf{s} = \mathbf{W}_{\text{fobi}}\mathbf{x} = \mathbf{U}^T\mathbf{v}$$

  as $\mathbf{W}_{\text{fobi}}$ can be represented as $\mathbf{W}_{\text{fobi}} = \mathbf{U}^T\boldsymbol{\mathcal{B}}$ for an orthogonal matrix $\mathbf{U}$; recall Step 3 of the original FOBI algorithm on p. 19.

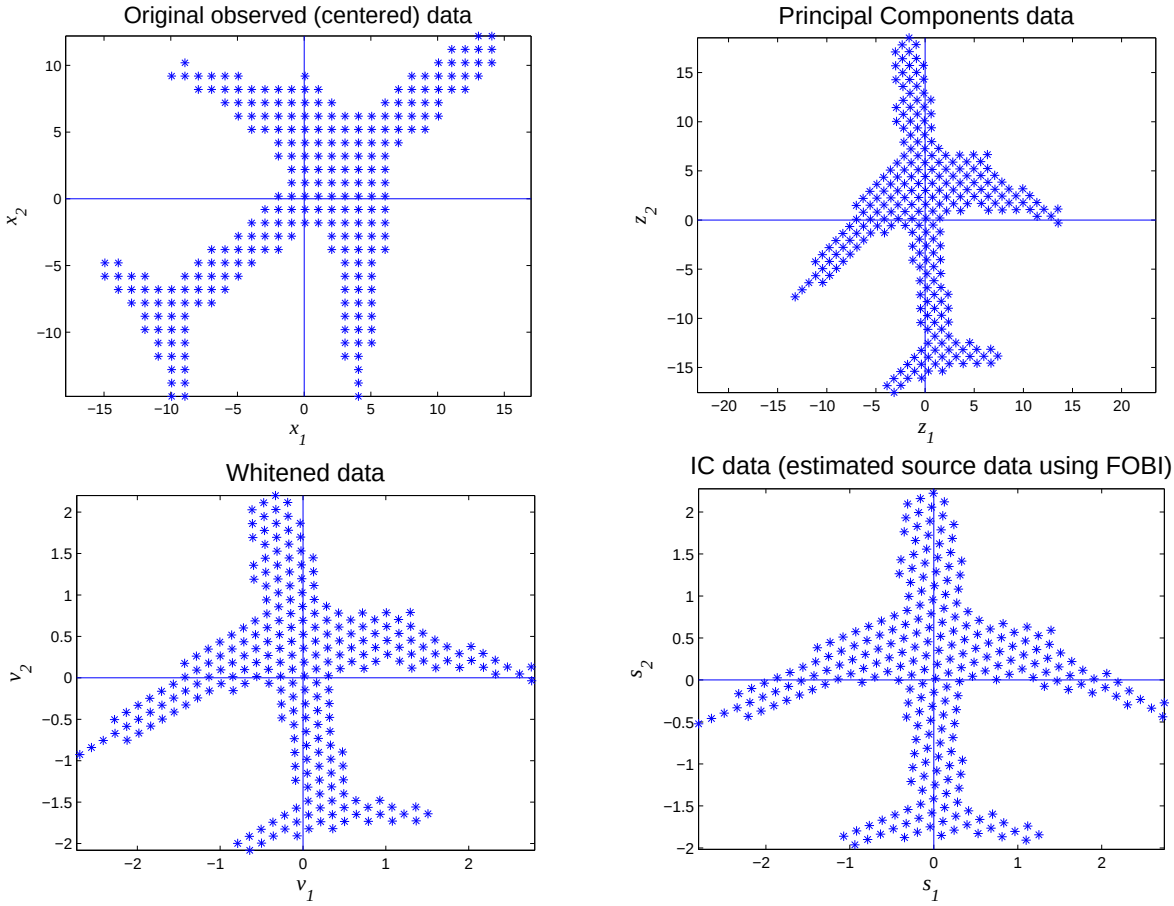


Figure 2.3: Scatter plots of the aircraft data and the transformed data.

Hence the above transforms are linear mappings related as

$$\mathbf{X} \overset{\mathbf{E}^T}{\longmapsto} \mathbf{Z} \overset{\mathbf{\Lambda}^{-1/2}}{\longmapsto} \mathbf{V} \overset{\mathbf{U}^T}{\longmapsto} \mathbf{S}.$$

These transforms are now illustrated with a simple image analysis example. Figure 2.3 depicts the picture of an aircraft that is segmented from the background using a simple thresholding method. As a result we have a sample of $n = 249$ bivariate observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Also depicted are the scatter plots of the corresponding (PC-, whitened, and IC-)transformed data based on sample statistics. Note the difference between the IC and whitening transform. While whitening translates the data to be uncorrelated, the IC-transform has the intuitive feature that the data appears as independent as possible. This is manifested by the fact that the head and the wings of the aircraft are now aligned with the axis.

2.4.2 Robustness concern illustrated

Now suppose that when the aircraft was segmented from the image, a single outlier $\mathbf{x}_0 = (-25, 20)^T$ that is not from the surface of the airplane remained unnoticed, as

shown in Figure 2.4. Also shown in the figure are the IC-transformed data using FOBI and a robust DOGMA estimator employing Tyler's and Huber's M-estimator of scatter with $q = 0.9$ as choices of scatter matrices. We recall that $0 < q < 1$ is a tuning constant that controls the robustness and efficiency of the Huber's M-estimator w.r.t. the nominal d-variate Gaussian distribution; the value $q = 0.9$ is commonly used and it can be seen as a compromise that provides sufficient robustness and yet a very small loss in efficiency at the normal model. As can be seen only the IC-transformed data based on FOBI is affected by the outlier. The outlier has affected the rotation: the aircraft is no longer aligned along the axis. IC-data using the robust DOGMA estimator is unaffected by the outlier and attains good alignment with the axis.

This example illustrates the importance of robust estimation in ICA as outliers can appear commonly in real-world data sets. We also point out that for the aircraft data set with an outlier, the FastICA algorithm employing standard nonlinearities did not converge. This feature illustrates the advantage that a method in the algebraic group of ICA methods (such as FOBI and DOGMA) can have over the methods in the optimization group (such as FastICA). It is a rather common feature for ICA methods based on optimization of a criterion function that they can experience convergence problems when the data contains spurious points or sample lengths are small.

2.5 Performance studies

2.5.1 Empirical influence functions

Let us denote by $\hat{\mathbf{w}}_j \equiv \hat{\mathbf{w}}_j(X_n)$ the finite sample estimator of the jth demixing vector $\mathbf{w}_j$ based on the sample $X_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. A popular tool to measure robustness of an estimator to an outlier is the *empirical influence function* (EIF) (also called sensitivity function [62]). The EIF is defined as

$$\mathrm{EIF}(\mathbf{x}_0; \hat{\mathbf{w}}_j, X_n) \triangleq (n+1)\{\, \hat{\mathbf{w}}_j(X_n \cup \{\mathbf{x}_0\}) - \hat{\mathbf{w}}_j(X_n)\,\}.$$

The EIF thus calculates the standardized effect (normalized by the mass $\frac{1}{n+1}$ of the contamination) of an additional observation at $\mathbf{x}_0$ on the demixing vector estimator. A robust estimator has an EIF that is a bounded function of the contamination point $\mathbf{x}_0$ which means that a large outlier $\mathbf{x}_0$ added to the given data set X_n cannot change the estimator dramatically. In most cases, EIF is a consistent estimator of the theoretical influence function [62, 63].

Figure 2.5 depicts the norms of the EIF's of the demixing vector estimates $\hat{\mathbf{w}}_1$ and $\hat{\mathbf{w}}_2$ in the case of two sources (s_1 and s_2 following a zero mean and unit variance t_6 and Uniform distribution, respectively) with no-mixing (*i.e.* $\mathbf{A} = \mathbf{I}$). In the simulations, we used the FastICA software [30] with its default settings. The sample length is $n = 2000$ and the EIF plot is averaged over 500 Monte-Carlo runs in order to obtain a smooth

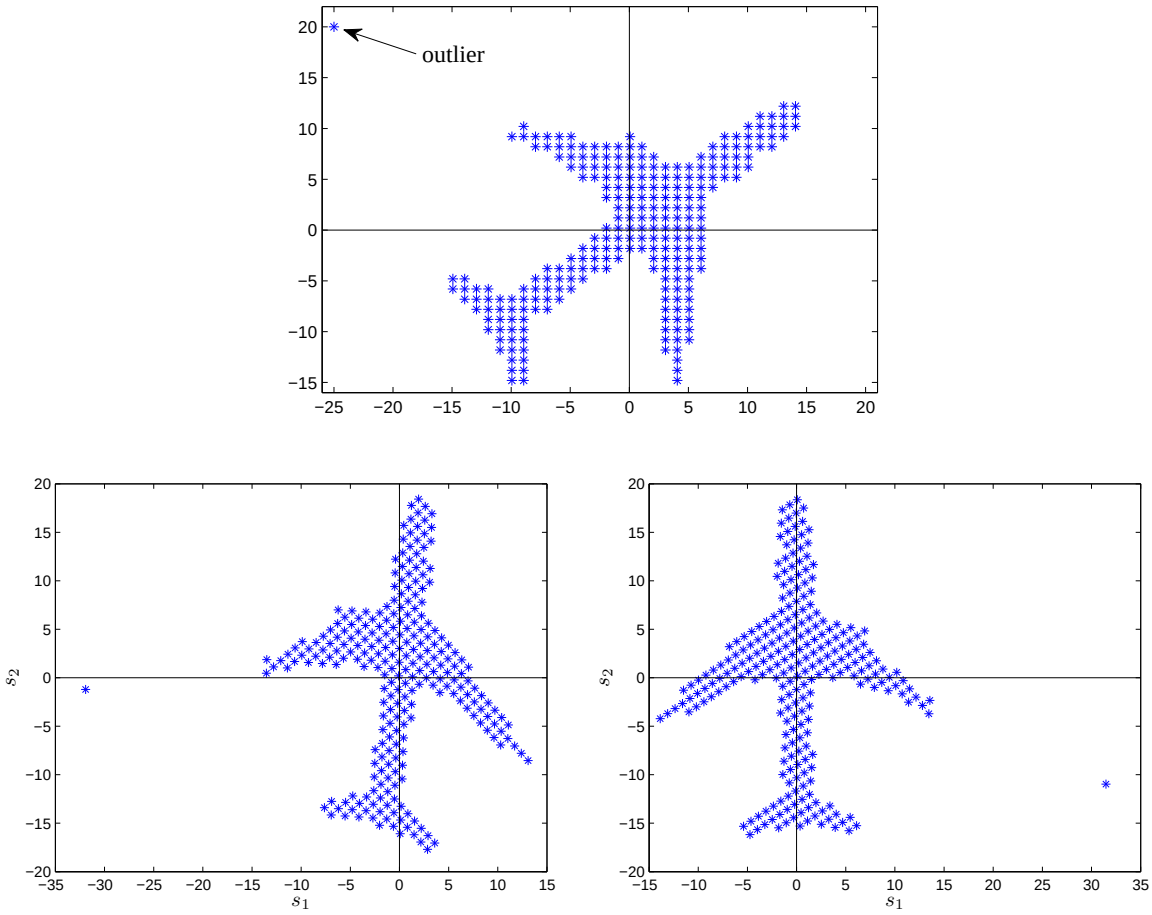
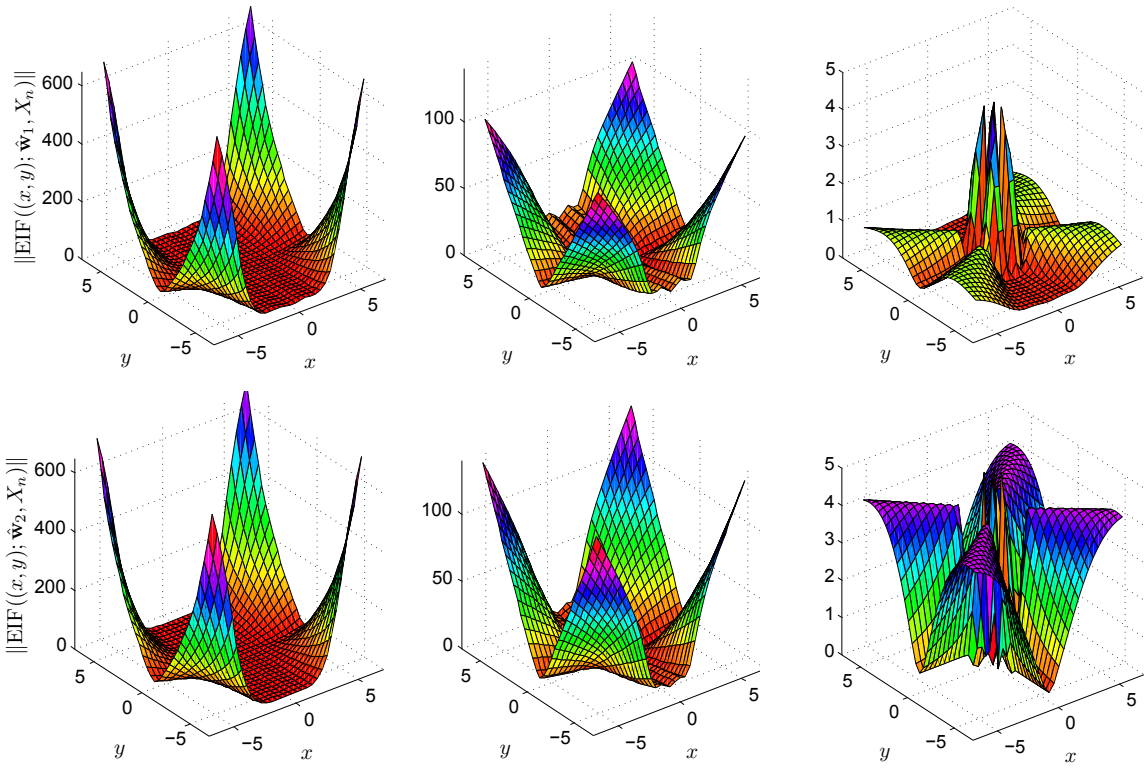


Figure 2.4: Upper plot: the aircraft data with an outlier $\mathbf{x}_0 = (-25, 20)^T$. Lower plot: IC-transformed aircraft data with an outlier based on FOBI (left) and a robust DOGMA method (right). The robust DOGMA method is unaffected whereas the orientation of the IC-data using FOBI changes due to the outlier.

surface. As can be seen, FastICA using *pow3* nonlinearity, is more influenced by an outlier than FastICA using *gaus* nonlinearity. The key finding, however, is that the EIF's of the FastICA estimators are unbounded (regardless of the used nonlinearity) but some "robustness" is offered by a robust choice of nonlinearity in the sense that the norm of the EIF then grows less rapidly with the norm of the contamination point. The EIF surfaces of the FastICA estimators also reveal that some observations are far more influential than others: namely, observation $\mathbf{x}_0 = (x, y)^T$ that lie on the diagonals of the plane (*i.e.* x and y coordinates have the same magnitude) has much larger impact than an observation of the same length lying outside the diagonals. In fact, in [37, 38] it was proved analytically based on the analysis of the theoretical influence function of the FastICA functional, that the most influential points are of the form

$$\mathbf{x}_0 = \mathbf{A}(r\boldsymbol{\ell}), \quad \boldsymbol{\ell} \triangleq (\pm 1, , \ldots, \pm 1)^T, r \in \mathbb{R}^+. \tag{2.8}$$



(a) *pow3* (b) *gaus* (c) DOGMA

Figure 2.5: The surface plots of the Euclidean norms of the empirical influence functions of $\hat{\mathbf{w}}_1$ (first row) and $\hat{\mathbf{w}}_2$ (second row) in the case of two sources (s_1 and s_2 following a zero mean and unit variance t_6 and Uniform distribution, respectively) with no-mixing ($\mathbf{A} = \mathbf{I}$). Sample length was $n = 2000$ and the EIF plots are averaged over 500 samples. Estimators used were (a) deflation-based FastICA with *pow3* and (b) *gaus* nonlinearities and (c) DOGMA utilizing Tyler's M-estimator and Huber's M-estimators with tuning constant $q = 0.9$ as the choice of scatter matrices. Only DOGMA has bounded EIF.

Since in the simulation study the 2×2 mixing matrix equals $\mathbf{A} = \mathbf{I}$, the most influential points (2.8) are indeed those that lie on the diagonals of the plane.

Let us now turn out attention to the EIF of the DOGMA estimator depicted in Figure 2.5(c). The key observation is that the EIF's are bounded. Hence an observation $\mathbf{x}_0$ (no matter how large) always has a limited influence on the estimator. First, we wish to highlight that due to the different standardization used by the above estimators, the calculated demixing vector estimates were post-processed to have unit norm. Hence the scales of the norms of the EIF's for different estimators shown in Figure 2.5 are comparable. As an example, if $\mathbf{x}_0 = (6, 6)^T$, then $\|\mathrm{EIF}(\mathbf{x}_0; \hat{\mathbf{w}}_1, X_n)\|$ equals 686, 104, and 0.86 in the case of FastICA with *pow3* and *gaus* nonlinearities and DOGMA, respectively. *This means that the influence of a point* $\mathbf{x}_0 = (6, 6)^T$ *on the 1st demixing vector estimator* $\hat{\mathbf{w}}_1$ *was about 700 times larger in the case of FastICA with pow3 nonlinearity as compared to the DOGMA estimator.*

2.5.2 A cautionary note

The quality of the separation can be assessed by calculating the *interference to signal ratio*

$$\mathrm{ISR}(\hat{\mathbf{V}}) = \sqrt{\frac{1}{d(d-1)}\Big\{\sum_{i=1}^{d}\Big(\sum_{j=1}^{d}\frac{(\hat{v}_{ij})^2}{\max_\ell(\hat{v}_{i\ell})^2}-1\Big)\Big\}}.$$

where $\hat{\mathbf{V}} = (\hat{v}_{ij}) = \hat{\mathbf{W}}\mathbf{A}$ and $\hat{\mathbf{W}}$ an estimator of the demixing matrix $\mathbf{W}$. The squared ISR (with different scaling) was called the ICI (inter-channel interference) in [64]. Alternatively, one could use the Amari performance index [65]. Perfect separation implies that $\hat{\mathbf{V}}$ is equal to a matrix with one non-zero entry in each row and each column, yielding $\mathrm{ISR}(\hat{\mathbf{V}}) = 0$. When the quality of separation degrades, the value of ISR increases and attains the maximum value of 1 when $\hat{\mathbf{V}}$ is non-singular with $|\hat{v}_{ij}|$ equal in *each* row $i = 1, \ldots, d$. Attaining the maximum value 1 is highly pathological situation and in practise an estimator never reaches it. More natural baseline of poor performance is the value of ISR for a random non-singular matrix $\hat{\mathbf{V}}$ (*i.e.* when $\hat{\mathbf{W}}$ represents a pure guess). Then maximal ISR value 1 or its vicinity would indicate pathological/defective (beyond poor) performance.

Calculating the plain guess baseline for $\mathrm{ISR}(\cdot)$ requires generating a random non-singular matrix which we calculate using the singular value decomposition (SVD) as $\hat{\mathbf{V}} = \mathbf{U}_1\mathbf{L}\mathbf{U}_2^T$, where $\mathbf{U}_1$ and $\mathbf{U}_2$ are randomly generated orthogonal matrices and $\mathbf{L}$ is a diagonal matrix with values from $Unif(0.01, 1)$ distribution. The baselines (calculated as the mean over 50000 randomly generated non-singular matrices) at dimensions $d = 2, 4, 8$ were $0.482, 0.487, 0.463$.

We consider bivariate ICA model where sources s_1 and s_2 have unit variance t_6-distribution and Uniform distribution, respectively. Sample length is $n = 2000$ and the number of samples is 3000. To study the robustness of FastICA, we add a point to the sample, and study its influence to the attained separation quality using the ISR. The added point is generated as

$$\mathbf{x}_0^{(i)} = \mathbf{A}\mathbf{s}_i \text{ with } \mathbf{s}_i = r(\cos(\vartheta_i), \sin(\vartheta_i))^T,\ i \in \{1, 2\}$$

where $\vartheta_1 \sim Unif(0, 2\pi)$, $\vartheta_2 = \frac{\pi}{4}$ and $r \in [0, \infty)$. The point $\mathbf{x}_0^{(2)}$ is of the most-influential type (2.8) whereas point $\mathbf{x}_0^{(1)}$ represents a uniform random vector having the same magnitude as $\mathbf{x}_0^{(2)}$; observe that $r = \|\mathbf{x}_0^{(1)}\| = \|\mathbf{x}_0^{(2)}\|$ and $(\cos(\vartheta_1), \sin(\vartheta_1))^T$ has a uniform distribution on the unit circle. Besides the deflation-based FastICA estimators, we also include in our study JADE and DOGMA estimator employing Tyler's and Huber's M-estimator of scatter matrix with tuning constant $q = 0.9$ as the choice of scatter matrices. We set $\mathbf{A} = \mathbf{I}$ in the simulations. Note that the choice of $\mathbf{A}$ is immaterial as FastICA and the DOGMA estimators are *equivariant* (Publication [II]) $\hat{\mathbf{V}}$ (and thus ISR) does not depend on the used mixing matrix $\mathbf{A}$.

For different methods, the effect of the added points $\mathbf{x}_0^{(1)}$ (random outlier) and $\mathbf{x}_0^{(2)}$ (most influential outlier) on the mean ISR (calculated over 3000 trials) as the function of $r = \|\mathbf{x}_0^{(1)}\| = \|\mathbf{x}_0^{(2)}\|$ are depicted at Figure 2.6. In case of random outlier, the FastICA estimates and JADE reach the plain guess baseline but they do not go beyond it. As can be seen, most influential point in the data set can render the FastICA estimates and also JADE seriously defective: the ISR values are consistently above the critical plain guess baseline for large enough magnitude r. In fact, *pow3* tends to the pathological maximum ISR value 1, reaching $\mathrm{ISR}(\hat{\mathbf{V}}) = 0.978$ at $r = 35$. Note that the robust DOGMA estimator remains unaffected no matter how large the magnitude r is. The fact that also JADE is prone to most influential points (2.8) indicates that this feature originally observed analytically for FastICA in [37, 38] may be due to the usage of the covariance matrix (for whitening) that is common to both methods.

Thus if the sample contains an observation that is a mixture of sources that possess similar magnitude, then the performance of FastICA and JADE can be highly unreliable. Naturally, most influential points (2.8) occur with probability zero if sources sequences are random samples from a continuous distribution. However, in real-world applications, most influential points can occur frequently. For example, consider two gray-scale images of different objects (e.g. persons) in a similar background that are mixed linearly. Since the backgrounds of the images are similar, there exists many source samples that have pixel values of similar magnitude. Hence when these images are mixed, FastICA and JADE may perform very poorly due to occurrence of many such influential points in the mixture. This suggest that one should try to remove such influential points (2.8) prior to the analysis.

2.6 Discussion

Despite the increased interest on ICA during the past two decades, not much attention has been paid to the robustness of the proposed estimators. In this thesis, we have developed demixing matrix estimators that are robust. The robustness stems from the robust matrix-valued statistics used in their construction. The main limitation of the DOGMA family is that it is not able to separate sources with the same distribution. Hence, an important task for future work, is to try to find approaches to circumvent this problem.

Also the asymptotics of ICA estimators has been widely neglected in ICA studies. In Publication [III], we derived a compact CRB expression for demixing vector estimation. Based on the maximum likelihood (ML) theory, the derived inverse of the Fisher information matrix also equals the asymptotic covariance matrix of the maximum likelihood estimator (MLE) $\hat{\mathbf{w}}_{ml,k}$ of the kth demixing vector $\mathbf{w}_k$, which obtains the following form under mild assumptions on the source distributions (see Publication

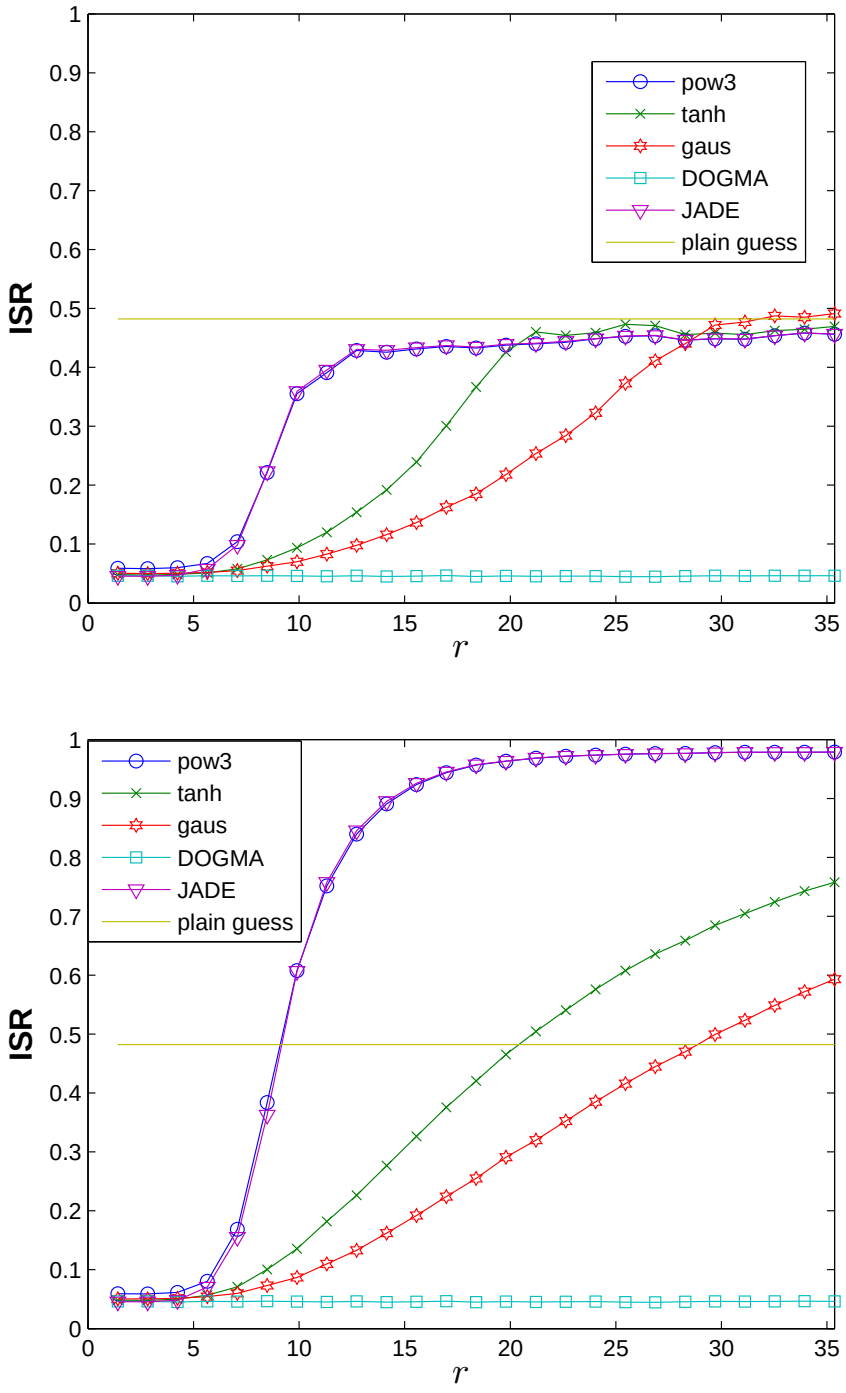


Figure 2.6: The effect of an added random point $\mathbf{x}_0^{(1)}$ (upper plot) and most influential point $\mathbf{x}_0^{(2)}$ (lower plot) on the mean ISR as the function of the magnitude $r = \|\mathbf{x}_0^{(1)}\| = \|\mathbf{x}_0^{(2)}\|$. Sample length $n = 2000$, number of trials 3000. In the presence of a most influential point, the non-robust FastICA and JADE can perform severely worse than a plain guess.

[III] for details)

$$\mathrm{ASV}(\hat{\mathbf{w}}_{ml,k}; F_{\mathbf{A}}) = \left(\frac{1}{\mathbb{E}[\varphi_k^2(s_k)s_k^2] - 1}\right)\mathbf{w}_k\mathbf{w}_k^T + \sum_{\substack{j=1 \\ j\neq k}}^{d}\left(\frac{\mathbb{E}[\varphi_j^2(s_j)]}{\mathbb{E}[\varphi_k^2(s_k)]\mathbb{E}[\varphi_j^2(s_j)] - 1}\right)\mathbf{w}_j\mathbf{w}_j^T,$$

where $\varphi_j(s) \triangleq -\frac{d}{ds}\log f_j(s)$ denotes the location score function for the jth source (and f_j denoting the p.d.f. of the jth source), $j = 1, \ldots, d$. Thus if an asymptotic covariance matrix of the demixing vector estimator is known, one can compute the asymptotic efficiency (e.g. as the ratios of the matrix trace) w.r.t. to the optimal MLE.

Chapter 3

Complex-valued signal processing

Complex-valued random signals play an increasingly important role in many diverse application areas such as biomedical sciences, physical sciences, communications, and related fields. In this section we briefly review some important tools, statistics, models and estimators that are useful for handling complex-valued random signals. The important problem of detecting circularity of complex random signals is also addressed.

3.1 Why complex-valued signal processing

Complex-valued random signals arise naturally in many application areas. For example, most practical modulation schemes (e.g. M-QAM, QPSK, 8-PSK) in communications are complex-valued and applications such as radar, sensor array processing [4], spectral analysis of time series [66] and magnetic resonance imaging [67, 68] lead to data that are inherently complex-valued. In some applications on the other hand, for example in statistical shape analysis [69], great simplifications are achieved by representing the observed 2-dimensional real-valued landmark data matrix as a complex vector and then conducting the statistical analysis in the complex domain. Functional magnetic resonance imaging (fMRI) data are originally acquired as complex-valued images while virtually all fMRI studies use only the magnitude of the data in the analysis and disregard the phase information. Recent studies [68, 70–72] have shown that fMRI analysis in the complex domain can offer several advantages. Complex weighted median filters has also been under active research; see e.g. [73] and references therein. The complex valued representation is also compact and simpler in notations and for algebraic manipulations, and convenient for calculations by computer. It is evident that the need of expertise in the analysis and statistical modelling and estimation of complex-valued multivariate data and phenomena are rapidly increasing.

Analysis in the complex domain presents a number of challenges since solid mathematical and statistical foundations, tools and algorithms for handling complex-valued signals are lacking, or, are simply too scattered in the literature. There appears to be a

need for concise, unified, and rigorous treatment of such topics. Several recent research papers have profoundly widened our knowledge and understanding of complex-valued random signals; see e.g. [16, 17, 74–83] and Publications [VIII,IX] to cite only a few. Recent forthcoming text-book [84] is also devoted to this topic.

Many methods are based on unnecessary simplifying assumptions that limits their usefulness, versatility and applicability in wider scenarios. Circular symmetry [77] of complex-valued signals is the most commonly made simplifying assumption in the statistical signal processing literature. Circular complex random variable possess the property that it is statistically uncorrelated with its complex-conjugate. In case the signals or noise are non-circular, we need to take the full 2nd-order statistics into account when deriving or applying signal processing algorithms. Consequently optimal estimation and detection techniques are different for circular and non-circular cases and recent research have elucidated that significant performance gains can be achieved by exploiting the circularity/non-circularity property of the complex-valued signals for example in designing wireless transceivers [85] or array processors such as beamformers, Direction-Of-Arrival algorithms [86, 87], blind source separation methods, etc. Also performance bounds can differ in circular and non-circular cases. Several authors have recently investigated the complex CRB theory; see Publication [IX] and [75, 88–92].

For example, widely linear processing [77, 79] can be advantageous for non-circular data. In complex-valued ICA and BSS, algorithms that explicitly exploit non-circularity statistics in their definition often give superior performance when the sources are non-circular; see [64, 81, 93–96] and Publications [I,II]. A virtue of complex-valued ICA is that it enables analysis of fMRI data in its native complex form [68].

3.2 Preliminaries

3.2.1 Complex field and functions

The set of complex numbers, denoted by $\mathbb{C}$, is the plane $\mathbb{R} \times \mathbb{R} = \mathbb{R}^2$ equipped with *complex addition* operator $+$ and *complex multiplication* operator $\cdot$ defined such that for all $z_1 = (x_1, y_1)$, $z_2 = (x_2, y_2) \in \mathbb{R} \times \mathbb{R} = \mathbb{C}$,

$$z_1 + z_2 = (x_1 + x_2, y_1 + y_2) \in \mathbb{C},$$
$$z_1 \cdot z_2 = (x_1 x_2 - y_1 y_2, x_1 y_2 + y_1 x_2) \in \mathbb{C},$$

making it the complex field $(\mathbb{C}, +, \cdot)$. For notational convenience, we write $z_1 z_2$ instead of $z_1 \cdot z_2$. We identify the set of real numbers $\mathbb{R}$ with the set $\mathbb{R} \times \{0\} \subset \mathbb{C}$ which forms a subfield of $\mathbb{C}$. Therefore we write $x = (x, 0)$ and in particular $0 = (0, 0)$ and $1 = (1, 0)$. The complex number $(0, 1)$, denoted by $\jmath$, is called the *imaginary unit* and it is the solution to the equation $z^2 = z \cdot z = -1$. With this notation every complex $z = (x, y)$

can represented in the form

$$z = x + \jmath y \tag{3.1}$$

since $z = (x, 0) + (0, 1)(y, 0)$. This will be the more commonly used notation for a complex number in this paper. The *complex conjugate* of $z = (x, y) = x + \jmath y \in \mathbb{C}$ is defined as $z^* \triangleq (x, -y) = x - \jmath y$. With this notation we can write the *real part* and the *imaginary part* of z as $\mathrm{Re}[z] \triangleq x = \frac{1}{2}(z + z^*)$ and $\mathrm{Im}[z] \triangleq y = \frac{\jmath}{2}(z^* - z)$ respectively. The *modulus* of $z = x + \jmath y$ is defined as the nonnegative real number $|z| = \sqrt{x^2 + y^2} = \sqrt{zz^*}$.

The *open disk* with center $c = a + \jmath b \in \mathbb{C}$ and radius $r > 0$ is defined as $B(c, r) \triangleq \{z \in \mathbb{C} : |z - c| < r\}$. Naturally, the open disk in $\mathbb{C}$ is equivalent to open 2-ball in $\mathbb{R}^2$ with center (a, b) and radius r. Throughout the paper $\mathcal{U}$ will stand for an *open set* in $\mathbb{C}$, *i.e.* for each $c \in \mathcal{U}$ there exists $r > 0$ such that $B(c, r) \subset \mathcal{U}$. A *function* f of the complex variable $z = x + \jmath y$ is a rule that assigns to each value z in $\mathcal{U}$ one and only one complex number $w = u + \jmath v \triangleq f(z)$. The real and imaginary part of the function $f(z)$ are real valued functions of real variables x and y, *i.e.* $u = u(x, y) \triangleq u(z) : \mathcal{U} \to \mathbb{R}$ and $v = v(x, y) \triangleq v(z) : \mathcal{U} \to \mathbb{R}$. Conversely, two such functions define a complex function $f(z) = u + \jmath v$ of $z = x + \jmath y$ over $\mathcal{U}$.

The *exponential function* is defined as the complex number $\exp(z) \triangleq \exp(x)\{\cos(y) + \jmath \sin(y)\}$, where $\exp(x)$ for real valued x denotes the usual exponential function. Any nonzero complex number has a polar representation, $z = |z| \exp(\jmath\theta)$, where $\theta = \arg(z) \in \mathbb{R}$ is called as the *argument* of z. The unique argument $\theta = \mathrm{Arg}(z)$ on the interval $-\pi \leq \theta < \pi$ is called as the *principal argument*. The *complex logarithm* of $z \neq 0$ is defined as the complex number $\log(z) \triangleq \log|z| + \jmath \mathrm{Arg}(z) + \jmath 2n\pi$ where n is an arbitrary integer. The particular value of the logarithm given by $\log|z| + \jmath \mathrm{Arg}(z)$ is called the *principal logarithm* and will be denoted by $\mathrm{Log}(z)$. With these definition of the complex logarithm and exponential one has the expected result that $\exp\{\log(z)\} = z$.

Since $az \in \mathbb{C}$ for all $a \in \mathbb{R}$ and $z \in \mathbb{C}$, the field $\mathbb{C}$ is also a vector space over the field $\mathbb{R}$. Therefore there are two kinds of linear mappings. A function $L : \mathbb{C} \to \mathbb{C}$ is $\mathbb{F}$-*linear* ($\mathbb{F} = \mathbb{C}$, or $\mathbb{R}$) if

$$L(az_1 + bz_2) = aL(z_1) + bL(z_2) \quad \forall\, z_1, z_2 \in \mathbb{C},\, a, b \in \mathbb{F}. \tag{3.2}$$

For example, the complex conjugation $z \mapsto z^*$ is $\mathbb{R}$-linear but not $\mathbb{C}$-linear. It is clear that function L is $\mathbb{C}$-linear if and only if $L(z) = \alpha z$ where $\alpha = L(1)$. The next theorem gives the explicit form of a $\mathbb{R}$-linear mapping.

Theorem 3.2.1 (Publication [VIII])**.** *Function $L : \mathbb{C} \to \mathbb{C}$ is $\mathbb{R}$-linear if and only if*

$$L(z) = \alpha z + \beta z^*, \tag{3.3}$$

where $\alpha = \dfrac{L(1) - \jmath L(\jmath)}{2}$ *and* $\beta = \dfrac{L(1) + \jmath L(\jmath)}{2}$.

Moreover, an $\mathbb{R}$-linear function L is invertible if and only if $|\alpha| \neq |\beta|$, and its inverse function $L^{-1} : \mathbb{C} \to \mathbb{C}$ is also an $\mathbb{R}$-linear function,

$$L^{-1}(z) = \alpha' z + \beta' z^*, \tag{3.4}$$

where $\alpha' = \dfrac{\alpha^*}{|\alpha|^2 - |\beta|^2}$ *and* $\beta' = \dfrac{-\beta}{|\alpha|^2 - |\beta|^2}$.

Proof. Clearly the function defined by eq. (3.3) is $\mathbb{R}$-linear. Therefore, we only need to show that if L is $\mathbb{R}$-linear then it may be written in the form (3.3). If L is $\mathbb{R}$-linear, then

$$L(z) = L(1 \cdot x + \jmath \cdot y) = L(1)x + L(\jmath)y$$

Substituting $(z + z^*)/2$ and $\jmath(z^* - z)/2$ in place of x and y yields the eq. (3.3). Denote $z' = \alpha z + \beta z^*$. Then observe that $L^{-1}(z') = z$, i.e. the function L^{-1} defined in (3.4) is the inverse of L which exists if $|\alpha| \neq |\beta|$. □

Corollary 1. *Function $L : \mathbb{C} \to \mathbb{C}$ is $\mathbb{C}$-linear if and only if L is $\mathbb{R}$-linear and $L(1) = -\jmath L(\jmath)$.*

We note that Theorem 3.2.1 can be generalized to multivariate mapping $\mathbb{C}^n \to \mathbb{C}^p$, and the form of $\mathbb{R}$-linear mappings remains the same: it is the sum of $\mathbb{C}$-linear mappings, the first one acting on the vector argument and the latter on its conjugate.

3.2.2 Complex derivatives

In this section we consider three different notion of derivatives of a complex function: directional derivative, complex partial derivatives, and complex derivative and their interrelations.

Definition 3.2.1. *The directional derivative of $f : \mathcal{U} \to \mathbb{C}$ at $c \in \mathcal{U}$ in the direction $t \in \mathbb{C}$ with $|t| = 1$, denoted by $D_{f,t}(t)$, is defined by the equation*

$$D_{f,c}(t) = \lim_{r \to 0} \frac{f(c + rt) - f(c)}{r}, \quad r \in \mathbb{R}$$

provided this limit exists.

Directional derivative can be viewed as the rate of change of $f(z)$ as z moves towards c *along the straight line* through c in the direction t. Directional derivative of $f = u + \jmath v$ can be related with the first partial derivatives of u and v. Namely, it is straightforward to verify that for a complex function $f = u + \jmath v : \mathcal{U} \to \mathbb{C}$ and $c \in \mathcal{U}$,

$$D_{f,c}(1) = \frac{\partial u}{\partial x}(c) + \jmath \frac{\partial v}{\partial x}(c) \triangleq \frac{\partial f}{\partial x}(c), \quad D_{f,c}(\jmath) = \frac{\partial u}{\partial y}(c) + \jmath \frac{\partial v}{\partial y}(c) \triangleq \frac{\partial f}{\partial y}(c), \tag{3.5}$$

provided that $D_{f,c}(1)$ and $D_{f,c}(\jmath)$ exist. Function f can also have complex partial derivatives (c.p.d.'s) defined next.

Definition 3.2.2. *Suppose that the complex function $f = u + \jmath v : \mathcal{U} \to \mathbb{C}$ is such that u and v possess first real partial derivatives at $c \in \mathcal{U}$. Then we define*

$$\frac{\partial f}{\partial z}(c) \triangleq \frac{1}{2}\left(\frac{\partial f}{\partial x}(c) - \jmath\frac{\partial f}{\partial y}(c)\right), \quad \frac{\partial f}{\partial z^*}(c) \triangleq \frac{1}{2}\left(\frac{\partial f}{\partial x}(c) + \jmath\frac{\partial f}{\partial y}(c)\right)$$

and call them complex partial derivatives of f w.r.t. z and z^* at c, respectively.

In [83, 97], the c.p.d.'s are called as the $\mathbb{R}$-derivative and the conjugate $\mathbb{R}$-derivative, respectively. The differential calculus based on these operators is known as Wirtinger calculus [96, 98], or, as we prefer, the $\mathbb{CR}$-calculus [97]. From (3.5) we observe that c.p.d.'s are related to directional derivatives as

$$\frac{\partial f}{\partial z}(c) = \frac{D_{f,c}(1) - \jmath D_{f,c}(\jmath)}{2}, \quad \frac{\partial f}{\partial z^*}(c) = \frac{D_{f,c}(1) + \jmath D_{f,c}(\jmath)}{2} \tag{3.6}$$

provided that $D_{f,c}(1)$ and $D_{f,c}(\jmath)$ exists.

If $u = \mathrm{Re}[f]$ and $v = \mathrm{Im}[f]$ possess first partial derivatives in some set $\mathcal{U}$, then $\partial f/\partial z$ and $\partial f/\partial z^*$ are complex functions from $\mathcal{U}$ to $\mathbb{C}$. Thus they themselves can have complex partial derivatives w.r.t z and z^* at $c \in \mathcal{U}$. There are called *higher-order c.p.d.'s*. For example, $\frac{\partial}{\partial z}(\frac{\partial f}{\partial z^*}) \triangleq \frac{\partial^2 f}{\partial z \partial z^*}$, is a 2nd-order c.p.d. (one among the four) and the total number of c.p.d.'s of order $k \geq 1$ which can be formed is 2^k.

The usefulness of the c.p.d.'s stems from an easily verifiable fact that they follow formally the same sum, product, and quotient rules as the ordinary partial derivatives. In particularly, $\frac{\partial}{\partial z}z = 1$, $\frac{\partial}{\partial z}z^* = 0$ and more generally, due to product rule and induction, one has the usual rules for polynomials

$$\frac{\partial}{\partial z}z^n z^{*m} = nz^{n-1}z^{*m} \quad \text{and} \quad \frac{\partial}{\partial z^*}z^n z^{*m} = mz^n z^{*m-1}.$$

However, it is easy to verify that the chain rule for the composition function $(f \circ g)(c) = f(g(c))$ is *not* of the regular form (c.f. Publication [VIII], [83, 97]), but of the form

$$\frac{\partial f \circ g}{\partial z}(c) = \frac{\partial f}{\partial z}\big(g(c)\big) \cdot \frac{\partial g}{\partial z}(c) \; + \frac{\partial f}{\partial z^*}\big(g(c)\big) \cdot \frac{\partial g^*}{\partial z}(c), \tag{3.7}$$

$$\frac{\partial f \circ g}{\partial z^*}(c) = \frac{\partial f}{\partial z}\big(g(c)\big) \cdot \frac{\partial g}{\partial z^*}(c) + \frac{\partial f}{\partial z^*}\big(g(c)\big) \cdot \frac{\partial g^*}{\partial z^*}(c). \tag{3.8}$$

Hence one should be cautious as simple and direct adaptation of the results derived for the real domain problems to complex domain can lead to wrong results and conclusions.

■ Example 3. Suppose that $(\partial f \circ g/\partial z)(c) = (\partial f/\partial z)\big(g(c)\big) \cdot (\partial g/\partial z)(c)$. If we choose $f(z) = z^*$ and $g(z) = z^*$, then $h(z) = (f \circ g)(z) = z$. Due to previous results, we have that $\partial h/\partial z = 1$, $\partial f/\partial z = 0$ and $\partial g/\partial z = 0$. If the regular form of the chain rule would hold, we would have $1 = 0 \cdot 0 = 0$, leading to contradiction. With the correct chain rule (3.7) one verifies that $1 = 0 \cdot 0 + 1 \cdot 1$. ■

As with usual partial derivatives, an important application of c.p.d.'s are related to optimization. It is known [74, 76] that both c.p.d.'s (and their multivariate extensions) vanish at stationary points of a function, but the (conjugate) c.p.d. $\partial/\partial z^*$ defines the direction of the maximum rate of change, *i.e.* it defines the complex gradient. In [83], c.p.d.'s were used in constructing the complex-valued Newton-Raphson iteration rule.

Cauchy-Riemann equations and the complex derivative, defined below, are the principal notions in the classical complex analysis [99–102].

Definition 3.2.3. *Let $f = u + \jmath v : \mathcal{U} \to \mathbb{C}$. Then the equations*

$$[\mathbf{a}]\quad \frac{\partial u}{\partial x} = \frac{\partial v}{\partial y},\quad \frac{\partial v}{\partial x} = -\frac{\partial u}{\partial y},\quad [\mathbf{b}]\quad \frac{\partial f}{\partial z^*} = 0,\quad [\mathbf{c}]\quad \frac{\partial f}{\partial z} = \frac{\partial f}{\partial x},\quad [\mathbf{d}]\quad \frac{\partial f}{\partial x} = -\jmath\frac{\partial f}{\partial y}$$

which are pairwise equivalent, are called Cauchy-Riemann (C-R) equations.

Definition 3.2.4. *Function $f : \mathcal{U} \to \mathbb{C}$ is said to have a derivative of f at $c \in \mathcal{U}$ if*

$$\lim_{h\to 0} \frac{f(c+h) - f(c)}{h} \in \mathbb{C}$$

exists; The value of the limit is denoted by $f'(c)$.

Definition is in appearance the same as that for real functions of one real variable. Despite of the apparent similarity, the complex case is remarkably different: h may approach c in any manner from any direction without affecting the value of the limit.

Next lemma shows that the derivative is related to directional derivatives, c.p.d.'s and C-R equations, but, particularly it demonstrates that possession of a derivative impose severe restrictions on the function.

Lemma 3.2.1. *If $f = u + \jmath v : \mathcal{U} \to \mathbb{C}$ possess a derivative at $c \in \mathcal{U}$, then f is continuous at c, $D_{f,c}(t)$ exists for all $t \in \mathbb{C}$ with $|t| = 1$, Cauchy-Riemann equations hold at c and*

$$f'(c) = t^* D_{f,c}(t) = \frac{\partial f}{\partial z}(c) = \frac{\partial f}{\partial x}(c).$$

Proof. A first consequence of the definition of derivative is that f is continuous at c [100, p. 38]. Suppose that h approaches c along the straight line through c in the direction $t \in \mathbb{C}$ with $|t| = 1$. Thus,

$$f'(c) = \lim_{h\to 0} \frac{f(c+h) - f(c)}{h} = \lim_{\substack{r\to 0 \\ r\in\mathbb{R}}} \frac{f(c+rt) - f(c)}{rt} = \frac{1}{t} \lim_{r\to 0} \frac{f(c+rt) - f(c)}{r},$$

which means that $f'(c) = t^* D_{f,c}(t)$ for all $t \in \mathbb{C}$ with $|t| = 1$ (observe that $t^{-1} = t^*$). This then means that $f'(c) = D_{f,c}(1) = -\jmath D_{f,c}(\jmath)$, which by (3.5) implies that Cauchy-Riemann equations hold at c and that $f'(c) = (\partial f/\partial x)(c) = (\partial f/\partial z)(c)$. □

3.2.3 Differentiability and Taylor's $\mathbb{R}$-theorem

Function $f : \mathcal{U} \to \mathbb{C}$ is said to be *holomorphic* or $\mathbb{C}$-*differentiable* on $\mathcal{U}$ if f has a derivative at every point $c \in \mathcal{U}$. Holomorphic functions form the hard core topic of the classical complex analysis [99–102]. If $f = u + \jmath v$ is holomorphic on $\mathcal{U}$, then it implies that f satisfies C-R equations in $\mathcal{U}$, f is infinitely $\mathbb{C}$-differentiable, u and v are harmonic functions (*i.e.* they satisfy Laplace's equations in $\mathcal{U}$), and for each $c \in \mathcal{U}$, the power series

$$\sum_{n=0}^{\infty} a_n (z-c)^n \tag{3.9}$$

with $a_n = f^{(n)}(c)/n!$ converges to $f(z)$ for all $z \in B(c,r) \subset \mathcal{U}$; see [99–101]. Conversely, if f is *analytic* in $\mathcal{U}$, *i.e.* if for every $c \in \mathcal{U}$ there is a power series of the form (3.9) that converges to $f(z)$ for all z in some neighbourhood of c, then f is holomorphic on $\mathcal{U}$. This is the reason why in many textbooks the terms holomorphic and analytic are used interchangeably. As an example, all polynomial functions in z with complex coefficients are holomorphic on the whole complex plane $\mathbb{C}$ and so is the exponential functions $\exp(z)$.

It is thus clear that holomorphic functions form a rather restricted class of complex functions. The concept of $\mathbb{C}$-differentiability being too stringent condition for many signal processing applications. For example, consider a real function of a complex variable, e.g. a cost function in optimization that arise naturally in a number of signal processing applications [74, 89], or, solving the maximum likelihood (ML-)estimate of a complex parameter where the problem is to maximize the (real-valued) likelihood function with respect to a complex parameter. However, a real function of a complex variable either has complex derivative at a point equal to zero, or else, the derivative does not exists; moreover, if the (real-valued) function is differentiable on $\mathcal{U}$, then the function is a constant. Also from the point of view of probability theory of complex random variables, it is the non-holomorphic functions that are of major importance. A less restrictive notion is $\mathbb{R}$-differentiability (Publication [VIII]).

Definition 3.2.5. *Function $f : \mathcal{U} \to \mathbb{C}$ is said to be $\mathbb{R}$-differentiable at $c \in \mathcal{U}$ if there exists an $\mathbb{R}$-linear function $L : \mathbb{C} \to \mathbb{C}$ such that*

$$f(c+h) - f(c) = L(h) + |h|\varepsilon(h) \quad \textit{and} \quad \lim_{|h|\to 0} \varepsilon(h) = 0. \tag{3.10}$$

The $\mathbb{R}$-linear function L is called the $\mathbb{R}$-differential of f at c and we denote it by $L_{f,c}$. The function f is said to be $\mathbb{R}$-differentiable if it is $\mathbb{R}$-differentiable at every point $c \in \mathcal{U}$.

Note that the differential $L_{f,c}$ is $\mathbb{C}$-linear if and only if f is $\mathbb{C}$-differentiable at c, and then $L_{f,c}(h) = f'(c)h$. Some important results are now collected in the following theorem illustrating that the c.p.d.'s play essential role.

Theorem 3.2.2. *Let $f = u + \jmath v : \mathcal{U} \to \mathbb{C}$ be $\mathbb{R}$-differentiable at $c \in \mathcal{U}$. Then*
[a] *$D_{f,c}(t)$ exists for all $t \in \mathbb{C}$ with $|t| = 1$ and $L_{f,c}(t) = D_{f,c}(t)$.*
[b] *$L_{f,c}$ is unique, first order partial derivatives of u and v exists at c and*

$$L_{f,c}(1) = \frac{\partial f}{\partial x}(c) \quad \text{and} \quad L_{f,c}(\jmath) = \frac{\partial f}{\partial y}(c).$$

[c] *for all $h \in \mathbb{C}$,*

$$L_{f,c}(h) = \frac{\partial f}{\partial z}(c) \cdot h + \frac{\partial f}{\partial z^*}(c) \cdot h^*.$$

Proof. [a] Due to (3.10) and since $L_{f,c}$ is $\mathbb{R}$-linear, we have

$$f(c + rt) - f(c) = L_{f,c}(rt) + |rt|\varepsilon(rt) = rL_{f,c}(t) + |r|\varepsilon(rt),$$

where $r \in \mathbb{R}$. Dividing by r and taking the limit when $r \to 0$ shows that $L_{f,c}(t) = D_{f,c}(t)$.

[b] If $L^\star_{f,c}$ is another $\mathbb{R}$-linear map for which eq. (3.10) holds, then $L^\star_{f,c}(1) = D_{f,c}(1) = L_{f,c}(1)$ and $L^\star_{f,c}(\jmath) = D_{f,c}(\jmath) = L_{f,c}(\jmath)$ by [a]-part of the theorem. This means that $L^\star_{f,c} = L_{f,c}$ due to Theorem 3.2.1, so $L_{f,c}$ is unique. The last statement follows since $L_{f,c}(1) = D_{f,c}(1)$ and $L_{f,c}(\jmath) = D_{f,c}(\jmath)$ and (3.5).

[c] Recalling eq. (3.6), the result follows immediately from [b]-part of the theorem and Theorem 3.2.1. □

■ Example 4. Consider the case $f(z) = |z|^2$. Then

$$f(c + h) - f(c) = (c + h)^*(c + h) - c^*c = c^*h + ch^* + |h|^2 = L(h) + |h|\varepsilon(h),$$

where $L(h) = c^*h + ch^*$ and $\varepsilon(h) = |h|$. First we observe that $\varepsilon(h) \to 0$ as $|h| \to 0$. Then we observe that L is $\mathbb{R}$-linear function since it is of the form (3.3). Furthermore, L is $\mathbb{C}$-linear if and only if $c = 0$. Thus, $f(z) = |z|^2$ is $\mathbb{R}$-differentiable at every point $c \in \mathbb{C}$ with $\mathbb{R}$-differential $L_{f,c}(h) = c^*h + ch^*$ and $\mathbb{C}$-differentiable *only* at the point $c = 0$ with the derivative $f'(0) = 0$. Furthermore, by Theorem 3.2.2[c], we observe the expected result that $(\partial f/\partial z)(c) = c^*$ and $(\partial f/\partial z^*)(c) = c$. ■

It is a celebrated result of complex analysis that $\mathbb{C}$-differentiable function possess a complex analogue of convergent Taylor series (3.9) and Taylor's formula. Let us call eq. (3.10) as the *first-order* Taylor's $\mathbb{R}$-formula. Functions with continuous c.p.d.'s of order m in $\mathcal{U}$ are denoted by $C^m(\mathcal{U})$.

Theorem 3.2.3 (Taylor's $\mathbb{R}$-theorem; Publication [VIII])**.** *Assume that $f = u + v \in C^{m+1}(\mathcal{U})$. Then,*

$$f(c + h) - f(c) = \sum_{p=1}^{m} \sum_{n=0}^{p} \frac{h^{*n} h^{p-n}}{n!(p-n)!} \frac{\partial^p f}{\partial z^{p-n} \partial z^{*n}}(c) + |h|^m \varepsilon(h) \tag{3.11}$$

and $\lim_{|h| \to 0} \varepsilon(h) = 0$.

An import special case occurs when $\frac{\partial}{\partial z^*} f \equiv 0$, *i.e.* f satisfies C-R equations in $\mathcal{U}$. This implies that f is holomorphic (and hence the power series (3.11) converges) and the Taylors series takes the usual form from the complex analysis: $f(c+h) - f(c) = \sum_{n=1}^{\infty} \frac{h^n}{n!} \frac{\partial^n f}{\partial z^n}(c)$.

3.3 The augmented signal model

Let us first recall the isomorphism between vector spaces $\mathbb{C}^d$ and $\mathbb{R}^{2d}$. Write

$$\bar{\mathbf{c}} \triangleq \begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix} \tag{3.12}$$

for the composite real-valued vector of $\mathbf{c} = \mathbf{a} + \jmath\mathbf{b} \in \mathbb{C}^d$. The mapping $\mathbf{c} \mapsto \bar{\mathbf{c}}$ is a group isomorphism between additive Abelian groups $\mathbb{C}^d$ and $\mathbb{R}^{2d}$.

The representation for complex random vectors exploited in the seminal works of [16, 17] is the so called *augmented model*, where a $2d$-variate complex-valued augmented vector

$$\hat{\mathbf{c}} \triangleq \begin{pmatrix} \mathbf{c} \\ \mathbf{c}^* \end{pmatrix}$$

is formed by stacking the complex vector and its complex conjugate $\mathbf{c}^*$. The two augmented models are related via invertible linear transform :

$$\bar{\mathbf{c}} = \mathbf{M}\hat{\mathbf{c}} \quad \Leftrightarrow \quad \hat{\mathbf{c}} = \mathbf{M}^{-1}\bar{\mathbf{c}} = 2\mathbf{M}^H\bar{\mathbf{c}} \tag{3.13}$$

where

$$\mathbf{M} \triangleq \frac{1}{2}\begin{pmatrix} \mathbf{I} & \mathbf{I} \\ -\jmath\mathbf{I} & \jmath\mathbf{I} \end{pmatrix}$$

with inverse $\mathbf{M}^{-1} = 2\mathbf{M}^H$. Above $(\cdot)^H$ denotes the Hermitian transpose, $(\mathbf{M})^H = (\mathbf{M}^*)^T$.

Let us next recall the following mapping from Publication [IX]. Observe that in [IX], the mapping was represented in a more general case that includes non-square matrices as well.

Definition 3.3.1 (Publication [IX])**.** *Define* $\langle\cdot\rangle_{\mathbb{C}} : \mathbb{R}^{2d\times 2d} \mapsto \mathbb{C}^{2d\times 2d}$ *as a mapping*

$$\langle\mathbf{G}\rangle_{\mathbb{C}} = 2\mathbf{M}^{-1}\mathbf{G}\mathbf{M} \tag{3.14}$$

that is,

$$\left\langle \begin{pmatrix} \mathrm{Re}[\mathbf{A}] & \mathrm{Re}[\mathbf{B}] \\ \mathrm{Im}[\mathbf{A}] & \mathrm{Im}[\mathbf{B}] \end{pmatrix} \right\rangle_{\mathbb{C}} = \begin{pmatrix} \mathbf{A} - \jmath\mathbf{B} & \mathbf{A} + \jmath\mathbf{B} \\ (\mathbf{A} + \jmath\mathbf{B})^* & (\mathbf{A} - \jmath\mathbf{B})^* \end{pmatrix}$$

for all $\mathbf{A} \in \mathbb{C}^{d\times d}$ *and* $\mathbf{B} \in \mathbb{C}^{d\times d}$.

Mapping $\langle\cdot\rangle_{\mathbb{C}}$ of $\mathbf{G} \in \mathbb{R}^{2d\times 2d}$ produces a complex $2d \times 2d$ matrix of the form

$$\langle \mathbf{G} \rangle_{\mathbb{C}} = \begin{pmatrix} \mathbf{C} & \mathbf{D} \\ \mathbf{D}^* & \mathbf{C}^* \end{pmatrix} \tag{3.15}$$

where $\mathbf{C}$ and $\mathbf{D}$ are complex $d \times d$ matrices. Hence we shall call matrix $\langle \mathbf{G} \rangle_{\mathbb{C}}$ as the *augmented matrix of* $\mathbf{C}$ *and* $\mathbf{D}$. Note that $\mathbf{G} = \frac{1}{2}\mathbf{M}\langle \mathbf{G} \rangle_{\mathbb{C}}\mathbf{M}^{-1}$.

Let PDH(d) and CS(d) denote the set of $d \times d$ positive definite Hermitian and complex symmetric matrices, respectively.

■ Example 5. We point out that the real quadratic form,

$$Q(\bar{\mathbf{z}}|\mathbf{\Gamma}) \triangleq \bar{\mathbf{z}}^T \bar{\mathbf{\Gamma}}^{-1} \bar{\mathbf{z}}, \tag{3.16}$$

where $\mathbf{\Gamma} \in \mathrm{PDS}(2d)$, possesses an equivalent complex representation

$$Q(\hat{\mathbf{z}}|\langle \mathbf{\Gamma} \rangle_{\mathbb{C}}) = \hat{\mathbf{z}}^H \langle \mathbf{\Gamma} \rangle_{\mathbb{C}}^{-1} \hat{\mathbf{z}}, \tag{3.17}$$

where $\langle \mathbf{\Gamma} \rangle_{\mathbb{C}} \in \mathrm{PDH}(2d)$ by Publication [IX]. This follows since

$$Q(\bar{\mathbf{z}}|\mathbf{\Gamma}) = \hat{\mathbf{z}}^H \mathbf{M}^H \mathbf{\Gamma}^{-1} \mathbf{M} \hat{\mathbf{z}} = \hat{\mathbf{z}}^H \tfrac{1}{4} \langle \mathbf{\Gamma}^{-1} \rangle_{\mathbb{C}} \hat{\mathbf{z}} = \hat{\mathbf{z}}^H \langle \mathbf{\Gamma} \rangle_{\mathbb{C}}^{-1} \hat{\mathbf{z}},$$

where we used (3.13) and the property that $\langle \mathbf{G}^{-1} \rangle_{\mathbb{C}} = 4\langle \mathbf{G} \rangle_{\mathbb{C}}^{-1}$ for $\mathbf{G}$ invertible; see Publication [IX]. ■

■ Example 6. Another useful property is that

$$\det(\mathbf{G}) = 2^{-2d} \det(\langle \mathbf{G} \rangle_{\mathbb{C}}). \tag{3.18}$$

which will come handy in the study of CES distributions. This property follows by observing that $\det(\mathbf{G}) = \det(\frac{1}{2}\mathbf{M}\langle \mathbf{G} \rangle_{\mathbb{C}}\mathbf{M}^{-1}) = 2^{-2d} \det(\mathbf{M}) \det(\mathbf{M})^{-1} \det(\langle \mathbf{G} \rangle_{\mathbb{C}})$. ■

3.4 Fundamentals of complex random vectors

3.4.1 Complex distribution

A complex r.v. $\mathbf{z} = \mathbf{x} + \jmath\mathbf{y} \in \mathbb{C}^d$ is comprised of a pair of real r.v.'s $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^d$. The distribution of $\mathbf{z}$ on $\mathbb{C}^d$ determines the joint real $2d$-variate distribution of $\mathbf{x}$ and $\mathbf{y}$ on $\mathbb{R}^{2d}$ and conversely due to isomorphism (3.12) between $\mathbb{C}^d$ and $\mathbb{R}^{2d}$. Hence the distribution of $\mathbf{z}$ is identified with the joint (real $2d$-variate) distribution of $\bar{\mathbf{z}}$,

$$F_{\mathbf{z}}(\mathbf{c}) \triangleq \Pr(\mathbf{x} \leq \mathbf{a}, \mathbf{y} \leq \mathbf{b}) \equiv \Pr(\bar{\mathbf{z}} \leq \bar{\mathbf{c}})$$

where $\mathbf{c} = \mathbf{a} + \jmath\mathbf{b} \in \mathbb{C}^d$. In a similar manner, the probability density function (p.d.f.) of $\mathbf{z} = \mathbf{x} + \jmath\mathbf{y}$ is identified with the joint p.d.f. $f(\bar{\mathbf{z}}) = f(\mathbf{x}, \mathbf{y})$ of $\mathbf{x}$ and $\mathbf{y}$. Hence $f(\mathbf{z})$ will be used as an equivalent alternative notation for $f(\bar{\mathbf{z}})$. It is worth pointing out

that in some applications (e.g. for optimization purposes [74, 76]) it is preferable to write the p.d.f. $f(\mathbf{z})$ in the form $f(\mathbf{z}, \mathbf{z}^*)$ that separates $\mathbf{z}$ and its conjugate $\mathbf{z}^*$ as if they were independent variates.

Now recall that the *mean* (or expectation) of a complex r.v. $\mathbf{z}$ is defined as $\mathbb{E}[\mathbf{z}] = \mathbb{E}[\mathbf{x}] + \jmath\mathbb{E}[\mathbf{y}]$. Recall that the expectation can be used to define important alternative characterization of the real r.v. $\bar{\mathbf{z}}$ via the concept of characteristic function (c.f.). The c.f. of the composite real r.v. $\bar{\mathbf{z}}$ is a function $\Phi_{\bar{\mathbf{z}}} : \mathbb{R}^{2d} \to \mathbb{C}$, defined as

$$\Phi_{\bar{\mathbf{z}}}(\bar{\mathbf{c}}) \triangleq \mathbb{E}\big[\exp\{\jmath(\bar{\mathbf{c}}^T\bar{\mathbf{z}})\}\big], \quad \bar{\mathbf{c}} \in \mathbb{R}^{2d}$$

which by utilizing complex notations takes the form

$$\Phi_{\mathbf{z}}(\mathbf{c}) = \mathbb{E}\big[\exp\{\jmath\mathrm{Re}(\mathbf{c}^H\mathbf{z})\}\big] = \mathbb{E}\big[\exp\big\{\tfrac{1}{2}(\mathbf{c}^H\mathbf{z} + \mathbf{c}^T\mathbf{z}^*)\big\}\big], \quad \mathbf{c} \in \mathbb{C}^d.$$

Relationship between moments and the characteristic function in the univariate case ($d = 1$) were established in Publication [VIII].

Characteristics of a complex r.v. can be described via symmetry properties of its distribution. The most commonly made symmetry assumption in the statistical signal processing literature is that of *circular symmetry*. See e.g. [77]. Circularity, or lack of it (non-circularity) is the fundamental concept differentiating complex signal analysis from the real case. Complex r.v. $\mathbf{z}$ is said to be *circular* or, to have a *circularly symmetric* distribution, if its distribution remains invariant under multiplication by any (complex) number on the unit complex sphere, *i.e.*

$$\mathbf{z} =_d e^{\jmath\theta}\mathbf{z}, \quad \forall\,\theta \in \mathbb{R},$$

where notation $=_d$ should be read "has same distribution as". A circular r.v. $\mathbf{z}$, in general, does not necessarily possess a density. However, if it does, then its p.d.f $f(\mathbf{z})$ satisfies

$$f(e^{\jmath\theta}\mathbf{z}) = f(\mathbf{z}) \quad \forall\,\theta \in \mathbb{R}.$$

In the univariate case ($d = 1$), this is equivalent to saying that the composite r.v. $(x, y)^T$ is spherically symmetric. The p.d.f $f(z) = f(x, y)$ is then a function of $|z|^2 = x^2 + y^2$ only, *i.e.* $f(z) = C{\cdot}g(|z|^2)$ for some non-negative function $g(\cdot)$ and normalizing constant C [103]. Hence the regions of constant contours are circles in the complex plane, thus justifying the name for this class of distributions. In the vector case, however, the term "circular" is a bit misleading since for $d \geq 2$, it does not imply that the regions of constant contours are spheres in complex Euclidean k-space. R.v. $\mathbf{z}$ is said to be *symmetric*, or to have a *symmetric distribution*, if $\mathbf{z} =_d -\mathbf{z}$. Naturally, circular symmetry implies symmetry.

3.4.2 Statistics of complex random vectors

Univariate case

Second-order moments. Important characteristics of a complex r.va. z can also be described via its moments. We recall that the *variance* of $z = x + \jmath y$,

$$\sigma^2(z) \triangleq \mathbb{E}[|z - \mathbb{E}[z]|^2] = \sigma^2(x) + \sigma^2(y)$$

does not carry information about the correlation between the real and the imaginary part of z, but this information can be retrieved from the *pseudo-variance* [104]

$$\tau(z) \triangleq \mathbb{E}[(z - \mathbb{E}[z])^2] = \sigma^2(x) - \sigma^2(y) + 2\jmath\mathbb{E}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])]$$

since the covariance between x and y can be obtained as $\mathrm{Cov}(x, y) = \frac{1}{2}\mathrm{Im}[\tau(z)]$. For simplicity of notation, we write τ and σ^2 if the r.va. z is understood from the context. Similar notation is adopted for other statistics defined below. The ratio of pseudo-variance and the variance,

$$\varrho(z) \triangleq \frac{\tau}{\sigma^2}$$

is called the *circularity quotient* of z; see Publication [IX]. If z is circular, then $\tau = \varrho = 0$. Hence a r.va. z with vanishing pseudo-variance is said to be *2nd-order circular*. Naturally, 2nd-order circularity does not imply that the distribution of the r.va. is circular.

A degree of circularity. The modulus of the circularity quotient, $|\varrho|$, is called as the *circularity coefficient* [81] of z and $\mathrm{Arg}(\varrho)$ as the *circularity angle* (Publication [VI]). Circularity coefficient measures the "degree of circularity" as it equals the squared eccentricity of the ellipse defined by the real covariance matrix of $\bar{\mathbf{z}} = (x, y)^T$; see Publication [VI] for details. Hence its maximum and minimum value are

$$|\varrho| = \begin{cases} 0, \text{ iff } x \text{ and } y \text{ are uncorrelated with equal variances} \\ 1, \text{ iff } x \text{ or } y \text{ is a constant, or } x \text{ is a linear function of } y. \end{cases}$$

Note that $|\varrho| = 1$ if z is purely real-valued such as BPSK modulated communication signal, or, if the signal lie on a line in the scatter plot (also called constellation or I/Q diagram) as is the case for BPSK, ASK, AM, or PAM-modulated communications signals. Asymptotic distribution of the MLE of the circularity coefficient was recently studied in [105] and [106].

Higher-order moments. A r.va. z has $p+1$ *pth-order moments*, namely $\alpha_{0;p}$, $\alpha_{1;p-1}$, $\alpha_{2;p-2}$, $\ldots$, $\alpha_{p;0}$ where

$$\alpha_{n;m}(z) \triangleq \mathbb{E}[z^n z^{*m}],$$

for $m, n \in \mathbb{N}_0 = \{0, 1, 2, \ldots\}$. Again for simplicity of notation we often write $\alpha_{n;m}$ if the r.va. z is understood from the context. Similarly for related quantities. Note that

symmetric moments are *redundant* in the sense that $\alpha_{m;n} = \alpha^*_{n;m}$. In an analogous fashion one can also define pth-order central moments; see Publication [VIII]. The relationship between moments and the c.f. $\Phi_z(c)$ of a complex r.va. z was established in Publication [VIII], namely,

$$\alpha_{n;m} = \left(\frac{2}{j}\right)^{m+n} \frac{\partial^{m+n}\Phi_z}{\partial c^m \partial c^{*n}}(0). \tag{3.19}$$

Equation (3.19) together with the Taylors $\mathbb{R}$-series (3.11) at zero then gives an expansion for the characteristic function of a r.va. z (Publication [VIII]),

$$\Phi_z(c) = 1 + \sum_{m=1}^{p} \left(\frac{j}{2}\right)^m \sum_{n=0}^{m} \frac{c^{*n} c^{m-n}}{n!(m-n)!}\, \alpha_{n;m-n} + o(|c|^p)$$

as $c \to 0$, provided that z has finite pth-order moment.

Kurtosis. There can be several different paths to generalize the notion of kurtosis for a complex r.va. z. Normalized 4th-order moment of a complex r.va. z can be defined as

$$\gamma(z) \triangleq \frac{\mathbb{E}[|z - \mathbb{E}[z]|^4]}{(\sigma^2)^2}. \tag{3.20}$$

Then the real-valued measure

$$\mathrm{kurt}(z) \triangleq \gamma - |\varrho|^2 - 2$$

is the most commonly used generalization of the kurtosis for a complex r.va. z (e.g. [64, 96, 107]). In [83] it was pointed out (based on complex 4th-order cumulants) that there exists in fact three natural measures of complex kurtosis. Note that if z is purely real r.va. (*i.e.* $y = \mathrm{Im}[z] = 0$ with probability one), then $\varrho(z) = 1$ and $\mathrm{kurt}(z)$ and $\gamma(z)$ coincide with the definition given in (2.2) for a real r.va. A complex Gaussian r.va. has kurtosis $\mathrm{kurt}(z) = 0$. It is not clear in the literature what the complex kurtosis really measures. In Publication [VII] some light was shed on this question by deriving a connection between the complex kurtosis of z and the (real) kurtosis of its real and imaginary part within the wide class of CES distributions. Namely, it was shown that if z has a CES distribution (e.g. complex Gaussian distribution), then $\mathrm{kurt}(z) = \frac{1}{3}(2 + |\varrho|^2)\mathrm{kurt}(x)$, where $\mathrm{kurt}(x) = \mathrm{kurt}(y)$ is the common kurtosis of the real and imaginary part of a r.va. z possessing a CES distribution. Since the complex kurtosis is simply a scaled version (the scaling factor obtaining values on the interval $[\frac{2}{3}, 1]$) of the real kurtosis, the usual "peakedness combined with heavy-tailedness" interpretation of the kurtosis applies for the complex kurtosis in this case.

Multivariate case

For simplicity of presentation, let us assume that the complex r.v. $\mathbf{z} \in \mathbb{C}^d$ has mean zero, *i.e.* $\mathbb{E}[\mathbf{z}] = \mathbf{0}$. R.v. $\mathbf{z}$ is further assumed to be non-degenerate in any subspace of $\mathbb{C}^d$.

Second-order moments A complete second-order description of complex r.v. $\mathbf{z}$ is given by its *covariance matrix*

$$\begin{aligned}\boldsymbol{\mathcal{C}}(\mathbf{z}) &\triangleq \mathbb{E}[\mathbf{z}\mathbf{z}^H] \\ &= \mathbb{E}[\mathbf{x}\mathbf{x}^T] + \mathbb{E}[\mathbf{y}\mathbf{y}^T] + \jmath(\mathbb{E}[\mathbf{y}\mathbf{x}^T] - \mathbb{E}[\mathbf{x}\mathbf{y}^T]) \in \mathrm{PDH}(d)\end{aligned}$$

and the *pseudo-covariance matrix* [104]

$$\begin{aligned}\boldsymbol{\mathcal{P}}(\mathbf{z}) &\triangleq \mathbb{E}[\mathbf{z}\mathbf{z}^T] \\ &= \mathbb{E}[\mathbf{x}\mathbf{x}^T] - \mathbb{E}[\mathbf{y}\mathbf{y}^T] + \jmath(\mathbb{E}[\mathbf{x}\mathbf{y}^T] + \mathbb{E}[\mathbf{y}\mathbf{x}^T]) \in \mathrm{CS}(d).\end{aligned}$$

Pseudo-covariance matrix is also called relation matrix in [77] or complementary covariance matrix in [80]. For simplicity of notation, we write $\boldsymbol{\mathcal{C}}$ and $\boldsymbol{\mathcal{P}}$ if the r.v. $\mathbf{z}$ is understood from the context. Similar notation is adopted for other quantities. R.v. $\mathbf{z}$ is said to be 2*nd-order circular* [77] or *proper* [104] if $\boldsymbol{\mathcal{P}} = \mathbf{0}$, or equivalently, if

$$\mathbb{E}[\mathbf{x}\mathbf{x}^T] = \mathbb{E}[\mathbf{y}\mathbf{y}^T] \;\text{ and }\; \mathbb{E}[\mathbf{x}\mathbf{y}^T] = -\mathbb{E}[\mathbf{y}\mathbf{x}^T]. \tag{3.21}$$

It is well-known (e.g. [17], Publication [IX]) that

$$\langle \boldsymbol{\mathcal{C}}(\bar{\mathbf{z}}) \rangle_{\mathbb{C}} = \begin{pmatrix} \boldsymbol{\mathcal{C}}(\mathbf{z}) & \boldsymbol{\mathcal{P}}(\mathbf{z}) \\ \boldsymbol{\mathcal{P}}(\mathbf{z})^* & \boldsymbol{\mathcal{C}}(\mathbf{z})^* \end{pmatrix} = \boldsymbol{\mathcal{C}}(\widehat{\mathbf{z}}), \tag{3.22}$$

i.e. operator $\langle\cdot\rangle_{\mathbb{C}}$ maps the covariance matrix of the composite real r.v. $\bar{\mathbf{z}}$ to the covariance matrix of the augmented r.v. $\widehat{\mathbf{z}}$. As we shall see later in Section 3.3 the assumption (3.21) on the covariance structure of the real part $\mathbf{x}$ and imaginary part $\mathbf{y}$ of $\mathbf{z}$ is crucial in writing the p.d.f. $f(\bar{\mathbf{z}})$ of the multivariate normal distribution using the complex notation so that it would resemble closely the real case; see [89, 108, 109].

Circularity matrix and circularity coefficients. There can be several different ways to extent the concept of circularity quotient ϱ to the vector case. Since $\varrho = [\sigma^2]^{-1}\tau$, one possible extension is

$$\boldsymbol{\varrho}(\mathbf{z}) \triangleq \boldsymbol{\mathcal{C}}^{-1}\boldsymbol{\mathcal{P}}, \tag{3.23}$$

referred to as the *circularity matrix* of $\mathbf{z}$. Furthermore, since the circularity coefficient is the modulus $|\varrho| = \sqrt{\varrho\varrho^*}$, one possible way to extent this concept to the vector case, is to call the square-roots of the eigenvalues of the matrix $\boldsymbol{\varrho}\boldsymbol{\varrho}^*$ as the circularity coefficients of $\mathbf{z}$. The eigenvalues of $\boldsymbol{\varrho}\boldsymbol{\varrho}^*$ are real-valued and take values on the interval $[0, 1]$ (See Publication [I, Theorem 2]). Hence, also in this sense, the square-roots of the eigenvalues are valid extensions of the circularity coefficient $|\varrho| \in [0, 1]$ to the multivariate case.

Strong Uncorrelating Transform. It is easy to show that circularity coefficients can also be calculated as the singular values of the symmetric matrix $\mathbf{K}(\mathbf{z}) \triangleq \boldsymbol{\mathcal{B}}\boldsymbol{\mathcal{P}}(\mathbf{z})\boldsymbol{\mathcal{B}}^T$, called as the *coherence matrix* [110], where $\boldsymbol{\mathcal{B}}$ is a whitening matrix of r.v. $\mathbf{z}$, *i.e.*

$\boldsymbol{\mathcal{C}}(\mathbf{z})^{-1} = \boldsymbol{\mathcal{B}}^H\boldsymbol{\mathcal{B}}$. This means that there exists a unitary matrix $\mathbf{U}$ such that symmetric matrix $\mathbf{K}(\mathbf{z})$ has a special form of SVD, called *Takagi factorization* [111], such that $\mathbf{K}(\mathbf{z}) = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$, where $\boldsymbol{\Lambda}$ is a diagonal matrix consisting of circularity coefficients. Thus, if we now define matrix $\mathbf{W} \in \mathbb{C}^{d\times d}$ as $\mathbf{W} = \boldsymbol{\mathcal{B}}^H\mathbf{U}$, where $\boldsymbol{\mathcal{B}}$ and $\mathbf{U}$ are defined as above, then it is easy to verify that the transformed data $\mathbf{s} = \mathbf{W}^H\mathbf{z}$ has strongly-uncorrelated components, *i.e.* $\boldsymbol{\mathcal{C}}(\mathbf{s}) = \mathbf{I}$ and $\boldsymbol{\mathcal{P}}(\mathbf{s}) = \boldsymbol{\Lambda}$. Hence the matrix $\mathbf{W}$ is called the *strong-uncorrelating transform* (SUT) [13, 81]. A more general concept, called the generalized uncorrelating transform (GUT), is obtained by utilizing any scatter matrix and pseudo-scatter matrix in place of the covariance matrix and pseudo-covarariance matrix above; see Publication [I] for details. SUT and GUT have found applications for example in complex-valued ICA.

Information and pseudo-information matrices. Let $f(\mathbf{z}|\boldsymbol{\theta})$ denote the p.d.f. of the r.v. $\mathbf{z} \in \mathbb{C}^d$ depending on the unknown complex parameter $\boldsymbol{\theta} \in \mathbb{C}^k$. Central to complex CRB theory are the *information matrix* and the *pseudo-information matrix*, defined as (Publication [IX]):

$$\boldsymbol{\mathcal{I}}_{\boldsymbol{\theta}} \triangleq \mathbb{E}[\nabla_{\boldsymbol{\theta}^*} \ln f(\mathbf{z};\boldsymbol{\theta})\{\nabla_{\boldsymbol{\theta}^*} \ln f(\mathbf{z};\boldsymbol{\theta})\}^H],$$
$$\boldsymbol{\mathcal{P}}_{\boldsymbol{\theta}} \triangleq \mathbb{E}[\nabla_{\boldsymbol{\theta}^*} \ln f(\mathbf{z};\boldsymbol{\theta})\{\nabla_{\boldsymbol{\theta}^*} \ln f(\mathbf{z};\boldsymbol{\theta})\}^T]$$

where the *complex gradient* [74] is defined as $\nabla_{\boldsymbol{\theta}^*} = (\partial/\partial\boldsymbol{\theta}^*)^T = (\partial/\partial\theta_1^*, \ldots, \partial/\partial\theta_k^*)^T$. Only if the pseudo-information matrix vanishes ($\boldsymbol{\mathcal{P}}_{\boldsymbol{\theta}} = \mathbf{0}$), then $\boldsymbol{\mathcal{C}}(\mathbf{t}) \geq \boldsymbol{\mathcal{I}}_{\boldsymbol{\theta}}^{-1}$ gives a CRB for an unbiased estimator $\mathbf{t}$ of $\boldsymbol{\theta}$, otherwise the bound depends on pseudo-information matrix also; See Publication [IX] for details. Above, notation $\mathbf{C} \geq \mathbf{D}$ means that the matrix $\mathbf{C} - \mathbf{D}$ is positive semidefinite.

3.5 A review of CES distributions

A complex r.v. $\mathbf{z} = \mathbf{x} + \jmath\mathbf{y}$ of $\mathbb{C}^d$ has d-variate (centered) *circular CN distribution* if $\bar{\mathbf{z}} = (\mathbf{x}^T, \mathbf{y}^T)^T$ has $2d$-variate real normal distribution with mean zero and $2d \times 2d$ real covariance matrix $\boldsymbol{\mathcal{C}}(\bar{z})$ of special form (3.21), *i.e.* $\boldsymbol{\mathcal{P}} = \mathbf{0}$. Since the introduction of the circular CN distribution in [108, 109], the assumption (3.21) seem to be commonly thought as essential - although it was based on application specific reasoning - in writing the normal p.d.f. into representative complex form with natural and interpretable complex-valued parameters. In fact, the prefix "circular" is often dropped when referring to circular CN distribution as it has due time become the commonly excepted complex normal distribution. In the seminal works [16, 17] an intuitive complex-valued expression for normal density was derived without the unnecessary 2nd-order circularity assumption (3.21). The essential key result used in the derivation was the complex augmented representation (3.22) of the real covariance matrix $\boldsymbol{\mathcal{C}}(\bar{\mathbf{z}})$.

A natural extension of the circular CN distribution is obtained by allowing $\mathbf{x}$ and $\mathbf{y}$ to possess a $2n$-variate real elliptically symmetric distribution (RES) with the re-

striction as in (3.21) on the scatter parameter. This class of distributions are called circular complex elliptically symmetric (CES) distributions, the properties of which are studied in [15, 112]. The extension of the RES distribution for the non-circular case was proposed and studied in Publication [X].

3.5.1 Complex normal distribution

For simplicity of presentations assume that the complex r.v. $\mathbf{z}$ has mean zero. A complex r.v. $\mathbf{z}$ is said to have a *CN distribution* if $\bar{\mathbf{z}}$ has $2d$-variate real normal distribution. By (3.22) and since the normal distribution is uniquely parametrized by the covariance matrix $\bar{\boldsymbol{\mathcal{C}}} \equiv \boldsymbol{\mathcal{C}}(\bar{\mathbf{z}})$, we express this by notation $\mathbf{z} \sim \mathrm{CN}_d(\boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{P}})$. The case of circular CN distribution (*i.e.* $\boldsymbol{\mathcal{P}} = \mathbf{0}$) is then denoted by $\mathrm{CN}_d(\boldsymbol{\mathcal{C}})$ for short.

The p.d.f. of the $2d$-variate real normal distribution

$$f_{CN}(\bar{\mathbf{z}}|\bar{\boldsymbol{\mathcal{C}}}) = (2\pi)^{-d} \det(\bar{\boldsymbol{\mathcal{C}}})^{-1/2} \exp(-\tfrac{1}{2}\, Q(\bar{\mathbf{z}}|\bar{\boldsymbol{\mathcal{C}}})\,)$$

can be written with complex notations using (3.17) and (3.18) and the augmented vector $\bar{\mathbf{z}}$ as

$$f_{CN}(\hat{\mathbf{z}}|\hat{\boldsymbol{\mathcal{C}}}) = \pi^{-d} \det(\hat{\boldsymbol{\mathcal{C}}})^{-1/2} \exp(-\tfrac{1}{2}\, Q(\hat{\mathbf{z}}|\hat{\boldsymbol{\mathcal{C}}})\,)$$

where $\hat{\boldsymbol{\mathcal{C}}} \equiv \langle \bar{\boldsymbol{\mathcal{C}}} \rangle_{\mathbb{C}}$ is the short hand notation for the covariance matrix (3.22) of $\hat{\mathbf{z}}$, *i.e.* the augmented covariance matrix. In the case of circular CN distribution, $\boldsymbol{\mathcal{P}} = \mathbf{0}$, we have that

$$\hat{\boldsymbol{\mathcal{C}}} = \begin{pmatrix} \boldsymbol{\mathcal{C}} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\mathcal{C}}^* \end{pmatrix},$$

which yields $Q(\hat{\mathbf{z}}|\hat{\boldsymbol{\mathcal{C}}}) = 2\mathbf{z}^H \boldsymbol{\mathcal{C}}^{-1}\mathbf{z}$ and $\det(\hat{\boldsymbol{\mathcal{C}}}) = \det(\boldsymbol{\mathcal{C}})^2$. Hence the p.d.f. of the circular CN distribution can be written in the form

$$f_{CN}(\mathbf{z}|\boldsymbol{\mathcal{C}}) = \pi^{-d} \det(\boldsymbol{\mathcal{C}})^{-1} \exp(-\mathbf{z}^H \boldsymbol{\mathcal{C}}^{-1}\mathbf{z}) \tag{3.24}$$

which closely resembles the classical real normal distribution.

3.5.2 Definition

Complex r.v. $\mathbf{z} = \mathbf{x} + \jmath\mathbf{y} \in \mathbb{C}^d$ has (centered) CES distribution if $\bar{\mathbf{z}}$ has (centered, or, symmetric about zero) $2d$-variate RES distribution, *i.e.* if its density function[1] has the form

$$f_{CE}(\bar{\mathbf{z}}|\boldsymbol{\Gamma}) = C \det(\boldsymbol{\Gamma})^{-1/2} g(\, Q(\bar{\mathbf{z}}|\boldsymbol{\Gamma})\,) \tag{3.25}$$

where $g(\cdot)$ is a non-negative function, called the *density generator*, $\boldsymbol{\Gamma} \in \mathrm{PDS}(2d)$ is the scatter parameter, and $Q(\bar{\mathbf{z}}|\boldsymbol{\Gamma})$ is the quadratic form (3.16). Above C is a normalizing

[1] RES distributions can also be defined more generally via the characteristic function (thus avoiding the assumption of the existence of a density function) [103]

constant that could be absorbed into function g, but with notation g can be independent of the dimension d. Write $\mathcal{G}^\ell \triangleq \{g : [0,\infty) \to [0,\infty) | \int_0^\infty t^{d+\ell-1} g(t) < \infty\}$. Then, any nonnegative function $g \in \mathcal{G}^0$ is a valid density generator of a d-variate CES distribution and $g \in \mathcal{G}^\ell$ indicates that the moments of order 2ℓ exists. If the covariance matrix exists ($g \in \mathcal{G}^1$) then $\mathbf{\Gamma}$ is equal up to multiplicative real positive scalar to the covariance matrix $\bar{\boldsymbol{\mathcal{C}}}$ of $\bar{\mathbf{z}}$. We note that CN distribution is obtained with $g(t) = \exp(-\frac{1}{2}t)$ (yielding $C = (2\pi)^{-d}$), the scatter parameter $\mathbf{\Gamma}$ in this case being exactly equal to the covariance matrix $\bar{\boldsymbol{\mathcal{C}}}$.

Similarly, as for the CN distribution, the p.d.f. can be written into natural complex form. First we note that the *augmented scatter*,

$$\hat{\mathbf{\Gamma}} \equiv \langle \mathbf{\Gamma} \rangle_{\mathbb{C}} = \begin{pmatrix} \mathbf{\Sigma} & \mathbf{\Omega} \\ \mathbf{\Omega}^* & \mathbf{\Sigma}^* \end{pmatrix},$$

is by construction a complex positive definite Hermitian $2d \times 2d$ matrix where the parameters $\mathbf{\Sigma} \in \mathrm{PDH}(d)$ and $\mathbf{\Omega} \in \mathrm{CS}(d)$ are called as the (complex) *scatter matrix* and *pseudo-scatter matrix* (Publication [IX]), respectively. Proceeding as earlier the p.d.f. (3.25) can be expressed by complex notations utilizing the augmented model as

$$f_{CE}(\hat{\mathbf{z}}|\hat{\mathbf{\Gamma}}) = C \det(\hat{\mathbf{\Gamma}})^{-1/2} g(\, Q(\hat{\mathbf{z}}|\hat{\mathbf{\Gamma}})\,). \tag{3.26}$$

Herein, for notation convenience, we have absorbed the constant 2^d resulting from the determinant relation (3.18) into the normalizing constant C. Hence, we shall write $\mathbf{z} \sim \mathrm{CE}_d(\mathbf{\Sigma}, \mathbf{\Omega}, g)$.

Observe the following indeterminacy: $\mathbf{\Gamma}$ and $g(\cdot)$ do not *uniquely* identify the $2d$-variate RES distribution without additional restriction on $g(\cdot)$ or on the scale of $\mathbf{\Gamma}$. Indeed, by writing $\mathbf{\Gamma}_0 = (1/c)\mathbf{\Gamma}$ and $g_0(t) = c^{-d} g(t/c)$ for any $c > 0$, the density (3.25) can be written in the form $C \det(\mathbf{\Gamma}_0)^{-1/2} g_0(\, Q(\bar{\mathbf{z}}|\mathbf{\Gamma}_0)\,)$. This ambiguity is easily avoided by restricting the function g in a suitable way, or, by restricting the scale of the parameter $\mathbf{\Gamma}$, e.g. that its matrix trace is equal to unity. However, if $g \in \mathcal{G}^1$, it is conventional to restrict g by requiring that

$$C \cdot \int_0^\infty t^d g(t) dt = \frac{2 \cdot \Gamma(d+1)}{\pi^d} \tag{3.27}$$

in which case $\mathbf{\Gamma}$ is equal to the covariance matrix $\bar{\boldsymbol{\mathcal{C}}}$, and consequently, $\hat{\mathbf{\Gamma}}$ is equal to the augmented covariance matrix $\hat{\boldsymbol{\mathcal{C}}}$ (so $\mathbf{\Sigma} = \boldsymbol{\mathcal{C}}$ and $\mathbf{\Omega} = \boldsymbol{\mathcal{P}}$). Therefore, if $g \in \mathcal{G}^1$ satisfies (3.27) (as is the case for CN distribution), we can write $\mathbf{z} \sim \mathrm{CE}_d(\boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{P}}, g)$. There are many widely-used CES distributions, however, which do not have finite 2nd-order moments, e.g. the multivariate complex Cauchy distribution.

3.5.3 Circular case

CES distribution with vanishing pseudo-scatter matrix, $\mathbf{\Omega} = \mathbf{0}$, is called circular CES distribution and denoted $\mathrm{CE}_d(\mathbf{\Sigma}, g)$ for short. In this case, the p.d.f. (3.26) becomes

$C\det(\boldsymbol{\Sigma})^{-1}g(2\mathbf{z}^H\boldsymbol{\Sigma}^{-1}\mathbf{z})$. Hence, if we define $g_0(t) = g(2t)$, we can write the p.d.f. (3.26) simply as

$$f_{CE}(\mathbf{z}|\boldsymbol{\Sigma}) = C \cdot \det(\boldsymbol{\Sigma})^{-1} g_0(\mathbf{z}^H\boldsymbol{\Sigma}^{-1}\mathbf{z}).$$

With the above notation, the circular CN distribution (3.24) for example, is obtained with $g_0(t) = \exp(-t)$ (and $C = \pi^{-d}$). For notational convenience, we now drop the subscript, and denote $g_0(t)$ simply by $g(t)$. This notation, however, is useful only in the case of circular CES distributions.

Recall that $\boldsymbol{\Sigma}$ is proportional to the complex covariance matrix $\boldsymbol{\mathcal{C}}(\mathbf{z})$ provided it exists. Hence MLE's of the scatter matrix $\boldsymbol{\Sigma} \in \mathrm{PDH}(d)$ can provide robust estimators of the complex covariance matrix $\boldsymbol{\mathcal{C}}(\mathbf{z})$. Let $\mathbf{z}_1, \ldots, \mathbf{z}_n$ be an i.i.d. sample ($n > d$) from a circular CES distribution $\mathrm{CE}_d(\boldsymbol{\Sigma}, g)$. The MLE of $\boldsymbol{\Sigma}$ is found by minimizing the negative of the log-likelihood function,

$$\begin{aligned} L_n(\boldsymbol{\Sigma}) &\triangleq -\prod_{i=1}^{n} \log f_{CE}(\mathbf{z}_i|\boldsymbol{\Sigma}) \\ &= n \log|\boldsymbol{\Sigma}| - \sum_{i=1}^{n} \log g(\mathbf{z}_i^H\boldsymbol{\Sigma}^{-1}\mathbf{z}_i), \end{aligned}$$

where we have omitted the constant term $\log(C)$ since it does not depend on the unknown parameter $\boldsymbol{\Sigma}$. By differentiating $L_n(\boldsymbol{\Sigma})$ with respect to $\boldsymbol{\Sigma}$ by using complex matrix differentiation rules [74] and equating to zero shows that the MLE is a solution of the estimating equation

$$\boldsymbol{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \psi_{ML}(\mathbf{z}_i^H\boldsymbol{\Sigma}^{-1}\mathbf{z}_i)\mathbf{z}_i\mathbf{z}_i^H, \tag{3.28}$$

where

$$\psi_{ML}(t) \triangleq -\frac{g'(t)}{g(t)} \tag{3.29}$$

is a weight function that depends on the density generator $g(\cdot)$ of the underlying circular CES distribution and $g'(t) = \frac{d}{dt}g(t)$ denotes the derivative of g. MLE $\hat{\boldsymbol{\Sigma}}$ solves the estimating equation (3.28) and thus can be interpreted as a weighted covariance matrix. Note, however, that equation (3.28) is implicit as the weights on the right hand side depends on $\boldsymbol{\Sigma}$. In general, the obtained MLE is robust if the corresponding weight function $\psi_{ML}(\cdot)$ descends to zero. This is needed so that small weights are given to observations $\mathbf{z}_i$ that are highly *outlying* in terms of the distance $\mathbf{z}_i^H\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{z}_i$.

■ Example 7. In the case of circular CN distribution, $g(t) = \exp(-t)$, which yields $\psi_{ML} \equiv 1$. This shows the well-known result that the *sample covariance matrix* (SCM) $\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_i\mathbf{z}_i^H$ is the MLE of the parameter $\boldsymbol{\Sigma}$ ($= \boldsymbol{\mathcal{C}}$ in this case). ■

■ Example 8. *d*-variate *circular t-distribution with ν degrees of freedom* (d.f), denoted $\mathbf{z} \sim \mathrm{CT}_{d,\nu}(\boldsymbol{\Sigma})$, is obtained with $g(t) = (1 + 2t/\nu)^{-(2d+\nu)/2}$. The value $\nu = 1$

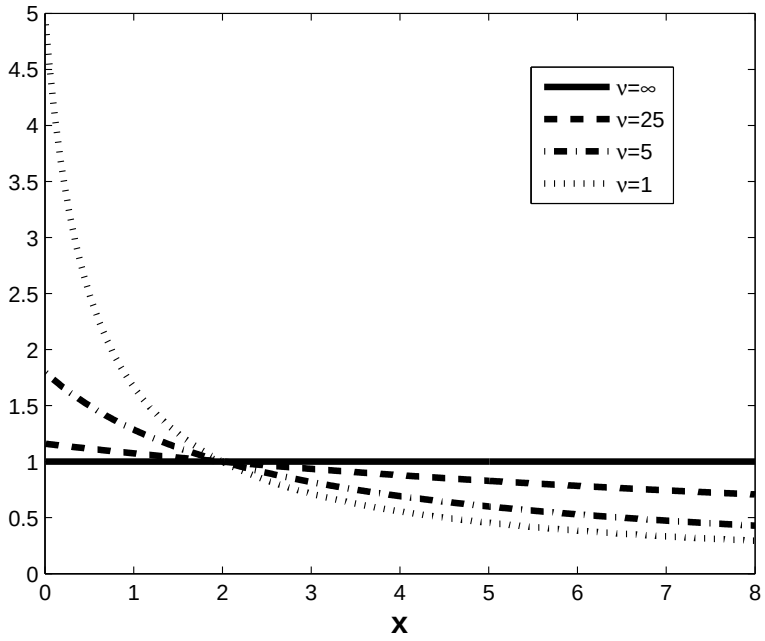


Figure 3.1: $\psi(x)$ of MLT(ν) estimators

gives the circular Cauchy distribution and the circular CN distribution is obtained at the limit as $\nu \to \infty$. We note that $g \in \mathcal{G}^1$ for $\nu \geq 3$, but for $\nu < 3$, the covariance matrix $\boldsymbol{\mathcal{C}}$ does not exist. Based on (3.29), the MLE of $\boldsymbol{\Sigma}$, labelled MLT(ν), is obtained with

$$\psi_{ML}(t) = \frac{2d + \nu}{\nu + 2t}. \tag{3.30}$$

Note that MLT(1) is the highly robust estimator corresponding to MLE of $\boldsymbol{\Sigma}$ for the complex circular Cauchy distribution, and that MLT(ν) $\to \mathbf{S}$ as $\nu \to \infty$. This means that the robustness of MLT(ν) estimators decrease with increasing values of ν (as expected). Figure 3.1 plots the weight function (3.30) of MLT(ν) estimators for selected values of ν. Note that the larger the value of the d.f. parameter ν is, the closer is the weight function to the unity weight $\psi_{ML} \equiv 1$ corresponding to the SCM $\mathbf{S}$ obtained when $\nu \to \infty$. ■

3.6 Detectors of circularity

In many applications it is not known a priori whether the source signals and/or noise are circular or non-circular. In such a case, one can resort to the decision (accept/reject) of a circularity test e.g. to guide the selection of the optimal array processors for further processing of the data since the optimal detection and estimation techniques are often different for circular and non-circular cases. Therefore circularity detectors have been

under active research in the recent literature; see [83, 110, 113–115] and Publications [VI,VII,X].

In Publication [X] we derived the generalized likelihood ratio test (GLRT) statistic assuming complex normal data. The same statistic was derived independently in [110] and was further studied in [114]. The deficiency of the GLRT of circularity is that it is sensitive to normality assumption - a feature that is common to most normal-theory based likelihood ratio (LR-)tests. However, a simple modification of the test, so called adjusted GLRT of circularity (Publication [VII]), is asymptotically robust with respect to departures from Gaussianity within the wide class of CES distributions with finite 4th-order moments. In the univariate case, circularity test based on characteristic functions were proposed [83], Wald's type circularity detectors under CES distributions were considered in [106] whereas [115] considered GLRT of circularity under complex generalized Gaussian distribution.

3.6.1 GLRT of circularity

In Publication [X] it was shown that the (logarithm of the) GLRT statistic for the hypothesis $H_0^N : \boldsymbol{\mathcal{P}} = \mathbf{0}$ against the general alternative $H_1^N : \boldsymbol{\mathcal{P}} \neq \mathbf{0}$ assuming that $\mathbf{z}_1, \ldots, \mathbf{z}_n$ are i.i.d. from $\mathrm{CN}_d(\boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{P}})$ is proportional to

$$\ell_n \triangleq -(n-d) \ln \det(\mathbf{I} - \hat{\boldsymbol{\varrho}} \hat{\boldsymbol{\varrho}}^*),$$

where $\hat{\boldsymbol{\varrho}} \triangleq \left(\sum_{i=1}^n \mathbf{z}_i \mathbf{z}_i^H \right)^{-1} \sum_{i=1}^n \mathbf{z}_i \mathbf{z}_i^T$ is the MLE of the circularity matrix $\boldsymbol{\varrho}$. Same test statistic was later derived independently in [110]. In Publications [VI,VII] and [107], it was shown using the general likelihood ratio theory [116] that ℓ_n possess an asymptotic chi-squared distribution with $p = d(d+1)$ degrees of freedom under the null hypothesis. Same result was shown independently in [114]. We note that the multiplier $(n-d)$ instead of n in ℓ_n serves as a small sample adjustment [114, Sect. VII-B]. The test that rejects H_0^N whenever ℓ_n exceeds the corresponding chi-square $(1-\alpha)$th quantile is thus GLRT with asymptotic level α. Note that in the scalar case ($d = 1$), the test statistic ℓ_n reduces to

$$\ell_n = -(n-1) \ln(1 - |\hat{\varrho}|^2), \tag{3.31}$$

where $\hat{\varrho}$ is the sample estimate of the circularity quotient ϱ. In this case, the GLRT of circularity is nothing but the well-know test of sphericity [VI, Section 4].

3.6.2 Adjusted GLRT of circularity

In Publication [VII], assuming that $\mathbf{z}_1, \ldots, \mathbf{z}_n$ is an i.i.d. random sample distributed as $\mathrm{CE}_d(\boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{P}}, g)$, $g \in \mathcal{G}^2$, we considered the hypothesis $H_0 : \boldsymbol{\mathcal{P}} = \mathbf{0}$ against the alternative $H_1 : \boldsymbol{\mathcal{P}} \neq 0$. Hence the purpose is to test the validity of circularity assumption when

sampling from an unspecified (not necessarily normal) CES distribution with finite 4th-order moments.

If $\mathbf{z} = (x_i + jy_i) \sim \mathrm{CE}_d(\boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{P}}, g)$ then one has that $\gamma(x_i) = \gamma(y_i)$, $\forall i = 1, \ldots, d$ and $\gamma(x_1) = \cdots = \gamma(x_d)$. *i.e.* the standardized 4th-order moments (2.2) (and also the kurtosis) of the real and imaginary parts of $\mathbf{z}$ are equal. Hence, let γ denote value of the common real normalized 4th-order moment of the marginals when $\mathbf{z} \sim \mathrm{CE}_d(\boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{P}}, g)$, $g \in \mathcal{G}^2$ and let $\hat{\gamma}$ denote any consistent estimate of γ. Then the *adjusted GLRT statistic* of circularity,

$$\ell_{n,adj} \triangleq (3/\hat{\gamma}) \cdot \ell_n, \tag{3.32}$$

has the same asymptotic χ^2_p-distribution with $p = d(d+1)$ under the more general null hypothesis H_0. Based on the asymptotic distribution, we reject the null hypothesis at (asymptotic) α-level if the P-value $P = 1 - F_{\chi^2_p}(\ell_{n,adj}) < \alpha$. A clever estimate $\hat{\gamma}$ of γ was proposed in Publication [IX]. In the scalar case ($d = 1$), using the sample estimate $\hat{\gamma}$ proposed in [VII], the adjusted GLRT statistic (3.32) becomes

$$\ell_{n,adj} = \frac{(2 + |\hat{\varrho}|^2)}{\hat{\gamma}_C} \cdot \ell_n$$

where $\hat{\gamma}_C$ is the sample estimate of the *complex* standardized 4th-order moment $\gamma(z)$ defined in (3.20) and ℓ_n is defined in (3.31)

■ Example 9. We now investigate the power of the adjusted GLRT test in detecting non-circularity in the scalar case by simulations. The sample $z_1, \ldots, z_n$ is generated from unit-variance ($\sigma^2 = 1$) CN distribution. For each generated sample, circularity coefficient $|\varrho| = |\tau|$ is kept fixed. Note that the larger is the circularity coefficient $|\varrho|$, the more non-circular is the sample. Figure 3.2 shows the detection performance of the GLRT test and the adjusted GLRT test at $\alpha = 0.05$ level (PFA, probability of false alarm) by depicting the proportion of correct rejections (observed probability of detection) as a function of $|\varrho|$. Sample size is $n = 1000$ and the number of generated samples (for each fixed $|\varrho|$) was 1000. Note that at $|\varrho| = 0$, the probability of rejection is close to the nominal 0.05 level. In the 2nd example, the setting is the same except that the sample is from (heavy-tailed) complex t-distribution with $\nu = 5$ degrees of freedom. Result are also shown in Figure 3.2. The GLRT test is not shown in this case as its performance becomes worse than a pure guess (due to vulnerability to normality assumption); see Publication [VII]. In the simulations, we used the GLRTcirc software [117]. ■

3.7 Discussion

Complex random signals play an increasingly important role in array, communications, and biomedical signal processing and related fields. The wider deployment of

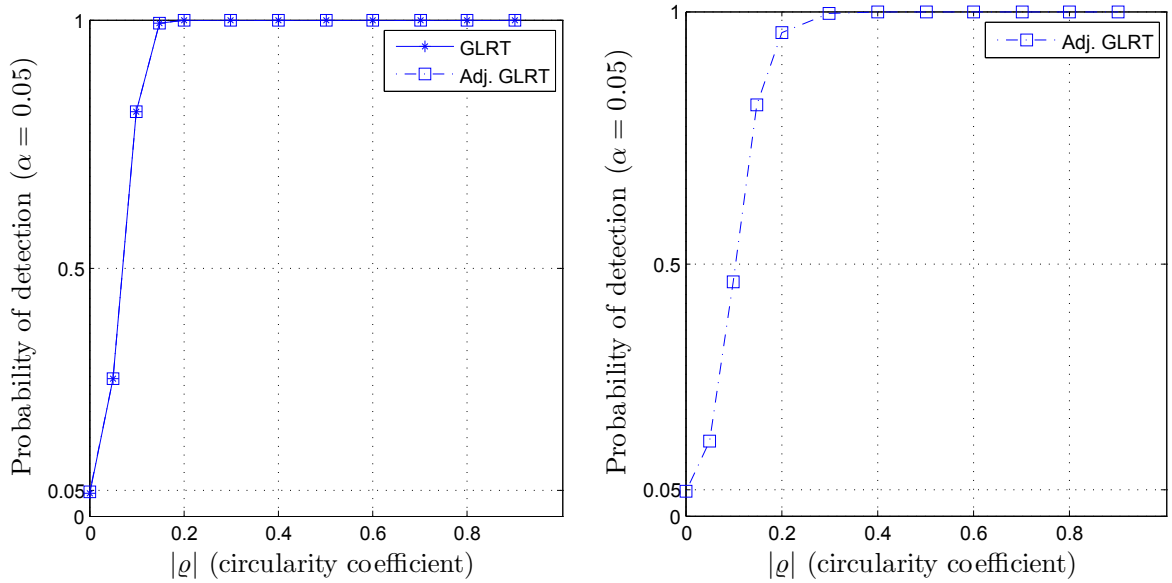


Figure 3.2: Observed probability of detection when sampling from unit-variance complex normal (left plot) and complex t-distribution with $\nu = 5$ d.f. (right plot) as a function $|\varrho|$. Sample size was $n = 1000$ and number of samples (for each fixed $|\varrho|$) was 1000.

complex-valued signal processing is often hindered by the fact that concepts, tools and algorithms for handling complex-valued signals are lacking, or, are simply too scattered in the literature. Due to extensive research in this area during the past few years (summarized in this section) these obstacles are no-longer existing, or, are at least less difficult to defeat. We also wish to highlight that circularity is a common hypothesis that is often assumed for simplicity of derivations. Since optimal methods for circular and non-circular cases are often different, detection of circularity of the complex-valued data is a highly important issue. The detectors developed herein are practical and easy to compute.

Chapter 4

Array signal processing

An important application area where complex-valued distributions and complex parameter estimation problems arise very naturally is array signal processing.

In this section we briefly review the common direction finding (DF) methods and MDL principle to estimate the number of sources. See [2–5, 89, 118] for in-depth account on DF methods. The aspect we emphasize is the lack of statistical robustness of the conventional array processors. Common to most DF methods is the need to estimate the unknown array covariance matrix. For example, the commonly used beamformers require the array covariance matrix to measure the power of the beamformer output as a function of the DOA. In addition, many high-resolution subspace-based DOA algorithms compute the noise or signal subspaces from the eigenvectors of the array covariance matrix and exploit the fact that signal subspace eigenvectors and the array steering matrix span the same subspace. Array covariance matrix is conventionally estimated from the array snapshots by the SCM. However, as we shall illustrate with numerous examples, employing a *r*obust scatter matrix such as a robust M-estimators of scatter adds robustness to the array processor against outliers and noise model deviations, yet the loss in efficiency when the conventional assumptions hold can be negligible. Thereby, we put special emphasis on the concept of scatter matrix and its applications to DF. Note however that there exists other approaches to robust DF as well; see [11, 119–125] to mention only a few.

4.1 The array model

The array consisting of m sensor elements receives d narrowband incoherent farfield plane-wave sources from a point source, where $m > d$. At discrete time t, the *array output* $\mathbf{z}(t) \in \mathbb{C}^m$ is a weighted linear combination of the signal waveforms $\mathbf{s}(t) = (s_1(t), \ldots, s_d(t))^T \in \mathbb{C}^d$ corrupted by additive noise $\mathbf{n}(t) \in \mathbb{C}^m$, that is,

$$\mathbf{z}(t) = \mathbf{A}(\boldsymbol{\theta})\mathbf{s}(t) + \mathbf{n}(t) \tag{4.1}$$

where $\mathbf{A} = \mathbf{A}(\boldsymbol{\theta})$ is the $m \times d$ complex array *steering matrix* parametrized by the vector $\boldsymbol{\theta} = (\theta_i)$ where θ_i is the distinct DOA of the ith source ($i = 1, \ldots, d$). Each column vector $\mathbf{a}_i$ of $\mathbf{A}$ represents a point in known array manifold (array transfer function, steering vector) $\mathbf{a}(\theta)$, *i.e.* $\mathbf{a}_i = \mathbf{a}(\theta_i)$. It is assumed that for any collection of d distinct θ_i the matrix $\mathbf{A}$ has full column rank. Identifying the steering matrix $\mathbf{A}$ is then equivalent with the problem of identifying the DOA's. The array manifold in the case of ULA with half a wavelength interelement spacing is $\mathbf{a}(\theta) = \begin{pmatrix} 1 & e^{-\jmath\pi\sin(\theta)} & \cdots & e^{-\jmath\pi(m-1)\sin(\theta)} \end{pmatrix}^T$.

Let us now drop the time index t for convenience. A common assumption is that the noise $\mathbf{n}$ is zero mean, spatially white (*i.e.* $\boldsymbol{\mathcal{C}}(\mathbf{n}) = \sigma_n^2\mathbf{I}$), and independent of the source vector $\mathbf{s}$ that is assumed to have zero mean and possess a full rank $d \times d$ covariance matrix $\boldsymbol{\mathcal{C}}(\mathbf{s})$. Starting point for most direction finding (DF) algorithms is the the array covariance matrix $\boldsymbol{\mathcal{C}}(\mathbf{z})$ which under the above assumptions can be represented as

$$\mathbf{A}\boldsymbol{\mathcal{C}}(\mathbf{s})\mathbf{A}^H + \sigma_n^2\mathbf{I}. \tag{4.2}$$

This decomposition of the array covariance matrix to low-rank signal subspace and full-rank noise subspace is exploited by high-resolution subspace methods [2]. Conventionally the array covariance matrix is estimated from the *array snapshots* $\mathbf{z}_1 = \mathbf{z}(t_1), \ldots, \mathbf{z}_n = \mathbf{z}(t_n)$ sampled at discrete time instants $t_1, \ldots, t_n$ by the SCM $\mathbf{S} \triangleq \frac{1}{n}\sum_{i=1}^n \mathbf{z}_i\mathbf{z}_i^H$.

The concept underlying the developments in this chapter is statistical robustness, which refers to robustness in the face of outliers (outliers occur in the array data e.g. due to measurements errors, or, heavy-tailed, impulsive noise such as man-made interference) or (slight/large) departures from nominal distributional assumptions. We wish to point out that in array processing literature, the word robust "robust" more commonly refers to (e.g. [126–128]):

- *Robustness to signal model errors*: signals may not be narrowband, they may not originate from a point source, emitters may not be in the far field and planewave assumption is not valid.
- *Robustness to steering errors*: imprecise knowledge of the array response may be due to uncertainty in array element locations, steering directions and calibration errors.
- *Robustness in the face of insufficient sample support* that may lead to rank deficient SCM or inaccurate estimates of the array covariance matrix.

For the last problem, the diagonal loading of the SCM is one of the most popular techniques to overcome the problem, *i.e.* to use $(\mathbf{S} + \gamma\mathbf{I})$, $\gamma \in \mathbb{R}$, in place of the sample covariance matrix $\mathbf{S}$, which may not be full rank and hence not invertible. For this type of robustness studies, see e.g. [126, 129–132] and references therein.

4.2 Scatter matrix

In many signal processing applications, the covariance matrix $\boldsymbol{\mathcal{C}}$ of the output $\mathbf{z} \in \mathbb{C}^d$ is unknown quantity that needs to be estimated from the sample $\mathbf{z}_1, \ldots, \mathbf{z}_n$. The most commonly used DF methods use the SCM in place of its true unknown quantity. Although statistical optimality can often be claimed for signal processors using the SCM under the circular CN assumption, they suffer from the lack of robustness in the face of outliers and signal or noise modelling errors. Furthermore, their efficiency for heavy-tailed non-Gaussian and impulsive noise environments are far from optimal. See e.g. [11, 119–125] for illustrations.

Simple and intuitive approach to robustify signal processors is then to use robust covariance matrix estimators instead of the conventional non-robust SCM. This objective leads to introduction of a more general notion of covariance, called the *scatter matrix* (Publications [I,II,IV,V]), which is best described as a generalization of the covariance matrix. Scatter matrix is a well-known concept in real-valued multivariate analysis; See page 22 for the definition in the real-valued case. Naturally, the concept of scatter matrix can easily be adapted to complex-valued case. A positive definite Hermitian $d \times d$ matrix $\mathbf{C}(\mathbf{z})$ is called a *scatter matrix* if it is equivariant in the sense that $\mathbf{C}(\mathbf{G}\mathbf{z}) = \mathbf{G}\mathbf{C}(\mathbf{z})\mathbf{G}^H$ for any nonsingular $d \times d$ matrix $\mathbf{G}$. Clearly the covariance matrix is a scatter matrix, but scatter matrix, by its definition, do not necessarily require the assumption of finite 2nd-order moments for its existence and is therefore capable in describing dependencies between complex random variables in more general settings than the covariance matrix.

More generally, any *weighted covariance matrix*, defined as

$$\mathbb{E}[\psi(\,\mathbf{z}^H\mathbf{C}(\mathbf{z})^{-1}\mathbf{z}\,)\mathbf{z}\mathbf{z}^H],$$

where $\psi(\cdot)$ is any real-valued weighting function on $[0, \infty)$ and $\mathbf{C}(\cdot)$ is any scatter matrix (e.g. the covariance matrix) is also a scatter matrix. The conventional covariance matrix is obtained with unit weight $\psi \equiv 1$. In the real-valued case, an improved and well-established idea of the weighted covariance matrices are M-estimators of scatter [14]. In fact, weighted covariance matrix can be thought of as "1-step M-estimator". M-estimators can also be generalized to complex case.

4.2.1 Complex M-estimators of scatter

In the real-valued case one of the first proposals of robust scatter matrix estimators were M-estimators of scatter due to Maronna [14]. Extension of M-estimators for complex-valued case were introduced in Publication [V]. As in the real case they can be defined by generalizing the MLE of the scatter parameter $\boldsymbol{\Sigma} \in \mathrm{PDH}(d)$ of circular CES distribution $\mathrm{CE}_d(\boldsymbol{\Sigma}, g)$.

We generalize the ML estimating equation (3.28), by defining *M-estimator of scatter*, denoted by $\hat{\mathbf{C}}_\psi$, as the choice of $\mathbf{C} \in \mathrm{PDH}(k)$ solving the estimating equation

$$\mathbf{C} = \frac{1}{n}\sum_{i=1}^{n} \psi(\mathbf{z}_i^H \mathbf{C}^{-1} \mathbf{z}_i)\mathbf{z}_i\mathbf{z}_i^H,$$

where ψ is any real-valued *weight function* on $[0, \infty)$. Hence M-estimators constitute a wide class of scatter matrix estimators that include the MLE's $\hat{\mathbf{\Sigma}}$ of the scatter parameter $\mathbf{\Sigma}$ of the circular CES distributions (discussed in Section 3.5.3) as important special cases. M-estimators can be calculated by a simple iterative algorithm; see Publications [IV,V] for details.

The theoretical (population) counterpart, the *M-functional of scatter*, denoted by $\mathbf{C}_\psi = \mathbf{C}_\psi(\mathbf{z})$, is defined analogously as the solution of an implicit equation

$$\mathbf{C}_\psi = \mathbb{E}[\psi(\mathbf{z}^H \mathbf{C}_\psi^{-1} \mathbf{z})\mathbf{z}\mathbf{z}^H].$$

It is easy to show that M-functional of scatter is equivariant under invertible linear transformation of the data in the sense required from the scatter matrix. Due to this equivariance property, M-functional is proportional to the scatter parameter $\mathbf{\Sigma}$ of the circular CES distribution $\mathrm{CE}_d(\mathbf{\Sigma}, g)$. In addition, since the scatter parameter $\mathbf{\Sigma}$ is proportional to the underlying covariance matrix $\boldsymbol{\mathcal{C}}$ provided it exists, we conclude that M-functional of scatter is also proportional to the covariance matrix in such instances (*i.e.* when $g \in \mathcal{G}^1$). For example, in many sensor array processing applications, covariance matrix is required only up to a constant scalar and hence M-functionals can be used to define a robust class of array processors [IV].

Some examples of M-estimators are given next; See [IV,V] for a more detailed description of these estimators. As explained earlier in Section 3.5.3, robust weight function should descend to zero.

■ Example 10. *Huber's M-estimator*, labelled HUB(q), is defined via weight

$$\psi(x) = \begin{cases} 1/b, & \text{for } x \le c^2 \\ c^2/(xb), & \text{for } x > c^2 \end{cases}$$

where c is a tuning constant defined so that $q = F_{\chi^2_{2k}}(2c^2)$ for a chosen q $(0 < q \le 1)$ and the scaling factor $b = F_{\chi^2_{2(d+1)}}(2c^2) + c^2(1-q)/d$. The choice $q = 1$ yields $\psi \equiv 1$, *i.e.* HUB(1) correspond to the SCM. In general, low values of q increase robustness but decrease efficiency at the nominal circular CN model (see Publication [IV]). Figure 4.1 depicts weight function of HUB(q) estimators for selected values of q. ■

■ Example 11. *Tyler's M-estimator* of scatter (Publication [V]) utilizes weight function

$$\psi(x) = d/x$$

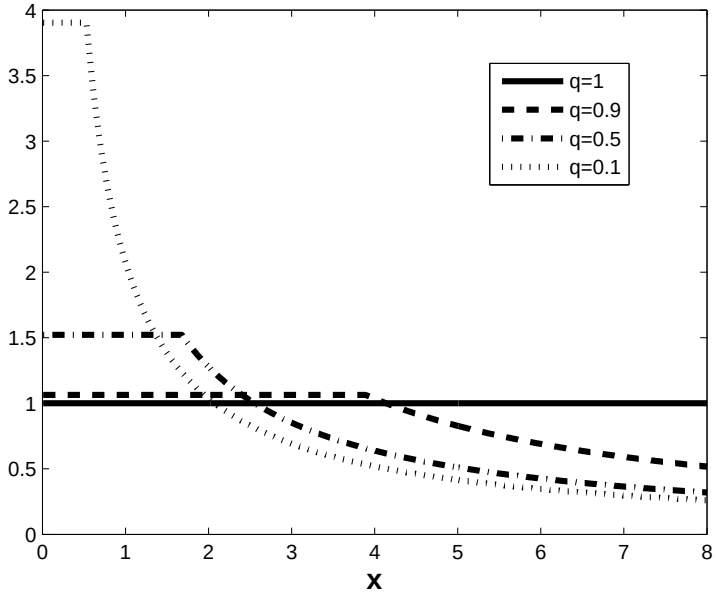


Figure 4.1: $\psi(x)$ function of HUB(q) estimators.

This M-estimator of scatter is also the MLE of the complex angular central Gaussian distribution [133]. ■

MATLAB functions to compute M-estimators is available at [134].

4.3 Beamformers

Beamforming is among the most important tasks in sensor array processing. Consequently, there exists a vast amount of research papers on beamforming techniques, see e.g. [3, 4, 135] for overviews.

Let us first recall the beamforming principles in narrowband applications. In receive beamforming, the *beamformer weight* vector $\mathbf{w} = \mathbf{w}(\theta_0) \in \mathbb{C}^m$ linearly transforms the output signal $\mathbf{z}$ of array of m sensors to form the *beamformer output* $y = \mathbf{w}^H\mathbf{z}$ with an aim of enhancing the signal-of-interest (SOI) from *look direction* (DOA of SOI) θ_0 and attenuating undesired signals (interferers) from other directions. The (look direction dependent) *beam response* or *gain* is defined as

$$b(\theta) \triangleq \mathbf{w}^H\mathbf{a}(\theta)$$

where $\mathbf{a}(\theta)$ is the array response (steering vector) to DOA θ. The modulus squared $|b(\theta)|^2$ as function of θ is called the *beampattern* or *antenna pattern*. Then, beamformer *output power*

$$P(\theta_0) \triangleq \mathbb{E}[|y|^2] = \mathbf{w}^H\boldsymbol{\mathcal{C}}(\mathbf{z})\mathbf{w} \tag{4.3}$$

should provide an indication of the amount of energy coming from the fixed look direction θ_0. Plotting $P(\theta)$ as a function of the look direction θ is called as the *spatial spectrum*. The d highest peaks (local maxima) of the spatial spectrum correspond to the beamformer DOA estimates of the d sources.

The beamformer weight vector $\mathbf{w}$ is chosen with an aim that *it is statistically optimum in some sense*. Naturally, different design objectives lead to different beamformer weight vectors. The classical beamformers [3, 4], namely the conventional (delay-and-sum) beamformer and the Capon MVDR beamformer are reviewed below.

4.3.1 Conventional beamformer

Suppose that a single SOI arrives from an angle θ_0. Assuming that the array model (4.1) holds, the array output is $\mathbf{z} = \mathbf{a}_0 s + \mathbf{n}$, where $\mathbf{a}_0 = \mathbf{a}(\theta_0)$ denotes the array response for look direction θ_0. Then maximizing the output power (4.3) (or, equivalently the SNR), $P(\theta_0) = \mathbb{E}[|s|^2]|\mathbf{w}^H\mathbf{a}_0|^2 + \sigma_n^2\|\mathbf{w}\|^2$, over all weight vectors of the same magnitude is equivalent to finding the maximizer of $|\mathbf{w}^H\mathbf{a}_0|^2$ yielding $\mathbf{w} = \mathbf{a}_0$ as the optimal weight vector. Hence the spectrum (4.3) becomes

$$P_{\mathrm{BF}}(\theta) \triangleq \mathbf{a}(\theta)^H \boldsymbol{\mathcal{C}}(\mathbf{z})\mathbf{a}(\theta). \tag{4.4}$$

Naturally, when multiple signals are present, they also contribute to the measured output power at each look direction. Then a local maxima (peak) of the spectrum can be shifted away from the true DOA of a weak signal by a strong interferer in the vicinity, or, two closely spaced signals can result in only one peak regardless of the available data amount or quality.

■ Example 12. We consider 5-element ULA with half a wavelength interelement spacing that receives two independent signals each having 10 dB SNR from 0 and 15 degrees. As can be seen from Figure 4.2, when the antenna array has formed a beam in the look direction 0^o of the SOI, it still exhibits significant gain in the direction of the signal from 15^o. Hence signal at 15^o contributes to the measured power at 0^o, resulting in poor angular resolution of average power at 0^o as revealed in the spectrum in Figure 4.2. We note that the spectrum above is idealized since we assumed infinite data records. The example illustrates that the conventional beamformer has poor resolution when two closely spaced source signals are present. For a ULA with half a wavelength interelement spacings, the beamforming resolution limit is approximately $2/m$ [3]. Hence in this example, it is about $2/5$ rad $\approx 22.2^o$. ■

4.3.2 MVDR beamformer

The classical Capon's [136] MVDR beamformer attempts to overcome the poor resolutions problems associated with the conventional beamformer by choosing the beam-

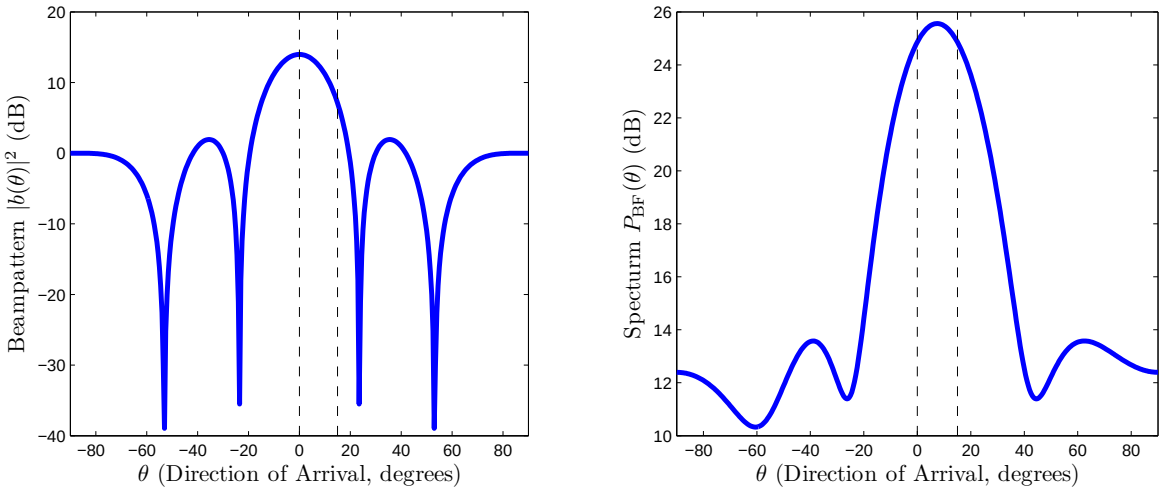


Figure 4.2: Antenna beampattern of the the conventional beamformer to the look direction 0^o and the respective spectrum assuming infinite data records for a 5-element ULA with $\lambda/2$ spacings receiving two 10 dB SNR signals from 0^o and 15^o (dashed lines).

former weight $\mathbf{w} \in \mathbb{C}^m$ as the minimizer of the output power $\mathbf{w}^H \boldsymbol{\mathcal{C}}(\mathbf{z})\mathbf{w}$ while constraining the beam response $b(\theta)$ along a specific look direction θ_0 of the SOI to be unity. The well-known solution to this constrained optimization problem is

$$\mathbf{w}_{\boldsymbol{\mathcal{C}}}(\mathbf{z}) \triangleq \frac{\boldsymbol{\mathcal{C}}(\mathbf{z})^{-1}\mathbf{a}_0}{\mathbf{a}_0^H \boldsymbol{\mathcal{C}}(\mathbf{z})^{-1}\mathbf{a}_0}. \tag{4.5}$$

Observe that Capon's beamformer weight vector is data dependent whereas the classical beamformer weight $\mathbf{w}_{\mathrm{BF}}$ is not, *i.e.* $\mathbf{w}_{\boldsymbol{\mathcal{C}}}(\cdot)$ is a statistical functional as its value depends on the distribution F of $\mathbf{z}$ via the covariance matrix $\boldsymbol{\mathcal{C}}(\mathbf{z})$. The spectrum (4.3) becomes

$$P_{\mathrm{CAP}}(\theta) \triangleq [\mathbf{a}(\theta)^H \boldsymbol{\mathcal{C}}(\mathbf{z})^{-1}\mathbf{a}(\theta)]^{-1}. \tag{4.6}$$

The implicit assumption here is that $\boldsymbol{\mathcal{C}}(\mathbf{z})$ is non-singular. Empirically, the MVDR beamformer is shown to possess superior performance to that of the conventional beamformer (see also Example below). Note that the MVDR beamformer do not make any assumption on the structure of the covariance matrix (unlike the subspace-methods of the next section) and hence can be considered as a "nonparametric method" [3].

■ Example 12 (cont'd). The performance improvement of the MVDR over the conventional beamformer is illustrated in the same environment as in Example 12. The spectrum in Figure 4.3 shows that the MVDR beamformer successfully resolves the two sources. Unlike the conventional beamformer, the MVDR beamformer succeeds because it attenuates the signal from 15^o while looking in the direction of the

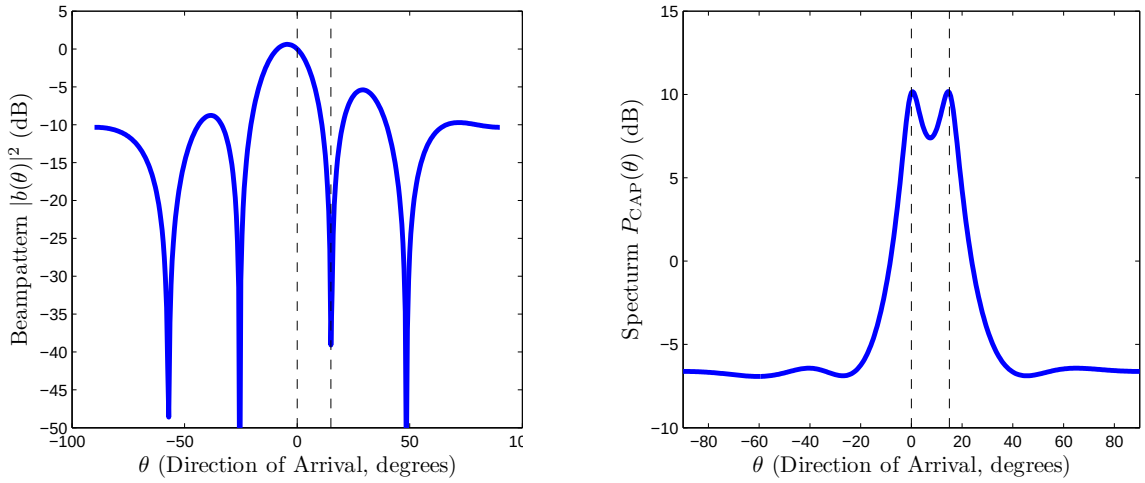


Figure 4.3: Antenna beampattern of the MVDR beamformer to the look direction 0^o and the respective spectrum assuming infinite data records for a 5-element ULA with $\lambda/2$ spacings receiving two 10 dB SNR signals from 0^o and 15^o (dashed lines).

signal from 0^o, as revealed by the antenna beampattern depicted in Figure 4.3. Beampattern/spectrum above are idealized since we assumed infinite data records. ■

■ Example 13. The environment is as in Example 12, but now the sources are only 10 degrees apart. As can be seen from the spectrum in Figure 4.5, the MVDR beamformer is no longer able to resolve the two sources, yielding only one broad peak in the middle of the signal DOA's. In general, the resolution of the MVDR depends upon the number of sources and on the SNR. MVDR method also fails if other signals that are correlated with the SOI are present. ■

In the examples above, we assumed infinite data records. In practice, of course, the DOA estimates for the classical beamformer and Capon's beamformer are calculated as the d highest peaks in the estimated spectrums $\hat{P}_{\mathrm{BF}}(\theta)$ and $\hat{P}_{\mathrm{CAP}}(\theta)$, where the true unknown covariance matrix is replaced by its conventional estimate, the SCM. An intuitive approach in obtaining robust beamformer DOA estimates is to use robust estimators instead of the SCM in (4.4) and (4.6), e.g. the M-estimators of scatter. Rigorous statistical robustness and efficiency analysis of MVDR beamformers based on M-estimators of scatter is presented in Publication [IV].

4.4 Subspace methods

As Example 13 demonstrated, the resolution capability of the MVDR beamformer is rather limited. Subspace methods can provide higher resolution for closely-spaced sources. Prior to introducing the methods we review the basic assumptions and intro-

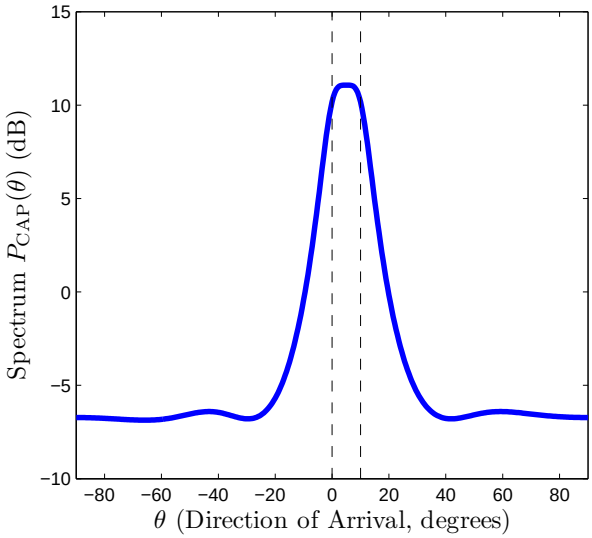


Figure 4.4: MVDR spectrum assuming infinite data records for a 5-element ULA with $\lambda/2$ spacings receiving two 10 dB SNR signals from 0^o and 10^o (dashed lines).

duce some terminology.

Assume that the array model (4.1) holds and the array covariance matrix $\boldsymbol{\mathcal{C}}$ obtains the decomposition (4.2). Due to the structure (4.2), the $m - d$ smallest eigenvalues of $\boldsymbol{\mathcal{C}}$ are equal to σ_n^2 and the corresponding eigenvectors $\mathbf{e}_{d+1}, \ldots, \mathbf{e}_m$ are orthogonal to the columns of the steering matrix $\mathbf{A}$. These eigenvectors span the *noise subspace* and the eigenvectors $\mathbf{e}_1, \ldots, \mathbf{e}_d$ corresponding to d largest eigenvalues span the *signal subspace* (the column space of $\mathbf{A}$).

The subspace DOA estimation methods are based on different properties of the signal/noise subspaces. Some subspace methods also impose additional assumptions on the array geometry (e.g. ESPRIT). Essentially, subspace methods need to solve the following two problems:

Prob1 Find an estimate $\hat{\mathbf{E}}_s$ of the signal subspace $\mathbf{E}_s = \begin{pmatrix} \mathbf{e}_1 & \cdots & \mathbf{e}_d \end{pmatrix}$ and/or estimate $\hat{\mathbf{E}}_n$ of the noise subspace $\mathbf{E}_n = \begin{pmatrix} \mathbf{e}_{d+1} & \cdots & \mathbf{e}_m \end{pmatrix}$.

Prob2 Find estimate $\hat{\boldsymbol{\theta}}$ of the DOA's which best optimizes the selected error criterion, for example, find $\hat{\boldsymbol{\theta}}$ such that distance between subspace $\mathbf{A}(\hat{\boldsymbol{\theta}})$ and the estimated subspace $\hat{\mathbf{E}}_s$ is minimal in some sense.

Commonly, the subspace methods differ only in how they approach Problem 2 since the estimates of signal and noise subspaces are calculated from the eigenvectors of the conventional, non-robust SCM. Intuitively speaking, it is evident that solving Problem 1 reliably, however, is more essential since no matter how clever criterion is used

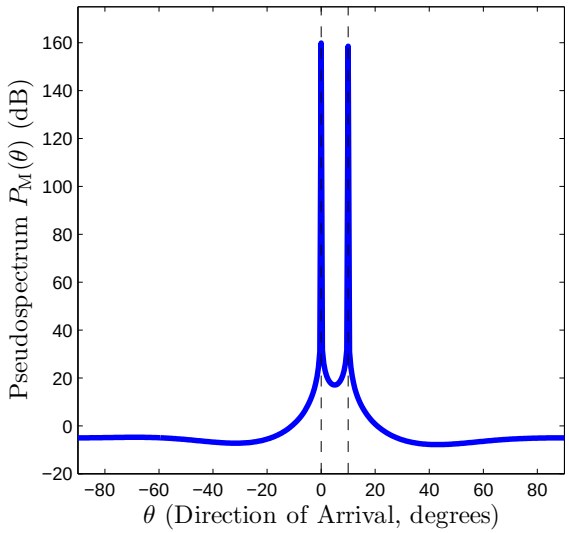


Figure 4.5: MUSIC pseudospectrum assuming infinite data records for a 5-element ULA with $\lambda/2$ spacings receiving two 10 dB SNR signals from 0^o and 10^o (dashed lines).

or how distances between subspaces are measured in Problem 2, the DOA estimates will be unreliable if the estimates of the subspaces are unreliable. In other words, accuracy and efficiency of the subspace method depends largely on the accuracy and efficiency of the estimates of the noise or signal subspaces. Again, to obtain robust subspace methods it is sensible to use estimates of noise or signal subspaces based on eigenvectors of the M-estimators of scatter for example.

4.4.1 MUSIC

The classical MUSIC (multiple signal classification) method [137] is based on the orthogonality of the signal and noise subspace and the fact that $\mathbf{A}$ and $\mathbf{E}_s$ span the same subspace. Due to the orthogonality of the signal and the noise subspace, $\mathbf{E}_n^H \mathbf{a}(\theta) = \mathbf{0}$, or equivalently, $\mathbf{a}(\theta)^H \mathbf{E}_n \mathbf{E}_n^H \mathbf{a}(\theta) = 0$, at the DOA's $\theta_1, \ldots, \theta_d$. Then, the MUSIC method finds DOA estimates as the d highest peaks of

$$P_M(\theta) \triangleq [\mathbf{a}(\theta)^H \hat{\mathbf{E}}_n \hat{\mathbf{E}}_n^H \mathbf{a}(\theta)]^{-1}$$

which is called as the *MUSIC pseudospectrum*. The resolution offered by MUSIC is much higher than that of conventional beamforming techniques [2].

■ Example 13 (cont'd). Figure 4.3 shows that the MUSIC method, unlike the MVDR beamformer, is able to resolve the two sources. The pseudospectrum is idealized since we assumed infinite data records. ■

Although MUSIC and subspace methods in general offer higher resolution than the classical beamformer techniques, yet they also suffer from poor robustness properties. Clearly, if the noise subspace $\mathbf{E}_n$ is unreliably estimated (e.g. via eigenvectors of the SCM when the noise is non-Gaussian or impulsive), then the obtained MUSIC DOA estimators are unreliable. For robust estimation of noise subspace one may use e.g. eigenvectors of M-estimators of scatter as in Publication [V], or, eigenvectors of the sample *(spatial) sign covariance matrix*

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{z}_i^H\mathbf{z}_i)^{-1}\mathbf{z}_i\mathbf{z}_i^H \tag{4.7}$$

as in [120].

■ Example 14. Four independent random signals, QPSK, 16-PSK, 32-QAM and BPSK signal of equal power σ_s^2, are impinging on an $k = 8$ element ULA with $\lambda/2$ spacing from DOA's $-10^o, 15^o, 10^o$ and 35^o. We consider two different noise environments. In the first setting, noise $\mathbf{n}$ has circular Gaussian distribution $\mathrm{CN}_m(\sigma_n^2\mathbf{I})$, and in the second setting noise has circular Cauchy distribution (recall Example 8) $\mathrm{CT}_{m,1}(\sigma_n^2\mathbf{I})$. Note that the Cauchy distribution does not have finite variance and σ_n^2 is the scale parameter of the distribution. In both simulation settings, the generalized signal to noise ratio (SNR) is $10\log_{10}(\sigma_s^2/\sigma_n^2) = 20$dB and the number of snapshots is $n = 300$. Recall that the Cauchy distribution does not have finite variance. Hence σ_n^2 in the Cauchy distribution case does not represent the variance but the squared scale parameter of the distribution. Hence the name generalized SNR. The number of signals ($d = 4$) is assumed to be known a priori. We then estimated the noise subspace $\mathbf{E}_n$ from eigenvectors of the SCM, sample sign covariance matrix (4.7) and MLT(1) estimator (recall Example 8). Typical MUSIC spectrums associated with different estimators are shown in Figure 4.6 for both the Gaussian and Cauchy noise settings. All the estimators are able to resolve the four sources correctly in the Gaussian noise case: in fact, the differences in the spectrums are very minor, *i.e.* they provide essentially the same DOA estimates. In the Cauchy noise case, however, MUSIC based on the classical sample estimator (*i.e.* the SCM) is not able to resolve the sources. The robust estimators, the sign covariance matrix and the MLT(1) estimator however yield reliable estimates of the DOA's. Based on the sharpness of the peaks, MLT(1) estimator is performing better than the sample sign covariance matrix. ■

4.4.2 Subspace fitting

There are plenty of subspace methods in addition to MUSIC, see e.g. [2, 3], such as subspace fitting methods. For example, in weighted signal subspace fitting (SSF) approach [2], one finds DOA's via criterion function

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \mathrm{Tr}[\Pi_{\mathbf{A}}^{\perp}\hat{\mathbf{E}}_s\mathbf{Y}\hat{\mathbf{E}}_s^H],$$

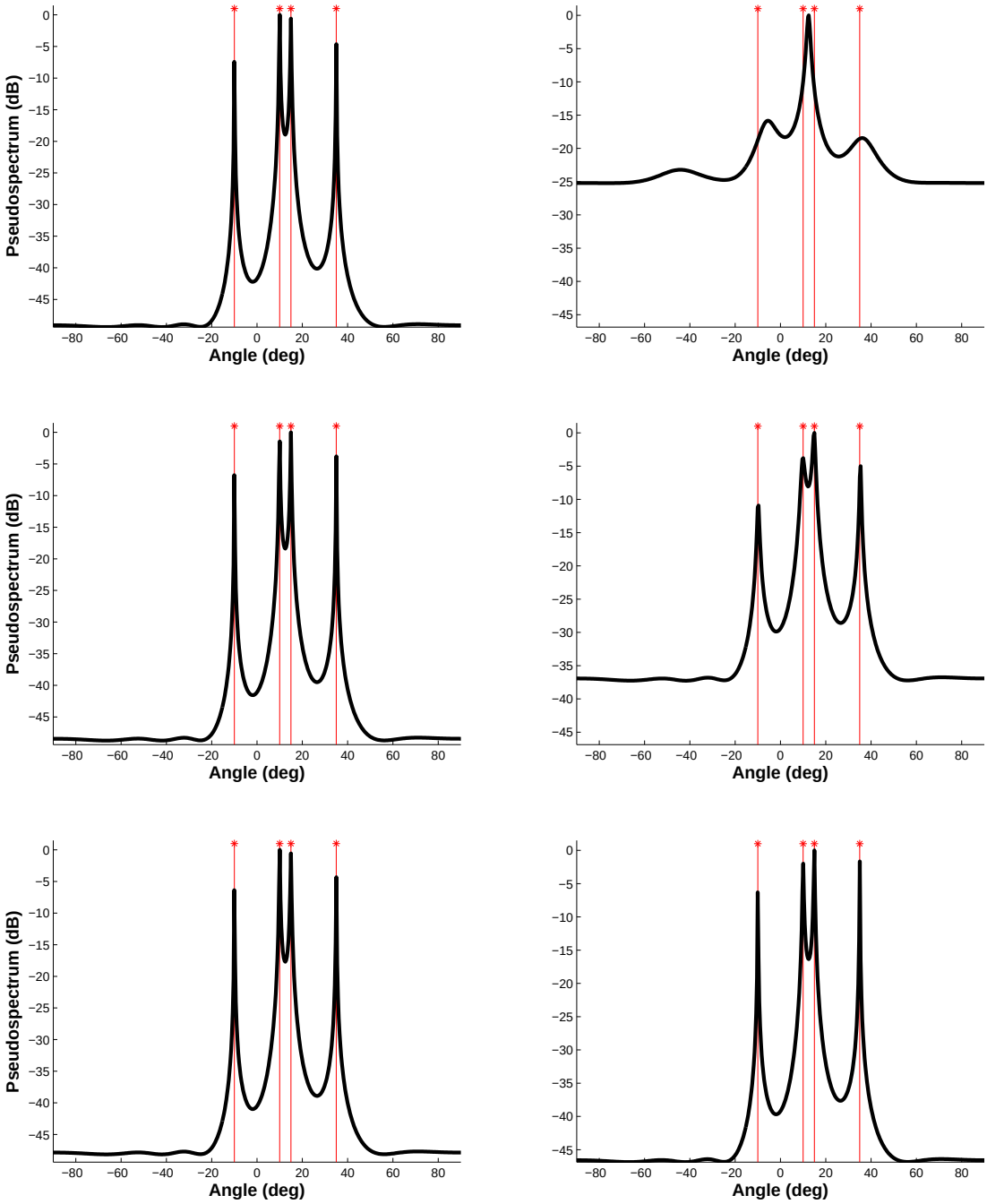


Figure 4.6: MUSIC spectrums when the noise subspace is estimated using SCM (first row), sample sign covariance matrix (second row) and MLT(1) estimator (third row) in circular Gaussian (first column) and Cauchy (second column) noise. Sources are independent random QPSK, 16-PSK, 32-QAM and BPSK signals that arrive at 8-element ULA from DOA's $-10^o, 15^o, 10^o$ and 35^o. The number of snapshots is $n = 300$.

where $\Pi_{\mathbf{A}}^{\perp} = \mathbf{I} - \mathbf{A}(\mathbf{A}^H\mathbf{A})^{-1}\mathbf{A}^H$ is a projection matrix onto the noise subspace and $\mathbf{Y}$ is some weighting matrix. The estimated optimal weighting matrix $\hat{\mathbf{Y}}_{opt}$ is a diagonal matrix, whose diagonal elements are certain functions of the estimated eigenvalues of the array covariance matrix. Hence, reliable and accurate estimation of DOA's via weighted SSF approach requires robust estimation of the signal subspace and eigenvalues of the covariance matrix. These can be obtained, for example, using eigenvectors and eigenvalues of robust M-estimators instead of the SCM.

4.4.3 Subspace DOA estimation for noncircular sources

We now describe the Root-MUSIC-like method presented in [138]. Assume that the signal $\mathbf{s}$ and noise $\mathbf{n}$ in the array model (4.1) are uncorrelated with zero-mean. The method further requires the following additional assumptions

Mu1 Array is ULA (in order to facilitate using polynomial rooting).

Mu2 Noise $\mathbf{n}$ is 2nd-order circular and spatially white, *i.e.* $\boldsymbol{\mathcal{C}}(\mathbf{n}) = \sigma_n^2\mathbf{I}$ and $\boldsymbol{\mathcal{P}}(\mathbf{n}) = \mathbf{0}$.

Mu3 Sources signals s_i, $i = 1, \ldots, d$ are uncorrelated: $\boldsymbol{\mathcal{C}}(\mathbf{s}) = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$ and $\boldsymbol{\mathcal{P}}(\mathbf{s}) = \mathrm{diag}(\tau_1, \ldots, \tau_d)$.

Under these assumptions,

$$\boldsymbol{\mathcal{C}}(\mathbf{z}) = \mathbf{A}\boldsymbol{\mathcal{C}}(\mathbf{s})\mathbf{A}^H + \sigma_n^2\mathbf{I}, \quad \boldsymbol{\mathcal{P}}(\mathbf{z}) = \mathbf{A}\boldsymbol{\mathcal{P}}(\mathbf{s})\mathbf{A}^H,$$

where as earlier $\mathbf{A} = \mathbf{A}(\boldsymbol{\theta})$ denotes the array response matrix. Further assume that

Mu4 $\boldsymbol{\mathcal{P}}(\mathbf{s}) = \boldsymbol{\mathcal{C}}(\mathbf{s})\boldsymbol{\Phi}$, where $\boldsymbol{\Phi} = \mathrm{diag}(e^{\jmath\phi_1}, \ldots, e^{\jmath\phi_d})$.

Assumption Mu4 means that circularity coefficient of the sources are equal to unity, *i.e.* $|\varrho(s_i)| = 1$ for $i = 1, \ldots, d$, which implies that the transmitted source signals s_i must be real-valued, such as AM or BPSK modulated signals, or the real part $\mathrm{Re}(s_i)$ of the transmitted signal is a linear function of the imaginary part $\mathrm{Im}(s_i)$. If Mu1-Mu4 holds, then the augmented covariance matrix is

$$\boldsymbol{\mathcal{C}}(\hat{\mathbf{z}}) = \begin{pmatrix} \mathbf{A} \\ \mathbf{A}^*\boldsymbol{\Phi}^* \end{pmatrix} \boldsymbol{\mathcal{C}}(\mathbf{s}) \begin{pmatrix} \mathbf{A} \\ \mathbf{A}^*\boldsymbol{\Phi}^* \end{pmatrix}^H + \sigma_n^2\mathbf{I}. \tag{4.8}$$

Then by computing the eigenvalue decomposition $\boldsymbol{\mathcal{C}}(\hat{\mathbf{z}})$ one may find d dimensional signal subspace and $2m - d$ dimensional orthogonal noise subspace. Thus Root-MUSIC-like direction finding algorithm can be designed; see [138] for details. By exploiting the noncircularity property one obtains extra degrees of freedom since noncircularity allows resolving more sources than sensors. Again, in the face heavy-tailed noise or outlying observations, a robust estimate of the array covariance matrix $\boldsymbol{\mathcal{C}}(\mathbf{z})$ and

pseudo-covariance matrix $\boldsymbol{\mathcal{P}}(\mathbf{z})$ can be used instead of the conventional non-robust sample estimators. We wish to point out, however, that the assumptions stated above are not necessary for all subspace DOA estimation methods for non-circular sources; see e.g. [87].

4.5 Estimating the number of sources

An equally important problem to the DOA estimation is the estimation of the number of sources. The subspace based methods introduced in the previous section usually assume that the number of source signals is known a priori. In practise, the number of sources d is often not known and needs to be estimated from the data. There are several criteria that can be used, see e.g. [139] for an overview. The commonly used Minimum Description Length (MDL) based information theoretic criterion [18], obtains the estimate $\hat{d}$ for the number of signals d as an integer $p \in \{0, 1, \ldots, m-1\}$ which minimizes the criterion [19]

$$\mathrm{MDL}(p) \triangleq -\log \left(\frac{\left(\prod_{i=p+1}^{m} l_i \right)^{1/(m-p)}}{\frac{1}{m-p} \sum_{i=p+1}^{m} l_i} \right)^{(m-p)n} + \frac{1}{2} p(2m-p) \log n,$$

where $l_1, l_2, \ldots, l_m$ denote the eigenvalues of the SCM arranged in descending order. An interesting alternative approach to MDL is the bootstrap-based detector proposed in [140]. Instead of using the eigenvalues of SCM, it is desirable for purposes of reliable estimation in non-Gaussian noise to employ eigenvalues of some robust estimator of covariance, e.g. M-estimator of scatter, instead of the SCM. We demonstrate this via simulation study.

■ Example 15. 8-element ULA with half a wavelength interelement spacing receives two uncorrelated Gaussian signals with equal power 20 dB from DOA's $\theta_1 = -5^o$ and $\theta_2 = 5^o$. The components of the additive noise $\mathbf{n}$ are modelled as i.i.d. with complex symmetric α-stable (SαS) distribution [119] with dispersion $\gamma = 1$ and values α ranging from $\alpha = 1$ (complex Cauchy noise) to $\alpha = 2$ (complex Gaussian noise). Simulations results are based on 500 Monte Carlo runs with $n = 300$ as the sample size. Figure 4.7 depicts the relative proportion of correct estimation results using MDL criterion, when the eigenvalues are obtained from SCM and robust MLT(1), HUB(0.9) and HUB(0.5) estimators (recall Example 7 and Example 8). The performance of the classical MDL employing the SCM is poor: it is able to estimate the number of signals reliably only for $\alpha = 2$, *i.e.* the Gaussian case. However, the robust M-estimator are able to estimate the number of sources reliably for large range of α-values. Among the robust M-estimators, MLT(1) has the best performance. ■

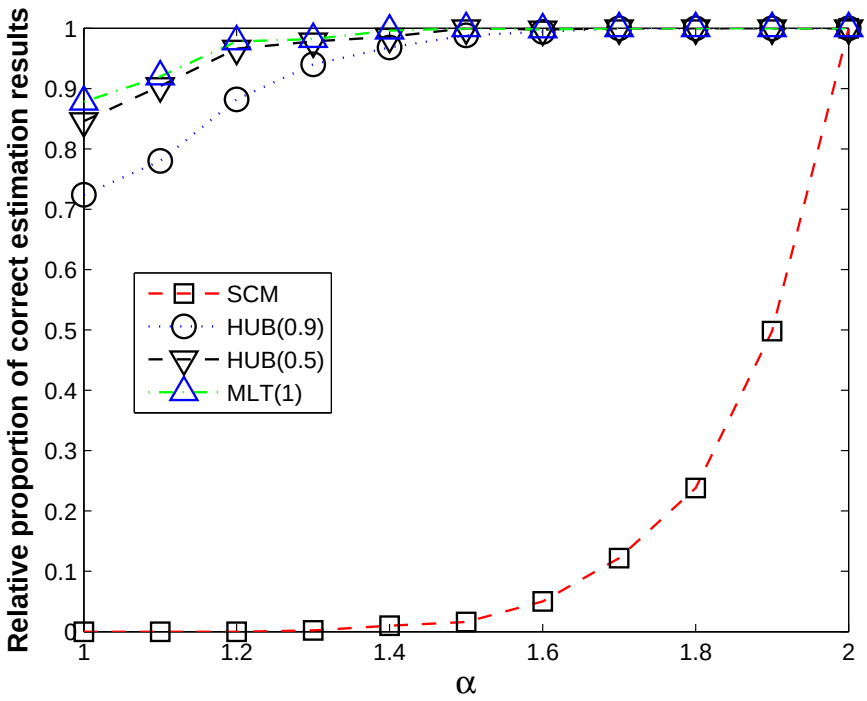


Figure 4.7: Simulation results for estimation of number of sources using the MDL criterion based on the SCM, HUB(0.9), HUB(0.5) and MLT(1)-estimators. There are $d = 2$ Gaussian source signals in SαS distributed noise for $1 \leq \alpha \leq 2$. The number of sensors is $d = 8$ and number of snapshot is $n = 300$.

4.6 Discussion

Common to most DF methods is the need to estimate the unknown array covariance matrix. Moreover, they often require covariance matrix $\boldsymbol{\mathcal{C}}$ only up to a constant scalar, and hence replacing the covariance matrix by a robust scatter matrix estimator provides a robust DF approach. Several examples were provided to illustrate the utility of this approach. The examples illustrated that DF methods based on scatter matrices, such as robust M-estimators have highly reliable performance regardless of the heavy-tailed nature of the noise distribution while having similar behaviour (resolution, accuracy) in nominal Gaussian conditions. Similar performance was observed when estimating the number of signals by MDL principle and robust scatter matrices.

Chapter 5

Conclusions

5.1 Summary

Array and multichannel signal processing are key technologies in wireless communications systems. Other application areas include radar, sonar and biomedicine. In this thesis new statistical procedures for array and multichannel signal processing were developed. Specifically, this thesis addressed the problem of independent component analysis, sensor array signal processing and fundamentals of complex-valued signal processing.

Complex-valued ICA model has attained increased interest recently as it is needed for source separation of complex-valued data arising e.g. in magnetic resonance imaging or antenna array signal processing of communications and radar signals. In this thesis we proposed two new classes of demixing matrix estimators for complex-valued ICA, called DOGMA (Publication [II]) and GUT (Publication [I]). The proposed methods are generalizations of Cardoso's FOBI [12] and Eriksson and Koivunen's SUT [13, 81] methods, respectively, which are included as special cases. Important benefit of the methods are their versatility as distinct estimators within the same class can have largely different statistical (robustness, efficiency) properties. Hence one can choose an estimator from the class that yields the best results to the specific application at hand. Another important benefit is that one can devise robust ICA estimators by choosing the required matrix-valued statistics properly. For example, a robust DOGMA demixing matrix estimator is obtained by choosing a pair of robust scatter matrices, e.g. robust M-estimators of scatter. The importance of robust estimation in ICA was amply demonstrated in Section 2.4 and Section 2.5. Both DOGMA and GUT are algebraic methods essentially requiring only a simple eigenvector decomposition. Hence numerical calculation and implementation of these methods is straightforward which is a remarkable benefit when compared to other ICA methods which are often based on optimization of a nonlinear function. Besides the aforementioned contributions to complex-valued ICA, we also derived a simple closed form expression of the Cramér-Rao

bound (CRB) for the demixing matrix estimation in real-valued ICA, thus filling an important gap that was still existing in the theoretical foundations of ICA. Usefulness of the derived bound was also shown with a simulation study.

Complex random signals arise naturally in many signal processing fields either directly (e.g. modulated signals in communications) or indirectly (e.g. spectral representations of real signals). Recently, there has been an increased awareness that simplistic adaptation of techniques developed for random real-valued signals to the complex-valued case may not be adequate, may lead to suboptimal results, or intractable calculations. Unfortunately, even fundamental results and tools for handling complex-valued random signals are scattered in the open literature.

In this thesis (Publications [VI-X]) useful tools, statistics and estimators/detectors were introduced and developed for proper processing of complex-valued signals. For example, we established properties of a measure of circularity of a complex random variable (called the circularity quotient) and derived the generalized likelihood ratio test (GLRT) of circularity under Gaussianity and an adjusted GLRT of circularity which is valid within the more general class of CES distributions. Detecting circularity is amongst the most important issues in complex-valued signal processing since optimal estimation methods and performance bounds are often different for circular and non-circular cases. Also some fundamental theory of complex-valued signal processing were developed. For example, a novel complex-valued extension of Taylor series was introduced and complex-valued cumulants were derived in a mathematically rigorous manner. In addition, the unconstrained and constrained CRB for complex-valued parameter estimation were derived and properties of an important class of complex multivariate distributions, called the CES distributions, were studied.

In the area of array signal processing, the work in Publications [IV,V] focused on robust beamforming, high-resolution DOA estimation and estimation of the number of sources. The conventional methods for these tasks rely heavily on the sample covariance matrix and often perform poorly in the face of outliers or if the noise and interference appearing in the measurements are non-Gaussian. The noise observed in real-world measurements can significantly deviate from the Gaussian assumption. Hence it is advisable to devise robust array processors that perform reliably also under non-Gaussian environments or in the presence of outliers. The robust methods developed were based on the concept of scatter matrix. Specifically, Maronna's celebrated M-estimators of scatter were extended to the complex-valued case and theoretical robustness and asymptotic results of scatter matrix based MVDR beamformers were derived. M-estimators are fast to compute by a simple iterative algorithm and hence easier to implement and apply successfully than many other robust methods in the literature (e.g. the computationally demanding SAGE algorithm [121, 122]). If the computation time is an important issue in the application at hand one can resort to k-step M-estimators of scatter where the iterative algorithm to compute the M-estimator

is stopped after k (for k small, e.g. $k = 3$) iterations. Experimental results showing reliable performance were given on all of the presented methods.

5.2 Future work

An important topic for the future work is to extend the theory presented in [VIII] for the univariate case for complex random vectors. Indeed the concepts and results of [VIII], such as $\mathbb{R}$-linear functions, Taylor's $\mathbb{R}$-theorem, complex cumulants can be generalized rather directly to the vector case also.

Another topic for future work is developing robust and non-parametric approaches for detecting circularity. The drawback of the proposed GLRT of circularity is its sensitivity to outliers and violations of Gaussian data assumption. The adjusted GRLT of circularity provides a remedy for the latter, but the former problem remains. There exists a demand for a robust and nonparametric tests of circularity. Robustness is desired especially in communications applications since the additive noise is often non-Gaussian and impulsive, e.g. man-made interference in out-door urban channels as well as interference in indoor channels.

For almost all ICA estimators the form of the asymptotic covariance matrix is unresolved. One notable exception is the deflation-based FastICA estimator for which the asymptotic covariance matrix was derived recently in [37, 38]. In our future work we plan to derive the asymptotic covariance matrices for the DOGMA and GUT family of estimators and the CRB theory for the complex-valued ICA model. Some preliminary definitive results have already been obtained.

So far almost all ICA research have focused on estimation, but almost nothing exists on testing. This is mainly due to the fact that the limiting distribution (*i.e.* asymptotic normality and covariance matrix) of most demixing matrix estimators are unresolved. This knowledge could be used to construct (asymptotically valid) tests about the structure of the mixing or demixing matrix, for example testing for H_0: $a_{ij} = 0$ against H_1: $a_{ij} \neq 0$ (*i.e.* that the jth source does not contribute (have statistically significant effect) to the ith observed variable). Such tests can have many potential applications; see e.g. [45]. Constructing (preferably robust) tests for such problems is one direction for future work.

Bibliography

[1] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing: Concepts and Techniques*. Prentice-Hall, 1993, 512 pages.

[2] H. Krim and M. Viberg, "Two decades of array signal processing: the parametric approach," *IEEE Signal Processing Mag.*, vol. 13, no. 4, pp. 67–94, 1996.

[3] P. Stoica and R. Moses, *Introduction to spectral analysis*. Upper Saddle River: Prentice-Hall, 1997.

[4] G. M. Manolakis, V. K. Ingle, and S. M. Kogon, *Statistical and adaptive signal processing*. Boston: Artech house, 2005, 816 pages.

[5] H. L. Van Trees, *Detection, Estimation and Modulation theory, Part IV: Optimum array processing*. New York: Wiley, 2002, 1456 pages.

[6] P. Comon, "Independent component analysis—a new concept?" *Signal Processing*, vol. 36, pp. 287–314, 1994.

[7] J. F. Cardoso, "High-order contrasts for independent component analysis," *Neural Computation*, vol. 11, no. 1, pp. 157–192, 1999.

[8] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: Wiley, 2001, 504 pages.

[9] A. Cichocki and S.-I. Amari, *Adaptive Blind Signal and Image Processing*. New York: Wiley, 2002, 586 pages.

[10] D. Middleton, "Man-made noise in urban environments and transportation systems: Models and measurements," *IEEE Trans. Commun.*, vol. 21, pp. 1232–1241, 1973.

[11] D. B. Williams and D. H. Johnson, "Robust estimation of structured covariance matrices," *IEEE Trans. Signal Processing*, vol. 41, no. 9, pp. 2891–2905, 1993.

[12] J. F. Cardoso, "Source separation using higher order moments," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'89)*, Glasgow, UK, 1989, pp. 2109–2112.

[13] J. Eriksson and V. Koivunen, "Complex-valued ICA using second order statistics," in *Proc. IEEE Workshop on Machine Learning for Signal Processing (MLSP'04)*, Sao Luis, Brazil, 2004.

[14] R. A. Maronna, "Robust M-estimators of multivariate location and scatter," *Ann. Statist.*, vol. 5, no. 1, pp. 51–67, 1976.

[15] P. R. Krishnaiah and J. Lin, "Complex elliptically symmetric distributions," *Comm. Statist. - Th. and Meth.*, vol. 15, pp. 3693–3718, 1986.

[16] A. van den Bos, "The multivariate complex normal distribution - a generalization," *IEEE Trans. Inform. Theory*, vol. 41, no. 2, pp. 537–539, 1995.

[17] B. Picinbono, "Second order complex random vectors and normal distributions," *IEEE Trans. Signal Processing*, vol. 44, no. 10, pp. 2637–2640, 1996.

[18] J. Rissanen, "Modeling by the shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[19] T. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 33, no. 2, pp. 387–392, 1985.

[20] J. Eriksson and V. Koivunen, "Identifiability, separability and uniqueness of linear ICA models," *IEEE Signal Proc. Letters*, vol. 11, no. 7, 2004.

[21] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. New York: Springer, 2002, 502 pages.

[22] L. De Lathauwer, B. De Moor, and J. Vandewalle, "Independent component analysis based on higher-order statistics only," in *Proc. IEEE Workshop on Statistical Signal and Array Processing*, Corfu, Greece, 1996.

[23] ——, "An introduction to independent component analysis," *Journal of Chemometrics*, vol. 14, pp. 123–149, 2000.

[24] M. Davies, "Identifiability issues in noisy ICA," *IEEE Signal Processing Lett.*, vol. 11, no. 5, pp. 470 – 473, 2004.

[25] R. J. Muirhead, *Aspects of Multivariate Statistical Theory*. New York: Wiley, 1982, 704 pages.

[26] N. Delfosse and P. Loubaton, "Adaptive blind separation of independent sources: A deflation approach," *Signal Processing*, vol. 45, pp. 59–83, 1995.

[27] J.-F. Cardoso, "Blind signal separation: statistical principles," *Proceedings of the IEEE*, vol. 9, no. 10, pp. 2009–2025, 1998.

[28] A. Hyvärinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural Computation*, vol. 9, no. 7, pp. 1483–1492, 1997.

[29] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999.

[30] *http://www.cis.hut.fi/projects/ica/fastica*.

[31] J. F. Cardoso and A. Souloumiac, "Blind beamforming for non-Gaussian signals," *IEE Proceedings-F*, vol. 140, no. 6, pp. 362–370, 1993.

[32] H. Oja, S. Sirkiä, and J. Eriksson, "Scatter matrices and independent component analysis," *Austrian Journal of Statistics*, vol. 35, no. 1&2, pp. 175–189, 2006.

[33] P. J. Huber, "Projection pursuit," *Ann. Statist.*, vol. 13, no. 2, pp. 435–475, 1985.

[34] J. H. Friedman, "Exploratory projection pursuit," *J. Amer. Stat. Assoc.*, vol. 82, no. 397, pp. 249–266, 1987.

[35] A. Yeredor, "Blind source separation via the second characteristic function," *Signal Processing*, vol. 80, no. 5, pp. 897–902, 2000.

[36] V. Zarzoso and P. Comon, "Robust ICA for blind source separation and extraction with applications in electrocardiography," in *Proc. 30th Int. EMBS Conf.*, Vancouver, Canada, Aug. 2008.

[37] E. Ollila, "On the robustness of the deflation-based FastICA estimator," in *Proc. IEEE Workshop on Statistical Signal Processing (SSP'09)*, Cardiff, Wales, Aug. 31– Sep. 3, 2009, pp. 673–676.

[38] ——, "The deflation-based FastICA estimator: statistical analysis revisited," *IEEE Trans. Signal Processing*, no. 3, pp. 1527–1541, 2010.

[39] B. Hong, G. D. Pearlson, and V. D. Calhoun, "Source density-driven independent component analysis approach for fMRI data," *Human Brain Mapping*, vol. 25, no. 3, pp. 297 – 307, 2005.

[40] J. Karvanen, J. Eriksson, and V. Koivunen, "Pearson system based method for blind separation," in *Proc. Int. Conf. Independent Component Analysis (ICA'00)*, 2000, pp. 585–590.

[41] S. Douglas, "On the convergence behavior of the FastICA algorithm," in *Proc. Int. Conf. Independent Component Analysis (ICA 2003)*, Kyoto, Japan, 2003, pp. 409–414.

[42] E. Oja and Z. Yuan, "The FastICA algorithm revisited: Convergence analysis," *IEEE Trans. Neural Networks*, vol. 17, no. 6, pp. 1370 – 1381, 2006.

[43] H. Shen, M. Kleinsteuber, and K. Huber, "Local convergence analysis of FastICA and related algorithms," *IEEE Trans. Neural Networks*, vol. 19, no. 6, pp. 1022 – 1032, 2008.

[44] A. Hyvärinen, "One-unit contrast functions for independent component analysis: A statistical analysis," in *IEEE Neural Networks for Signal Processing (NNSP)*, Amelia Island, FL, 1997, pp. 388–397.

[45] S. Shimizu, A. Hyvarinen, K. Yutaka, P. Hoyer, and A. J. Kerminen, "Testing significance of mixing and demixing coefficients in ICA," in *Proc. Int. Conf. Independent Component Analysis (ICA 2006)*, Charleston, SC, USA, 2006, pp. 901 – 908.

[46] G. Brys, M. Hubert, and P. Rousseeuw, "A robustification of independent component analysis," *Journal of Chemometrics*, vol. 19, no. 5–7, pp. 364 – 375, 2006.

[47] K. Nordhausen, H. Oja, and E. Ollila, "Multivariate models and the first four moments," in *Festschrift in Honour of Professor Thomas P. Hettmansperger*. (to be published), 2010.

[48] J.-F. Cardoso and B. Laheld, "Equivariant adaptive source separation," *IEEE Trans. Signal Processing*, vol. 44, no. 12, pp. 3017 – 3030, 1996.

[49] J. Eriksson and V. Koivunen, "Characteristic-function-based independent component analysis," *Signal Processing*, vol. 83, pp. 2195–2208, 2003.

[50] L. Molgedey and H. G. Schuster, "Separation of a mixture of independent signals using time delayed correlations," *Physical Review Letters*, vol. 72, no. 23, pp. 3634 – 3637, 1994.

[51] A. Beloucharani, K. Abed-Merain, J. F. Cardoso, and E. Moulines, "A blind source separation technique using second-order statistics," *IEEE Trans. Signal Processing*, vol. 45, no. 2, pp. 434–444, 1997.

[52] D. T. Pham, "Joint approximate diagonalization of positive definite hermitian matrices," *SIAM Journal on Matrix Analysis and Applications*, vol. 22, no. 4, pp. 1136 – 1152, 2001.

[53] A. Yeredor, "Non-orthogonal joint diagonalization in the least-squares sense with application in blind source separation," *IEEE Trans. Signal Processing*, vol. 50, no. 7, pp. 1545 – 1553, 2002.

[54] P. J. Rousseeuw, *Mathematical Statistics and Applications*. Reidel, Dordrecht, 1985, ch. Multivariate Estimation With High Breakdown Point.

[55] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*. New York: Wiley, 1987.

[56] P. L. Davies, "Asymptotic behavior of S-estimates of multivariate location parameters and dispersion matrices," *Ann. Statist.*, vol. 15, pp. 1269–1292, 1987.

[57] H. P. Lopuhaä, "Multivariate τ-estimators for location and scatter," *Canadian Journal of Statistics*, vol. 19, pp. 307–321, 1991.

[58] J. T. Kent and D. E. Tyler, "Constrained M-estimation for multivariate location and scatter," *Ann. Statist.*, vol. 24, no. 3, pp. 1346–1370, 1996.

[59] K. S. Tatsuoka and D. E. Tyler, "On the uniqueness of S-functionals under non-elliptical distributions," *Ann. Statist.*, vol. 28, pp. 1219–1243, 2000.

[60] S. Visuri, V. Koivunen, and H. Oja, "Sign and rank covariance matrices," *J. Statist. Plann. Inference*, vol. 91, pp. 557–575, 2000.

[61] E. Ollila, H. Oja, and C. Croux, "The affine equivariant sign covariance matrix: asymptotic behavior and efficiencies," *J. Mult. Anal.*, vol. 87, no. 2, pp. 328–355, 2003.

[62] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, *Robust Statistics: The Approach Based on Influence Functions*. New York: Wiley, 1986, 536 pages.

[63] C. Croux, "Limit behavior of the empirical influence function of the median," *Statistics & Probability Letters*, vol. 37, pp. 331–340, 1998.

[64] S. C. Douglas, "Fixed-point algorithms for the blind separation of arbitrary complex-valued non-gaussian signal mixtures," *EURASIP J. Advances in Signal Processing*, vol. 2007, no. 1, pp. 83–83, 2007.

[65] S. I. Amari, A. Cichocki, and H. H. Yang, "A new learning algorithm for blind source separation," in *Advances in Neural Information Processing Systems 8*, D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, Eds. Cambridge, MA: MIT Press, 1996, pp. 757–763.

[66] D. Brillinger, *Time series: Data analysis and theory*. New York: Holt, Rinehart and Winston Inc., 1975.

[67] N. A. Lazar, *The Statistical Analysis of Functional MRI Data*. Springer, 2008, 299 pages.

[68] T. Adali and V. Calhoun, "Complex ICA of brain imaging data," *IEEE Signal Processing Magazine*, pp. 136–139, Sept. 2007.

[69] I. L. Dryden and K. V. Mardia, *Statistical shape analysis.* Chichester: Wiley, 1998.

[70] F. G. C. Hoogenraad, J. R. Reichenbach, E. M. Haacke, S. Lai, K. Kuppusamy, and M. Sprenger, "In vivo measurement of changes in venous blood-oxygenation with high resolution functional MRI at 0.95 tesla by measuring changes in susceptibility and velocity," *Magnetic Resonance in Medicine*, vol. 39, no. 1, pp. 97–107, 1998.

[71] D. B. Rowe and B. R. Logan, "A complex way to compute fMRI activation," *NeuroImage*, vol. 23, pp. 1078–1092, 2004.

[72] W. X. Xiong, Y. Li, H. Li, T. Adali, and V. D. Calhoun, "On ICA of complex-valued fMRI: advantages and order selection," in *Proc. Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP'08)*, 2008, pp. 529–532.

[73] S. Hoyos, Y. Li, J. Bacca, and G. Arce, "Weighted median filters admitting complex-valued weights and their optimization," *IEEE Trans. Signal Processing*, vol. 52, no. 10, pp. 2776 – 2787, 2004.

[74] D. H. Brandwood, "A complex gradient operator and its applications in adaptive array theory," *IEE Proc. F and H*, vol. 1, pp. 11–16, 1983.

[75] A. van den Bos, "A Cramér-Rao lower bound for complex parameters," *IEEE Trans. Signal Processing*, vol. 42, no. 10, p. 2859, 1994.

[76] ——, "Complex gradient and Hessian," *IEE Proc.-Vis. Image Signal Process.*, vol. 141, no. 6, pp. 380–382, 1994.

[77] B. Picinbono, "On circularity," *IEEE Trans. Signal Processing*, vol. 42, no. 12, pp. 3473–3482, 1994.

[78] P. Amblard, M. Gaeta, and L. Lacoume, "Statistics for complex variables and signals - part I: variables," *Signal Processing*, vol. 53, no. 1, pp. 1–13, 1996.

[79] B. Picinbono and P. Chevalier, "Widely linear estimation with complex data," *IEEE Trans. Signal Processing*, vol. 43, no. 8, pp. 2030–2033, 1995.

[80] P. J. Schreier and L. L. Scharf, "Second-order analysis of improper complex random vectors and processes," *IEEE Trans. Signal Processing*, vol. 51, no. 3, pp. 714–725, 2003.

[81] J. Eriksson and V. Koivunen, "Complex random vectors and ICA models: Identifiability, uniqueness and seperability," *IEEE Trans. Inform. Theory*, vol. 52, no. 3, pp. 1017–1029, 2006.

[82] A. Hjorungnes and D. Gesbert, "Complex-valued matrix differentiation: techniques and key results," *IEEE Trans. Signal Processing*, vol. 55, pp. 2740–2746, 2007.

[83] J. Eriksson, E. Ollila, and V. Koivunen, "Processing of complex random signals: fundamentals revisited," *IEEE Trans. Signal Processing*, 2009 (submitted).

[84] P. Schreier and L. Scharf, *Statistical Signal Processing of Complex-Valued Data: the theory of improper and non-circular signals.* Cambridge University Press, Feb. 2010.

[85] M. Valkama, M. Renfors, and V. Koivunen, "Blind signal estimation in conjugate signal models with application to I/Q imbalance compensation," *IEEE Signal Proc. Lett.*, no. 11, pp. 733 – 736, 2005.

[86] M. Haardt and F. Römer, "Enhancements of unitary ESPRIT for non-circular sources," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP'04)*, Montreal, Canada, May 2004.

[87] H. Abeida and J.-P. Delmas, "MUSIC-like estimation of direction of arrival for noncircular sources," *IEEE Trans. Signal Processing*, vol. 54, no. 7, pp. 2678–2690, 2006.

[88] S. F. Yau and Y. Bresler, "A compact Cramér-Rao bound expression for parametric estimation of superimposed signals," *IEEE Trans. Signal Processing*, vol. 40, pp. 1226–1230, 1992.

[89] S. M. Kay, *Fundamentals of Statistical Signal Processing.* New Jersey: Prentice-Hall, 1993.

[90] E. de Carvalho, J. Cioffi, and D. T. M. Slock, "Cramér-Rao bounds for blind multichannel estimation," in *Proc. IEEE Global telecommunications conference*, San Francisco, USA, Nov. 27 – Dec. 1, 2000.

[91] A. K. Jagannatham and B. D. Rao, "Cramér-Rao lower bound for constrained complex parameters," *IEEE Signal Proc. Lett.*, vol. 11, no. 11, pp. 875–878, 2004.

[92] S. T. Smith, "Statistical resolution limits and the complexified Cramér-Rao bound," *IEEE Trans. Signal Processing*, vol. 53, no. 5, pp. 1597 – 1609, 2005.

[93] L. De Lathauwer and B. De Moore, "On the blind separtion of non-circular sources," in *Proc. 11th European Signal Processing Conference(EUSIPCO 2002)*, Toulouse, France, Sept. 2002.

[94] J. Eriksson, A. M. Seppola, and K. V., "Complex ICA for circular and non-circular sources," in *Proc. 13th European Signal Processing Conference (EUSIPCO'05)*, Antalya, Turkey, 2005.

[95] M. Novey and T. Adali, "On extending the complex FastICA algorithm to non-circular sources," *IEEE Trans. Signal Processing*, vol. 56, no. 5, pp. 2148–2154, 2008.

[96] H. Li and T. Adali, "A class of complex ICA algorithms based on the kurtosis cost function," *IEEE Trans. Neural Networks*, vol. 19, no. 3, pp. 408–420, 2008.

[97] K. Kreutz-Delgado, "The complex gradient operator and the cr-calculus, Lecture Notes Supplement [online]," 2007.

[98] R. Remmert, *Theory of Complex Functions*. Springer, 1991.

[99] L. Ahlfors, *Complex analysis*. New York: McGraw-Hill, 1953.

[100] W. Kaplan, *Introduction to Analytic Functions*. Reading, MA: Addison-Wesley, 1966.

[101] S. G. Krantz, *Handbook of Complex Variables*. Boston: Birkhauser, 1999.

[102] S. Lang, *Complex Analysis*. New York: Springer-Verlag, 1999.

[103] K.-T. Fang, S. Kotz, and K. W. Ng, *Symmetric multivariate and related distributions*. London: Chapman and hall, 1990.

[104] F. D. Neeser and J. L. Massey, "Proper complex random processes with applications to information theory," *IEEE Trans. Inform. Theory*, vol. 39, no. 4, pp. 1293–1302, 1993.

[105] J.-P. Delmas and H. Abeida, "Asymptotic distribution of circularity coefficients estimate of complex random variables," *Signal Processing*, vol. 89, no. 12, pp. 2670–2675, 2009.

[106] E. Ollila, J. Eriksson, and V. Koivunen, "Complex elliptically symmetric random variables – generation, characterization and circularity tests," *IEEE Trans. Signal Processing*, submitted Dec., 2009.

[107] E. Ollila and V. Koivunen, "Robust estimation techniques for complex-valued random vectors," in *Adaptive Signal Processing: Next Generation Solutions*, T. Adali and S. Haykin, Eds. Wiley, 2010.

[108] R. A. Wooding, "The multivariate distribution of complex normal variables," *Biometrika*, vol. 43, pp. 212–215, 1956.

[109] N. R. Goodman, "Statistical analysis based on certain multivariate complex Gaussian distribution (an introduction)," *Annals Math. Statist.*, vol. 34, pp. 152–177, 1963.

[110] P. J. Schreier, L. L. Scharf, and A. Hanssen, "A generalized likelihood ratio test for impropriety of complex signals," *IEEE Signal Processing Letters*, vol. 13, no. 7, pp. 433–436, 2006.

[111] R. A. Horn and C. A. Johnson, *Matrix Analysis*. Cambridge: Cambridge University Press, 1985.

[112] C. G. Khatri and C. D. Bhavsar, "Some asymptotic inferential problems connected with complex elliptical distribution," *J. Mult. Anal.*, vol. 35, pp. 66–85, 1990.

[113] P. J. Schreier, L. L. Scharf, and C. T. Mullis, "Detection and estimation of improper complex random signals," *IEEE Trans. Inform. Theory*, vol. 51, no. 1, pp. 306–312, 2005.

[114] A. T. Walden and P. Rubin-Delanchy, "On testing for impropriety of complex-valued gaussian vectors," *IEEE Trans. Signal Processing*, vol. 57, no. 3, pp. 835–842, 2009.

[115] M. Novey, T. Adali, and A. Roy, "Circularity and Gaussianity detection using the complex generalized Gaussian distribution," *IEEE Signal Processing Lett.*, vol. 16, no. 11, pp. 993–996, 2009.

[116] D. E. Tyler, "Robustness and efficiency of scatter matrices," *Biometrika*, vol. 70, pp. 411–420, 1983.

[117] *http://cc.oulu.fi/∼esollila*.

[118] S. Visuri, *Array and multichannel signal processing using nonparametric statistics*. Espoo: Ph.D. Thesis, Helsinki University of Technology, 2000.

[119] P. Tsakalides and C. L. Nikias, "The robust covariation based MUSIC (roc-MUSIC) algorithm for bearing estimation in impulsive noise environments," *IEEE Trans. Signal Processing*, vol. 44, no. 7, pp. 1623–1633, 1995.

[120] S. Visuri, H. Oja, and V. Koivunen, "Subspace-based direction of arrival estimation using nonparametric statistics," *IEEE Trans. Signal Processing*, vol. 49, no. 9, pp. 2060–2073, 2001.

[121] R. J. Kozick and B. M. Sadler, "Maximum-likelihood array processing in non-Gaussian noise with Gaussian mixtures," *IEEE Trans. Signal Processing*, vol. 48, no. 12, pp. 3520–3535, 2000.

[122] ——, "Robust subspace estimation in non-Gaussian noise," in *Proc. IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP'00)*, Istanbul, Turkey, 2000, pp. 3818 – 3821.

[123] T.-H. Liu and J. M. Mendel, "A subspace-based direction finding algorithm using fractional lower order statistics," *IEEE Trans. Signal Processing*, vol. 49, pp. 1605–1613, 2001.

[124] Y. Yardimci, A. E. Cetin, and J. A. Cadzow, "Robust direction-of-arrival estimation in non-Gaussian noise," *IEEE Trans. Signal Processing*, vol. 46, no. 5, pp. 1443–1451, 1998.

[125] C.-H. Lim, S. C.-M. See, A. M. Zoubir, and B. P. Ng, "Robust adaptive trimming for high-resolution direction finding," *IEEE Signal Proc. Lett.*, vol. 16, no. 7, pp. 580 – 583, 2009.

[126] A. Gershman, "Robust adaptive beamforming: an overview of recent trends and advances in the field," in *Proc. 4th International Conference on Antenna Theory and Techniques*, Sept. 9–12, 2003, pp. 30–35.

[127] ——, "Robustness issues in adaptive beamforming and high-resolution direction finding," in *High-resolution and Robust Signal Processing*, Y. Hua, A. Gershman, and Q. Cheng, Eds. Marcel Dekker, 2003, 550 pages.

[128] L. Godara, *Smart Antennas*. CRC Press, 2004, 472 pages.

[129] H. Cox, R. M. Zeskind, and M. M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 35, no. 10, pp. 1365 – 1376, 1987.

[130] B. D. Carlson, "Covariance matrix estimation errors and diagonal loading in adaptive arrays," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 24, no. 4, pp. 397 – 401, 1988.

[131] J. Li, P. Stoica, and Z. Wang, "On robust Capon beamforming and diagonal loading," *IEEE Trans. Signal Processing*, vol. 51, no. 7, pp. 1702 – 1715, 2003.

[132] W. Zhang and B. D. Rao, "Robust broadband beamformer with diagonally loaded constraint matrix and its application to speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP'06)*, vol. I, 2006, pp. 785–788.

[133] J. T. Kent, "Data analysis for shapes and images," *J. Statist. Plann. Inference*, vol. 57, pp. 181–193, 1997.

[134] *http://wooster.hut.fi/∼esollila/MVDR.*

[135] B. Van Veen and K. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP magazine*, April 1988.

[136] J. Capon, "High resolution frequency-wavenumber spectral analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[137] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propagat.*, vol. 34, no. 3, pp. 276–280, 1986.

[138] P. Chargé, Y. Wang, and J. Saillard, "A non-circular sources direction finding methods using polynomial rooting," *Signal Processing*, vol. 81, pp. 1765–1770, 2001.

[139] P. Stoica and Y. Selen, "Model-order selection: a review of information criterion rules," *IEEE Signal Proc. Mag.*, vol. 21, no. 4, pp. 36 – 47, 2004.

[140] R. F. Brcich, A. M. Zoubir, and P. Pelin, "Detection of sources using bootstrap techniques," *IEEE Trans. Signal Processing*, vol. 50, no. 2, pp. 206–215, 2002.