Helsinki University of Technology Signal Processing Laboratory
Teknillinen korkeakoulu Signaalinkäsittelytekniikan laboratorio
Espoo 2004 Report 47

CONTRIBUTIONS TO THEORY AND ALGORITHMS OF INDEPENDENT COMPONENT ANALYSIS AND SIGNAL SEPARATION

Jan Eriksson

Dissertation for the degree of Doctor of Science in Technology to be presented with due permission of the Department of Electrical and Communications Engineering for public examination and debate in Auditorium S4 at Helsinki University of Technology (Espoo, Finland) on the 20th of August, 2004, at 12 noon.

Helsinki University of Technology
Department of Electrical and Communications Engineering
Signal Processing Laboratory

Teknillinen korkeakoulu
Sähkö- ja tietoliikennetekniikan osasto
Signaalinkäsittelytekniikan laboratorio

Distribution:
Helsinki University of Technology
Signal Processing Laboratory
P.O. Box 3000
FIN-02015 HUT
Tel. +358-9-451 3211
Fax. +358-9-452 3614
E-mail: Mirja.Lemetyinen@hut.fi

ISBN 951-22-7224-5 (Printed)
ISBN 951-22-7227-X (Electronic)
ISSN 1458-6401

Otamedia Oy
Espoo 2004

Abstract

This thesis addresses the problem of blind signal separation (BSS) using independent component analysis (ICA). In blind signal separation, signals from multiple sources arrive simultaneously at a sensor array, so that each sensor array output contains a mixture of source signals. Sets of sensor outputs are processed to recover the source signals or to identify the mixing system. The term blind refers to the fact that no explicit knowledge of source signals or mixing system is available. Independent component analysis approach uses statistical independence of the source signals to solve the blind signal separation problems. Application domains for the material presented in this thesis include communications, biomedical, audio, image, and sensor array signal processing.

In this thesis reliable algorithms for ICA-based blind source separation are developed. In blind source separation problem the goal is to recover all original source signals using the observed mixtures only. The objective is to develop algorithms that are either adaptive to unknown source distributions or do not need to utilize the source distribution information at all. Two parametric methods that can adapt to a wide class of source distributions including skewed distributions are proposed. Another nonparametric technique with desirable large sample properties is also proposed. It is based on characteristic functions and thereby avoids the need to model the source distributions. Experimental results showing reliable performance are given on all of the presented methods.

In this thesis theoretical conditions under which instantaneous ICA-based blind signal processing problems can be solved are established. These results extend the celebrated results by Comon of the traditional linear real-valued model. The results are further extended to complex-valued signals and to nonlinear mixing systems. Conditions for identification, uniqueness, and separation are established both for real and complex-valued linear models, and for a proposed class of non-linear mixing systems.

Preface

The work constituting this thesis was carried out in the Signal Processing Laboratory at Helsinki University of Technology during years 2000–2004. The research group is a member of SMARAD, Center of Excellence of the Academy of Finland.

I wish to express my deep gratitude to my supervisor Prof. Visa Koivunen for his continuous encouragement, guidance and support during the course of this work. I especially admire his excellent work ethic and dedication to the science. It has been a real pleasure to work with him.

I would like to thank the thesis pre-examiners, Prof. Christian Jutten and Dr. Arie Yeredor, for their constructive and helpful comments that improved the manuscript a lot. Prof. Iiro Hartimo, the director of Graduate School in Electronics, Telecommunications and Automation (GETA) and the GETA coordinator Marja Leppäharju are also highly acknowledged. Many thanks go to our laboratory secretaries Mirja Lemetyinen and Anne Jääskeläinen for the help with all the practical issues and arrangements.

I am grateful to all the colleagues in the lab and especially co-workers Dr. Annaliisa Kankainen and Dr. Juha Karvanen. Prof. Risto Wichman, Dr. Samuli Visuri, Dr. Mihai Enescu, Dr. Marius Sirbu, Dr. Esa Ollila, Maarit Melvasalo, Timo Roman and Traian Abrudan are also acknowledged for our many interesting discussions, not only research related. I would like to give my warmest thanks to my parents, Terttu and Trygve, and to my sister, Tiia, for all their support. I am also thankful to all my friends, in particular Jarkko and Ville, for the quality leisure time and memorable moments.

The financial support from Academy of Finland, GETA, Nokia Foundation and Jenny and Antti Wihuri Foundation is gratefully acknowledged.

Finally, the work put into this thesis was worthwhile due to Sanna, the foremost source of happiness in my life. Thank you so much.

Espoo, August 2004

Jan Eriksson

Contents

List of original publications

(I) J. Eriksson, J. Karvanen, and V. Koivunen. Source distribution adaptive maximum likelihood estimation of ICA model. In *Proc. of Second International Workshop on Independent Component Analysis and Blind Source Separation (ICA 2000)*, pp. 227–232, Helsinki, Finland, June 2000.

(II) J. Karvanen, J. Eriksson, and V. Koivunen. Pearson system based method for blind separation. In *Proc. of Second International Workshop on Independent Component Analysis and Blind Source Separation (ICA 2000)*, pp. 585–590, Helsinki, Finland, June 2000.

(III) J. Karvanen, J. Eriksson, and V. Koivunen. Adaptive Score Functions for Maximum Likelihood ICA. *Journal of VLSI Signal Processing*, 32: 83–92, 2002.

(IV) J. Eriksson and V. Koivunen. Blind separation using characteristic function based criterion. In *Proc. of 35th Conference on Information Sciences and Systems (CISS 2001)*, vol. 2, pp. 781–785, Baltimore, Maryland, March 2001.

(V) J. Eriksson, A. Kankainen, and V. Koivunen. Novel characteristic function based criteria for ICA. In *Proc. of Third International Conference on Independent Component Analysis and Blind Source Separation (ICA 2001)*, pp. 108–113, San Diego, California, December 2001.

(VI) J. Eriksson and V. Koivunen. Blind identifiability of class of nonlinear instantaneous ICA models. In *Proc. of the XI European Signal Processing Conference (EUSIPCO 2002)*, vol. 2, pp. 7–10, Toulouse, France, September 2002.

(VII) J. Eriksson and V. Koivunen. Characteristic-function-based independent component analysis. *Signal Processing*, 83: 2195-2208, 2003.

(VIII) J. Eriksson and V. Koivunen. Identifiability, separability, and uniqueness of linear ICA models. *IEEE Signal Processing Letters*, vol. 11, No. 7, pp. 601–604, July 2004.

(IX) J. Eriksson and V. Koivunen. Complex Random Vectors and ICA Models: Identifiability, Uniqueness and Separability. Submitted to *IEEE Transactions on Information Theory*, March 2004. Material presented also in the technical report: Identifiability, Separability, and Uniqueness of Complex ICA Models. *Signal Processing Laboratory Technical Report 44*, Helsinki University of Technology, 2004, ISBN 951-22-6774-8, ISSN 1458-6401.

Abbreviations and symbols

Abbreviations

BSP	blind signal processing
BSS	blind source/signal separation
c.f.	characteristic function
CHESS	characteristic function enabled source separation
d.f.	distribution function
e.c.f.	empirical characteristic function
EGLD	extended generalized lambda distribution
GBD	generalized beta distribution
GLD	generalized lambda distribution
i.i.d.	independent identically distributed
ICA	independent component analysis
JECFICA	Jacobi optimized empirical characteristic function ICA
m.g.f.	moment generating function
m.i.	mutual information
MIMO	multiple-input multiple-output
p.d.f.	probability density function
PCA	principal component analysis
PNL	post-nonlinear
r.v.	random variable
r.vc.	random vector
s.c.f.	second characteristic function

Symbols

$\boldsymbol{A}$, $\boldsymbol{B}$	mixing matrix
$\boldsymbol{\alpha}$	column of a matrix, real vector
a, b	scalar constant
$\mathbb{C}$	field of complex numbers
$\mathrm{E}_{(\cdot)}\{\cdot\}$	expectation operator
$F_{\bar{x}}$	distribution function of a random variable $\bar{x}$

$f_{\bar{x}}$	probability density function of a random variable $\bar{x}$
$\boldsymbol{\mathcal{F}}\{\cdot\}$	system operator
$\boldsymbol{\mathcal{F}}(\cdot)$, $\boldsymbol{\mathcal{G}}(\cdot)$, $\boldsymbol{\mathcal{H}}(\cdot)$	vector-valued function
$\mathcal{F}(\cdot)$, $\mathcal{G}(\cdot)$, $\mathcal{H}(\cdot)$	scalar-valued function
$\mathrm{H}(\bar{\boldsymbol{x}})$	entropy of a random vector $\bar{\boldsymbol{x}}$
$\boldsymbol{A}^H$	hermitian transpose of a matrix $\boldsymbol{A}$
$\boldsymbol{I}$	identity matrix
$\jmath$	imaginary unit, $\jmath^2 = 1$
k, l, n	discrete index
$\mathrm{KL}\big(f_{\bar{\boldsymbol{x}}}, f_{\bar{\boldsymbol{s}}}\big)$	Kullback-Leibler divergence between probability density functions $f_{\bar{\boldsymbol{x}}}$ and $f_{\bar{\boldsymbol{s}}}$
$\mathrm{MI}(\bar{\boldsymbol{x}})$	mutual information of a random vector $\bar{\boldsymbol{x}}$
m	number of sources
$\bar{\boldsymbol{n}}[\cdot]$	discrete-time noise signal
$\mathbb{N}$	set of natural numbers $0, 1, 2, \ldots$
p	number of mixtures
$\boldsymbol{P}$	permutation matrix
$\mathcal{P}(\cdot)$	polynomial
$\mathbb{R}$	field of real numbers
$\boldsymbol{s}[\cdot]$	discrete-time source signal
$\bar{\boldsymbol{s}}$, $\bar{\boldsymbol{r}}$	random vector (sources)
$\bar{s}$, $\bar{r}$	random variable (source)
$\boldsymbol{A}^T$	transpose of a matrix $\boldsymbol{A}$
$\boldsymbol{t}$, $\boldsymbol{x}$	real vector
t, x, y, u, v	real variable
$\boldsymbol{U}$, $\boldsymbol{V}$	orthonormal matrix
$\boldsymbol{W}$	separating matrix
$\boldsymbol{x}[\cdot]$	discrete-time mixture signal
$\bar{\boldsymbol{x}}$, $\bar{\boldsymbol{y}}$	random vector
$\bar{x}$, $\bar{y}$	random variable
$\boldsymbol{z}$	complex vector
$\boldsymbol{\Lambda}$	diagonal matrix
$\phi_{\bar{x}}$	characteristic function of a random variable $\bar{x}$
$\hat{\phi}_{\bar{x}}$	empirical characteristic function of a random variable $\bar{x}$
$\psi_{\bar{x}}$	second characteristic function of a random variable $\bar{x}$
$\varphi_{\bar{x}}$	score function of a random variable $\bar{x}$
$\boldsymbol{\varphi}_{\bar{\boldsymbol{x}}}$	vector of score functions of the marginals of a random vector $\bar{\boldsymbol{x}}$
$\nabla \boldsymbol{\mathcal{F}}$	gradient of a function $\boldsymbol{\mathcal{F}}$
$\nabla_{\mathrm{N}} \boldsymbol{\mathcal{F}}$	natural gradient of a function $\boldsymbol{\mathcal{F}}$

Chapter 1

Introduction

1.1 Motivation

Multichannel observations are often encountered in signal processing application areas. For example, there may be several microphones recording the sound waves in audio and speech signal processing, in biomedical signal processing there are several sensors measuring e.g. heart or brain activity, and in communication systems multiple antennas are receiving the communication signals. Multiple time series may be describing the same phenomenon, e.g. consumer confidence, in econometrics, and several images, e.g. in different wave lengths or from different angles, may be available from the same object in image processing. These measurements can be obtained over a period of time (e.g. in communications), or over a certain surface area (e.g. in image processing). Such measurements can be then discretized such that there is a collection of observation vectors, each vector representing all measurements from different sensors at a specific observation time instant or spatial location.

Often each of these several measurements is a *mixture* of some source signals. For instance, each communication antenna receives the superposition of all communication signals transmitted in the same carrier frequency, and each sensor measuring the brain activity records the brain activity signals coming from different areas of the brain. Although different sensors may be recording a mixture of the same source signals, the measurements may be different in all sensors due to the differences in e.g. positions of the sensors. For instance, each antenna receives each transmitted signal from a different angle with different amplitude depending the relative positioning of the transmitter and the antenna. Systems with multiple signal emitters and multiple sensors used for observing them are generally called multiple-input multiple-output (MIMO) systems. An example of a MIMO system is illustrated in Figure 1.1.

In *blind source separation* (BSS) the observations from a MIMO system

Figure 1.1: A MIMO system. Three signals produced by sources (speakers) are recorded by three sensors (microphones). Each microphone records slightly different mixture of speech due to the differences in the relative distances and angles between the mouths and the microphones, and due to the multipath propagation, i.e. the same waveform propagates via different paths.

are processed in order to recover initial multiple source signals. The term *blind* refers to the fact that there is no, or little, explicit information about the mixing process or about the source signals, and statistical or structrual properties of the signals are used instead. Similarly, in blind signal extraction the objective is to estimate a selective number of source signals with desired properties from a possibly large number of mixed original signals. In *blind system identification* the interest lies in estimating the mixing system instead of the original signals. All these and related blind estimation problems are generically called *blind signal processing* (BSP) problems [CA02].

Many BSP problems can be solved by stochastically modeling the source signals, i.e. modeling each observation vector as an outcome of a random process at a specific time instance, and assuming that the signals are statistically independent. This approach is known as *independent component*

analysis (ICA). Recent textbooks [CA02, HKO01, Hay00] provide a good introduction to ICA-based BSP.

1.2 Scope of the thesis

The purpose of this thesis is to further develop the theory of ICA-based blind signal processing, and derive algorithms for practical ICA-based blind source separation. This thesis considers only the situation where there is no time dependency in the mixing system, i.e. the system is instantaneous, nor any time dependency in the source signals themselves, i.e. each source signal is considered as an outcome of a random process that consist of independent identically distributed (i.i.d.) random variables. Some results on ICA-based methods utilizing time structure in the system, e.g. convolutive mixing, or in the source signals, e.g. autocovariance, can be found from [CA02, HKO01, Hay00].

The first goal of this thesis is to develop reliable algorithms for ICA-based blind source separation. The objective is to develop algorithms that are either adaptive to unknown source distributions or do not need to utilize the source distribution information at all. The algorithms should perform reliably for all type of source signals, and their performance should improve as there are more observations available (consistency).

The second goal of this thesis is to establish theoretical conditions under which instantaneous ICA-based BSP problems can be solved. This includes the traditional linear real-valued model and its extensions to complex-valued signals and to nonlinear mixing systems. The conditions should be established for identification, extraction, and separation type of BSP problems.

1.3 Contribution of the thesis

The contributions of this thesis include the following.

- Extended Generalized Lambda Distribution (EGLD) is proposed as an adaptive score function model in the ICA separation problem.
- The use of L-moments is proposed for the estimation of the Generalized Lambda Distribution (GLD) and the use of Pearson system in the ICA separation problem is introduced in co-operation with the co-authors.
- Characteristic function-based method is proposed for the ICA separation problem. The connection of the criterion to nonlinear correlations-based independence characterization is derived.
- Consistent characteristic function-based objective function is proposed for the ICA separation problem. Consistency being established in

[CB04]. Novel minimization algorithm for the consistent objective function is developed, and connection to the cumulant-based methods is established.

- The concept "uniqueness" is introduced to the ICA models. Separability conditions for the real-valued linear ICA model are extended to necessary conditions by removing earlier requirement of finite second-order moments. Conditions for identifiability and uniqueness are established.

- A class of nonlinear ICA models is introduced, and the separability of the class is established. A generic algorithm for separation of these models is proposed. A nonlinear ICA model is applied to an image enhancement problem.

- Conditions for separability, identifiability, and uniqueness of complex-valued ICA models are established. The condition for separability is shown to be sufficient and necessary.

- A characterization of second-order structure of circular and non-circular complex-valued random vectors, a decomposition of complex normal random vector, and extensions of the Darmois-Skitovich theorem to complex-valued random variables are established.

1.4 Summary of publications

This thesis consists of nine publications and a summary. The summary part of the thesis is organized as follows: Chapter 2 introduces the general ICA model and reviews the related probabilistic concepts and optimization algorithms. The theoretical conditions for linear real and complex-valued ICA models are summarized in Chapter 3. Chapter 4 introduces a class of nonlinear ICA models and reviews are related nonlinear model. Source adaptive ICA separation methods are reviewed and Pearson and EGLD-based methods are summarized in Chapter 5. Chapter 6 reviews characteristic function-based ICA methods and summarizes the JECFICA algorithm. Chapter 7 provides a brief summary and outlines future research.

In Publication I, a EGLD-based BSS method is introduced. An ICA method utilizing the method of moments for fitting the EGLD, the mutual information contrast, and the fixed point algorithm [H^+] is proposed. The good performance of the algorithm is demonstrated in simulations.

In Publication II, a Pearson system-based BSS method is introduced. An ICA method utilizing the method of moments for finding the parameters of the Pearson system, the mutual information contrast, and the fixed point algorithm [H^+] is proposed. The simulation examples demonstrate that the method can separate both super- and sub-Gaussian sources.

In Publication III the methods proposed in Publication I and in Publication II are further studied and compared. It is demonstrated in simulations that the standard BSS methods may perform poorly in the cases where the sources have asymmetric distributions. Due to source adaptation the EGLD and Pearson system based methods reliably separate the sources. Additionally the method of L-moments is proposed for the estimation of GLD parameters.

Publication IV proposes the use of characteristic functions to solve reliably the ICA-based BSS problem without need to model the source distributions. The method utilizing empirical characteristic functions, pairwise processing, and golden section search is proposed. The reliable performance is demonstrated in simulations.

Publication V improves the method of Publication IV by introducing the consistent objective functions. The consistency of such objective functions was recently shown in [CB04]. Additionally, a connection between the two dimensional criterion introduced in Publication IV and the independence characterization by nonlinear correlations is established.

Publication VI introduces a class of nonlinear ICA models. The conditions for separability are established, and a generic method for separation is proposed. An application to image enhancement is presented.

In Publication VII the characteristic function-based methods of Publication IV and of Publication V are further developed. A fast minimization of the two dimensional objective function based on Fourier series representation is proposed. A connection to cumulant-based methods is established, and criteria for noisy ICA are proposed. Further extensive simulations highlight the highly reliable performance of the proposed method.

The sufficient conditions for separability, identifiability, and uniqueness of real-valued linear ICA models are established in Publication VIII. The conditions for separability are proved to be necessary thereby extending the seminal results by Comon [Com94].

The sufficient conditions for separability, identifiability, and uniqueness of complex-valued linear ICA models are established in Publication IX. The conditions for separability are also found to be necessary. A novel characterization of the second-order structure of complex random vectors, and a decomposition of complex normal random vectors is derived. This is applied to derive the characteristic functions and the entropy of complex normal random vector. Also extensions of the Darmois-Skitovich theorem to complex-valued random variables are presented.

The author of this thesis proposed the use of EGLD system and derived the analytical results in Publication I. The experiments were performed and the writing was done in co-operation with co-authors.

In Publication II, the idea of using the Pearson system for the score function modeling is due to Juha Karvanen. Also the idea of using L-moments for estimating the EGLD parameters, proposed in Publication III, is due to

him. The implementation of the ICA method, the experiment designs, and writing were done by Jan Eriksson in co-operation with the co-authors.

The results in publications IV–IX were derived independently by the author of this thesis. Visa Koivunen contributed in steering the research, and helped in writing and structuring the publications. Annaliisa Kankainen, the second author of Publication V, helped providing useful references.

Chapter 2

Overview of ICA

2.1 Problem formulation and general assumptions

The general discrete-time *blind source separation* (BSS) system model is described by the input-output relationship

$$\boldsymbol{x}[k] = \boldsymbol{\mathcal{F}}\{\boldsymbol{s}[k], \bar{\boldsymbol{n}}[k]\}, \tag{2.1}$$

where k is the discrete time index, $\boldsymbol{x}[k]$ is the observed multidimensional *mixture signal* and $\bar{\boldsymbol{n}}[\cdot]$ is the *noise*. The goal of BSS is to reconstruct the original *source signal* $\boldsymbol{s}[k]$ and/or identify the system operator $\boldsymbol{\mathcal{F}}\{\cdot\}$ from the observed mixture signal without explicit knowledge of the source signals nor the system operator. The term blind is used to distinguish the problem from the cases, where there is some explicit information available, e.g. a known training source signal.

The general problem of Eq. (2.1) is naturally ill-posed, and further constraints are needed. A common approach is to model the signals stochastically, that is, to view the source signal as a realization of a random process. Then one can exploit statistical properties of the processes e.g. in time or in space to obtain the goal. A relatively recent approach is to try to obtain the solution based on the assumption that the source signals are stochastically mutually independent. This approach is known as *independent component analysis* (ICA). If the system is memoryless, the model is called *instantaneous* ICA, and the problems where noise is taken into account the model is known as *noisy* ICA. If the signals have no time structure, then they are typically assumed to be vector random processes of independent identically distributed (i.i.d.) random vectors (r.vc.'s). Thus the general *instantaneous time-independent noiseless ICA* model is described by the equation

$$\bar{\boldsymbol{x}} = \boldsymbol{\mathcal{F}}(\bar{\boldsymbol{s}}), \tag{2.2}$$

where $(\bar{s}_1, \ldots, \bar{s}_m)^T = \bar{\boldsymbol{s}}$ are unknown mutually independent non-degenerate random variables (r.v.'s) called *sources*, $\boldsymbol{\mathcal{F}}(\cdot)$ is an unknown *mixing function*,

and $\bar{\boldsymbol{x}} = (\bar{x}_1, \ldots, \bar{x}_p)^T$ are *mixtures*, i.e., the observed r.vc. (sensor array output). In the traditional setting, the mixing functions are constrained to be matrices, i.e. linear mappings, and the ICA model is described by the equation

$$\bar{\boldsymbol{x}} = \boldsymbol{A}\bar{\boldsymbol{s}}, \tag{2.3}$$

and the model is known as *linear ICA model*. When this property does not hold, the corresponding model is naturally called *nonlinear ICA model*. It should be noted that the methods to solve linear ICA model-based BSP problems usually require *nonlinear methods*, e.g. higher order statistics.

The models with the form (2.2) are studied in this thesis. The ICA models with system memory, e.g. convolutive mixtures, have also found some applications, for instance in audio and speech separation and wireless communications. In convolutive mixtures, the system operator $\boldsymbol{\mathcal{F}}$ is multidimensional linear time-invariant filter (usually also with finite impulse response). For recent reviews of such ICA models, ICA models with time structured processes, or with noise, the reader is referred to the recent textbooks [CA02, Chapters 8 and 9][HKO01, Chapters 15 and 18] [Hay00, Chapter 8 and 9].

2.2 Anatomy of ICA methods

An ICA *method* is an algorithm which, given a realization $\boldsymbol{x}[\cdot]$ of the process (2.2), estimates or learns the system function $\boldsymbol{\mathcal{F}}(\cdot)$ and/or the source signals $\boldsymbol{s}[\cdot]$. Since ICA is based on the crucial assumption on mutual independence of sources, the first step in derivation of an ICA method is to formulate a *criterion* which serves as a measure of independence. After that there is essentially two ways to proceed, as noted for instance in [APJ03]:

1. Derive an algorithm to optimize the criterion with respect to the space of mixing functions, and then replace the theoretical quantities in the optimization algorithm by their estimates obtained from the mixture signal $\boldsymbol{x}[\cdot]$.

2. Derive an *objective* function, which estimates the criterion from the mixture signal $\boldsymbol{x}[\cdot]$, and then derive an algorithm to optimize the objective function with respect to the space of mixing functions.

The most of commonly used ICA methods as well as the methods proposed in Chapter 5 of this thesis are of the first type whereas the methods proposed in Section 6.3 belong to the second class. The difference between the two approaches comes from the optimizing algorithms which usually include differential operators. Thus the first approach places smoothness constraints on the criterion thereby possibly restricting the class of allowed r.v.'s. On the other hand, this may lead to easier computations. In many cases, however, the two approaches may lead to the same final method.

From the optimization point of view, it is convenient that a criterion attains its maximum if and only if the components of a r.vc. are mutually independent. Furthermore the criterion should be invariant with respect to permutation and scaling. Such criteria are generally called *contrast functions* [Com94]. Finally, ICA methods that work on the entire data set are called *batch* methods, whereas the methods that work sequentially one sample at the time are referred to as *on-line* methods.

2.3 Measuring stochastic independence

The stochastic independence is a property of the underlying probability space. Since in ICA modeling one is not working directly with the probability space but with r.v.'s (i.e. mappings of the space), the definition of independence should be given in terms of functions of r.vc.'s in order to be useful for formulating a criterion. The definition is usually given in terms of *distribution functions* (d.f.'s) as follows. A r.vc. $\bar{\boldsymbol{x}}$ has independent components if and only if its joint d.f. $F_{\bar{\boldsymbol{x}}}$ factors to a product of the marginal d.f.'s $F_{\bar{x}_k}$, i.e.

$$F_{\bar{\boldsymbol{x}}}(\boldsymbol{t}) = \prod_{k=1}^{p} F_{\bar{x}_k}(t_k) \tag{2.4}$$

for all $\boldsymbol{t} = (t_1, \ldots, t_p)^T \in \mathbb{R}^p$. This definition could be used [BZP00] as a basis of forming a criterion. However, finding the factorization of Eq. (2.4) is numerically cumbersome. Therefore, alternative characterizations of independence are considered in this section.

2.3.1 Correlation

A natural starting point in finding independent components is to require that r.v.'s are uncorrelated, i.e.

$$\mathrm{E}_{(\bar{s},\bar{r})^T}\big\{(\bar{s} - \mathrm{E}_{\bar{s}}\big\{\bar{s}\big\})(\bar{r} - \mathrm{E}_{\bar{r}}\big\{\bar{r}\big\})\big\} = 0, \tag{2.5}$$

where $\mathrm{E}_{\bar{\boldsymbol{s}}}\big\{\cdot\big\}$ denotes the expectation with respect to r.vc. $\bar{\boldsymbol{s}}$ and T denotes the transpose of a vector. Thus, the *covariance matrix* measures the pairwise linear correlation. Correlations are quantified by off-diagonal elements of the covariance matrix. Although uncorrelateness implies independence for jointly Gaussian r.vc.'s, r.vc.'s in general can have uncorrelated but dependent marginals. However, uncorrelateness is a necessary condition for independence whenever the covariance exists (i.e. for r.v.'s with finite variance).

Any r.vc. with finite marginal variances can be linearly transformed such that the resulting r.vc. has uncorrelated components with equal (unit) variance. Such a r.vc. is called *white*, and procedure of making a r.vc.

white is referred as *whitening transform*. This is essentially the same as performing the well-known *principal component analysis* (PCA) or the discrete Karhunen-Lóeve transform in signal processing jargon. It is easily shown that whitening a r.vc. in a linear ICA structure (Eq. (2.3) with at least as many mixtures as sources results to a r.vc. with *orthonormal* mixing of the sources. Consequently the matrix $\boldsymbol{A}$ in Eq. (2.3) becomes orthonormal. Since by the singular value decomposition any matrix $\boldsymbol{A}$ can be decomposed as $\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^T$, where $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthonormal matrices and $\boldsymbol{\Lambda}$ is a diagonal matrix, PCA reduces the number of unknown system parameters to half. This technique is used in many practical ICA algorithms, and it can be said that "PCA reduces approximately the number of unknown systems parameters to half". For details and algorithms performing the whitening transform in ICA context, see [HKO01, Chapter 6][CA02, Chapter 3].

Independence can be guaranteed with *nonlinear correlations*. Two r.v.'s $\bar{s}$ and $\bar{r}$ are independent if and only if

$$\mathrm{E}_{(\bar{s},\bar{r})^T}\big\{\mathcal{F}(\bar{s})\mathcal{G}(\bar{r})\big\} = \mathrm{E}_{\bar{s}}\big\{\mathcal{F}(\bar{s})\big\}\,\mathrm{E}_{\bar{r}}\big\{\mathcal{G}(\bar{r})\big\} \tag{2.6}$$

for all functions $\mathcal{F}$ and $\mathcal{G}$ ranging over a *separating class of functions* [Bre92, Feu93]. A well-known separating class consists of the functions $\cos(tx), \sin(tx),\ t \geq 0$. This requires infinite number of correlation values which is impractical. However, one may try to rely on a few well-chosen functions and hope that they guarantee independence. Incidentally, the pioneering work [Jut87] in ICA was essentially based on this idea (see [HKO01, Chapter 12.2]).

2.3.2 Mutual Information

If it assumed that the source r.v.'s have d.f.'s which are absolutely continuous with respect to Lebesque measure, i.e. r.v.'s with the *probability density function* (p.d.f.), the condition for independence may be also stated as

$$f_{\bar{\boldsymbol{x}}}(\boldsymbol{t}) = \prod_{k=1}^{p} f_{\bar{x}_k}(t_k) \tag{2.7}$$

for all $\boldsymbol{t} \in \mathbb{R}^p$ by employing the joint p.d.f. $f_{\bar{\boldsymbol{x}}}$ and the marginal p.d.f.'s $f_{\bar{x}_k}$. Since this is again a functional relationship, one might try to average the function values over $\mathbb{R}^p$. This leads to the expression

$$\mathrm{MI}(\bar{\boldsymbol{x}}) \triangleq \int f_{\bar{\boldsymbol{x}}}(\boldsymbol{t}) \log \frac{f_{\bar{\boldsymbol{x}}}(\boldsymbol{t})}{\prod_{k=1}^{p} f_{\bar{x}_k}(t_k)} d\boldsymbol{t} = \mathrm{E}_{\bar{\boldsymbol{x}}}\big\{\log \frac{f_{\bar{\boldsymbol{x}}}(\bar{\boldsymbol{x}})}{\prod_{k=1}^{p} f_{\bar{x}_k}(\bar{x}_k)}\big\} \tag{2.8}$$

known as the *mutual information* (m.i.), which is the *Kullback-Leibler divergence* [CT91]

$$\mathrm{KL}\big(f_{\bar{\boldsymbol{x}}}, f_{\bar{\boldsymbol{s}}}\big) \triangleq \int f_{\bar{\boldsymbol{x}}}(\boldsymbol{t}) \log \frac{f_{\bar{\boldsymbol{x}}}(\boldsymbol{t})}{f_{\bar{\boldsymbol{s}}}(\boldsymbol{t})} d\boldsymbol{t} = \mathrm{E}_{\bar{\boldsymbol{x}}}\big\{\log \frac{f_{\bar{\boldsymbol{x}}}(\bar{\boldsymbol{x}})}{f_{\bar{\boldsymbol{s}}}(\bar{\boldsymbol{x}})}\big\}$$

between the joint p.d.f. and the product of the marginal p.d.f.'s. Mutual information is also the difference between the sum of entropies of the marginals and the *entropy* of a r.vc. $\bar{\boldsymbol{x}}$

$$\mathrm{H}(\bar{\boldsymbol{x}}) \triangleq -\int f_{\bar{\boldsymbol{x}}}(\boldsymbol{t}) \log f_{\bar{\boldsymbol{x}}}(\boldsymbol{t}) d\boldsymbol{t} = -\mathrm{E}_{\bar{\boldsymbol{x}}}\big\{\log f_{\bar{\boldsymbol{x}}}(\bar{\boldsymbol{x}})\big\}, \tag{2.9}$$

i.e.

$$\mathrm{MI}(\bar{\boldsymbol{x}}) = \mathrm{KL}\big(f_{\bar{\boldsymbol{x}}}, \prod_{k=1}^{p} f_{\bar{x}_k}\big) = \sum_{k=1}^{p} \mathrm{H}(\bar{x}_k) - \mathrm{H}(\bar{\boldsymbol{x}}). \tag{2.10}$$

Kullback-Leibler divergence is a semi-distance, and therefore mutual information is always non-negative and zero only for independent marginals, i.e. $-\mathrm{MI}(\cdot)$ is a contrast function [Com94]. For further properties, see [CT91, Gra90]. It should be also recognized that Kullback-Leibler divergence can be defined more generally [Kul68, Gra90], but this is rarely used in the BSS context.

2.3.3 Characteristic function, generating functions, moments and cumulants

A *characteristic function* (c.f.) of a p-dimensional r.vc. $\bar{\boldsymbol{x}}$ is defined (see [Ush99, Cup75, LO77]) as

$$\phi_{\bar{\boldsymbol{x}}}(\boldsymbol{t}) \triangleq \int_{\mathbb{R}^p} e^{\jmath <\boldsymbol{t},\boldsymbol{x}>} dF_{\bar{\boldsymbol{x}}}(\boldsymbol{x}) = \mathrm{E}_{\bar{\boldsymbol{x}}}\big\{e^{\jmath <\boldsymbol{t},\bar{\boldsymbol{x}}>}\big\},$$

where $\boldsymbol{t} = (t_1, \ldots, t_p)^T \in \mathbb{R}^p$, $<\cdot,\cdot>$ denotes the standard vector inner product, and $\jmath$ is the imaginary unit, $\jmath^2 = -1$. Characteristic functions are connected to d.f.'s via Fourier transform, they always exist, and are unique, i.e. there is one-to-one correspondence between c.f.'s and d.f.'s. Therefore by (2.4), a r.vc. $\bar{\boldsymbol{x}}$ has independent components if and only if the joint c.f. $\phi_{\bar{\boldsymbol{x}}}(\boldsymbol{t})$ factorizes to the product of the marginal c.f.'s, i.e. for all $\boldsymbol{t} = (t_1, \ldots, t_p)^T \in \mathbb{R}^p$,

$$\phi_{\bar{\boldsymbol{x}}}(\boldsymbol{t}) = \prod_{k=1}^{p} \phi_{\bar{x}_k}(t_k), \tag{2.11}$$

where $\phi_{\bar{x}_k}$ denotes the c.f. of the kth marginal distribution. Thus dependence may be quantified by deviations of the difference

$$\Delta_{\phi_{\bar{\boldsymbol{x}}}}(\boldsymbol{t}) \triangleq \phi_{\bar{\boldsymbol{x}}}(\boldsymbol{t}) - \prod_{k=1}^{p} \phi_{\bar{x}_k}(t_k) \tag{2.12}$$

from zero. The independence characterization (2.11) is also obtained from the nonlinear correlation characterization (2.6), since complex exponentials form a separating class of functions [Bre92].

The c.f. $\phi_{\bar{x}}$ of a r.v. $\bar{x}$ is said to be *analytic* [Luk70, Ush99], if there exists a function of the complex variable such that the function is analytic in a neighborhood of zero, and the restriction of the complex function to the real line coincides with $\phi_{\bar{x}}$ in the neighborhood. This is extended naturally to c.f.'s of r.vc.'s. Analytic c.f.'s have the appealing property that all c.f. factors of an analytic c.f. are also analytic c.f.'s. The r.v.'s $\bar{x}$ with analytic c.f.'s are exactly those r.v.'s for which the *moment generating function* (m.g.f.) $M_{\bar{x}}(t) \triangleq \mathrm{E}_{\bar{x}}\{e^{t\bar{x}}\}$ exists, i.e. $M_{\bar{x}}(t)$ is convergent for all t belonging to an interval containing the origin. The *moments* $\mathrm{E}_{\bar{x}}\{\bar{x}^k\}$ of a r.v. $\bar{x}$ with the m.g.f. $M_{\bar{x}}(t)$ are finite for all $k \in \mathbb{N}$, and they are obtained as the kth derivative of $M_{\bar{x}}$ evaluated at zero. Independence can be also characterized with m.g.f.'s by an identity analogous to (2.11).

Since a c.f. is continuous and equals unity at zero, there exists a neighborhood of the origin where the principal branch of logarithm of the c.f. is uniquely defined. The function $\psi_{\bar{\boldsymbol{x}}} \triangleq \log \phi_{\bar{\boldsymbol{x}}}$ is called the *second characteristic function* (s.c.f.) of a r.vc. $\bar{\boldsymbol{x}}$. A necessary condition for independence by Eq. (2.11) is given by the identity

$$\psi_{\bar{\boldsymbol{x}}}(\boldsymbol{t}) = \sum_{k=1}^{p} \psi_{\bar{x}_k}(t_k) \tag{2.13}$$

for all $\boldsymbol{t}$ in the neighborhood of the origin where the s.c.f.'s are defined. If two analytic c.f.'s coincide for all $\boldsymbol{t}$ in a neighborhood of the origin, they are the same. Thus the condition (2.13) is also sufficient for r.vc.'s with analytic c.f.'s. The coefficients of the differentials at $\boldsymbol{t} = \boldsymbol{0}$ of s.c.f. are called *cumulants*. Because of this property s.c.f. is sometimes called the cumulant generating function. However, this is somewhat misleading, since s.c.f. exists even when the cumulants do not. Therefore, it is better to reserve the name *cumulant generating function* for the logarithm of m.g.f. From Eq. (2.13) it may be seen that the cross-cumulants should vanish for a r.vc. with independent marginals.

2.3.4 Complex random variables

In many applications signals are complex-valued in Eq. (2.1). Therefore, for these type of applications, complex-valued random quantities are needed in Eq. (2.2). A p-variate *complex r.vc.* $\bar{\boldsymbol{x}}$ is defined as a r.vc. of the form

$$\bar{\boldsymbol{x}} = \bar{\boldsymbol{x}}_R + \jmath \bar{\boldsymbol{x}}_I, \tag{2.14}$$

where $\bar{\boldsymbol{x}}_R$ and $\bar{\boldsymbol{x}}_I$ are p-variate real r.vc.'s. Due to the separability of the complex space, the probabilistic structure of the r.vc. $\bar{\boldsymbol{x}}$ of (2.14) is equivalent to the structure of $2p$-variate real r.vc. $(\bar{\boldsymbol{x}}_R^T \; \bar{\boldsymbol{x}}_I^T)^T$. Therefore, the

probabilistic tools for complex r.vc.'s can be defined through this equivalence. For instance, the c.f. of the complex r.vc. $\bar{\boldsymbol{x}}$ is given as

$$\phi_{\bar{\boldsymbol{x}}}(\boldsymbol{z}) \triangleq \phi_{\bar{\boldsymbol{x}}_{\mathbb{R}}}(\boldsymbol{z}_{\mathbb{R}}) = \mathrm{E}_{\bar{\boldsymbol{x}}_{\mathbb{R}}}\big\{\exp\big(\jmath < \boldsymbol{z}_{\mathbb{R}}, \bar{\boldsymbol{x}}_{\mathbb{R}} >\big)\big\} = \mathrm{E}_{\bar{\boldsymbol{x}}}\big\{\exp\big(\jmath\,\mathrm{Re}\big\{< \boldsymbol{z}, \bar{\boldsymbol{x}} >\big\}\big)\big\}, \tag{2.15}$$

where $\boldsymbol{z} \in \mathbb{C}^p$ and the operator $()_{\mathbb{R}}$ gives for a p-variate complex vector the $2p$-dimensional real representation described above.

Although the probabilistic structure of the complex r.vc.'s can be easily described by their real counterparts, the operational structure can not. This is due to the fact that the p-dimensional complex space is not equivalent to the $2p$-dimensional real space as an inner product space. Thus r.vc.'s with values in complex space have distinct properties, and they need to be studied separately from their real counterparts. Some work on complex r.vc.'s can be found from [AGL96a, AGL96b, NM93, Pic96, PB97, SS03, VK96].

The difference between real and complex r.vc.'s is evident already from the second-order properties studied in Publication IX. Since the real normal r.vc.'s are fully specified by the mean and the second-order structure, this leads to the characterization of complex normal r.vc.'s. A complex r.vc. $\bar{\boldsymbol{x}}$ is naturally called *complex normal* if $\bar{\boldsymbol{x}}_{\mathbb{R}}$ is multivariate real normal r.vc. Let $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)^T$ be a vector such that $1 \geq \lambda_1 \geq \cdots \geq \lambda_p \geq 0$, and let $\mathrm{diag}\big(\boldsymbol{\lambda}\big)$ denote the $p \times p$ square matrix with $\boldsymbol{\lambda}$ on its main diagonal and zeros elsewhere. Now it is shown in Publication IX that any complex normal r.vc. $\bar{\boldsymbol{n}}$ with mean $\boldsymbol{\mu}$ can be decomposed as

$$\bar{\boldsymbol{n}} = \boldsymbol{C}(\bar{\boldsymbol{n}}_R + \jmath\bar{\boldsymbol{n}}_I) + \boldsymbol{\mu}, \tag{2.16}$$

where $\boldsymbol{C}$ is a nonsingular complex matrix and independent zero mean multinormal real r.vc.'s $\bar{\boldsymbol{n}}_R$ and $\bar{\boldsymbol{n}}_I$ have the covariance matrices $\mathrm{diag}\big(\frac{1+\boldsymbol{\lambda}}{2}\big)$ and $\mathrm{diag}\big(\frac{1-\boldsymbol{\lambda}}{2}\big)$ (respectively) for some $\boldsymbol{\lambda}$. The vector $\boldsymbol{\lambda}$ is called the *spectrum* of the r.vc. $\bar{\boldsymbol{n}}$, and the elements of a spectrum $\boldsymbol{\lambda}$ are called *spectral coefficients*. The decomposition (2.16) together with the spectrum $\boldsymbol{\lambda}$ allows a complete characterization of complex normal r.vc.'s, and an easy derivation of some fundamental theorems. For instance, the entropy (see Eq. (2.9)) of a complex normal r.vc. $\bar{\boldsymbol{n}}$ is given as (see Publication IX)

$$\mathrm{H}(\bar{\boldsymbol{n}}) = \mathrm{H}(\bar{\boldsymbol{n}}_{\mathbb{R}}) = \log\big(\det(\pi e \boldsymbol{C}\boldsymbol{C}^H)\big) + \frac{1}{2}\sum_{k=1}^{p}\log(1-\lambda_k^2),$$

where $\boldsymbol{C}$ is given by Eq. (2.16), $\boldsymbol{C}\boldsymbol{C}^H = \mathrm{E}_{\bar{\boldsymbol{n}}}\big\{(\bar{\boldsymbol{n}} - \mathrm{E}_{\bar{\boldsymbol{n}}}\big\{\bar{\boldsymbol{n}}\big\})(\bar{\boldsymbol{n}} - \mathrm{E}_{\bar{\boldsymbol{n}}}\big\{\bar{\boldsymbol{n}}\big\})^H\big\}$ is the *covariance matrix* of the r.vc. $\bar{\boldsymbol{n}}$, and $(\lambda_1, \ldots, \lambda_p)^T$ denotes its spectrum.

The spectrum is an inherited property of any complex r.vc. $\bar{\boldsymbol{x}}$ with second order statistics. Indeed, it can be shown that the spectrum is invariant with respect to nonsingular linear transformations. There always exists a nonsingular transformation $\boldsymbol{C}$ such that the r.vc. $\bar{\boldsymbol{y}} = \boldsymbol{C}\bar{\boldsymbol{x}}$ has the covariance matrix equal to the identity matrix and the *pseudo-covariance matrix*

$\mathrm{E}_{\bar{\boldsymbol{y}}}\{(\bar{\boldsymbol{y}} - \mathrm{E}_{\bar{\boldsymbol{y}}}\{\bar{\boldsymbol{y}}\})(\bar{\boldsymbol{y}} - \mathrm{E}_{\bar{\boldsymbol{y}}}\{\bar{\boldsymbol{y}}\})^T\}$ is a diagonal matrix with the spectrum in its diagonal. Such a transform is called a *strong-uncorrelating transform*. If the spectrum of a r.vc. is distinct, then the strong uncorrelating transform is essentially unique. For the proofs of these statements, see Publication IX.

If the spectrum of a r.vc. $\bar{\boldsymbol{x}}$ consists of zeros, the r.vc. is called *circular* (or proper). Most of the result concerning complex r.vc.'s in the literature are derived for circular r.vc.'s. Therefore, the results allowing any spectrum contain the results for circular r.vc.'s as special cases.

The characterization of the complex normal r.vc. and the second order statistics is used in Publication IX to extend some independence related theorems of the real-valued r.vc.'s to the complex-valued r.vc.'s. These theorems, including the extension of the celebrated Darmois-Skitovich theorem [KLR73], allow the extension of results concerning real-valued ICA models to be extended to the complex field. This is done in Section 3.2.

2.3.5 Discussion

It is important to realize that although all the expressions of independence in this section are necessary conditions for independence, only the conditions (2.4), (2.6), and (2.11) are also *sufficient for all r.v.'s*. This means that if other expressions are used, then there exist r.vc.'s such that either the condition is satisfied but the marginals are dependent or the expression is not defined. If such an expression, e.g. mutual information, is used as the basis for an ICA criterion, the corresponding methods, e.g. those considered in Chapter 5, are only guaranteed to work with the subset of r.v.'s for which the expression is also sufficient for independence. Consequently, such ICA methods are limited to a subclass of problems. In the case of mutual information, the model is limited to r.v.'s that have p.d.f.'s. Moreover, these quantitities may be hard to estimate. For instance, there does not even exist an umbiased estimator for mutual information [Pan03].

Although the independence related expressions remain essentially the same for complex-valued r.vc.'s, the behavior of these quantities is different from their real counterparts due to the difference in multiplication structure. Therefore, the properties of functions of complex r.vc.'s should be studied separately. The second-order properties of complex r.vc.'s are studied in Publication IX. Additionally, some well-known theorems for real r.vc.'s are extended to complex-valued r.vc.'s. These theorems cover both circular and non-circular complex r.vc.'s. The publication is also referred for references for literature about complex r.vc.'s.

2.4 Review of optimization algorithms

Every ICA method needs an optimization method for minimizing or maximizing the employed ICA criterion. In these optimization techniques there

is nothing specifically ICA related, although some new development has taken place while researching the BSS problem. Optimization algorithms have been studied for a long time, and there are plenty of algorithms (e.g. [Lue73, PTVF92, GL89]) to choose from. Thus only the basic methods used in ICA context (for details, see [Hay00, HKO01, CA02]) are reviewed in this Section.

2.4.1 Gradient descent methods

The most classical method of minimizing a function with respect to parameter vector $\boldsymbol{\alpha}$ is *gradient descent*. It is based on the observation that the gradient $\nabla\mathcal{F}$ of the function $\mathcal{F}$ gives the direction of the greatest increase of the function value. Thus the gradient descent algorithm starts with an initial guess, and calculates the gradient at this point. Then the parameter vector is updated to the opposite direction of the gradient by a small amount (step size). The gradient is evaluated at the new parameter value, the parameter vector is updated, and so on. This is repeated until the minimum is found. The algorithm has a direct generalization to matrix valued parameters with the matrix gradient (see e.g. [HKO01, Chapter 3]). The gradient descent algorithm for functions involving r.v.'s is called *stochastic gradient descent* [Hay96, p. 11], since the true gradient depends on the unknown distribution, and the corresponding method is an approximate or a stochastic implementation of the true descent procedure. The ICA methods proposed in Chapter 5 use enhanced versions of this type of optimization idea to be described in the following.

The main problem with the steepest descent method is the selection of the correct step size, which is crucial for the convergence even to a local minimum as well as for the convergence speed. This problem can be avoided by the use of the inverse of second derivative of the function, i.e. the inverse of the Hessian in the multidimensional parameter case. Such an algorithm is known as *Newton's method*. Since the direct calculation of the inverse of the Hessian may be computationally demanding, there are a number of algorithms that make a trade-off between the convergence speed and computational complexity. A popular ICA method known as *FastICA* [HO97, H$^+$] uses quasi-Newton optimization specifically tailored to the ICA model (for different versions and details, see [HKO01, Chapter 8]).

The ordinary gradient gives the direction of the deepest ascent for parameters in the Euclidean orthogonal coordinate system. However, this is not true if the coordinate system is non-Euclidean. For example, this is the case when the parameter space is the space of all invertible matrices or the space of all orthonormal matrices. If the parameter space is Riemannian, the deepest ascent direction is given by *natural gradient* ∇_{N} [ACY96] or *relative gradient* [CL96] first introduced in [CUR94]. The corresponding gradient descent optimization algorithm is known as natural gradient learning. In-

Figure 2.1: The shortest path between Vardø, Norway, and Sagres, Portugal, on the map (dotted line) and in reality (solid arc). The gradients towards the shortest path are given using the map geometry (Euclidean) and the true geometry (Riemannian).

tuitive difference between natural and conventional gradients is understood by the well-known map analogy: the shortest path between two places on earth is given by an arc on a map. The natural gradient gives the direction of the arc while the ordinary gradient gives the direction of the straight line between the two points on the map. See Figure 2.1 for an illustration. A detailed discussion of natural gradient methods for different type of ICA and BSS problems can be found from [CA02, Chapter 6][Hay00, Chapters 2 and 3]. See [Man02, Dou00, DAK00, EAS98] for related algorithms with orthogonality constraints and for further references.

2.4.2 Jacobi algorithms

Another popular optimization method type in ICA context are algorithms, which are based on a classical idea in multivariate numerical analysis [GL89]: instead of optimizing directly with respect to all dimensions of data, one optimizes the function iteratively in a *pairwise* manner. This means that instead of hopefully going directly towards the global optimum, one finds

the optimum of two dimensional data, transforms the data accordingly, and hopes that by iterating this several times to each pair finally leads to the global optimum. This is especially useful approach if the parameter space consists of orthonormal matrices since each two dimensional orthonormal matrix can be parameterized with a single parameter. Such algorithms are known as *Jacobi algorithms* [Car99, Com94].

The advantage of the Jacobi algorithms is that the two-dimensional optimization surface and function may be relatively simple compared to the full multidimensional case. The main drawback is that the computational complexity is quadratic with respect to the number of dimensions. The ICA methods presented in Chapter 6 use Jacobi type of optimization.

Jacobi optimization has been extensively used [Car99, Com94] in the ICA context when the function to be optimized can be viewed as a set of matrices whose off-diagonals should vanish for the correct parameter (usually an orthonormal matrix). Since the matrices are also estimated, the off-diagonals of different matrices do not usually vanish simultaneously for any parameter value, and one is left to find a parameter vector that minimizes the off-diagonals e.g. in mean square sense. An example of such a procedure is called *joint diagonalization* [Car96]. This optimization procedure is behind the well-known *JADE* algorithm [CS93, Car], where a set of forth-order cumulant matrices are orthogonally jointly diagonalized. In the the *CHESS* method [Yer00] described in Section 6.2 the matrices are obtained from the second derivative of the s.c.f. In the SOBI method [BAMCM97] , a set of autocovariance matrices are unitary diagonalized, and a method for unitary diagonalizing cumulants of any order greater or equal to three is proposed in [Mor01]. See also [Yer02] for a recent extension of joint-diagonalization to non-orthonormal matrices.

2.5 Discussion

The most important part for the success of an ICA method is naturally the correct model structure, i.e. if the system is linear, memoryless, convolutive etc. If the model approximately holds, then the criterion the ICA method is based on becomes crucial. If the criterion does not quantify the independence well or it is hard to estimate, one can not expect a good performance from the method no matter which type of optimization is used. The mutual information is a relatively good independence measure, but it is extremely hard to estimate. The characteristic function-based criteria allow good quantification of the dependence, and are also relatively easily estimated. However, the computational costs of the estimation seems to be higher, especially if compared to the methods approximating mutual information.

Gradient-based methods usually offer a fast optimization for ICA meth-

ods if the gradient is easily computed. This is usually the case if the method is based on the mutual information or a related criterion. However, when the gradient is computationally cumbersome, as it seems to be the case with characteristic function or cumulants -based criteria, the Jacobi algorithms offer usually a decent optimization alternative.

Chapter 3

Identifiability, separability, and uniqueness of linear ICA models

In a BSS problem, the goal is to reconstruct the original source signals and/or identify the mixing system (or its inverse) only from the observations. In the ICA model there is an underlying assumption that the original source r.v.'s are mutually independent. To what extend and with what restrictions this is possible, i.e. when the model is well-defined, is considered is this chapter.

The traditional real-valued ICA model is considered in Section 3.1. The original results for this model were established by Comon [Com94] stating essentially the separability and the system identification conditions for models with at least as many mixture r.v.'s as source r.v.'s with finite variance. These results are extended to cases where there are more sources than mixtures and also the requirement for finite variances is relaxed in Publication VIII. The extension to complex-valued variables introduced in Publication IX is treated in Section 3.2.

3.1 Real-valued linear instantaneous ICA model

The *linear instantaneous real-valued ICA model* is obtained from the general model (2.2) by allowing only linear mixing functions, i.e.

$$\bar{\boldsymbol{x}} = \boldsymbol{A}\bar{\boldsymbol{s}}, \tag{3.1}$$

where $(\bar{s}_1, \ldots, \bar{s}_m)^T = \bar{\boldsymbol{s}}$ are unknown real-valued independent r.v.'s, i.e. *sources*, $\boldsymbol{A}$ is a constant $p \times m$ unknown *mixing matrix*, $p \geq 2$, and $\bar{\boldsymbol{x}} = (\bar{x}_1, \ldots, \bar{x}_p)^T$ are *mixtures*, i.e., the observed r.vc. (sensor array output). The couple $(\boldsymbol{A}, \bar{\boldsymbol{s}})$ is called a *representation* of r.vc. $\bar{\boldsymbol{x}}$. It is said

that a representation is *reduced* if no two columns in the mixing matrix are *colinear*, that is, the columns in the mixing matrix are not pairwise linearly dependent. Finally, a reduced representation $(\boldsymbol{B}, \bar{\boldsymbol{r}})$ of $\bar{\boldsymbol{x}}$ is *proper*, if it satisfies the same assumptions as $(\boldsymbol{A}, \bar{\boldsymbol{s}})$ in Eq. (3.1), i.e. the representation $(\boldsymbol{B}, \bar{\boldsymbol{r}})$ satisfies the assumptions for the model.

When there are more sources than mixtures, the ICA model is termed *overcomplete ICA*, and the problem is *underdetermined* source separation. Since $\bar{\boldsymbol{x}} = \boldsymbol{A}\bar{\boldsymbol{s}} = (\boldsymbol{A}\boldsymbol{\Lambda}\boldsymbol{P})(\boldsymbol{P}^{-1}\boldsymbol{\Lambda}^{-1}\bar{\boldsymbol{s}})$, where $\bar{\boldsymbol{r}} = \boldsymbol{P}^{-1}\boldsymbol{\Lambda}^{-1}\bar{\boldsymbol{s}}$ has independent components, for scaling by any diagonal matrix $\boldsymbol{\Lambda}$ with nonzero diagonals and for any permutation matrix $\boldsymbol{P}$, the mixtures $\bar{\boldsymbol{x}}$ can never have completely unique representation. These ambiguities are called the *fundamental indeterminacy*.

If one does not assume reduced representations, it is hard to obtain any type of uniqueness. Indeed, if any two columns, say $\boldsymbol{\alpha}_k$ and $\boldsymbol{\alpha}_l$, of $\boldsymbol{A}$ are colinear, i.e. $\boldsymbol{\alpha}_k = a\boldsymbol{\alpha}_l$ for some constant $a \in \mathbb{R}$, then $\bar{\boldsymbol{x}}$ has also a representation with only $m-1$ source r.v.'s by combining the kth and lth source to a single source $a\bar{s}_k + \bar{s}_l$. Furthermore, suppose that colinear columns were allowed. Then if any of the source r.v.'s has a *divisible distribution*, then $\bar{\boldsymbol{x}}$ would have representations also for some dimension $\hat{m} > m$. A divisible distribution means that the c.f. of a r.v. $\bar{r}$ can be written as product of n c.f.'s for some positive integer $n > 1$ [Luk70], i.e. the r.v. $\bar{r}$ can be presented as a sum $\sum_{k=1}^{n} \bar{r}_k$ of independent r.v.'s $\bar{r}_k$, $k = 1, \ldots, n$. A r.v. can even have an infinitely divisible distribution, and then the mixture $\bar{\boldsymbol{x}}$ would have representations for any given dimension $\hat{m} \geq m$. *Infinitely divisible distributions* include normal, Cauchy, Poisson, Gamma and all α-stable distributions (for the use of heavy-tailed distributions in signal processing, see e.g. [NS95]).

One is trying to solve the BSP problem based on a priori assumptions about the model. The solutions not satisfying these assumptions have little meaning. Thus, alternative non-proper representations of $\bar{\boldsymbol{x}}$ should not be considered. In many cases, a mixture has infinitely non-proper representations as the following example shows.

Example 1. *Let $\bar{s}_1$ and $\bar{s}_2$ be non-normal independent r.v.'s and let $\bar{n}_1$, $\bar{n}_2$ and $\bar{n}_3$ be independent and standard normal (also independent of $\bar{s}_1$ and of $\bar{s}_2$). Then the mixture*

$$\bar{\boldsymbol{x}} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \bar{s}_1 + \bar{n}_1 + \bar{n}_2 \\ \bar{s}_2 + \bar{n}_1 - \bar{n}_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cos(t) & \sin(t) \\ 0 & 1 & -\sin(t) & \cos(t) \end{pmatrix} \begin{pmatrix} \bar{s}_1 \\ \bar{s}_2 \\ \sqrt{2}\bar{n}_1 \\ \sqrt{2}\bar{n}_2 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 & \cos(t) & \sin(t) & \sin(t) \\ 0 & 1 & -\sin(t) & \cos(t) & \cos(t) \end{pmatrix} \begin{pmatrix} \bar{s}_1 \\ \bar{s}_2 \\ \sqrt{2}\bar{n}_1 \\ \bar{n}_2 \\ \bar{n}_3 \end{pmatrix},$$

where $t \in \mathbb{R}$ is an arbitrary angle, has a unique (up to the fundamental indeterminacy) representation among representations with no normal r.v.'s. However, it has infinitely non-trivially different reduced representations with four r.v.'s and (non-reduced) representations for more than four source r.v.'s.

The different underlying assumptions required for the ICA model are defined below, and studied in detail in the following subsections. The first definition formalizes the concept of the traditional source separation in the ICA context.

Definition 1. *The model (3.1) is called* separable*, if for every matrix $\boldsymbol{W}$ such that $\boldsymbol{W}\bar{\boldsymbol{x}}$ has m independent components, we have $\boldsymbol{\Lambda P}\bar{\boldsymbol{s}} = \boldsymbol{W}\bar{\boldsymbol{x}}$ for some diagonal matrix $\boldsymbol{\Lambda}$ (with nonzero diagonals) and permutation matrix $\boldsymbol{P}$. Moreover, such a* separating *matrix $\boldsymbol{W}$ has to always exist.*

Separating matrices are also called *solution* matrices. It is seen that a separating matrix is unique up to the scaling and permutation of its rows for a given separable mixture.

Sometimes we can not separate the sources in the sense of Definiton 1, but the goal may be the blind identification of the system or some probabilistic treatmeat of the source signals. These ideas are formalized in the following definitions.

Definition 2. *The model (3.1) is called* identifiable*, if in every proper representations $(\boldsymbol{A}, \bar{\boldsymbol{s}})$ and $(\boldsymbol{B}, \bar{\boldsymbol{r}})$ of $\bar{\boldsymbol{x}}$, every column of $\boldsymbol{A}$ is colinear with a column of $\boldsymbol{B}$ and vice versa.*

Definition 3. *The model (3.1) is called* unique*, if the model is identifiable and further source r.vc.'s $\bar{\boldsymbol{s}}$ and $\bar{\boldsymbol{r}}$ in different proper representations have the same distribution for some permutation up to changes of location and scale.*

3.1.1 Identifiability

Identifiability states the conditions when it is possible to identify the mixing system up to the fundamental indeterminacy. This is formulated in the following theorem proved in Publication VIII.

Theorem 3.1.1 (Identifiability of Linear ICA). *The model of Eq. (3.1) is identifiable, if*

(i) all source r.v.'s are non-normal, or

(ii) the mixing matrix $\boldsymbol{A}$ is of full column rank and at most one source r.v. is normal.

Since independent normal r.v.'s are independent for any orthogonal transformation, the mixing matrix associated with Gaussian source r.v.'s can not be identified (see Example 1 in Publication VIII). For overcomplete ICA, even single normal r.v. is too much (see Example 3 in Publication VIII).

It also follows from the assumption on reduced representations that identifiability does not only guarantee that columns in different representations are necessarily linearly dependent but also that the number of sources, or the *model order*, is the same. For more information on ICA methods identifying underdetermined mixtures, see e.g. [HKO01, Tal01a, LMVC03].

3.1.2 Separability

Separability addresses the problem of whether (and if so, under what conditions) it is possible to reconstruct the original source signals up to the fundamental indeterminacy. It is seen that separation is possible in general if there are at least equal number of mixtures as sources, and at most one normal source. This is formulated in the following theorem originally established in [Com94] and extended in Publication VIII by removing the requirement for finite variances. Separation of infinite-variance sources was considered in [SYM01].

Theorem 3.1.2 (Separability of Linear ICA). *The model of Eq. (3.1) is separable if and only if the mixing matrix $\boldsymbol{A}$ is of full column rank and at most one source r.v. is normal.*

It should be noticed that although the scaling is unavoidable ambiguity in ICA, the location can be recovered (up to scaling) in a separable model. Furthermore, the separation is only possible if there are at least as many mixtures as sources. Some separation methods are presented in chapters 5 and 6 of this thesis.

3.1.3 Uniqueness

Separation in the sense of Definition 1 is only possible if there are at least as many mixtures as sources. However, even if there are fewer mixtures than sources, the model may be identifiable. In these cases, it would be valuable if also the distribution of the sources could be uniquely determined. Then it might be possible to reconstruct the original sources in a *non-deterministic way*, e.g. by maximizing the *a posteriori* distribution, as for instance in [PK97, GC98, LLS99, Com04]. Thus, uniqueness essentially means separability for overcomplete mixtures.

Except for the last condition, the following theorem is proved in Publication VIII. In the theorem $\odot$ denotes the Khatri-Rao product, i.e. the columnwise Kronecker product of matrices. It is said in the following that

a c.f. $\phi_1(t)$ has an exponential factor with a polynomial $\mathcal{P}(t)$, if $\phi_1(t)$ can be written as $\phi_1(t) = \phi_2(t)\exp(\mathcal{P}(t))$ for a c.f. $\phi_2(t)$.

Theorem 3.1.3 (Uniqueness of Linear ICA). *The model of Eq. (3.1) is unique if any of the following properties holds.*

(i) The model is separable.

(ii) All c.f.'s of source r.v.'s are analytic (or all c.f.'s are non-vanishing), and none of the c.f.'s has an exponential factor with a polynomial of degree at least 2*.*

(iii) All source r.v.'s are non-normal with non-vanishing c.f.'s, and $\operatorname{rank}\big[\boldsymbol{A} \odot \boldsymbol{A}\big] = m$*.*

(iv) All source r.v.'s have non-vanishing c.f.'s without exponential factors with a polynomial of degree k*,* $1 < k \leq q$*, and* $\operatorname{rank}\big[(\boldsymbol{A}\odot)^q \boldsymbol{A}\big] = m > \operatorname{rank}\big[(\boldsymbol{A}\odot)^{q-1}\boldsymbol{A}\big]$*.*

(v) All source r.v.'s are non-normal with finite variances and non-vanishing c.f.'s, and $\operatorname{rank}\big[(\boldsymbol{A}\odot)^3 \boldsymbol{A}\big] = m$*.*

Proof of Case (v). The model is identifiable by Theorem 3.1.1. Without loss of generality we may assume that sources have zero mean and unit variance, since the means can be removed and there is the scaling ambiguity anyway. Then uniqueness follows from the main theorem in [SR00], which states that the distribution of the sources is uniquely determined if the mixing matrix satisfies the assumed rank condition, given the first two moments, and assuming c.f.'s are non-vanishing. □

By Marcinkiewicz theorem all c.f.'s of the form $\exp\big(\mathcal{P}(t)\big)$ are normal (or degenerate). Further, all c.f. factors of an analytic c.f.'s are analytical. Therefore, it seems [1] that if analytical s.c.f.'s differ by a polynomial, then the polynomial must be of degree at most two. Thus it is *conjectured* that the analytic part of Case (ii) could be reformulated in a compact form as

(ii') All source r.v.'s have m.g.f.'s and none has a normal component.

R.v. $\bar{x}$ is said to have a *normal component*, if it allows the decomposition $\bar{x} = \bar{s} + \bar{n}$ for some independent r.v. $\bar{s}$ and a normal r.v. $\bar{n}$.

In Case (ii) the number of sources for any given number of mixtures is unlimited. In other cases, the number of sources is limited. However, it grows fast for the last three cases as the number of mixtures gets larger. Additionally, the cases (iii) and (iv) impose no restrictions on the moments

[1] According to prof. Nikolai Ushakov from NTSU, Norway, it is not known if there exist analytical c.f.'s $\phi_1(t)$ and $\phi_2(t)$ and a polynomial $\mathcal{P}(t)$ of degree greater than 2 such that $\phi_1(t) = \phi_2(t)\exp\big(\mathcal{P}(t)\big)$ (personal communication).

of the source r.v.'s. Especially, they apply to all non-normal α-stable distributions. This follows from the canonical representation of the stable distributions (e.g. Theorem 5.7.3. in [Luk70]).

In Case (v) the rank condition holds if $\binom{p+2}{3} \geq m$ with some trivial exceptions, and surely does not hold if $\binom{p+2}{3} < m$ [SR00]. This means that the existence of the variances of the sources can guarantee the uniqueness of the model up to, for instance, 220 sources for only ten mixtures.

The importance of the non-analytical c.f.'s to be non-vanishing is demonstrated in the following example.

Example 2 (Adapted from Remark 3 in [SR00]). *Let* $\bar{\boldsymbol{x}} = \left(\begin{smallmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{smallmatrix}\right) \left(\begin{smallmatrix} \bar{s}_1 \\ \bar{s}_2 \\ \bar{r}_1 \end{smallmatrix}\right)$ *with the following Pólya-type c.f.'s (see e.g. [Ush99]):*

$$\phi_{\bar{s}_1}(t) = \phi_{\bar{s}_2}(t) = \begin{cases} 1 - |t|, & |t| \leq 1 \\ 0, & |t| > 1 \end{cases},$$

and

$$\phi_{\bar{r}_1}(t) = \begin{cases} 1 - \frac{|t|}{2}, & |t| \leq 2 \\ 0, & |t| > 2 \end{cases}.$$

Define further a c.f. $\phi_{\bar{r}_2}(t) = \phi_{\bar{r}_1}(t)$, $|t| \leq 2$, *and* $\phi_{\bar{r}_2}(t+4) = \phi_{\bar{r}_2}(t)$. *But now* $\phi_{\bar{\boldsymbol{x}}}\big((t_1 \ t_2)^T\big) = \phi_{\bar{s}_1}(t_1)\phi_{\bar{s}_2}(t_2)\phi_{\bar{r}_1}(t_1 + t_2) = \phi_{\bar{s}_1}(t_1)\phi_{\bar{s}_2}(t_2)\phi_{\bar{r}_2}(t_1 + t_2)$, *and thus r.vc.* $\bar{\boldsymbol{x}}$ *has another representation* $\bar{\boldsymbol{x}} = \left(\begin{smallmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{smallmatrix}\right) \left(\begin{smallmatrix} \bar{s}_1 \\ \bar{s}_2 \\ \bar{r}_2 \end{smallmatrix}\right)$. *Since r.v.'s* $\bar{r}_1$ *and* $\bar{r}_2$ *differ more than just by the change of location and scale, it is seen that the corresponding model can not be unique.*

Another example of identifiable but non-unique representation is given in Example 2 of Publication VIII. It is based on the property that r.v.'s have a normal component, which means that the corresponding s.c.f.'s have a polynomial term of degree two. The example can be generalized to r.v.'s, which do not have a normal component, but the s.c.f. has a polynomial term $\mathcal{P}(t)$ of any degree. This is shown for a third order polynomial in the following example.

Example 3. *Let* $\phi_1(t) = \exp(\jmath t^3)\phi_2(t)$, *and let independent r.v.'s* $\bar{s}_k$, $k = 1\ldots,6$, *and* $\bar{r}_k$, $k = 1\ldots,6$, *have non-vanishing c.f.'s* ϕ_1 *and* ϕ_2 *(respectively). Existence of such c.f.'s follows from Lemma 3 in [SR00]. Since* $2t_1^3 - 108t_2^3 + (t_1 + 7t_2)^3 - (t_1 - 2t_2)^3 - (t_1 + 3t_2)^3 - (t_1 + 6t_2)^3 \equiv 0$, *it follows that the c.f. of a r.vc.*

$$\bar{\boldsymbol{x}} = \begin{pmatrix} 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 7 & -2 & 3 & 6 \end{pmatrix} \begin{pmatrix} \sqrt[3]{2}\bar{s}_1 & -3\sqrt[3]{4}\bar{s}_2 & \bar{s}_3 & -\bar{s}_4 & -\bar{s}_5 & -\bar{s}_6 \end{pmatrix}^T$$

equals to the c.f. of a r.vc.

$$\begin{pmatrix} 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 7 & -2 & 3 & 6 \end{pmatrix} \begin{pmatrix} \sqrt[3]{2}\bar{r}_1 & -3\sqrt[3]{4}\bar{r}_2 & \bar{r}_3 & -\bar{r}_4 & -\bar{r}_5 & -\bar{r}_6 \end{pmatrix}^T,$$

i.e. $\bar{\boldsymbol{x}}$ *has two representations that are not equivalent.*

C.f.'s with a factor $\exp\big(\mathcal{P}(t)\big)$ are not limited by moment properties, since it is known [Gol73] that for any even polynomial $\mathcal{P}(t)$, $\mathcal{P}(0) = 0$, there exist c.f.'s ϕ_1 and ϕ_2 such that $\phi_1(t) = \phi_2(t)\exp\big(\mathcal{P}(t)\big)$, and the corresponding r.v.'s possess moments of all orders. Further, it is known [SR00] that the c.f.'s ϕ_1 and ϕ_2 can be non-vanishing for polynomials of all orders. Thus Examples 2 and 3 show that requirement of non-vanishing c.f.'s without exponential polynomial factors can not be in general avoided, i.e. there is not much room for improvement in Theorem 3.1.3.

3.1.4 Discussion

M.g.f.'s have most of the nice properties of c.f.'s, and are further real-valued. Therefore, in the light of Conjecture 3.1.3(ii'), they might offer a natural framework for developing algorithms that try to reconstruct the original signals in underdetermined ICA. This is also supported by the fact that the existing identification algorithms are all based on either cumulants (e.g. [LMVC03]) or on c.f.'s (e.g. [Tal01a]).

3.2 Complex linear instantaneous signal model

The real-valued model of Eq. (3.1) can be extended to complex domain, i.e.

$$\bar{\boldsymbol{x}} = \boldsymbol{A}\bar{\boldsymbol{s}}, \tag{3.2}$$

where $(\bar{s}_1, \ldots, \bar{s}_m)^T = \bar{\boldsymbol{s}}$ are unknown complex-valued independent non-constant r.v.'s and $\boldsymbol{A}$ is a constant $p \times m$ unknown complex-valued matrix, $p \geq 2$. As noted in subsection 2.3.4, the operator structure of complex r.vc.'s is different from real r.vc.'s. Therefore, it is not a priori clear under what conditions the model (3.2) is well-defined. Considering two dimensional real r.vc.'s it may actually appear that the model is not identifiable under any general conditions as the following example shows.

Example 4. *Let* $\bar{r}_k$, $k = 1\ldots, 4$, *be independent real r.v.'s, and let* $\boldsymbol{A}_1$, $\boldsymbol{A}_2$, $\boldsymbol{B}_1$, *and* $\boldsymbol{B}_2$ *be* 2×2 *nonsingular real matrices. Define* $\bar{\boldsymbol{s}}_1 = \boldsymbol{A}_1(\bar{r}_1\ \bar{r}_2)^T$ *and* $\bar{\boldsymbol{s}}_2 = \boldsymbol{A}_2(\bar{r}_3\ \bar{r}_4)^T$. *Now* $\bar{\boldsymbol{s}}_1$ *and* $\bar{\boldsymbol{s}}_2$ *are independent, but so are also* $\bar{\boldsymbol{y}}_1$ *and* $\bar{\boldsymbol{y}}_2$,

$$\begin{pmatrix} \bar{\boldsymbol{y}}_1 \\ \bar{\boldsymbol{y}}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{B}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B}_2 \end{pmatrix} \boldsymbol{P} \begin{pmatrix} \boldsymbol{A}_1^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A}_2^{-1} \end{pmatrix} \begin{pmatrix} \bar{\boldsymbol{s}}_1 \\ \bar{\boldsymbol{s}}_2 \end{pmatrix},$$

for any permutation matrix $\boldsymbol{P}$.

However, it turns out that theorems similar to the real case (Section 3.1) hold also in complex domain as it was shown in Publication IX. This is considered next.

3.2.1 Identifiability, separability, and uniqueness

The definitions of different degrees of unambiguity, i.e. identifiability, separability, and uniqueness, do not require any changes for the complex model (3.2). However, notice that in uniqueness by scaling it is now meant complex scaling, i.e. multiplication by a complex number. The relevant theorems are all stated in this subsection. For the proofs and some examples, see Publication IX.

The identifiability theorem is completely analogous to the real-case.

Theorem 3.2.1 (Identifiability of Linear Complex ICA). *The model of Eq. (3.2) is identifiable, if*

(i) no source r.v. is complex normal, or

(ii) the model is separable.

The separability is different from the separability of real r.v.'s. Namely, it turns out that mixtures of some complex normal r.v.'s can be separated (see Example 5 in Publication IX).

Theorem 3.2.2 (Separability of Linear Complex ICA). *The model of Eq. (3.2) is separable if and only if the complex mixing matrix $\boldsymbol{A}$ is of full column rank and there are no two complex normal source variables with the same spectral coefficient.*

Some work on separation algorithms for complex ICA can be found from [Car93, Com94, BH00, Fio03, ASM03]. Notice also that if the source r.vc. has a distinct spectrum, then the separation can be achieved simply by applying the strong-uncorrelating transform (see Publication IX), i.e. using the second order statistics only.

Finally, the first two cases of the real uniqueness theorem (Theorem 3.1.3) can be extended also to the complex case. The rest of the cases depend on the structure of the coefficients in the mixing matrix, and further work is needed to see if those theorems also have a complex counterpart.

Theorem 3.2.3 (Uniqueness of Linear Complex ICA). *The model of Eq. (3.2) is unique if either of the following properties hold.*

(i) The model is separable.

(ii) All c.f.'s of source r.v.'s are analytic (or all c.f.'s are non-vanishing), and none of the c.f.'s has an exponential factor with a polynomial of degree at least two, i.e. no source r.v. has the c.f. ϕ_1 such that $\phi_1(z) = \phi_2(z)\exp(\mathcal{P}(z, z^*))$ for some polynomial $\mathcal{P}(z, z^*)$ of degree at least two.

3.3 Discussion

Identifiability considers the reconstruction of the mixing system. It is rather striking that theoretically we can obtain the mixing matrix from only two sensors for whatever number of non-Gaussian sources. It would be rather interesting to know what are the true practical limitations. This is ultimately connected to the convergence rate of the central limit theorem. The separation, that is the exact reconstruction of the source signals by a transformation, is only possible if there are at least as many sensors as sources. However, it may turn out that there are some other ways of reconstructing the source signals in underdetermined unique models, since the full probabilistic information can be recovered. Developing such algorithms may be an interesting future research topic.

There are some differences between real- and complex-valued ICA models. First, the separability of some complex mixtures with more than a single normal r.v. is rather surprising. Second, some complex mixtures can be separated by the second order statistics only. This may have interesting applications in areas where one can affect the statistics of the source signals, e.g. in communications.

Chapter 4

Nonlinear instantaneous ICA models

Both the models introduced in the previous chapter were subclasses of the general instantaneous ICA model of Eq. (2.2) with the restriction that the mixing functions are linear. It can be fairly easily shown (see e.g. [HP99, TJ99b]) that there are infinite number of mixing functions that produce independent marginals but are still mixtures of original independent sources. Thus, the general model is not *separable* in the sense [Tal01b] that if a function produces independent components from the mixture, each component is necessarily a transformation of a source signal.

However, one may still find *classes of functions* and r.v.'s such that the corresponding model is separable. This type of ICA model is called *nonlinear* instantaneous ICA. Some interesting nonlinear models are reviewed in this section. A fairly general class of nonlinear models stemming from the addition theorem [Acz66] is considered in Section 4.1. These models are introduced in Publication VI. A overview of the well-known post-nonlinear model is given in Section 4.2. Recent reviews of nonlinear general ICA models and methods are given in [JK03, JBZH04]. Some nonlinear ICA modeling approaches, especially Bayesian methods, can be found from [HKO01, Chapter 17]. A general framework and relative gradient-based method for separation of nonlinear mixtures whose parameters form a Lie group, is presented in [Tal02].

4.1 Models implied by Addition Theorem

Symmetry is a fundamental property of the physics of nature [HL]. Mathematically symmetry is described by the group theory (see [HL] for a nice discussion on symmetry in nature and the group theory). Therefore, from the application point of view, it is natural to require that the mixing functions in Eq. (2.2) satisfy the group axioms with respect to source r.v.'s. This

motivates the following construction.

A closed operation $\circ$ on a set forms a *group* [Sco87], if the operation is associative, there exists a unit element, and every element has an inverse element. A group is called *Abelian* if additionally the group operation is commutative. It is straightforward to check that any continuous and strictly monotonic (i.e. invertible) function $\mathcal{G}: \mathbb{R} \to \mathbb{U}$ gives an Abelian group on an open interval $\mathbb{U} \subseteq \mathbb{R}$ by defining for all $u, v \in \mathbb{U}$,

$$u \circ v \triangleq \mathcal{G}(\mathcal{G}^{-1}(u) + \mathcal{G}^{-1}(v)). \tag{4.1}$$

By denoting $x = \mathcal{G}^{-1}(u)$, $y = \mathcal{G}^{-1}(v)$, and $\mathcal{H}_2\big(u, v\big) = u \circ v$, Eq. (4.1) can be written as

$$\mathcal{G}(x + y) = \mathcal{H}_2\big(\mathcal{G}(x), \mathcal{G}(y)\big). \tag{4.2}$$

This type of equations are called *addition theorems* [Acz66]. It can be shown [Acz66] that given an Abelian operation $\circ$ on an open interval $\mathbb{U}$, Eq. (4.2) is satisfied for a unique (up to constant multiplication of its argument) continuous strictly monotonic function $\mathcal{G}$. The converse is also true. For a fixed continuous function $\mathcal{G}$ satisfying (4.2) with some function $\mathcal{H}_2\big(\cdot, \cdot\big)$, the operation $\mathcal{H}_2\big(\cdot, \cdot\big)$ necessarily defines an Abelian group and $\mathcal{G}$ is strictly monotonic.

Since $\mathcal{G}$ has the inverse function $\mathcal{G}^{-1}$, it follows by (4.2)

$$\mathcal{G}(kx) = \mathcal{G}(x + x + \cdots + x) = \mathcal{G}(x) \circ \mathcal{G}(x) \circ \cdots \circ \mathcal{G}(x) \triangleq k \star \mathcal{G}(x) = k \star u,$$

which gives a new multiplication operation $\star$ for an integer k. This extends uniquely by continuity to all reals a by defining

$$a \star u \triangleq \mathcal{G}(ax) = \mathcal{G}(a\mathcal{G}^{-1}(u)). \tag{4.3}$$

Using operators defined by Eq. (4.1) and Eq. (4.3) one can define a nonlinear function

$$\mathcal{F}: \mathbb{U}^m \to \mathbb{R}, \quad \mathcal{F}(u_1, \ldots, u_m) = (a_1 \star u_1) \circ (a_2 \star u_2) \circ \cdots \circ (a_m \star u_m), \tag{4.4}$$

where the parenthesis could be dropped since $\circ$ is associative. Using these type of functions as mixing functions, a nonlinear ICA model of Eq. (2.2) can be written as

$$\bar{\boldsymbol{x}} = \boldsymbol{\mathcal{F}}(\bar{\boldsymbol{s}}) = \begin{bmatrix} \mathcal{F}_1(\bar{s}_1, \ldots, \bar{s}_m) \\ \mathcal{F}_2(\bar{s}_1, \ldots, \bar{s}_m) \\ \vdots \\ \mathcal{F}_p(\bar{s}_1, \ldots, \bar{s}_m) \end{bmatrix} = \begin{bmatrix} a_{11} \star \bar{s}_1 \circ a_{12} \star \bar{s}_2 \circ \cdots \circ a_{1m} \star \bar{s}_m \\ a_{21} \star \bar{s}_1 \circ a_{22} \star \bar{s}_2 \circ \cdots \circ a_{2m} \star \bar{s}_m \\ \vdots \\ a_{p1} \star \bar{s}_1 \circ a_{p2} \star \bar{s}_2 \circ \cdots \circ a_{pm} \star \bar{s}_m \end{bmatrix}. \tag{4.5}$$

These models were introduced in Publication VI. It should be emphasized that *any* continuous strictly monotonic function introduces a model of the form (4.5) by Eq.'s (4.1) and (4.3).

Now, let $\boldsymbol{\alpha}_k = (a_{1k}\ a_{2k}\ \cdots\ a_{pk})^T$. Since the location parameter of a separable linear ICA model can be recovered as noted on page 22, the following separability result holds (see Publication VI and Theorem 3.1.2).

Theorem 4.1.1 (Separability of Addition Theorem models). *Suppose the model (4.5) holds such that at most one of the r.v.'s $\mathcal{G}^{-1}(\bar{s}_k)$, $k = 1, 2, \ldots, m$ is normal, where $\mathcal{G}$ is the function defined by the operator $\circ$, and that vectors $\boldsymbol{\alpha}_k$, $k = 1, 2, \ldots, m$, define a full column rank matrix $\boldsymbol{A} = \begin{pmatrix} \boldsymbol{\alpha}_1 & \boldsymbol{\alpha}_2 & \cdots & \boldsymbol{\alpha}_m \end{pmatrix}$. Then the model (4.5) is separable up to arbitrary permutation and a constant $\star$-multiplication (defined in Eq. (4.3)) of each source.*

Theorem 3.1.1 and Theorem 3.1.3 (results from Publication VIII) are similarly extended to addition theorem models. An important consequence of Theorem 4.1.1 is that if $\mathcal{G}$ does not preserve normality, then it is possible to separate normal r.v.'s contrary to the linear model.

A function is called *algebraic*, if it can be formed by using finite number additions, subtractions, multiplications, divisions, and root taking. All analytic algebraic mixing functions of Eq. (4.4) are obtained from the following fundamental theorem from the theory of elliptic functions, see [Acz66] for references to the original papers.

Theorem 4.1.2. *Every single-valued analytic function $\mathcal{G}$ with an algebraic addition theorem is either a rational function of x, or a rational function of e^{ax} or a doubly periodic function (a rational function of Weierstrass' p-function and its derivative).*

However, there are some non-algebraic and non-analytic mixing functions that are of engineering interest. The operator $\circ$ may be described by common non-algebraic operations such as the exponential function or the trigonometric functions. Furthermore, the function $\mathcal{G}$ could be defined piecewise. This type of models arise, if the system is homogeneous within certain interval of signal values but treats differently values from two different intervals. In the following example the model has two "subsystems" (for negative and positive signals), which are both linear, but the overall model is not.

Example 5. *By Eq. (4.1) the piecewise linear function*

$$\mathcal{G}(x) = \begin{cases} x & x \geq 0 \\ bx & x < 0 \end{cases}, \tag{4.6}$$

where b is a positive constant, introduces the group operation

$$u \circ v = \begin{cases} u + v, & uv \geq 0 \\ \frac{1}{b}\min(u, v) + \max(u, v), & uv < 0,\ b \geq -\frac{\min(u,v)}{\max(u,v)} \\ \min(u, v) + b\max(u, v), & uv < 0,\ b < -\frac{\min(u,v)}{\max(u,v)} \end{cases},$$

which is seen to be bilinear in first and third quadrants of $\mathbb{R}^2$*. The multiplication operation is given by*

$$a \star u = \begin{cases} \frac{1}{b}au, & a, u < 0 \\ au, & a \geq 0 \\ bau, & a < 0,\ u \geq 0 \end{cases} .$$

Notice that by Theorem 4.1.1, normal r.v.'s can be separated in the ICA model (4.5) defined by the function (4.6) (assuming $b \neq 1$*).*

A separation method for the addition theorem models (4.5) with known structure (i.e. function $\mathcal{G}$ is known) is obtained from any ICA method for the linear model (3.1) with the following generic procedure (see Publication VI):

1. Transform every value of the observed signal $\boldsymbol{x}[k]$, $k = 1, \ldots, N$, with the function $\mathcal{G}^{-1}$.

2. Use the linear ICA method for the transformed signal.

3. Transform the output of the linear method with the function $\mathcal{G}$.

Such a method is applied in Publication VI to the reduction of light reflection from specular object surfaces in natural images. Also a toy example of separation of Gaussian signals is presented for illustration purposes.

4.2 Post-nonlinear model

Another nonlinear ICA model (2.2) is described by

$$\bar{\boldsymbol{x}} = \boldsymbol{\mathcal{F}}(\boldsymbol{A}\bar{\boldsymbol{s}}), \tag{4.7}$$

where $\boldsymbol{\mathcal{F}}(\boldsymbol{t}) = (\mathcal{F}_1(t_1), \ldots, \mathcal{F}_p(t_p))^T$ is a *component-wise* nonlinear function. The model is known as *post-nonlinear* (PNL) model [TJ99c], and can be viewed as linear mixing followed by a nonlinear distortion at each sensor. Since $\boldsymbol{\mathcal{F}}(\cdot)$ is also unknown, the PNL model provides a wide variety of nonlinear mixing systems.

The *separating structure* of Eq. (4.7) is a pair $(\boldsymbol{\mathcal{G}}, \boldsymbol{W})$, where $\boldsymbol{\mathcal{G}}$ is also a component-wise nonlinear function, such that $\boldsymbol{W}\boldsymbol{\mathcal{G}}(\bar{\boldsymbol{x}})$ has m mutually independent components. The model of Eq. (4.7) is then *separable* if for every separating structure $(\boldsymbol{\mathcal{G}}, \boldsymbol{W})$ the r.vc. $\boldsymbol{W}\boldsymbol{\mathcal{G}}(\bar{\boldsymbol{x}})$ is a permuted, scaled, and translated version of the source $\bar{\boldsymbol{s}}$. The separability of the PNL model (4.7) is shown in [TJ99c] under the following conditions:

(i) The number of mixtures is equal to the number of sources, i.e. $p = m$, and the matrix $\boldsymbol{A}$ is invertible.

(ii) The matrix $\boldsymbol{A}$ has at least two nonzero entries per row or per column.

(iii) Functions $\mathcal{F}_k$, $k = 1, \ldots, p$, of Eq. (4.7) , are differentiable and invertible functions such that the derivative of the composition of $\mathcal{G}_k$ and $\mathcal{F}_k$ does not vanish.

(iv) Each source $\bar{s}_k$, $k = 1, \ldots, m$, has a p.d.f. such that it vanishes at least at a single point.

For slightly more restricting conditions, but for a simpler proof, see [BZ02]. Especially conditions (iii) and (iv) are restrictive. Further research is needed to determine if the conditions could be relaxed.

The PNL model does not describe truly nonlinear mixing but rather nonlinear distortion after linear mixing. However, if the components $\mathcal{F}_k$ of $\boldsymbol{\mathcal{F}}$ are the same and $p = m$, i.e. $\boldsymbol{\mathcal{F}}(\boldsymbol{t}) = (\mathcal{F}(t_1), \ldots, \mathcal{F}(t_m))^T$, then by writing $\bar{\boldsymbol{s}} = \boldsymbol{\mathcal{F}}^{-1}(\bar{\boldsymbol{r}})$ and using the addition theorem (4.2), the model of Eq. (4.7) reduces to an addition theorem model (4.5) in source r.v.'s $\bar{r}_k$, $k = 1, \ldots, m$. Therefore, the PNL mixture can be viewed as an outcome of a truly nonlinear mixing. If the mixing functions of Eq. (4.5) satisfy conditions for PNL mixture, the statement may be formulated another way around: the addition theorem models belong to the class of PNL models with post-processing.

Methods for separating the PNL mixtures based on mutual information and score function estimation are proposed in [TJ99c, TJ99a]. A geometric algorithm based on the border detection for bounded PNL models is proposed in [BZJN02, BZ02]. Methods essentially relying on the c.f. difference condition of Eq. (2.12) are proposed in [APJ03, Ach03].

4.3 Discussion

So far, there have been relatively few applications of nonlinear ICA models. However, it is hard to imagine any physical mixing system that would *not* satisfy the group operations. Furthermore, many of PNL outputs can actually be viewed as results of a truly nonlinear mixing described by the addition theorem. It is probably fair to say, at this state of development of nonlinear ICA models and algorithms, that there either do not exist many blindly invertible truly nonlinear mixing systems in nature, or that they are yet to be found. However, they are likely to belong to the class of models described by the addition theorem because of the inherited symmetry in nature.

It is stated in [JK03, JBZH04] that the general addition theorem separation method described on page 31 is *not blind*, since the function $\mathcal{G}$ has to be known. This statement is inaccurate requiring some clarification. The

proposed method is blind in exactly the same way as the usual linear ICA: *unknown* mixing coefficients are obtained from mixtures of *unknown* source signals. Indeed, the linear ICA model of Eq. (3.1) is just a special case of the model of Eq. (4.5) with $\mathcal{G}(x) = x$.

Naturally, there might be room for additional degrees of freedom (i.e. blindness) in addition theorem models. That is, instead of considering a model described by a known function $\mathcal{G}$, one could try to find a parameterized family of functions such that each of them describes an addition theorem model, and the parameters are also unknown. An example of a such class of functions is obtained from Example 5 by taking the constant b as a parameter. However, with such an extension, one should show that only the true parameter values b and $\boldsymbol{A}$ lead to independent components for any allowed r.v.'s. This might be an interesting topic for a future research, and one should especially look for families where the sources could be recovered up to scaling (no non-linear distortion). Since this extension can be viewed as a PNL mixture, the separation conditions proved for PNL mixtures give sufficient but rather restrictive conditions for some function classes. There exists also an interesting connection to nonlinear filtering theory. Namely, taking $\mathcal{G}(x) = \exp(x)$ the function of Eq. (4.4) defines a *homomorphic system* [OS89].

Chapter 5

Source adaptive methods for blind separation

When the gradient based optimization is applied to the mutual information criterion of Eq. (2.8), it turns out that the optimal gradient depends on the unknown distribution of the sources. ICA methods that try to calculate the optimal gradient by also estimating the source distribution, or a function related to it, are generally called *source adaptive* ICA methods. In sections 5.2 and 5.3 two such ICA methods, originally introduced in Publication I and in Publication II respectively, are reviewed. These methods are applicable to separating the linear instantaneous model of Eq. (3.1). Source adaptive methods are first briefly reviewed in Section 5.1. For convenience, it is assumed that the number of sources is equal to the number of mixtures, i.e. $\boldsymbol{A}$ is a square matrix of rank m.

5.1 Overview of mutual information-based separation methods

Let

$$\bar{\boldsymbol{y}} = \boldsymbol{W}\bar{\boldsymbol{x}}. \tag{5.1}$$

Then the ICA separating methods based on m.i. (2.8) criterion minimize the m.i. of $\bar{\boldsymbol{y}}$, i.e. the separating matrix is obtained as

$$\hat{\boldsymbol{W}} = \operatorname*{argmin}_{\boldsymbol{W}} \mathrm{MI}(\boldsymbol{W}\bar{\boldsymbol{x}}). \tag{5.2}$$

It is possible to use the second type approach (see Section 2.2) in deriving the ICA method (e.g. [DT00, MI03]). However, since the p.d.f.'s are unknown, the direct estimation of m.i. is hard [Pan03]. It is also possible to approximate m.i. by series approximations. However, the resulting approximation is in terms of moments, and moments and cumulants are related.

Therefore, this type of approximation methods are essentially the same as cumulant methods treated in Section 6.1. A more traditional way is to adopt the first type of approach (see Section 2.2), and directly solve (5.2) using a stochastic gradient descent method. By Eq. (2.10) and the relation (5.1),

$$\mathrm{MI}(\bar{\boldsymbol{y}}) = \sum_{k=1}^{p} \mathrm{H}(\bar{y}_k) - \mathrm{H}(\bar{\boldsymbol{x}}) - \log|\det(\boldsymbol{W})|,$$

and a direct calculation gives the matrix gradient

$$\nabla\mathrm{MI}(\bar{\boldsymbol{y}}) = \mathrm{E}_{\bar{\boldsymbol{y}}}\{\boldsymbol{\varphi}_{\bar{\boldsymbol{y}}}(\bar{\boldsymbol{y}})\bar{\boldsymbol{x}}^T\} - (\boldsymbol{W}^T)^{-1} = \mathrm{E}_{\bar{\boldsymbol{y}}}\{\boldsymbol{\varphi}_{\bar{\boldsymbol{y}}}(\bar{\boldsymbol{y}})\bar{\boldsymbol{y}}^T\}(\boldsymbol{W}^T)^{-1} - (\boldsymbol{W}^T)^{-1}, \tag{5.3}$$

where $\boldsymbol{\varphi}_{\bar{\boldsymbol{y}}}(\boldsymbol{t}) = \big(\varphi_{\bar{y}_1}(t_1), \ldots, \varphi_{\bar{y}_p}(t_p)\big)^T$ is the vector of the *score functions*

$$\varphi_{\bar{y}}(t) \triangleq -\frac{d}{dt}\log f_{\bar{y}}(t) = -\frac{f'_{\bar{y}}(t)}{f_{\bar{y}}(t)} \tag{5.4}$$

of the marginal r.v.'s of the r.vc. $\bar{\boldsymbol{y}}$. The associated natural gradient is given by

$$\nabla_{\mathrm{N}}\mathrm{MI}(\bar{\boldsymbol{y}}) = \nabla\mathrm{MI}(\bar{\boldsymbol{y}})\boldsymbol{W}^T\boldsymbol{W} = \big(\mathrm{E}_{\bar{\boldsymbol{y}}}\{\boldsymbol{\varphi}_{\bar{\boldsymbol{y}}}(\bar{\boldsymbol{y}})\bar{\boldsymbol{y}}^T\} - \boldsymbol{I}\big)\boldsymbol{W}, \tag{5.5}$$

see [CA02, Hay00, HKO01, Car98, Pha96] for detailed derivation and historical references. Eq. (5.3) shows that the gradient is a function of the updated output $\bar{\boldsymbol{y}}$ and the previous estimate of $\boldsymbol{W}$ in every iteration of the learning algorithm. Essentially the same result is obtained with other methods relying on Eq. (2.7), e.g. *information maximization* [BS95], *(quasi)-maximum likelihood* approach [PG97], and *maximization of non-Gaussianity* [DL95], see [Hay00, HKO01, CA02, Car98] for detailed discussion and references.

In order to have a complete ICA method, an estimate of $\mathrm{E}_{\bar{\boldsymbol{y}}}\{\boldsymbol{\varphi}_{\bar{\boldsymbol{y}}}(\bar{\boldsymbol{y}})\bar{\boldsymbol{y}}^T\}$ is needed. The expectation operator can be replaced by the realization sample average, and therefore the only remaining problem is the unknown vector of score functions $\boldsymbol{\varphi}_{\bar{\boldsymbol{y}}}(\cdot)$. Perhaps the simplest solution is to use a score vector comprised of *fixed* nonlinear functions $\varphi(\cdot)$. The resulting methods can be interpreted from the nonlinear correlation (2.6) viewpoint. Rather surprisingly, some well-chosen nonlinear functions seem to work well in many practical situations, see [Hay00, Appendix 3.3 and Section 4.6][HKO01, Section 14.3][CA02, Chapter 6] for some local stability theorems. However, in the view of the necessity part of Eq. (2.6), it comes as no surprise that methods relying on a fixed nonlinearity are bound to fail for some r.vc.'s even in the case, where it is assumed that the sources have symmetric p.d.f.'s [MD02].

Another, theoretically more justified method is to adapt to the source distributions by estimating the score vector $\boldsymbol{\varphi}_{\bar{\boldsymbol{y}}}(\cdot)$. This type of source adaptive ICA is essentially *adaptive score function modeling*. Two parametric solutions, originally introduced in Publication I and in Publication II, are

considered in the following sections of this chapter. Another method relying on generalized Gaussian distribution is presented in [CCA00], a method using the t-distribution is proposed in [CM99], a nonparametric method based on linear combination of basis function is found from [PG97], and a kernel density estimation based methods are proposed in [BPR04, LY03]. Both the generalized Gaussian distribution and the t-distribution are symmetrical.

5.2 Extended generalized lambda distribution

By the relation (5.4), a parametric model for the score function is obtained from any twice differentiable parametric model for d.f.'s. This approach was taken in Publication I, where the *extended generalized lambda distribution* (EGLD) was used as the d.f. model. This distribution family is a generalization of Tukey's lambda distribution [Tuk60], and it has been mainly used in fitting a distribution to the empirical data. The latest extension, the EGLD family [KDM96], is a combination of *generalized lambda distribution* (GLD) and *generalized beta distribution* (GBD). See Figure 1 in Publication I for the characterization of the EGLD family and some common distributions in terms of the third and the fourth central moments.

A r.v. $\bar{x}$ belongs to the GLD family if its inverse d.f. can be expressed in the form

$$F_{\bar{x}}^{-1}(u) = \lambda_1 + \frac{u^{\lambda_3} - (1-u)^{\lambda_4}}{\lambda_2}, \tag{5.6}$$

where $0 \leq u \leq 1$ and λ_1, λ_2, λ_3 and λ_4 are the parameters of the distribution. The other part of EGLD, the GBD family is characterized by the p.d.f.

$$f_{\bar{x}}(t) = a{\beta_2}^{-(\beta_3+\beta_4+1)}(t-\beta_1)^{\beta_3}(\beta_1+\beta_2-t)^{\beta_4}, \tag{5.7}$$

where a is a constant and β_1, β_2, β_3 and β_4 are the parameters of the distribution. The score functions of the EGLD distributions (5.6) and (5.7) can be solved in a closed form, see Publication III for correct formulas.

Finally, one needs to estimate the parameters of distributions. This can be achieved [KDM96] by the *method of moments*, which means that the estimated moment values are substituted in the place of the theoretical moments directly calculated from the formulas describing the family. In the case of GBD, the parameters can be given in a closed form in terms of the sample moments. On the other hand, the corresponding GLD equations do not allow simple formulas, and computationally complex numerical solutions are needed. These solutions are tabulated in [DK96], and they have been used in the EGLD based method in Publication I and in Publication III.

A more convenient way to estimate the parameters of GLD distribution is proposed in Publication III. It is based on *L-moments*, which are linear combinations of order statistics. In this case, the parameters of GLD are

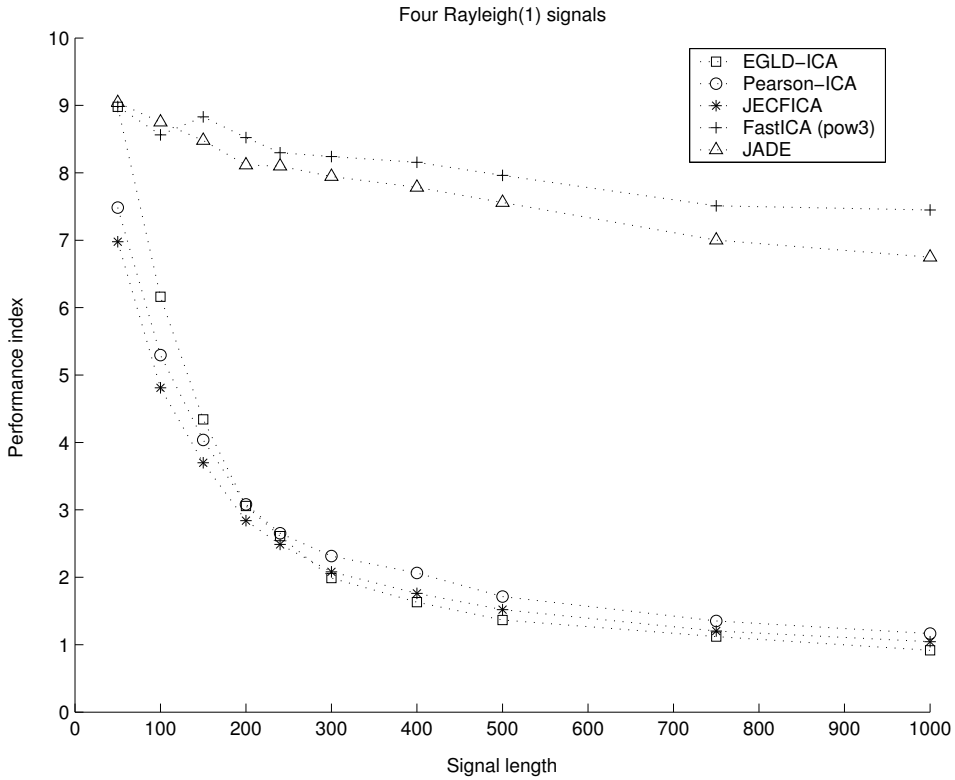

Figure 5.1: Performance of different ICA methods for four Rayleigh(1) distributed signals with varying signal length. Compared methods: EGLD-ICA (Section 5.2), Pearson-ICA (Section 5.3), JECFICA (Section 6.3), FastICA [H$^+$], and JADE [Car].

obtained in a simpler form, and thus the extensive use of the tabulated values is avoided.

Comparison of the performance of the conventional moment based ICA method (*EGLD-ICA*) and four other methods is presented in Fig. 5.1. In the simulation four signals with skewed distributions (Rayleigh(1) distributed) were mixed with varying signal length. For each of 500 realization, the quality of separation was measured by the performance index [ACY96]

$$J \triangleq \sum_{k=1}^{p}\Big(\sum_{l=1}^{p}\frac{|\boldsymbol{P}_{kl}|}{\max_n|\boldsymbol{P}_{kn}|}-1\Big)+\sum_{l=1}^{p}\Big(\sum_{k=1}^{p}\frac{|\boldsymbol{P}_{kl}|}{\max_n|\boldsymbol{P}_{nl}|}-1\Big), \tag{5.8}$$

where the matrix $\boldsymbol{P}=(\boldsymbol{P}_{kl}) \triangleq \boldsymbol{W}\boldsymbol{A}$ for the solution matrix estimate $\boldsymbol{W}$ and the true mixing matrix $\boldsymbol{A}$. The zero value corresponds to the perfect separation. The average performance as function of signal length is presented in Fig. 5.1 and the detailed analysis of the performance for the signal length 500 is plotted in Fig. 5.2. For longer signals, EGLD-ICA shows the best

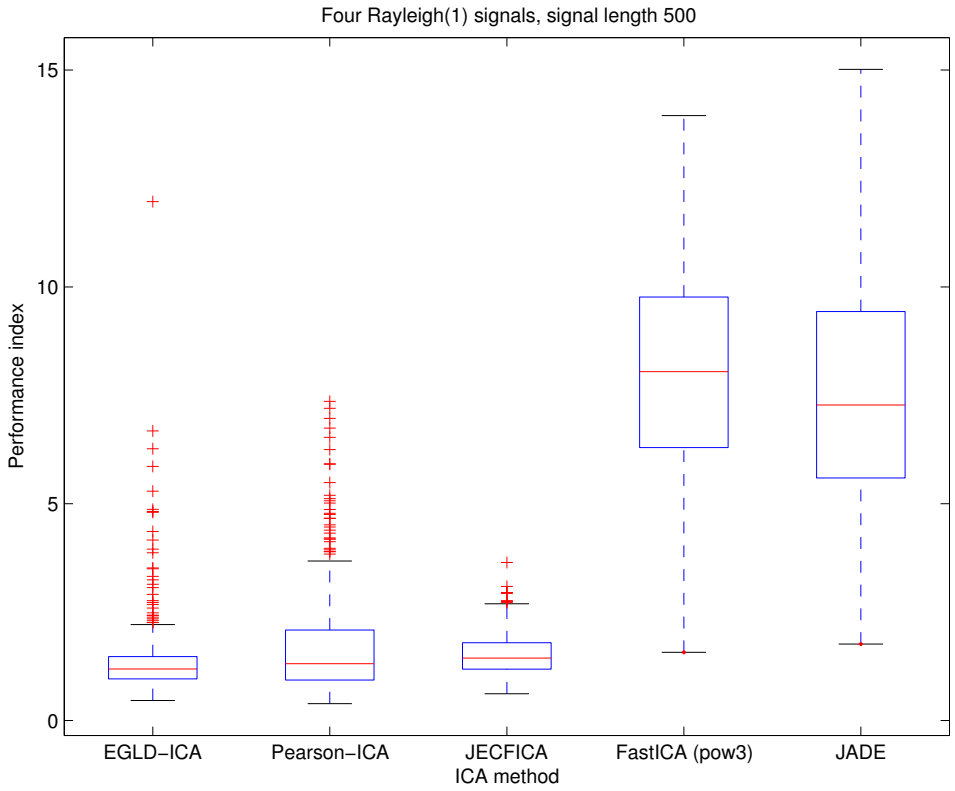

Figure 5.2: Performance of different ICA methods for four Rayleigh(1) distributed signals with signal length 500. Compared methods: EGLD-ICA (Section 5.2), Pearson-ICA (Section 5.3), JECFICA (Section 6.3), FastICA [H$^+$], and JADE [Car].

performance among tested methods. Another simulation with four discrete Poisson(1) signals is presented in figures 6.1 and 6.2. Also there the EGLD-ICA method shows relatively good performance although it was designed for absolutely continuous distributions. Extensive simulation examples showing the reliable performance of the EGLD methods are presented in Publication III. For some new estimation procedures of EGLD parameters and for further references, see [KM99].

5.3 Pearson system

Another parametric model for the score function is proposed in Publication II. A r.v. $\bar{x}$ belonging to the *Pearson system* is described through the differential equation

$$f'_{\bar{x}}(t) = \frac{(t-a)f_{\bar{x}}(t)}{b_0 + b_1 t + b_2 t^2}, \tag{5.9}$$

where a, b_0, b_1, and b_2 are parameters of the distribution. In the view of the equations (5.4) and (5.9), the Pearson system is convenient for modeling the score functions. The score may be easily found to be

$$\varphi_{\bar{x}}(t) = -\frac{(t-a)}{b_0 + b_1 t + b_2 t^2}. \tag{5.10}$$

Since only the the score functions and not the densities are actually needed in gradients, the simplicity of the score (5.10) makes the Pearson system particularly appealing for a source adaptive ICA method. The parameters of the Pearson system can be found using the method of moments, and the corresponding equations can be solved in a closed form, see Publication II. Such an ICA method is called *Pearson-ICA*.

The simulations results presented in figures 5.1 and 5.2 show that Pearson-ICA has only marginally worse average performance than EGLD-ICA for longer signals, but it shows better performance for short signals. The same phenomenon is also apparent in the simulation results presented in figures 6.1 and 6.2. Further simulations are presented in Publication II, in Publication III, and in [BPR04]. It is seen that the Pearson system based ICA works reliably, and also separates skewed distributions with high fidelity. All distributions modeled by Eq. (5.9) are unimodal. For the extension of the Pearson system to multi-modal distributions, and also for some additional simulation examples, see [KK02].

5.4 Discussion

In his 1998 review paper [Car98], J-F. Cardoso states that the design of practical algorithms achieving source adaptivity is the first open question in ICA. The methods proposed in sections 5.2 and 5.3 provide feasible solutions to this problem. Both the proposed methods can model asymmetric distributions unlike other parametric models proposed for ICA in the literature. The EGLD family covers slightly wider class of distributions than the Pearson system, and it can be therefore considered a bit more general. However, the computational simplicity allowed by the direct relationship to the score function makes the Pearson system an appealing choice for a source adaptive ICA method.

Both EGLD and Pearson methods are able to separate significantly larger family of distributions than any conventional method with any fixed nonlinear score function. Furthermore, the computational cost is only marginally higher. Therefore, these adaptive methods should be preferred over the traditional fixed nonlinearity methods unless one has some specific a priori knowledge about the source distributions. Furthermore, both proposed methods can exploit the skewness information of the distribution. Therefore, they may be also preferable to the traditional nonparametric

methods such as JADE [Car] implicitly assuming that the distributions are symmetric.

Chapter 6

Characteristic function-based methods for blind separation

The mutual information-based separation methods discussed in Chapter 5 lead essentially to the estimation of the score function. However, the distribution and the score of the sources are *nuisance parameters* for separable models. One may ask if it is really necessary to estimate them. Furthermore, the use of m.i. limits the methods to class of the source r.v.'s where p.d.f.'s exist, and the direct estimation of m.i. is hard. A solution for these problems may be obtained by using the independence characterization of Eq. (2.11) instead of m.i. (2.8) as a criterion. Cumulant-based methods are reviewed in Section 6.1, and a method based on derivatives of s.c.f. is reviewed in Section 6.2. Finally, a c.f.-based method originally introduced in Publication IV and further improved in Publication V and in Publication VII is proposed in Section 6.3.

6.1 Cumulant-based methods

An alternative criterion for independence can be constructed based on Eq. (2.13) showing that all cross-cumulants should vanish for independent marginals. Additionally, the Taylor expansion of Eq. (2.13) gives the optimal relative weighting between different order cumulants to be minimized. However, cumulants need to be estimated in any cumulant-based ICA method, thus factors like estimation error variance should be also considered in designing such an ICA method.

Popular methods such as JADE [Car] use prewhitening to obtain an orthogonal mixture, and then find the final orthonormal matrix by minimizing a criterion based on equally weighted fourth order cumulants (see [HKO01]). This implicitly assumes that the source distributions are symmetrical. Some more advanced methods (see [PM01, Mor01]) incorporate third order cumulants with different weighting between third and fourth order cumulants at

the price of increased computational complexity. Computationally attractive method optimally combining second and fourth order cumulants was recently proposed in [SY04].

Since there is one-to-one correspondence between cumulants and moments, the methods based on polynomial approximations, such as Gram-Charlier (see [HKO01]), of m.i. (2.8) can be viewed as cumulant-based methods. Furthermore, cumulants are usually estimated through sample moments.

6.2 Characteristic function enabled source separation

Instead of approximating Eq. (2.13) with cumulants, an alternative method based on the same equation and called *CHaracteristic-function Enabled Source Separation* (CHESS) was proposed in [Yer00]. If the 2nd order moments of a r.vc. $\bar{\boldsymbol{x}}$ exist, the Hessian of s.c.f. exist. Now the identity $\phi_{\boldsymbol{A}\bar{\boldsymbol{s}}}(\boldsymbol{t}) = \phi_{\bar{\boldsymbol{s}}}(\boldsymbol{A}^T\boldsymbol{t})$ implies $\frac{\partial^2}{\partial \boldsymbol{t}^2}\psi_{\boldsymbol{A}\bar{\boldsymbol{s}}}(\boldsymbol{t}) = \boldsymbol{A}\big(\frac{\partial^2}{\partial \boldsymbol{t}^2}\psi_{\bar{\boldsymbol{s}}}(\boldsymbol{t})\big)\boldsymbol{A}^T$. Since $\frac{\partial^2}{\partial \boldsymbol{t}^2}\psi_{\bar{\boldsymbol{s}}}(\boldsymbol{t})$ is diagonal for independent marginals, one could separate the mixture of Eq. (3.1) with the following method:

1. Estimate the matrix $\frac{\partial^2}{\partial \boldsymbol{t}^2}\psi_{\bar{\boldsymbol{x}}}(\boldsymbol{t})\big)$ at points $\boldsymbol{t} \in \{\boldsymbol{t}_1, \ldots, \boldsymbol{t}_K\}$.

2. Jointly diagonalize (see Section 2.4.2) the K matrices.

The matrices are easily estimated from the realization vectors $\boldsymbol{x}[k]$, $k = 1, \ldots, N$, using *empirical characteristic functions* (e.c.f.'s). Namely, the value of c.f. $\phi_{\bar{\boldsymbol{x}}}$ of r.vc. $\bar{\boldsymbol{x}}$ can be directly estimated at any point $\boldsymbol{t}$ by the e.c.f. defined [Csö81, Ush99] as

$$\hat{\phi}_{\bar{\boldsymbol{x}}}(\boldsymbol{t}) \triangleq \frac{1}{N}\sum_{k=1}^{N} e^{i\langle \boldsymbol{t}, \boldsymbol{x}[k]\rangle}. \tag{6.1}$$

At every *fixed* point $\boldsymbol{t}$, the empirical characteristic function is an unbiased estimator of the corresponding characteristic function, and the estimator converges both almost surely and in the mean square sense which implies consistency. Further, it is almost surely uniformly consistent on each bounded subset in $\mathbb{R}^p$.

Additionally, it was proposed in [Yer00] to evaluate the second derivative matrices $\frac{\partial^2}{\partial \boldsymbol{t}^2}\psi_{\bar{\boldsymbol{x}}}(\boldsymbol{z})$ also with complex arguments $\boldsymbol{z} \in \mathbb{C}^p$. However, the benefits of this extension are unclear.

6.3 Jacobi optimized empirical characteristic function ICA

The advantage of the difference $\Delta_{\phi_{\bar{\boldsymbol{x}}}}(\boldsymbol{t})$ of Eq. (2.12) compared to the analogous difference obtained from Eq. (2.4) is that the former allows direct estimate of its value at any point using the e.c.f.'s. The square

$$|\Delta_{\phi_{\bar{\boldsymbol{x}}}}(\boldsymbol{t})|^2 = |\phi_{\bar{\boldsymbol{x}}}(\boldsymbol{t}) - \prod_{k=1}^{p} \phi_{\bar{x}_k}(t_k)|^2 \tag{6.2}$$

of the difference (2.12) is always nonnegative, and zero for all $\boldsymbol{t} \in \mathbb{R}^p$ iff r.vc. $\bar{\boldsymbol{x}}$ has independent components. Thus an objective function (see Section 2.2) is obtained if c.f.'s in Eq.(6.2) are replaced by their empirical counterparts (6.1), and the function is evaluated at a point $\boldsymbol{t}_0$. Since the theoretical difference can then vanish at a point even if the marginals are dependent, a better objective function is obtained by summing the square values evaluated at different points $\boldsymbol{t}_k$. This e.c.f.-based ICA method is introduced in Publication IV. Minimization is done by orthogonal Jacobi algorithm (see Section 2.4.2), where the required two dimensional minimization is achieved numerically by using the golden section search [PTVF92]. See also Publication V for a derivation based on the separating class of Eq. (2.6).

Since the difference (2.12) can vanish even for dependent marginals, it should be evaluated at every point to guarantee independence. One way to do this is to integrate the nonnegative quadratic function (6.2) over $\mathbb{R}^p$, i.e. to consider independence criteria of the form

$$\int_{\mathbb{R}^p} |\Delta_{\bar{\boldsymbol{x}}}(\boldsymbol{t})|^2 \boldsymbol{g}_p(\boldsymbol{t}) d\boldsymbol{t}, \tag{6.3}$$

where $\boldsymbol{g}_p(\cdot)$ is an appropriate nonnegative weight function. This functional form is considered in [Kan95, KU98] for testing the independence, and introduced in Publication V as an ICA criterion. The empirical counterpart of Eq. (6.3) (see Eq. (9) in Publication VII) has a desirable statistical large sample property. Namely it is a consistent estimator of independence for all r.vc.'s with an appropriate weight function. One such weight function is the product of standard normal p.d.f.'s which was used in [Kan95] and in Publication V. Another appropriate weight function is the product of standard Laplace p.d.f.'s, which was introduced in Publication VII. The objective functions associated with these weight functions are called *consistent objective functions*. It was recently proved [CB04] that such objective functions are indeed consistent, i.e. the matrix minimizing the objective function converges to the true demixing matrix (with fundamental indeterminacy eliminated) as the signal length approaches infinity. If the p.d.f.'s exist, the criteria (6.3) is related through Parseval's theorem to the quadratic independece criteria formed from the independece characterization (2.7) [APJ03].

This suggests connection to *Kernel-ICA* [BJ02]. See also [APJ03, Ach03] for additional weight functions, which are derived as the Fourier transform of a summable kernel function.

Minimization of the consistent objective functions can be done similarly to the objective function corresponding to Eq. (6.2). However, the evaluation of the two-dimensional consistent objective function has high computational complexity. In order to minimize the number of evaluations needed, an alternative minimization method was proposed in Publication VII. It is based on the property that the objective functions under two-dimensional orthonormal transformations are periodic and have a very good Fourier series approximation with only very few coefficients. The ICA method based on this minimization of a consistent objective function is called *Jacobi optimized Empirical Characteristic Function ICA* (JECFICA). The simulation results presented in figures 5.1 and 5.2 show reliable performance for skewed signals already with a relatively short signal length. JECFICA was also tested for discrete distributions with four Poisson(1) distributed signals using the performance index (see Eq. (5.8)) as the performance measure. Its superior performance among tested methods is apparent from figures 6.1 and 6.2. See Publication VII for additional simulations showing that JECFICA is an extremely reliable ICA method. Finally, see Publication VII for extension of the difference (2.12) for noisy ICA.

6.4 Discussion

The JECFICA method proposed in Publication VII is the only existing ICA method so far that directly minimizes a criterion quantifying independence for all r.v.'s. Furthermore, the consistency property [CB04] of its objective functions establishes that the method has desirable large sample properties that remain to be found for other ICA methods. Methods that do not use the c.f. criterion are either based on a criterion that do not guarantee independence for all r.v.'s, or they need approximations for the corresponding criterion. The other methods based on the c.f. criterion do not take the full advantage of its possibilities. The cumulant-based methods essentially approximate the criterion through Taylor series, and CHESS relies on a few points in the criterion. Despite extensive simulations, we are yet to face a mixture with reasonable signal length where JECFICA fails. The main drawback of the method is the computaional complexity, since the number of operations grows in square with respect to the number of source signals. Moreover, the number of operations for the introduced consistent objective functions also grows in square with respect to the number of signal samples.

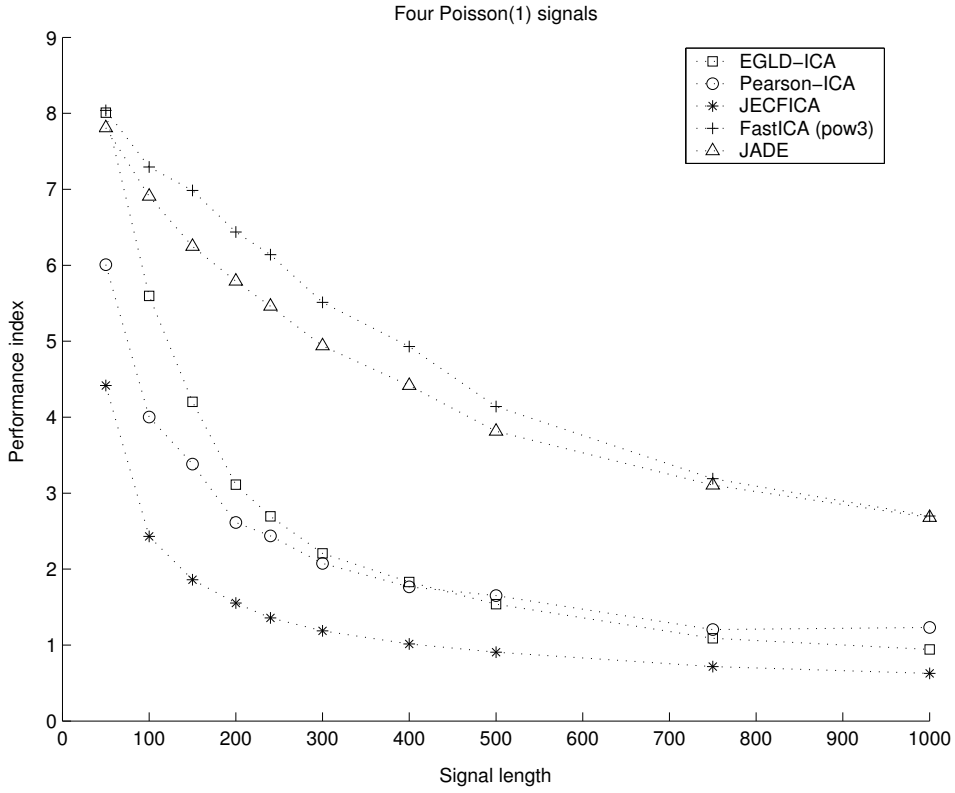

Figure 6.1: Performance of different ICA methods for four Poisson(1) distributed signals with varying signal length. Compared methods: EGLD-ICA (Section 5.2), Pearson-ICA (Section 5.3), JECFICA (Section 6.3), FastICA [H$^+$], and JADE [Car].

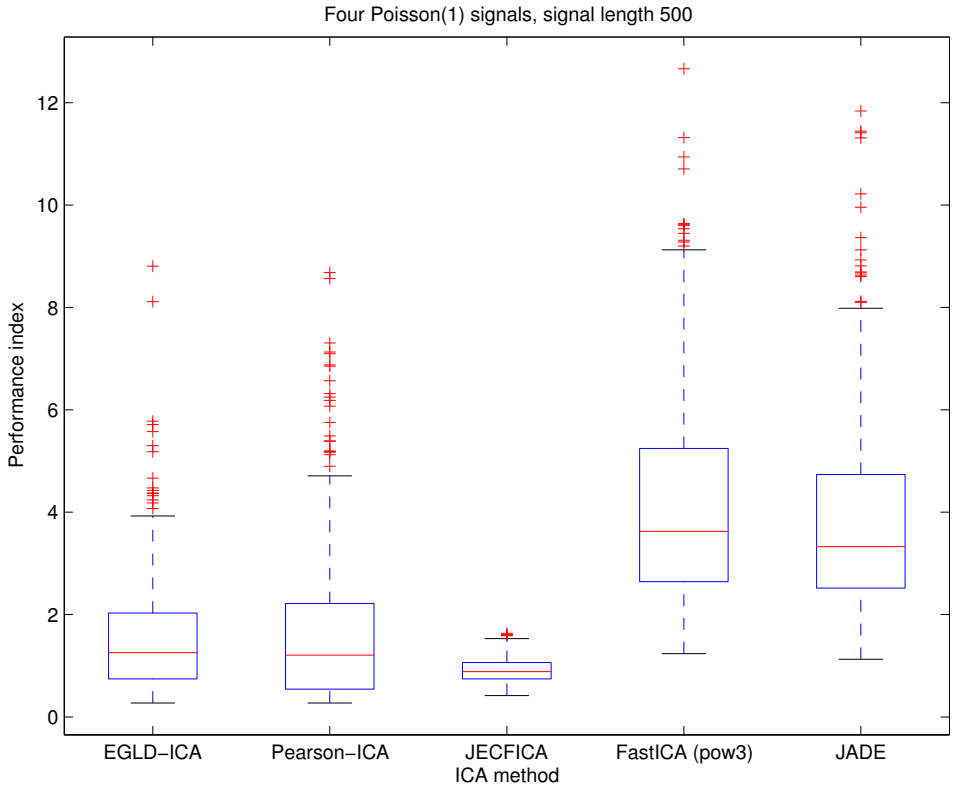

Figure 6.2: Performance of different ICA methods for four Poisson(1) distributed signals with signal length 500. Compared methods: EGLD-ICA (Section 5.2), Pearson-ICA (Section 5.3), JECFICA (Section 6.3), FastICA [H$^+$], and JADE [Car].

Chapter 7

Conclusion

7.1 Summary

This thesis addressed the problem of blind signal separation using independent component analysis (ICA). These methods allow solving many difficult signal processing problems in different application domains. Blind techniques are very important in applications for which there are sets of recorded data or observations, but exact details of the system which produced the observations are unknown. The ICA-based methods have shown to be very successful blind approach.

In this thesis, the theoretical conditions for the traditional linear ICA model to qualify as blind separation model were established. It was shown that it is both sufficient and necessary that the mixing matrix is of full column rank and that there is at most one Gaussian source signal. It was also shown that for identifying the mixing matrix it is enough to have no Gaussian signals no matter how many sources signals there are for a given number of sensors. This gives conditions for the linear ICA techniques to be used as a blind system identification methods. Additionally, a stronger concept of uniqueness applying also to the cases where there are more source signals than sensors was introduced. Uniqueness quantifies the situations where ICA models could be used to develop blind signal extraction techniques when there are more sources than sensors. Sufficient conditions were also found for uniqueness.

The results of the real-valued linear ICA models were extended into two directions. First, the conditions for separability, identifiability, and uniqueness were derived for the models with complex-valued signals. Separability condition was again shown to be both sufficient and necessary. Surprisingly, in a complex case, some Gaussian signal mixtures were shown to be separable. In order to establish this result, a novel characterization of second-order statistics of complex random vectors and decomposition of complex normal random vectors were derived. Second, the real-valued linear models were

extended to nonlinear mixing models by introducing a class of nonlinear models. These models were shown to be separable, and a generic separation technique was proposed. An instance of this model class was applied successfully to an image enhancement problem.

Finally, two different types of techniques for blind source separation based on ICA were proposed in this thesis. In the first approach, source adaptation was proposed as a method for separation. The EGLD family of distributions and the Pearson system were used as a flexible models for source distributions. The strength of these parametric families is that they model a wide class of source distributions, including asymmetric distributions, while the estimation of the parameters is still a relatively simple procedure. In the second approach, the idea was to measure directly stochastic independence of the signals instead of trying to model the sources. It was shown that characteristic functions provide a criterion quantifying the independence for all distributions. Furthermore, the terms in the expression are relatively easily estimated. Based on this expression, an objective function was proposed. It was recently shown to be consistent [CB04] and thereby it guarantees a good performance. Finally, an algorithm for relatively fast minimization of the consistent objective function was derived.

7.2 Future work

The linear real-valued ICA model as a tool for blind separation has been extensively studied over the past fifteen years. The methods presented in this thesis are highly reliable for all type of sources while still being relatively fast. Thus, it is natural to consider this basic separation of instantaneous mixtures of real-values sources a relatively well-understood problem with feasible solutions. However, there are relatively few ICA techniques for blind identification and even fewer for blind source extraction for the underdetermined case. These techniques should have a huge demand in some application areas, e.g. in communications. The development of these methods ought to get more attention in the future. A possible option is to base these methods on moment generating functions as noted in the end of Chapter 3.

Surprisingly, only few authors have considered directly the linear instantaneous ICA model in the complex domain. Since it was shown in this thesis that the complex-valued ICA models are also well-defined, a reliable separation of complex mixtures should be possible. An interesting option would be an extension of the characteristic function based techniques to complex signals.

Nonlinear ICA models have so far found only few applications. This is more likely due to the fact that nonlinear methods are relatively unknown and difficult to analyse than due to the lack of possible applications. Nonlin-

ear models proposed in this thesis should cover many situations of practical interest. However, they might be more appealing if the methods developed would be more flexible to the exact knowledge of the system structure. As noted in the end of Chapter 4, a possibility is to form hierarchical parametric models from individual addition theorem models. The development of separation methods for such models is a challenging task. However, one should benefit greatly from ideas used to develop separating algorithms for PNL mixtures.

Bibliography

[Ach03] S. Achard. *Mesures de dépendance pour la séparation aveugle de sources*. PhD thesis, L'Université Joseph Fourier (Grenoble I), 2003. (In French).

[ACY96] S. Amari, A. Cichocki, and H. Yang. A new learning algorithm for blind signal separation. In D. Touretsky, M. Mozer, and H. Hasselmo, editors, *Advances in Neural Information Processing Systems 8*, pages 757–763. MIT Press, 1996.

[Acz66] J. Aczél. *Lectures on Functional Equations and Their Applications*. Academic Press, New York, NY, 1966.

[AGL96a] P.O. Amblard, M. Gaeta, and J.L. Lacoume. Statistics for complex variables and signals – Part I: Variables. *Signal Processing*, 53:1–13, 1996.

[AGL96b] P.O. Amblard, M. Gaeta, and J.L. Lacoume. Statistics for complex variables and signals – Part II: signals. *Signal Processing*, 53:15–25, 1996.

[APJ03] S. Achard, D.T. Pham, and C. Jutten. Quadratic dependence measure for nonlinear Blind Source Separation. In *Proc. of the 4th Int. Symposium on ICA and BSS*, pages 263–268, Nara, Japan, April 2003.

[ASM03] J. Anemüller, T.J. Sejnowski, and S. Makeig. Complex independent component analysis of frequency-domain electroencephalographic data. *Neural Networks*, 16(9):1311–1323, November 2003.

[BAMCM97] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines. A blind source separation technique using second-order statistics. *IEEE Trans. Signal Processing*, 45(2):434, February 1997.

[BH00] E. Bingham and A. Hyvärinen. A fast fixed-point algorithm for independent component analysis of complex valued signals. *Int. J. Neural Systems*, 10(1):1–8, February 2000.

[BJ02] F.R. Bach and M.I. Jordan. Kernel independent component analysis. *J. Machine Learning Research*, 3:1–48, July 2002.

[BPR04] R. Boscolo, H. Pan, and V.P. Roychowdhury. Independent component analysis based on nonparametric density estimation. *IEEE Trans. Neural Networks*, 15(1):55–65, January 2004.

[Bre92] L. Breiman. *Probability*. Classics in Applied Mathematics. SIAM, 1992.

[BS95] A.J. Bell and T.J. Sejnowski. An information maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, November 1995.

[BZ02] M. Babaie-Zadeh. *On blind source separation in convolutive and nonlinear mixtures*. PhD thesis, INPG, Grenoble, France, 2002.

[BZJN02] M. Babaie-Zadeh, C. Jutten, and K. Nayebi. A geometric approach for separating post non-linear mixtures. In *Proc. of XI European Signal Processing Conference*, pages 11–14, Toulouse, France, September 2002.

[BZP00] Y. Blanco, S. Zazo, and J. Principe. Alternative statistical Gaussianity measure using the cumulative density function. In *Proc. of the 2nd Int. Workshop on ICA and BSS*, pages 537–542, Helsinki, Finland, June 2000.

[CA02] A. Cichocki and S. Amari. *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*. John Wiley & Sons, 2002.

[Car] J.-F. Cardoso. JADE algorithm. Matlab code with references. http://www.tsi.enst.fr/∼cardoso/stuff.html.

[Car93] J-F. Cardoso. An efficient technique for the blind separation of complex sources. In *Proc. of the IEEE Signal Proc. Workshop on Higher-Order Statistics*, pages 275–279, South Lake Tahoe, CA, June 1993.

[Car96] J.-F. Cardoso. Jacobi angles for simultaneous diagonalization. *SIAM J. Matrix Anal. Appl.*, 17(1):161–164, January 1996.

[Car98] J.-F. Cardoso. Blind signal separation: Statistical principles. *Proc. IEEE*, 86(10):2009–2025, October 1998.

[Car99] J.-F. Cardoso. Higher-order contrast for independent component analysis. *Neural Computation*, 11(1):157–192, January 1999.

[CB04] A. Chen and P.J. Bickel. Consistent independent component analysis and prewhitening. *IEEE Trans. Signal Processing*, February 2004. Submitted.

[CCA00] S. Choi, A. Cichocki, and S. Amari. Flexible independent component analysis. *J. VLSI Signal Processing*, 26(1/2):25–38, 2000.

[CL96] J.-F. Cardoso and B. H. Laheld. Equivariant adaptive source saparation. *IEEE Trans. Signal Processing*, 44(12):3017–3030, December 1996.

[CM99] J. Cao and M. Murata. A stable and robust ica algorithm based on t-distribution and generalized gaussian distribution models. In *Proc. of the IEEE Signal Proc. Society Workshop on Neural Networks for Signal Processing*, pages 283–292, Madison, WI, August 1999.

[Com94] P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, April 1994.

[Com04] P. Comon. Blind identification and source separation in 2x3 under-determined mixtures. *IEEE Trans. Signal Processing*, 52(1):11–22, January 2004.

[CS93] J.-F. Cardoso and A. Souloumiac. Blind beamforming for non-gaussian signals. *IEE Proceedings-F*, 140(6):362–370, December 1993.

[Csö81] S. Csörgő. Multivariate empirical characteristic functions. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 55:203–229, 1981.

[CT91] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 1991.

[Cup75] R. Cuppens. *Decomposition of Multivariate Probabilities*, volume 29 of *Probability and Mathematical Statistics*. Academic Press, New York, NY, 1975.

[CUR94] A. Cichocki, R. Unbehauen, and E. Rummert. Robust learning algorithm for blind source separation of signals. *Electronics Letters*, 30(17):1386–1387, August 1994.

[DAK00] S.C. Douglas, S. Amari, and S.-Y. Kung. On gradient adaptation with unit-norm constraints. *IEEE Trans. Signal Processing*, 48(6):1843–1847, June 2000.

[DK96] E.J. Dudewicz and Z.A. Karian. The extended generalized lambda distribution (EGLD) system for fitting distributions to data with moments, II: Tables. *Amer. J. of Math. Management Sci.*, 16(3 & 4):271–332, 1996.

[DL95] N. Delfosse and P. Loubaton. Adaptive blind separation of independent sources: a deflation approach. *Signal Processing*, 45(1):59–83, July 1995.

[Dou00] S.C. Douglas. Self-stabilized gradient algorithms for blind source separation with orthogonality constraints. *IEEE Trans. Neural Networks*, 11(6):1490–1497, November 2000.

[DT00] G.A. Darbellay and P. Tichavský. Independent component analysis through direct estimation of the mutual information. In *Proc. of the 2nd Int. Workshop on ICA and BSS*, pages 69–74, Helsinki, Finland, June 2000.

[EAS98] A. Edelman, T.A. Arias, and S.T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.*, 20(2):303–353, 1998.

[Feu93] A. Feuerverger. A consistent test for bivariate dependence. *International Statistical Review*, 61(3):419–433, 1993.

[Fio03] S. Fiori. Extended hebbian learning for blind separation of complex-valued sources. *IEEE Trans. Circuits Syst. II*, 50(4):195–202, April 2003.

[GC98] O. Grellier and P. Comon. Blind separation of discrete sources. *IEEE Signal Processing Lett.*, 5(8):212–214, August 1998.

[GL89] G.H. Golub and C.F.V. Loan. *Matrix computations*. The Johns Hopkins University Press, 2nd edition, 1989.

[Gol73] A.A. Goldberg. On a problem of Ju. V. Linnik. *Soviet Math. Dokl.*, 14(4):950–953, 1973.

[Gra90] R.M. Gray. *Entropy and Information Theory*. Springer-Verlag, New York, NY, 1990.

[H$^+$] A. Hyvärinen et al. FastICA algorithm. Matlab code with references. http://www.cis.hut.fi/projects/ica/fastica/.

[Hay96] S. Haykin. *Adaptive Filter Theory*. Prentice Hall, 3rd edition, 1996.

[Hay00] S. Haykin, editor. *Unsupervised Adaptive Filtering: Vol. 1: Blind Source Separation*. Wiley, 2000.

[HKO01] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley & Sons, 2001.

[HL] C.T. Hill and M.L. Lederman. Teaching symmetry in the introduction physics curriculum. Online article. http://www.emmynoether.com/.

[HO97] A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483–1492, 1997.

[HP99] A. Hyvärinen and P. Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, April 1999.

[JBZH04] C. Juttan, M. Babaie-Zadeh, and S. Hosseini. Three easy ways for separating nonlinear mixtures? *Signal Processing*, 84(2):217–229, February 2004.

[JK03] C. Jutten and J. Karhunen. Advances in nonlinear blind source separation. In *Proc. of the 4th Int. Symposium on ICA and BSS*, pages 245–256, Nara, Japan, April 2003.

[Jut87] C. Jutten. *Calcul neuromimétique et traitment du signal, analyse en composantes indépendantes*. PhD thesis, INPG, University of Grenoble, France, 1987. (In French).

[Kan95] A. Kankainen. *Consistent Testing of Total Independence Based on Empirical Characteristic Function*. PhD thesis, University of Jyväskylä, Finland, 1995.

[KDM96] Z.A. Karian, E.J. Dudewicz, and P. McDonald. The extended generalized lambda distribution system for fitting distributions to data: history, completion of theory, tables, applications, the "final word" on moment fits. *Comm. in Statistics: Simulation and Computation*, 25(3):611–642, 1996.

[KK02] J. Karvanen and V. Koivunen. Blind separation methods based on pearson system and its extensions. *Signal Processing*, 82(4):663–673, April 2002.

[KLR73] A.M. Kagan, Yu.V. Linnik, and C.R. Rao. *Characterization Problems in Mathematical Statistics*. Probability and Mathematical Statistics. John Wiley & Sons, New York, NY, 1973.

[KM99] R.R. King and H. MacGillivray. A starship method for the generalized lambda distributions. *Australian New Zealand J. Statistics*, 41(3):353–374, 1999.

[KU98] A. Kankainen and N.G. Ushakov. Consistent modification of a test for independence based on the empirical characteristic functions. *Journal of Mathematical Sciences*, 89(5):1486–1493, 1998.

[Kul68] S. Kullback. *Information Theory and Statistics*. Dover Publications, Inc., New York, NY, reprint edition, 1968.

[LLS99] T-W. Lee, M.S. Lewicki, and T.J. Sejnowski. Blind source separation of more sources than mixtures using overcomplete representations. *IEEE Signal Processing Lett.*, 6(4):87–90, April 1999.

[LMVC03] L. De Lathauwer, B. De Moor, J. Vandewalle, and J.-F. Cardoso. Independent component analysis of largely underdetermined mixtures. In *Proc. of the 4th Int. Symposium on ICA and BSS*, pages 29–33, Nara, Japan, April 2003.

[LO77] Yu.V. Linnik and I.V. Ostrovskiĭ. *Decomposition of Random Variables and Vectors*, volume 48 of *Translation of Mathematical Monographs*. AMS, 1977.

[Lue73] D.G. Luenberger. *Introduction to Linear and Nonlinear Programming*. Addison-Wesley, Reading, MA, 1973.

[Luk70] E. Lukacs. *Characteristic Functions*. Griffin, London, Great Britain, second edition, 1970.

[LY03] Y. Lomnitz and A. Yeredor. A blind-ML scheme for blind source separation. In *Proc. of the 2003 IEEE Workshop on Statistical Signal Processing*, pages 581–584, St. Louis, MO, October 2003.

[Man02] J.H. Manton. Optimization algorithms exploiting unitary constraints. *IEEE Trans. Signal Processing*, 50(3):635–650, March 2002.

[MD02] H. Mathis and S.C. Douglas. On the existence of universal nonlinearities for blind source separation. *IEEE Trans. Signal Processing*, 50(5):1007–1016, May 2002.

[MI03] E.G. Miller and J.W. Fisher III. Independent components analysis by direct entropy minimization. Report No. UCB/CSD-3-1221, University of California, Berkeley, CA, January 2003.

[Mor01] E. Moreau. A generalization of joint-diagonalization criteria for source separation. *IEEE Trans. Signal Processing*, 49(3):530–541, March 2001.

[NM93] F.D. Neeser and J.L. Massey. Proper complex random processes with applications to information theory. *IEEE Trans. Inform. Theory*, 39(4):1293–1302, July 1993.

[NS95] C.L. Nikias and M. Shao. *Signal Processing with α-Stable Distributions and Applications*. Wiley, New York, NY, 1995.

[OS89] A.V. Oppenheim and R.W. Schafer. *Discrete-Time Signal Processing*. Prentice Hall, Englewood Cliffs, NJ, 1989.

[Pan03] L. Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253, June 2003.

[PB97] B. Picinbono and P. Bondon. Second-order statistics of complex signals. *IEEE Trans. Signal Processing*, 45(2):411–420, February 1997.

[PG97] D.T. Pham and P. Garat. Blind separation of mixture of independent sources through a quasi-maximum likelihood approach. *IEEE Trans. Signal Processing*, 45(7):1712–1725, July 1997.

[Pha96] D.T. Pham. Blind separation of instantaneous mixture of sources via an independent component analysis. *IEEE Trans. Signal Processing*, 44(11):2768–2779, November 1996.

[Pic96] B. Picinbono. Second-order complex random vectors and normal distributions. *IEEE Trans. Signal Processing*, 44(10):2637–2640, October 1996.

[PK97] P. Pajunen and J. Karhunen. Blind separation of binary sources with less sensors than sources. In *Proc. of the 1997 Int. Conference on Neural Networks*, pages 1994–1997, Houston, TX, June 1997.

[PM01] J-C. Pesquet and E. Moreau. Cumulant-based independence measures for linear mixtures. *IEEE Trans. Inform. Theory*, 47(5):1947–1956, July 2001.

[PTVF92] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, second edition, 1992.

[Sco87] W.R. Scott. *Group Theory*. Dover Publications, Inc., New York, NY, reprint edition, 1987.

[SR00] G.J. Székely and C.R. Rao. Identifiability of distributions of independent random variables by linear combination and moments. *Sankhyā, Series: A*, 62(2):193–202, 2000.

[SS03] P.J. Schreier and L.L. Scharf. Second-order analysis of improper complex random vectors and processes. *IEEE Trans. Signal Processing*, 51(3):714–725, March 2003.

[SY04] A. Smekhov and A. Yeredor. Optimization of JADE using a novel optimally weighted joint diagonalization approach. In *Proc. of the 2004 European Conference on Signal Processing*, Vienna, Austria, September 2004. to appear.

[SYM01] Y. Shereshevski, A. Yeredor, and H. Messer. Super-efficiency in blind signal separation of symmetric heavy-tailed source. In *Proc. of the 2001 IEEE Workshop on Statistical Signal Porcessing (SSP2001)*, pages 78–81, Singapore, August 2001.

[Tal01a] A. Taleb. An algorithm for the blind identification of n independent signals with 2 sensors. In *Proc. of International Symposium on Signal Processing and Its Applications*, pages 5–8, Kuala Lumpur, Malaysia, August 2001.

[Tal01b] A. Taleb. Source separation in structured nonlinear models. In *Proc. of the 26th Int. Conference on Acustics, Speech, and Signal Processing*, pages 3513–3516, Salt Lake City, UT, May 2001.

[Tal02] A. Taleb. A generic framework for blind source separation in structured nonlinear models. *IEEE Trans. Signal Processing*, 50(8):1819–1830, August 2002.

[TJ99a] A. Taleb and C. Jutten. Batch algorithm for source separation in post-nonlinear mixtures. In *Proc. of the First Int. Workshop on ICA and Signal Separation*, pages 155–160, Aussois, France, March 1999.

[TJ99b] A. Taleb and C. Jutten. On underdetermined source separation. In *Proc. of the 24th Int. Conference on Acustics, Speech, and Signal Processing*, volume 3, pages 1445–1448, Phoenix, AZ, March 1999.

[TJ99c] A. Taleb and C. Jutten. Source separation in post-nonlinear mixtures. *IEEE Trans. Signal Processing*, 47(10):2807–2820, October 1999.

[Tuk60] J.W. Tukey. The practical relationship between the common transformations of percentages of counts and of amounts. Technical Report 36, Statistical Techniques Research Group, Princeton University, 1960.

[Ush99] G. Ushakov. *Selected Topics in Characteristic Functions*. Modern Probability and Statistics. VSP VB, Utrect, The Netherlands, 1999.

[VK96] N.N. Vakhania and N.P. Kandelaki. Random vectors with values in complex Hilbert spaces. *Theory Probab. Appl.*, 41(1):116–131, February 1996.

[Yer00] A. Yeredor. Blind source separation via the second characteristic function. *Signal Processing*, 80(5):897–902, May 2000.

[Yer02] A. Yeredor. Non-orthogonal joint diagonalization in the least-squares sense with application in blind source separation. *IEEE Trans. Signal Processing*, 50(7):1545–1553, July 2002.