Helsinki University of Technology Signal Processing Laboratory

Teknillinen korkeakoulu Signaalinkäsittelytekniikan laboratorio

Espoo 2002 Report 35

ADAPTIVE METHODS FOR SCORE FUNCTION MODELING IN BLIND SOURCE SEPARATION

Juha Karvanen

Helsinki University of Technology Signal Processing Laboratory
Teknillinen korkeakoulu Signaalinkäsittelytekniikan laboratorio
Espoo 2002 Report 35

ADAPTIVE METHODS FOR SCORE FUNCTION MODELING IN BLIND SOURCE SEPARATION

Juha Karvanen

Dissertation for the degree of Doctor of Science in Technology to be presented with due permission for public examination and debate in Auditorium S4 at Helsinki University of Technology (Espoo, Finland) on the 26th of August, 2002, at 12 o'clock 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 2486
Fax. +358-9-460 224
E-mail: anne.jaaskelainen@hut.fi

ISBN 951-22-5990-7
ISSN 1458-6401

Otamedia Oy
Espoo 2002

Abstract

In signal processing and related fields, multichannel measurements are often encountered. Depending on the application, for instance, multiple antennas, multiple microphones or multiple biomedical sensors are used for the data acquisition. Such systems can be described using Multiple-Input Multiple-Output (MIMO) system models. In many cases, several source signals are present at the same time and there is only limited knowledge of their properties and how they contribute to each sensor output. If the source signals and the physical system are unknown and only the sensor outputs are observed, the processing methods developed for recovering the original signals are called blind.

In Blind Source Separation (BSS) the goal is to recover the source signals from the observed mixed signals (mixtures). Blindness means that neither the sources nor the mixing system is known. Separation can be based on the theoretically limiting but practically feasible assumption that the sources are statistically independent. This assumption connects BSS and Independent Component Analysis (ICA). The usage of mutual information as a measure of independence leads to iterative estimation of the score functions of the mixtures.

The purpose of this thesis is to develop BSS methods that can adapt to different source distributions. Adaptation makes it possible to separate sources without knowing the source distributions or even the characteristics of source distributions. Special attention is paid to methods that allow also asymmetric source distributions. Asymmetric distributions occur in important applications such as communications and biomedical signal processing. Adaptive techniques are proposed for the modeling of score functions or estimating functions. Three approaches based on the Pearson system, the Extended Generalized Lambda Distribution (EGLD) and adaptively combined fixed estimating functions are proposed. The Pearson system and the EGLD are parametric families of distributions and they are used to model the distributions of the mixtures. The strength of these parametric families is that they

contain a wide class of distributions, including asymmetric distributions with positive and negative kurtosis, while the estimation of the parameters is still a relatively simple procedure. The methods may be implemented using existing ICA algorithms.

The reliable performance of the proposed methods is demonstrated in extensive simulations. In addition to symmetric source distributions, asymmetric distributions, such as Rayleigh and lognormal distribution, are utilized in simulations. The score adaptive methods outperform commonly used methods due to their ability to adapt to asymmetric distributions.

Preface

Ja minä käänsin sydämeni tutkimaan viisautta ja tietoa, mielettömyyttä ja tyhmyyttä,
ja minä tulin tietämään, että sekin oli tuulen tavoittelemista. Saarn. 1:17

The work reported in this thesis was carried out in the Signal Processing Laboratory, Helsinki University of Technology during the years 1999–2002. The period included three months in the Signal Processing Laboratory at University of Pennsylvania, Philadelphia, USA.

I wish to express my gratitude to my supervisor Prof. Visa Koivunen. I would like to thank my thesis reviewers Dr. Aapo Hyvärinen and Dr. Jyrki Möttönen for their constructive comments. I wish to thank Prof. Saleem Kassam for my visit to University of Pennsylvania. I am also grateful to my colleagues and co-workers, especially Jan Eriksson, Dr. Yinglu Zhang and Dr. Charles Murphy. In addition, I would like to thank my parents Toivo and Marja-Liisa for their support.

The research was funded by the Academy of Finland and the Graduate School in Electronics, Telecommunications and Automation (GETA). Additional financial support was provided by Elektroniikkainsinöörien säätiö, the Nokia Foundation and Tekniikan edistämissäätiö. These organizations and foundations are greatly acknowledged for making this work possible.

Espoo, Finland
May 13, 2002

Juha Karvanen

Contents

List of original publications

I J. Karvanen, J. Eriksson and V. Koivunen. Pearson System based Method for Blind Separation. In *Proc. of The Second International Workshop on Independent Component Analysis and Blind Signal Separation, ICA2000*, pages 585–590, June 2000.

II J. Eriksson, J. Karvanen and V. Koivunen. Source Distribution Adaptive Maximum Likelihood Estimation of ICA Model. In *Proc. of The Second International Workshop on Independent Component Analysis and Blind Signal Separation, ICA2000*, pages 227–232, June 2000.

III J. Karvanen, J. Eriksson and V. Koivunen. Maximum Likelihood Estimation of ICA model for Wide Class of Source Distributions. In *Proc. of the 2000 IEEE Workshop on Neural Networks for Signal Processing X*, pages 445–454, December 2000.

IV J. Karvanen and V. Koivunen. Blind Separation of Communication Signals Using Pearson System Based Method. In *Proc. of The Thirty-Fifth Annual Conference on Information Sciences and Systems*, Volume II, pages 764–767, March 2001.

V J. Karvanen and V. Koivunen. Blind Separation Methods Based on Pearson system and its Extensions. *Signal Processing* Volume 82, Issue 4, pages 663–673, April 2002.

VI J. Karvanen, J. Eriksson and V. Koivunen. Adaptive Score Functions for Maximum Likelihood ICA. *Journal of VLSI Signal Processing*, Volume 32, pages 83–92, 2002.

VII J. Karvanen and V. Koivunen. Blind Separation using Absolute Moments Based Adaptive Estimating Function. In *Proc. of the Third International Conference on Independent Component Analysis and Signal Separation, ICA2001*, pages 218–223, December 2001.

List of abbreviations and symbols

Abbreviations

BER	bit error rate
BSS	blind source separation
cdf	cumulative distribution function
EGLD	extended generalized lambda distribution
FIR	finite impulse response
GBD	generalized beta distribution
GLD	generalized lambda distribution
GMSK	gaussian mean shift keying
ICA	independent component analysis
i.i.d.	independent identically distributed
I-MIMO	instantaneous multi-input multi-output
ISI	intersymbol interference
MIMO	multiple-input multiple-output
MSE	mean square error
PCA	principal component analysis
pdf	probability density function
SIR	signal-to-interference ratio

Symbols

$\mathbf{x}$	vector of output signals
$\mathbf{A}$	mixing matrix
$\mathbf{s}$	vector of source signals
m	number of sensors, dimension of the data
$\mathbf{y}$	vector of source estimates
$\mathbf{W}$	demixing matrix
$\mathbf{A}^T$	Transpose of matrix $\mathbf{A}$
$\mathbf{A}^{-1}$	Inverse of matrix $\mathbf{A}$
$\det(\mathbf{A})$	determinant of matrix $\mathbf{A}$
$\mathbf{I}$	identity matrix
f	probability density function
f_G	Gaussian density
F	cumulative distribution function
$K(f(\mathbf{y}), g(\mathbf{y}))$	Kullback-Leibler divergence between densities $f(\mathbf{y})$ and $g(\mathbf{y})$
$K\big(\mathbf{y} \,\|\, \mathbf{s}\big)$	Kullback-Leibler divergence between densities of random variables $\mathbf{y}$ and $\mathbf{s}$
Φ_{ML}	maximum likelihood contrast
Φ_{MI}	mutual information contrast
$\Psi(\mathbf{x}, \mathbf{W})$	matrix-valued estimating function
$\varphi_{\mathbf{y}}(\mathbf{y})$	score function of $\mathbf{y}$
$\varphi(y_i)$	(one-unit) estimating function
ϕ	characteristic function of a distribution
t	time index
$\mathcal{T}$	number of observations
$f'(x)$	derivate of $f(x)$
$\mu_1, \mu_2, \mu_3, \ldots$	central moments
$\kappa_1, \kappa_2, \kappa_3, \ldots$	cumulants
$L_1, L_2, L_3, \ldots$	L-moments
$\nu_1, \nu_2, \nu_3, \ldots$	absolute moments
$\nu_1^*, \nu_2^*, \nu_3^*, \ldots$	skewed absolute moments
$G_0, G_1, G_2, \ldots$	Gaussian moments

$\kappa_3^\circ, \kappa_4^\circ$	cumulant based skewness and kurtosis
τ_3, τ_4	L-moment based skewness and kurtosis
ν_2°, ν_3°	absolute moments based skewness and kurtosis
$\gamma_0^\circ, \gamma_1^\circ$	Gaussian moments based skewness and kurtosis
$E\{x\}$	expected value of x
$a_0, a_1, \dots, a_p$	numerator polynomial parameters of the Pearson system
$b_0, b_1, \dots, b_q$	denominator polynomial parameters of the Pearson system
$\mathbf{a}, \mathbf{b}$	vectors of the Pearson system parameters
$\mathbf{M}, \mathbf{Q}$	matrices containing central moments
$\lambda_1, \lambda_2, \lambda_3, \lambda_4$	parameters of the GLD
$\beta_1, \beta_2, \beta_3, \beta_4$	parameters of the GBD
$\omega_1, \omega_2, \dots$	weighting parameters
E_1	Performance index
ϑ_i	Local stability
ξ_i	variance of stability solution
Υ	BSS efficacy
ζ	kernel function
$\iota\mathcal{T}$	bin-width parameter

Chapter 1

Introduction

1.1 Motivation

In signal processing and related fields, multichannel measurements are often encountered. The obtained data can be represented as multivariate time series. Depending on the application, for instance, multiple antennas, multiple microphones or multiple biomedical sensors are used for the data acquisition. Such systems can be described using Multiple-Input Multiple-Output (MIMO) system models. The observed sensor outputs are different because the sensors have different properties, e.g. separate locations. On the other hand, the sensor outputs are related because the sensors are observing the same source signals. In many cases, several source signals are present at the same time and there is only limited knowledge of their properties and how they contribute to each sensor output. If the source signals and the physical system are unknown and only the sensor outputs are observed, the processing methods developed for recovering the original signals are called blind. An illustration of an instantaneous mixing MIMO-model is presented in Figure 1.1.

In Blind Source Separation (BSS, also known as Blind Signal Separation) the goal is to recover the source signals from the observed mixed signals. Blindness means that neither the sources nor the mixing system is known. Separation can be based on the theoretically limiting but practically feasible assumption that the sources are statistically independent. This assumption connects BSS and Independent Component Analysis (ICA). The terms BSS and ICA are often used exchangeable but the basic difference is that in BSS the goal is

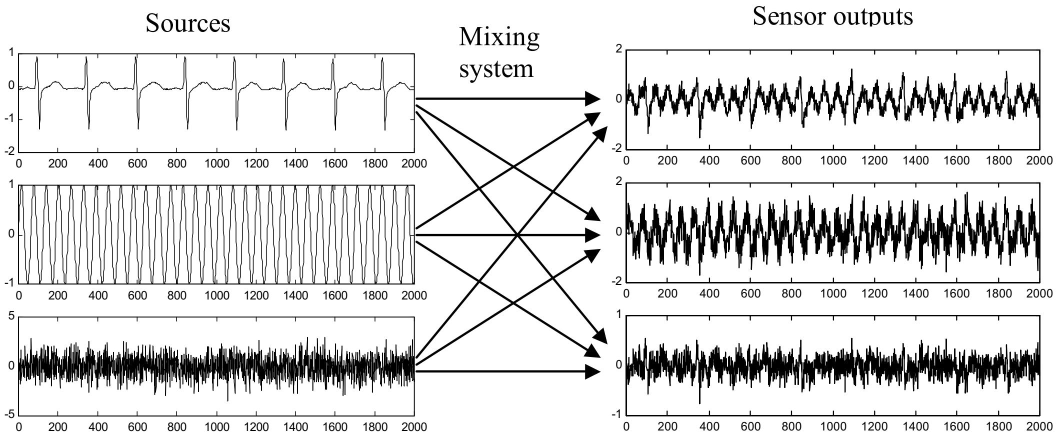


Figure 1.1: An illustration of an instantaneous noise-free mixing system. The system and the sources are unknown and only the sensor outputs are observed.

to separate certain transmitted signals whereas in ICA the goal is to find some components that are statistically as independent as possible. Thus, ICA can be seen as a tool to solve the BSS problem.

BSS and ICA have been applied, for example, in the following application domains

- Audio and speech signal separation e.g. [113, 102, 62]
- Multiple-Input Multiple-Output (MIMO) communications systems e.g. [60, 19, 42, 10, 29, 123, 116]
- Biomedical signal processing e.g. [82, 85, 66, 24, 118]
- Image processing and feature extraction e.g. [9, 59, 23]
- Econometrics and financial applications e.g. [8, 76, 49]

During the last ten years, a considerable amount of work has been focused on BSS/ICA. Conferences and special sessions concentrating on ICA have been organized. The theoretical background has been established and various algorithms have been proposed. Several recent textbooks and tutorial papers provide a good introduction to the field [60, 49, 50, 80, 17, 4].

1.2 Scope of the Thesis

The purpose of this thesis is to develop ICA methods that can adapt to different source distributions. Adaptation makes it possible to separate sources without knowing the source distributions or even the characteristics of source distributions. Special attention is paid to methods that allow not only symmetric but also asymmetric source distributions. Asymmetric distributions occur in key application areas, such as, communications and biomedical signal processing.

The ICA model has two groups of parameters: the mixing system and the source distributions. It has been shown that if the source distributions are known, optimal separation algorithms may be derived [17]. This is done by utilizing the score functions of the sources. Blindness means, however, that no explicit knowledge on the source distributions is available. It follows that the better the sources or the score functions of the sources are estimated the better separation result we can expect.

The first goal of this thesis is to develop methods for learning the source distributions. In practice, it is adequate to concentrate on the approaches that capture the essential properties of the source distributions. The second goal is to find efficient implementations of the proposed methods. This includes the choice of the optimization algorithm, robustness considerations and simulation studies of practical performance. The objective is to show that adaptive estimation methods are necessary and on the other hand, show that the price paid for the increased flexibility is not too high.

1.3 Contribution of the Thesis

The contributions of this thesis are in developing new methods for ICA. Adaptive techniques are proposed for the modeling of score functions or estimating functions. Score functions are modeled using parametric families. The methods may be incorporated into existing ICA algorithms. The contributions can be summarized as follows:

- The relationship between score adaptive estimation and minimization of mutual information is established.

- Pearson system is proposed as a flexible score model.

- An extended Pearson system model allowing multimodal distributions is introduced. The case of bimodal distributions is considered in more detail. The obtained score functions are bounded and defined everywhere. The parameters can be estimated using the method of moments.
- The use of the Extended Generalized Lambda Distribution (EGLD) in the ICA problem is introduced in co-operation with the co-authors [Publication II].
- The method of L-moments is proposed for the estimation of the parameters of the Generalized Lambda Distribution (GLD).
- The optimal weighting is derived for the adaptive estimating functions comprised of two fixed components using the concept of BSS efficacy.
- Absolute moments are proposed as estimating functions.
- The performance of the proposed methods is studied quantitatively and qualitatively in simulations. The simulations demonstrate the reliable performance of the methods.

1.4 Summary of Publications

This thesis consists of 7 publications and a summary. The summary part of the thesis is organized as follows: Chapter 2 introduces the basic concepts and methods of BSS. Chapter 3 contains an overview of the existing methods for the source adaptive ICA. In Chapter 4 the main contribution of the thesis is summarized and methods of Pearson-ICA, EGLD-ICA and adaptive estimating functions are presented. Chapter 5 provides a brief summary and outlines future research.

In Publication I a Pearson system based BSS method is introduced. An algorithm using the method of moments is proposed for finding the parameters of the Pearson system. The actual separation is performed using fixed point algorithm [58]. The simulation examples demonstrate that the method can separate super- and sub-Gaussian sources and even non-Gaussian sources with zero kurtosis.

In Publication II an EGLD based BSS method is introduced. An algorithm utilizing the inverse of cumulative distribution function, method of moments and fixed point algorithm is proposed. The good performance of the algorithm is demonstrated in simulations.

In Publication III the algorithms proposed in Publications II and I 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.

In Publication IV the applications of Pearson-ICA are considered from the viewpoint of telecommunications. Separation of binary sources and instantaneous mixing of Rayleigh or lognormal faded signals are used as examples. Simulation results are provided.

In Publication V the use of the Pearson system is further developed. The different types of distributions in Pearson family are studied in ICA context. It is shown using the results by Pham [97] that the minimization of the mutual information contrast leads to iterative use of score functions as estimation functions. An extension of the Pearson system that can model multimodal distributions is introduced. The applicability of the Pearson system based method is demonstrated in simulation examples, including blind equalization of GMSK signals.

Publication VI is an extended version of Publication III. The performance of the proposed methods is studied in more detail. The additional contribution is the method of L-moments proposed for the estimation of GLD parameters. It is argued that the L-moments are a more natural way to estimate the GLD parameters than the conventional sample moments. Additionally, the L-moments have attractive theoretical properties, including lower sample variance compared to the sample moments.

Publication VII considers the problem of adaptive score estimation from a different viewpoint. The proposed estimating functions comprised of symmetric and asymmetric part can capture the essential features of the source distributions. The optimal weighting between the symmetric and asymmetric part is derived using the concept of BSS efficacy. General results are derived and absolute moment based estimating functions are presented as an example.

Author derived all the equations, performed all the simulations and was mainly responsible for writing in Publications I, III, IV, V, VI and VII. The co-authors contributed in steering the research, in designing experiments and in writing the papers.

The first author was mainly responsible for writing Publication II. The idea of using the EGLD is originally proposed by him. Derivation of the score function and the implementation of the ICA algorithm were done by this author in co-operation with the co-authors.

The EGLD model was also utilized in Publications III and VI.

Chapter 2

Blind Source Separation

2.1 Overview

This chapter provides a short overview to blind source separation (BSS) and independent component analysis (ICA). The key concepts and assumptions needed in ICA and BSS are described. Basic ICA model and its extensions are considered. The elements and principles of an ICA method are explained. More extensive overviews are given in several books and tutorial articles [60, 49, 50, 80, 17, 4].

2.2 Independent Component Analysis Model

2.2.1 The basic ICA model

In this thesis we consider the noiseless instantaneous ICA model

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

where $\mathbf{s} = [s_1, s_2, \ldots, s_m]^T$ is an unknown source vector and matrix $\mathbf{A}_{m \times m}$ is an unknown real-valued mixing matrix. The observed mixtures $\mathbf{x} = [x_1, x_2, \ldots, x_m]^T$ are sometimes called as sensor outputs. The following assumptions for the model to be identifiable are needed according to [27, 68]

1. The sources are statistically mutually independent.

2. At most one of the sources has Gaussian distribution.

3. Mixing matrix $\mathbf{A}$ is invertible.

4. Moments of the sources exist up to the necessary order.

Separation means that we find a separating matrix $\mathbf{W}$ that makes the components of

$$\mathbf{y} = \mathbf{W}\mathbf{x} \tag{2.2}$$

mutually independent. The ith row vector of $\mathbf{W}$ is marked by $\mathbf{w}_i$. It is possible to find solution up to some scaling and permutation. If $\mathbf{W}$ is a separating matrix, any matrix $\mathbf{\Lambda P W}$, where $\mathbf{\Lambda}$ is a diagonal matrix and $\mathbf{P}$ is a permutation matrix (in permutation matrix exactly one element on every row and column is 1 and the other elements are 0), is also a separating matrix [27].

The ICA model has two types of parameters: the mixing coefficients in $\mathbf{A}$ and the source densities. Usually, we are interested in the mixing matrix $\mathbf{A}$ or the actual source values, and the source densities are treated as nuisance parameters. Without any additional assumptions, the estimation of the densities is considered as a nonparametric problem. Together with the parametric estimation of the mixing matrix, the estimation of the ICA model is referred to as a semiparametric problem [3].

2.2.2 Extensions of the basic ICA model

The basic ICA model may be extended several different ways. The noisy ICA model is expressed as

$$\mathbf{x} = \mathbf{A}\mathbf{s} + \mathbf{n}, \tag{2.3}$$

where $\mathbf{n}$ is a Gaussian noise vector independent from the sources. Adding the noise makes the model more realistic because there is always noise in physical sensor measurements. If the noise variances are small compared to the output variances, the methods for noiseless ICA can be utilized with good results. At the presence of heavy noise additional methods are needed to remove the noise from the separated signals. Methods for noisy ICA are considered e.g. in [60, 77, 31, 36].

Complex-valued sources and mixing matrices occur especially in communication problems. There exist ICA methods developed for the complex-valued problem [15, 11]. Sometimes the problem reduces to the real-valued problem, e.g. [123].

In some applications the mixing matrix $\mathbf{A}$ is not a square matrix. The case where the number of mixtures is higher than the number of sources we essentially have the basic problem with extra information. The rank of the model or the number of sources might be unknown and should be also estimated e.g. [22]. The case where the number of the mixtures is lower than the number of the sources is a difficult problem. Since the mixing is not invertible the identification of the mixing matrix and the recovery of the sources are individual problems. Generally, the sources cannot be recovered without additional assumptions. The problem has been considered in [108, 28, 20, 32, 79, 63].

In convolutive mixing, the observed discrete-time signals $x_i(t)$, $i = 1, \ldots, m$ are generated from the model

$$x_i(t) = \sum_{j=1}^{m} \sum_{k} a_{ikj} s_j(t-k). \tag{2.4}$$

This is a Finite Impulse Response Multi-input Multi-output (FIR-MIMO) model, whereas the basic instantaneous mixing model (2.1) can be seen as an instantaneous MIMO (I-MIMO) system. In model (2.4) each FIR filter (for fixed indices i and j) is defined by the coefficients a_{ijk}. Convolutive models are considered e.g. in [6, 114, 93, 45, 46].

Nonlinear ICA model is given by

$$\mathbf{x} = \mathbf{h}(\mathbf{s}), \tag{2.5}$$

where $\mathbf{h}$ is an unknown m-component mixing function. If the space of the nonlinear functions $\mathbf{h}$ is not limited there exist an infinity of solutions [61, 40]. Recently, the interest towards nonlinear ICA has increased. The uniqueness problems are avoided using Bayesian approach [117], regularization techniques [1] or structured models [109, 40]. An important special case of the general nonlinear model (2.5) is post-nonlinear mixture model [111]

$$x_i = h_i\left(\sum_{j=1}^{m} a_{ij} s_j\right), \quad i = 1, \ldots, m, \tag{2.6}$$

where nonlinear functions h_i, $i = 1, \dots, m$, are applied to the linear mixtures.

2.3 Anatomy of an ICA Method

In this thesis ICA methods are studied in the following framework. An ICA method consist of three parts:

1. Measure for independence (theoretical contrast)
2. Estimator of the measure, or objective function
3. Algorithm for optimization

These parts are considered in the following sections. We make a distinction between theoretical measures of independence and the estimators of independence calculated from the data. From the theoretical point of view the linear instantaneous ICA problem is solved: independent components are found when the chosen measure for independence is minimized. However, the great number of the proposed ICA methods shows that there is work to do with estimators and algorithms.

2.4 Measures of Independence

Mutual independence of random variables $\mathbf{y} = [y_1, y_2, \dots, y_m]^T$ means that the joint distribution can be factorized and presented as a product of the marginals. The factorization can be defined using cumulative distribution functions

$$F(\mathbf{y}) = F_1(y_1)F_2(y_2)\dots F_m(y_m), \tag{2.7}$$

probability densities

$$f(\mathbf{y}) = f_1(y_1)f_2(y_2)\dots f_m(y_m), \tag{2.8}$$

or characteristic functions

$$\phi(\mathbf{t}) = \phi_1(t_1)\phi_2(t_2)\dots \phi_m(t_m) \tag{2.9}$$

where characteristic function is defined by

$$\phi(t) = \int e^{\jmath ty} dF = \int e^{\jmath ty} f(y) dy, \tag{2.10}$$

where $\jmath$ is the imaginary unit. These definitions characterize independence but they do not directly tell how to measure dependencies. A natural way to do this is to construct a measure using, for instance, the difference between the joint characteristic functions and the product of marginal characteristic functions [122, 38] or alternatively, the difference between the joint cdf and the product of the marginal cdfs, e.g. Kolmogorov-Smirnov [47] test statistics

$$\Phi_{KS} = \sup_{\mathbf{x}} |F(\mathbf{x}) - F_1(x_1)F_2(x_2)\ldots F_m(x_m)|. \tag{2.11}$$

A contrast function or briefly a contrast is one of the key terms in ICA. A contrast is a function to be minimized in order to separate the sources. Formally a contrast function is defined as [27]

Definition 1 *A contrast is a mapping Φ from the set of densities $\{f_{\mathbf{y}}, \mathbf{y} \in \mathbb{R}^m\}$ to $\mathbb{R}$ satisfying the following three requirements*

1. $\Phi(f_{\mathbf{Py}}) = \Phi(f_{\mathbf{y}}), \quad \forall\, \mathbf{P}$ *permutation,*

2. $\Phi(f_{\Lambda\mathbf{y}}) = \Phi(f_{\mathbf{y}}), \quad \forall\, \Lambda$ *diagonal invertible,*

3. *If* $\mathbf{y}$ *has independent components, then* $\Phi(f_{\mathbf{Ay}}) \leq \Phi(f_{\mathbf{y}}), \quad \forall \mathbf{A}$ *invertible.*

According to Definition 1 a contrast is a function of densities. Under the assumption that the densities are correctly estimated, a contrast becomes a function of the current mixture $\mathbf{y}$ or equivalently a function of the separating matrix $\mathbf{W}$.

Two fundamental ICA contrasts, the maximum likelihood contrast and the mutual information contrast, are based on Kullback-Leibler divergence. The Kullback-Leibler divergence between the random variables $\mathbf{y}_1$ and $\mathbf{y}_2$ is defined as

$$K(\mathbf{y}_1||\mathbf{y}_2) = \int f_1(\mathbf{y}) \log \frac{f_1(\mathbf{y})}{f_2(\mathbf{y})} d\mathbf{y}, \tag{2.12}$$

where f_1 and f_2 are the density functions of $\mathbf{y}_1$ and $\mathbf{y}_2$, respectively. The maximum likelihood contrast can be defined as the Kullback-Leibler divergence between $\mathbf{y}$ and $\mathbf{s}$

$$\Phi_{ML}(\mathbf{y}) = K(\mathbf{y} \,||\, \mathbf{s}) \tag{2.13}$$

and the mutual information contrast can be defined as

$$\Phi_{MI}(\mathbf{y}) = K(\mathbf{y} \,||\, \tilde{\mathbf{y}}), \tag{2.14}$$

where $\tilde{\mathbf{y}}$ denotes the vector with independent entries with each entry distributed as the corresponding marginal of $\mathbf{y}$. Now the connection between mutual information and likelihood can be written as

$$K(\mathbf{y} \,||\, \mathbf{s}) = K(\mathbf{y} \,||\, \tilde{\mathbf{y}}) + K(\tilde{\mathbf{y}} \,||\, \mathbf{s}), \tag{2.15}$$

Mutual information is a sufficient statistics in ICA [17]. Likelihood is a sum of mutual information and a nuisance term that gives the marginal mismatch between the output and the assumed sources.

2.5 Objective Functions and Estimating Functions

An estimator of a contrast function is often called as objective function, criterion function or cost function. In addition, the term contrast is sometimes used also for the estimator calculated from the data. It should be mentioned that the meaning of contrast in ICA differs from the meaning contrast has in statistics [103]. The ICA terminology may be confusing here but the basic idea is that we have a measure of independence and an estimator for it.

The derivative of an objective function may be called an estimating function. Estimating functions are sometimes also called separating functions or activation functions. Since the objective functions must be minimized numerically, the estimating functions have an essential role in practical ICA algorithms. Formally, the estimating function [3] can be defined as a matrix-valued function $\Psi(\mathbf{x}, \mathbf{W})$ such that

$$E\{\Psi(\mathbf{x}, \mathbf{W}^*)\} = 0, \tag{2.16}$$

where $\mathbf{W}^*$ is the true separating matrix. A typical form of estimating function (2.16) is

$\Psi(\mathbf{x}, \mathbf{W}) = \mathbf{I} - \varphi(\mathbf{y})\mathbf{y}^T$ where $\varphi(\mathbf{y}) = [\varphi_1(y_1), \varphi_2(y_2), \dots, \varphi_m(y_m)]^T$ is a vector of one-unit estimating functions. The term 'estimating function' is commonly used to refer to these one-unit estimating functions, as done also in this thesis. This definition of the estimating function is related to the projection pursuit [53, 65] and the deflation approach [33, 54] where one-unit objective functions are used to extract the sources one by one.

If the source distributions are known, the maximum likelihood principle leads to the estimating functions that are the score functions of the sources [17]:

$$\varphi_{\mathbf{y}}(\mathbf{y}) = -\frac{d}{d\mathbf{y}} \log f_{\mathbf{y}}(\mathbf{y}). \tag{2.17}$$

This is a fundamental result but it applies only when the source densities are positive everywhere. For example, if uniformly distributed sources are mixed we cannot use score functions in separation because the score functions are zero in a finite interval and undefined elsewhere.

In practice, the source distributions are not known. The maximum likelihood contrast can be employed with some pre-chosen densities for the sources. An equivalent approach is to choose directly a suitable nonlinear function as estimating function. We use the notation where $\varphi_{\mathbf{y}}$ refers to the true score function of random variable $\mathbf{y}$, as defined in equation (2.17). The notation without the subindex φ refers to an estimating function, or to the estimated score function. This emphasizes the close relationship between the score function modeling and the nonlinearity selection. If estimating function φ is used, we observe that the following expression for the assumed source density is obtained

$$f(y_i) = \frac{\exp(-\int \varphi(y_i) dy_i)}{\int_{-\infty}^{\infty} \exp(-\int \varphi(y_i) dy_i) dy_i}. \tag{2.18}$$

It should be noted that (2.18) is not always a valid density in traditional sense. For some typical choices of the estimating function, the denominator in (2.18) tends to infinity. This can be avoided making a working assumption that y_i belongs to a finite interval and evaluating the integrals over this interval.

In linear ICA accurate estimation of source distributions is not always crucial. However, better separation may be achieved if the source distributions are estimated. This becomes obvious when the number of the sources increases and source distributions are challenging (e.g. skewed distributions close to Gaussian distribution).

Cumulants have been used as objective functions since the early days of blind separation [67, 27, 15, 33]. Cumulants are employed also in some recent works; see e.g. [86] and [30]. Cumulants $\kappa_1,\ \kappa_2,\ \kappa_3, \ldots$ are defined via characteristic function (2.10) by the identity

$$\exp\left(\sum \kappa_q (\jmath t)^q / q!\right) = \phi(t) \tag{2.19}$$

Cumulants may be estimated from the sample moments of same and lower orders. The estimates of the sample moments (central moments) are obtained as follows

$$\begin{aligned} \bar{x} = \sum_{t=1}^{\mathcal{T}} x(t)/\mathcal{T} && (2.20) \\ \hat{\mu}_2 \quad = \quad \hat{\sigma}^2 = \sum_{t=1}^{\mathcal{T}} (x(t) - \bar{x})^2/\mathcal{T} && (2.21) \\ \hat{\mu}_3 \quad = \quad \sum_{t=1}^{\mathcal{T}} (x(t) - \bar{x})^3/\mathcal{T} && (2.22) \\ \hat{\mu}_4 \quad = \quad \sum_{t=1}^{\mathcal{T}} (x(t) - \bar{x})^4/\mathcal{T}, && (2.23) \end{aligned}$$

where $\mathcal{T}$ is the number of observations. In this thesis, notation $\mu_1, \mu_2, \mu_3, \ldots$ is used for both the theoretical sample moments and their estimators. The cumulant-based skewness and kurtosis may be defined as follows

$$\kappa_3^\circ(y_i) = \frac{\kappa_3(y_i)}{\kappa_2(y_i)^{3/2}} = E\left\{\left(\frac{y_i - \mu_{y_i}}{\sigma_{y_i}}\right)^3\right\} \tag{2.24}$$

$$\kappa_4^\circ(y_i) = \frac{\kappa_4(y_i)}{\kappa_2(y_i)^2} = E\left\{\left(\frac{y_i - \mu_{y_i}}{\sigma_{y_i}}\right)^4\right\} - 3, \tag{2.25}$$

where μ_{y_i} and σ_{y_i} are the expected value and the standard deviation of y_i, respectively. The separation can be based on the fact that for Gaussian distribution the higher order cumulants equal to zero. Maybe the simplest technique to separate the sources is to maximize or minimize kurtosis.

When sample variance and robustness to outliers (in noisy ICA model) are of concern, bounded nonlinear functions may be more advisable than cumulants. However, the practical performance also depends on the underlying source distributions. In Table 2.1 some typical one-unit objective functions and the corresponding estimating functions are presented.

Objective function	Estimating function
kurtosis $\Phi(y_i) = y_i^4$	cubic $\varphi(y_i) = y_i^3$
skewness $\Phi(y_i) = y_i^3$	$\varphi(y_i) = y_i^2$
$\Phi(y_i) = \log(\cosh(y_i))$	hyperbolic tangent $\varphi(y_i) = \tanh(y_i)$
Gaussian moments, e.g. $\Phi(y_i) = e^{-y_i^2/2}$	$\varphi(y_i) = -y_i e^{-y_i^2/2}$
3rd absolute moment $\Phi(y_i) = \|y_i\|^3$	$\varphi(y_i) = y_i\|y_i\|$

Table 2.1: Some typical one-unit objective functions and the corresponding estimating functions. The scaling constants are omitted.

These simple estimating functions are good benchmark for more advanced methods: they are easy to implement and they successfully separate most of typical sources.

The objective functions in Table 2.1, expect the skewness, employ only even moments or symmetric properties of the source distributions. This means that there is an implicit assumption that the sources have a symmetric distribution. The explicit connection can be found using the equation (2.18). In Publication VII adaptive methods for finding objective functions with the optimal weighting between the symmetric and asymmetric properties have been proposed and they will be considered in Section 4.4 of this thesis.

2.6 Mutual Information and Source Adaptation

The ICA methods proposed in this thesis are based on direct minimization of mutual information. The direct minimization of mutual information leads to the adaptive estimation of the score functions of the mixtures as shown in [97] and in Publication V. Starting from mutual information contrast $\Phi_{MI}(\mathbf{W})$ defined as a function of $\mathbf{W}$, the following gradient (called as relative gradient in [97]) is obtained

$$\Phi'_{MI}(\mathbf{W}) = \int \varphi_{\mathbf{y}}(\mathbf{y})\mathbf{y}^T f_{\mathbf{x}}(\mathbf{x})d\mathbf{x} - \mathbf{I}. \tag{2.26}$$

Using the relation $\mathbf{y} = \mathbf{W}\mathbf{x}$, where $\mathbf{W}$ is orthogonal, we can write (2.26) in the form

$$\Phi'_{MI}(\mathbf{W}) = \int \varphi_{\mathbf{y}}(\mathbf{y})\mathbf{y}^T f_{\mathbf{y}}(\mathbf{y})d\mathbf{y} - \mathbf{I}. \tag{2.27}$$

If $\mathbf{y}(t)$ is an ergodic random process, where the individual samples are distributed according to $f_{\mathbf{y}}(\mathbf{y})$, we obtain the following estimator

$$\hat{\Phi}'_{MI}(\mathbf{W}) = \frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} \hat{\varphi_{\mathbf{y}}}(\mathbf{y}(t))\mathbf{y}(t)^T - \mathbf{I}, \tag{2.28}$$

where $\hat{\varphi_{\mathbf{y}}}$ is an estimator for the score function of $\mathbf{y}$ and $\mathcal{T}$ is the sample size. In the case of mutual information contrast, the estimating function is the score function of $\mathbf{y}$. Because the output $\mathbf{y}$ changes on every iteration of the optimization algorithm, the optimal estimating functions also change in each iteration.

A procedure for parametric minimization of mutual information may be given as follows: After the choice of model family and some suitable algorithm, such as natural gradient (2.29) or fixed point algorithm (2.30), the following steps are repeated until the convergence:

1) Appropriate sample statistics (e.g. moments) are computed from the current data $\mathbf{y}_k = \mathbf{W}_k\mathbf{x}$.

2) The parameters of score function are estimated for each component using the sample statistics.

3) The score functions are utilized as estimating functions in the ICA algorithm performing the separation.

2.7 Algorithms

Numerical methods are needed in order to optimize an ICA objective function. In general, the choice of the algorithm is independent from the choice of the objective function. Of course, there may be differences in the computational complexity. It is commonly assumed that the data is centered and whitened (zero mean, uncorrelated, unit variance) prior to the actual separation. After whitening the separating matrix is (asymptotically) orthogonal and the number of parameters to be estimated is smaller. Prewhitening improves the convergence but is not necessary for the algorithms to work. The three basic types of algorithms are reviewed in the following.

2.7.1 Natural gradient algorithm

A basic principle of the gradient type optimization methods is to move to the direction of (negative) gradient. In ICA, the gradient can be adjusted to correspond to the geometry of the problem. This leads to natural gradient [2] or relative gradient [17] algorithm. The updating rule for the separating matrix is the following

$$\mathbf{W}_{k+1} = \mathbf{W}_k + \eta \left(\mathbf{I} - \varphi(\mathbf{y})\mathbf{y}^T\right) \mathbf{W}_k, \tag{2.29}$$

where $\varphi(\mathbf{y}) = [\varphi_1(y_1), \varphi_2(y_2), \dots, \varphi_m(y_m)]^T$ is the vector of estimating functions and η is the learning rate.

2.7.2 Fixed-point algorithm

Fixed-point algorithm [57, 58] can be seen as a computationally more efficient version of natural gradient algorithm. The update rule can be expressed as

$$\mathbf{W}_{k+1} = \mathbf{W}_k + \mathbf{D}\left(E\{\varphi(\mathbf{y})\mathbf{y}^T\} - \mathrm{diag}(E\{\varphi(y_i)y_i\})\right) \mathbf{W}_k, \tag{2.30}$$

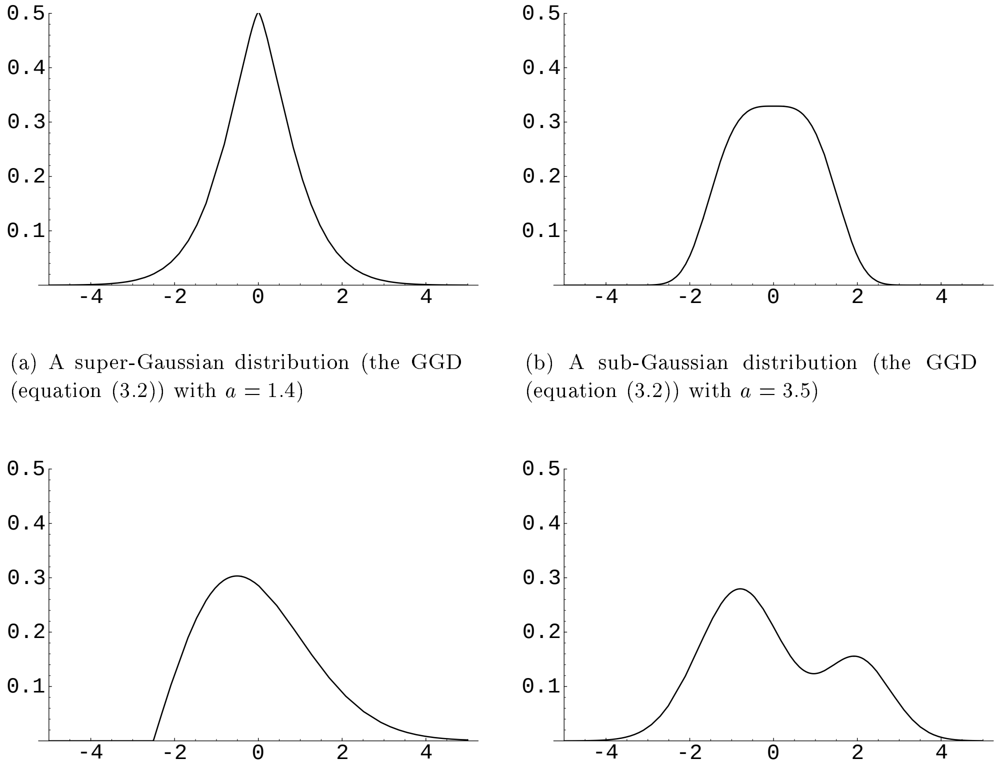
where $\mathbf{D} = \mathrm{diag}\left(1/(E\{\varphi(y_i)y_i\} - E\{\varphi^{'}(y_i)\}\right)$. After every iteration, the separating matrix is projected to the set of orthogonal matrices (in the case of prewhitened data) using symmetric orthogonalization $\mathbf{W}_{orth} = (\mathbf{W}\mathbf{W}^T)^{1/2}\mathbf{W}$. The algorithm converges when $||\mathbf{W}_{k+1} - \mathbf{W}_k|| < \varepsilon$ with e.g. $\varepsilon = 0.0001$.

2.7.3 Jacobi algorithms

Jacobi-type algorithms are based on the theorem [27] stating that in the case of the linear ICA model, pairwise independence implies mutual independence. This leads to the algorithms where pairwise cost functions are sequentially optimized. Such algorithms converge when all the pairs are optimized in the limits of some predetermined converge criterion. The best-known Jacobi type algorithm is probably Joint Approximate Diagonalization of Eigenmatrices (JADE) [15] where the eigenmatrices of the fourth order cumulant tensors are jointly diagonalized.

2.8 Characterization of Source Distributions

In many applications the nature of the source signals is known even if the exact source distributions are unknown. Commonly, distributions are divided to super- and sub-Gaussian distributions. A symmetric zero mean distribution $f(x)$ is super-Gaussian (respectively sub-Gaussian) if $\exists\, x_0 > 0 \,|\, \forall\, x \geq x_0,\ f(x) > f_G(x)$ ($f(x) < f_G(x)$ for sub-Gaussian) , where $f_G(x)$ is the normalized Gaussian pdf. In the case of unimodal symmetric sources the sign of kurtosis (2.25) depends on super- and sub-Gaussianity [83]. The concept of super- and sub-Gaussianity is not very informative in the case of asymmetric or multimodal distributions. Measures of both the skewness and the kurtosis are needed to describe asymmetric distributions. Multimodal distributions may be characterized by the locations of the modes. Examples on the different types of pdf are provided in Figure 2.1.



(a) A super-Gaussian distribution (the GGD (equation (3.2)) with $a = 1.4$)

(b) A sub-Gaussian distribution (the GGD (equation (3.2)) with $a = 3.5$)

(c) An asymmetric distribution (Centered Rayleigh(2) distribution)

(d) An asymmetric bimodal distribution (mixture of two Gaussian distributions)

Figure 2.1: Examples of different types of distributions

2.9 Discussion

In this chapter the ICA models and terminology were reviewed. Measures for independence, their estimators and optimization algorithms were considered. When the family of the possible source distributions is expanded from symmetric unimodal distributions to asymmetric and multimodal distributions the need for the source adaptation becomes obvious. The connection with the source adaptation and minimization of mutual information is established. This suggests the adaptive estimation of the score functions of the mixtures. Methods applying the score adaptation are considered in the following chapters.

Chapter 3

Review of source adaptive ICA methods

3.1 Overview

As discussed in Chapter 2, the optimal separation requires that the source distributions are known. In practice, the source distributions are not known and need to be estimated reliably. In the pure maximum likelihood approach the prior knowledge on the sources is refined to a density model or an estimating function. In the adaptive maximum likelihood approach or mutual information approach, densities or score functions are iteratively estimated from the data. In this chapter, methods for modeling and estimating the source distributions in ICA are reviewed. The estimation methods may be divided into three classes:

- nonparametric methods, e.g. kernel estimation,
- parametric models for densities and score functions,
- models for estimating functions.

In this chapter, models and methods suitable for the source adaptive approach are reviewed. The chapter provides a background for the score adaptive models that are presented in Chapter 4 and in the original publications.

3.2 Kernel estimation of densities

An overview of kernel estimation and related nonparametric techniques is given in [104]. A separation method with kernel estimates for the source densities is proposed in [97]. Kernel estimation of densities is also applied to nonlinear ICA problem [110, 111]. The kernel density estimate [104] is defined by

$$\hat{f}_i(u) = \frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} \frac{1}{\iota_{\mathcal{T}}} \zeta \left(\frac{u - y_i(t)}{\iota_{\mathcal{T}}} \right), \tag{3.1}$$

where ζ is the kernel function and $\iota_{\mathcal{T}}$ is a bin-width parameter depending on the number of observations $\mathcal{T}$. To guarantee that $\hat{f}_i(u)$ is a density, it suffices to take ζ a density itself. The bin-width parameter affects on the smoothness of the estimate. Pham [97] provides a detailed theoretical analysis on the use of kernel estimates in ICA.

Some computational problems need to be solved in order to apply kernel estimation. The integrals in the gradient of mutual information contrast must be discretized by choosing the spacing for the estimation grid, i.e. the points u where estimator (3.1) is computed. The computation can be made faster using Fast Fourier Transform (FFT) [104]. The kernel-based method is further developed in some recent papers [119, 12].

3.3 Parametric models

3.3.1 Distribution families

The main contributions of this thesis are in using parametric families of distributions for modeling the score functions. These methods are considered in Chapter 4 and in Publications I-VI. Different parametric families for ICA are also employed in [21, 14, 41]. The models used in these papers are the Generalized Gaussian Distribution (GGD) and t-distribution. Both are families of symmetric distributions with shape depending on the parameters. The pdf of the GGD is defined as

$$f(y; a, \lambda_a) = \frac{a\lambda_a}{2\Gamma(\frac{1}{a})} \exp(-|\lambda_a y|^a), \tag{3.2}$$

where a is the parameter of the distribution, λ_a a scaling factor and $\Gamma(x)$ is Gamma function given by

$$\Gamma(x) = \int_0^\infty u^{x-1} \exp(-u) du. \tag{3.3}$$

The parameter a controls the peakiness of the distribution. If $a = 2$, the distribution is reduced to Gaussian distribution; if $a < 2$, the distribution is super-Gaussian; and if $a > 2$, the distribution is sub-Gaussian. Examples are presented in Figure 2.1. The parameter λ_a is a scaling factor controlling the variance. The score function of the GGD is given by

$$\varphi(y_i) = a\lambda_a \operatorname{sign}(y_i)|\lambda_a y_i|^{a-1}. \tag{3.4}$$

The parameters of the GGD can be solved from the following moment equations

$$\kappa_4 = \frac{\Gamma(\frac{5}{a})\Gamma(\frac{1}{a})}{\Gamma^2(\frac{3}{a})} - 3, \tag{3.5}$$

$$\lambda_a = \sqrt{\frac{\Gamma(\frac{3}{a})}{\mu_2 \Gamma(\frac{1}{a})}}, \tag{3.6}$$

where κ_4 is the kurtosis and μ_2 is the second order moment. In practice, to estimate the parameters, the sample kurtosis is calculated from the data and the values of the parameters a and λ_a are solved numerically.

Another model, t-distribution, is familiar from t-test [107, 106]. The pdf of t-distribution with b degrees of freedom and the scaling factor λ_b is

$$f(y; b, \lambda_b) = \frac{\lambda_b \Gamma(\frac{b+1}{2})}{\sqrt{\pi b}\Gamma\left(\frac{b}{2}\right)} \left(1 + \frac{\lambda_b^2 y^2}{b}\right)^{-\frac{1}{2}(b+1)}. \tag{3.7}$$

The score function of t-distribution can be written as

$$\varphi(y_i) = \frac{(1-b)y_i}{y_i^2 - \frac{b}{\lambda_b^2}}. \tag{3.8}$$

The parameters of t-distribution can be solved from the following moment equations

$$\kappa_4 = \frac{3\Gamma(\frac{b-4}{2})\Gamma(\frac{b}{2})}{\Gamma^2(\frac{b-2}{2})} - 3, \tag{3.9}$$

$$\lambda_b = \sqrt{\frac{b\Gamma(\frac{b-2}{2})}{2\mu_2\Gamma(\frac{b}{2})}}. \tag{3.10}$$

A simpler way to estimate the parameters of t-distribution using the Pearson system is presented later in Section 4.2.

In [21], only the GGD is employed as a model for the sources. In [14] the choice between the GGD and t-distribution is done based on the sample kurtosis

$$\varphi(y_i) = \begin{cases} a\lambda_a \operatorname{sign}(y_i)|\lambda_a y_i|^{a-1}, & \text{if } \hat{\kappa}_4(y_i) \leq 0 \\ \frac{(1-b)y_i}{y_i^2 - \frac{b}{\lambda_b^2}}, & \text{if } \hat{\kappa}_4(y_i) > 0. \end{cases} \tag{3.11}$$

3.3.2 Mixture of Densities

Mixture of Gaussians model (MOG) is employed as the model of source densities especially in Bayesian approach [7, 117, 78]. The density model is the following

$$f(x) = \frac{\sum_j \omega_j \frac{1}{\sigma_j\sqrt{2\pi}} \exp(-(x-\mu_j)^2/2\sigma_j^2)}{\sum_j \omega_j}, \tag{3.12}$$

where μ_i and σ_i^2 are mean and variance and ω_j is a weighting parameter. Mixtures of Gaussians can approximate virtually any continuous source distribution but the number of required Gaussians depends on the source distribution. For instance, several Gaussians are needed to approximate uniform density. The expectation-maximization (EM) algorithm [34] is often used in the learning of the MOG parameters. Due to computational complexity of MOG-based ICA, the number of Gaussians is usually fixed to some small number. This may limit the performance in some cases even though the performance of the method is generally good.

Mixture of densities models are also proposed in [120, 43, 48, 81]. In [120] a mixture of Gaussian or logistic densities is proposed. In [43] a closely related method of adaptive activation function neurons is studied. In [48] and [81] MOG and hyperbolic-Cauchy distribution are used. These approaches are related to the basis functions approach presented in Section 3.4.2.

3.4 Adaptive nonlinearities

The methods presented in Sections 3.2 and 3.3 started from the estimation of densities. The methods presented in this section approach the problem from a different viewpoint. Instead of densities, estimating functions are directly worked out. As mentioned in Section 2.5 these two approaches are theoretically equivalent. In practice, adaptive nonlinearities may have some appealing computational properties although ad hoc adaptation rules are often needed.

3.4.1 Polynomial expansions

Edgeworth and Gram-Charlier expansions [106] provide approximations for densities in the vicinity of a Gaussian density. The expansions can be used to obtain approximations for negentropy [60]

$$\Phi_{Neg} = \frac{1}{12}\kappa_3^2 + \frac{1}{48}\kappa_4^2, \tag{3.13}$$

or for the score function of a symmetric density

$$\varphi(s) = s - \frac{\kappa_4}{6}(s^3 - 3s), \tag{3.14}$$

where κ_3 and κ_4 are the third and fourth cumulant, respectively. Polynomial expansions are considered e.g. in [65, 27, 5, 121]. The approximation of entropy can be also based on other functions than polynomials as proposed in [56]. For instance, Gaussian density and its derivatives may be employed. These approximations are usually more exact and more robust than the approximations based on polynomials.

3.4.2 Basis functions

Quasi-maximum likelihood approach employing a set of arbitrary basis functions is proposed by Pham [98](see [17] for a brief summary). The score function is approximated by a linear combination

$$\varphi(y_i) = \sum_{n=1}^{N} \omega_n \varphi_n(y_i) \tag{3.15}$$

of a fixed set $\{\varphi_1, \varphi_2, \ldots, \varphi_N\}$ of arbitrary basis functions. It turns out that the weighting parameters $\omega_1, \omega_2, \ldots, \omega_N$ can be solved without knowing the true score function. Mean square error between the true score function and its approximation is minimized when

$$\varphi(y_i) = (E\{R^{'}(y_i)\})^T (E\{R(y_i)R(y_i)^T\})^{-1} R(y_i), \tag{3.16}$$

where $R(y_i) = [\varphi_1(y_i), \varphi_2(y_i), \ldots, \varphi_N(y_i)]$ is the $N \times 1$ column vector of basis functions and $R^{'}(y_i)$ is the column vector of their derivatives. In practice, the expectations are replaced by sample averages.

Algorithms where the nonlinearities are adaptively chosen on the basis of sub/super-Gaussianity are used e.g. in [35, 49, 80]. Typically, the nonlinearities are based on functions such as $tanh(y)$ and y^3 and the sign of the nonlinearity is chosen adaptively.

3.4.3 Threshold functions and quantizers

Very simple algorithms can be constructed using adaptive threshold functions. A threshold activation function [84] is defined as

$$\varphi(y_i) = \begin{cases} 0, & |y_i| < b_i, \\ a_i \operatorname{sign}(y_i), & |y_i| \geq b_i, \end{cases} \tag{3.17}$$

where a_i and b_i are data dependent parameters. The threshold b_i may be chosen so that the local stability is maximized. However, this maximization requires knowledge of the source distribution. As a practical solution, the authors in [84] propose the following updating rules

$$a_i(t+1) = a_i(t) - \eta_a(1 - \hat{\sigma}^2(y_i, t)), \tag{3.18}$$

$$b_i(t+1) = b_i(t) - \eta_b \hat{\kappa}_4^{\circ}(y_i, t), \tag{3.19}$$

where $\hat{\sigma}^2(y_i, t)$ is the sample variance of y_i after t observations, $\hat{\kappa}_4^{\circ}(y_i, t)$ is the sample kurtosis and η_a and η_b are the learning rates. Additionally, the values of b_i are forced to the interval $[0, 1.5]$.

The simple threshold function can be generalized introducing more thresholds and levels. This leads to piecewise constant estimating functions that are also called as quantizers [73]. Optimal quantizer can be found if the source distributions are known. The main advantage of quantizers and threshold functions is that they can be easily implemented in digital signal processing.

3.5 Discussion

The presented estimation methods illustrate the trade-off between generality and simplicity. The nonparametric estimation is apparently the most flexible concept. However, a certain implementation with a fixed kernel is already a more restricted model. The critical part of kernel estimation is the choice of the kernel function and the bin-width parameter. There exist opposing opinions on the complexity and the computational cost of kernel estimation in ICA [97, 60]. The speed requirement depends of course on the particular application but it seems that kernel estimation is relatively complex method when compared to other methods.

The flexibility of parametric estimation depends on the chosen distribution family. Problems may occur if the chosen distribution family cannot model the essential features of the actual distribution. On the other hand, if an appropriate parametric model is used, the methods work efficiently.

The advantage of the adaptive nonlinearities is that they are computationally simple and easy to implement. The performance depends on the source distributions. Successful separation is expected if the nonlinearities can react to the essential features of the distributions. Otherwise, the performance may be poor.

Chapter 4

Adaptive Score Models

4.1 Overview

In this chapter we introduce methods for estimating score functions adaptively. The parametric models employed are the Pearson system and Generalized Lambda Distribution. Additionally, adaptive estimating functions using iterative weighting are presented. The guidelines used for choosing an appropriate parametric model are

1. The model should adapt to asymmetric or multimodal sources, but the performance should not degrade in the case of unimodal symmetric source distributions.

2. The parameters of the model should be easy to estimate from the data.

3. The functional form of the score function should be easy to compute and robust against outliers.

Asymmetric and multimodal source distributions are considered because blindness means that we cannot restrict to symmetric sources. Asymmetric and multimodal source distributions also occur in the key application areas, such as, telecommunications and biomedical signal processing. The requirement of easy parameter estimation is natural from the point of computational efficiency and simplicity of the concept. A suitable functional form of the score function is important to ensure the numerical stability of the practical algorithm.

4.2 Pearson System

The Pearson system is a four parametric family of distributions defined by the differential equation

$$f^{'}(x) = \frac{(x-a)f(x)}{b_0 + b_1 x + b_2 x^2}, \tag{4.1}$$

where $f(x)$ is a density function and a, b_0, b_1 and b_2 are the parameters of the distribution.

The Pearson system has been extensively studied in statistics. Overviews are given in [90] and in [106]. The distribution family is named after Karl Pearson [94, 95]. The estimation of the Pearson parameters is considered e.g. in [105, 13, 25, 26, 51, 87, 74, 96]. Some related distributions are presented in [25, 88, 64, 89, 112].

An alternative parameterization is

$$f^{'}(x) = \frac{(a_1 x - a_0)f(x)}{b_0 + b_1 x + b_2 x^2}, \tag{4.2}$$

where a_0, a_1, b_0, b_1 and b_2 are the parameters of the distribution. Both parameterizations (4.1) and (4.2) characterize the same distributions but the expression (4.2) has the advantage that a_1 can be zero and the values of the parameters are bound when the fourth cumulant exists. Thus, we use the parameterization (4.2). The score function of the Pearson system is easily solved from (4.2)

$$\varphi(x) = -\frac{f^{'}(x)}{f(x)} = -\frac{a_1 x - a_0}{b_0 + b_1 x + b_2 x^2}. \tag{4.3}$$

The derivative of the score function is

$$\varphi^{'}(x) = -\frac{a_1 b_0 + a_0 b_1 + 2a_0 b_2 x - a_1 b_2 x^2}{(b_0 + b_1 x + b_2 x^2)^2}. \tag{4.4}$$

Several well-known distributions belong to the Pearson family. For instance, for Gaussian distribution with mean μ and variance σ^2 the values of the parameters are $a_0 = 12(\sigma^2)^3\mu$, $a_1 = 12(\sigma^2)^3$, $b_0 = -12(\sigma^2)^4$, $b_1 = 0$ and $b_2 = 0$. Also Gamma, Beta and Student's t-distribution belong to the Pearson family. This is illustrated in Figure 4.1.

The distributions in Pearson family can be defined everywhere (type (iii)), they may be bounded from left or right (type (ii)), or defined in a finite interval (type (i)). For the ICA

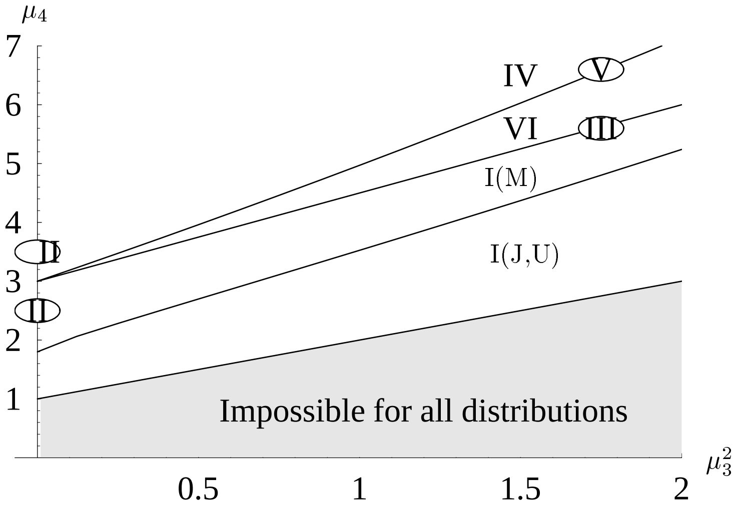


Figure 4.1: An illustration of the Pearson system in (μ_3^2, μ_4)-plane. The limit for all distributions is line $\mu_4 = \mu_3^2 + 1$. The Latin numbers refer to the traditional classification of Pearson distributions. Types I and II are beta distributions of first kind. The notation I(J,U) refers to J- and U-shaped distributions and I(M) to unimodal distribution. The boundary between I(J,U) and I(M) is curve $4(4\mu_4 - 3\mu_3^2)(5\mu_4 - 6\mu_3^2 - 9)^2 = \mu_3^2(\mu_4 + 3)^2(8\mu_4 - 9\mu_3^2 - 12)$ Type III is Gamma distribution for which $\mu_4 = \frac{3}{2}\mu_3^2 + 3$. Type VI is the beta distribution of second kind. Type V is characterized by curve $\mu_3^2(\mu_4 + 3)^2 = 4(4\mu_4 - 3\mu_3^2)(2\mu_4 - 3\mu_3^2 - 6)$. Type IV is the case where the equation $b_0 + b_1 + b_2 x^2 = 0$ has complex roots. Type VII is the Student's t-distribution.

problem this classification is more useful than the traditional classification (types I-VII) [106]. The classification is presented and discussed in Publication V.

Pearson system based blind separation algorithm, Pearson-ICA [71], was originally proposed in Publication I and further improved in Publication V. The implementation is based on the FastICA algorithm [55].

4.2.1 Estimation of the Pearson system parameters

The parameters of the Pearson system can be estimated using method of moments [106]. The moment equations are derived directly from the definition (4.2)

$$x^n(b_0 + b_1 x + b_2 x^2)f^{'}(x) = x^n(a_1 x - a_0)f(x). \tag{4.5}$$

When the left side is integrated by parts, (4.5) leads to a recursion formula

$$\begin{aligned} -nb_0\mu_{n-1} - (n+1)b_1\mu_n - (n+2)b_2\mu_{n+1} = \\ a_1\mu_{n+1} - a_0\mu_n, \end{aligned} \tag{4.6}$$

where μ_n is nth theoretical central moment. When this recursion formula is successively applied for values $n = 0, 1, 2, 3$, the following relationship between the parameters a_0, a_1, b_0, b_1 and b_2 and the theoretical central moments $\mu_{-1} \equiv 0$,$\mu_0 \equiv 1$, $\mu_1 = 0$, μ_2, μ_3 and μ_4 arises

$$a_1 = |10\mu_4\mu_2 - 12\mu_3^2 - 18\mu_2^3| \tag{4.7}$$

$$a_0 = b_1 = -\mu_3(\mu_4 + 3\mu_2^2) \tag{4.8}$$

$$b_0 = -\mu_2(4\mu_2\mu_4 - 3\mu_3^2) \tag{4.9}$$

$$b_2 = -2(\mu_2\mu_4 - 3\mu_3^2 - 6\mu_2^3). \tag{4.10}$$

When the theoretical central moments are replaced by the sample moments, the moment estimators for the parameters a_0, a_1, b_0, b_1 and b_2 are obtained. The number of the parameters actually reduces to three because $b_1 = a_0$ and a_1 is a scaling term.

If the approximated density is symmetric (i.e. $\mu_3 = 0$) the estimated score reduces to

$$\varphi(x) = -\frac{(5\mu_4 - 9\mu_2^2)x}{-2\mu_2\mu_4 - (\mu_4 - 6\mu_2^2)x^2} \tag{4.11}$$

It can be easily checked that when $\mu_4 \geq 3$ this corresponds t-distribution defined in (3.8), (3.9) and (3.10).

The type of the distribution, (i), (ii) or (iii), must be recognized after the model is estimated. For types (i) and (ii) it is possible that the estimated density is not exactly correct and thus some observations lay outside the domain. In the ICA problem we are only interested in finding the score function, which makes it easy to heuristically solve this problem. First, the sample minimum and maximum can be utilized in the estimation. Alternatively, saturated score functions (the values of the score function are bounded between suitable chosen minimum and maximum) can be used. These, as well other practical algorithmic issues are considered in Publications I and V.

4.2.2 Extensions of the Pearson system

The estimation of the Pearson system parameters can be based on sample statistics other than the first four moments. For instance, in [87] the parameter estimation is based on the mean, the variance, the skewness and the left (or right) boundary.

The differential equation defining the Pearson system may also be generalized. A natural generalization is

$$\frac{f^{'}(x)}{f(x)} = \frac{\mathbf{a}(x)}{\mathbf{b}(x)} \tag{4.12}$$

where $\mathbf{a}(x) = a_0 + a_1x + a_2x^2 + \ldots + a_px^p$ and $\mathbf{b}(x) = b_0 + b_1x + b_2x^2 + \ldots + b_qx^q$ are some polynomials of x. Some generalizations of this kind are considered in [25] and briefly discussed in Publication V.

In Publication V we propose a multimodal generalization of the Pearson system defined as follows

$$\frac{f^{'}(x)}{f(x)} = \frac{a_3x^3 + a_2x^2 + a_1x + a_0}{x^4 + 1} \tag{4.13}$$

where a_0, a_1, a_2 and a_3 are the parameters of the system. The third order polynomial in the numerator enables modeling bimodal distributions. The fourth order polynomial in the denominator makes sure that the score function behaves robustly when outliers are encountered by bounding their influence. Since the denominator is always positive, the score function does not have points of discontinuity.

The method of moments can be used to estimate the parameters of (4.13). This leads to the use of the fifth and the sixth order sample moments that are very sensitive to outliers. Fortunately, some simple heuristic solutions exist for stabilizing the estimates of the fifth and the sixth moments. One can simply set maximum values for the higher order moments used. In addition, the influence of each individual observation can be made bounded. These simple modifications result to sensible parameter values in practice.

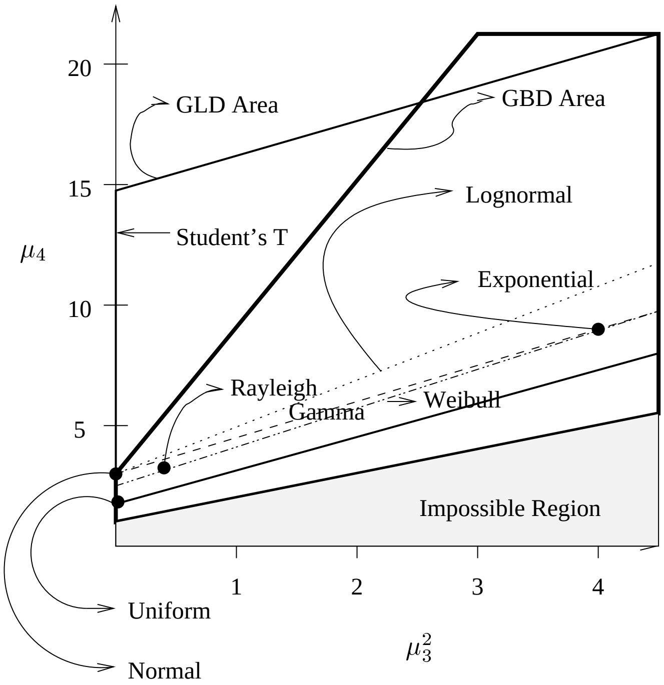


Figure 4.2: Characterization of some standardized distributions by their third and fourth moments. The EGLD family covers the area above the shaded region, which is not valid for any distribution. The skewness and the kurtosis of many distributions occurring in the engineering applications are pointed out

4.3 Extended Generalized Lambda Distribution

The Extended Generalized Lambda Distribution (EGLD) is a large family of distributions covering the whole space of the third and the fourth moment. The lambda distribution was presented by Tukey [115] in 1960. The concept was generalized in 70's [100, 101, 99]. Its main use has been in fitting a distribution to the empirical data, and in the computer generation of different distributions. The latest extension of the family by Karian and Dudewicz in 1996 [70] is a combination of Generalized Lambda Distribution (GLD) and Generalized Beta Distribution (GBD). The space of (μ_3, μ_4) values, which is covered by the EGLD distribution family, includes the values for all the most important distribution including normal, uniform, gamma and beta distributions as illustrated in Figure 4.2.

The Generalized Lambda Distribution is defined by the inverse distribution function

$$F^{-1}(p) = \lambda_1 + \frac{p^{\lambda_3} - (1-p)^{\lambda_4}}{\lambda_2}, \tag{4.14}$$

where $0 \leq p \leq 1$ and λ_1, λ_2, λ_3 and λ_4 are the parameters of the distribution. Karian and Dudewicz [70] showed that GLD is a valid distribution if and only if

$$\frac{\lambda_2}{\lambda_3 p^{\lambda_3 - 1} + \lambda_4 (1-p)^{\lambda_4 - 1}} \geq 0. \tag{4.15}$$

The alternative Freimer-Mudholkar-Kollia-Lin (FMKL) parameterization [44] is given by

$$F^{-1}(p) = \lambda_1 + \left(\frac{p^{\lambda_3} - 1}{\lambda_3} - \frac{(1-p)^{\lambda_4} - 1}{\lambda_4} \right) \Big/ \lambda_2. \tag{4.16}$$

The FMKL-parameterization seems to have some advantages over the parameterization in equation (4.14) but so far it has not been used for fitting the distribution to data.

The EGLD based blind separation algorithm, EGLD-ICA [39], was originally proposed in Publication II. The L-moment based estimation was proposed in Publication VI. The implementation is similar to Pearson-ICA expect for the score function calculation and parameter estimation.

4.3.1 Parameter estimation via sample moments

Estimation of the GLD parameters using the method of moments is proposed in [70]. The relationship between the parameters λ_1, λ_2, λ_3 and λ_4 and the moments μ_1, μ_2, μ_3 and μ_4 is established by four nonlinear equations [70] that can be solved numerically. However, due to the intricacy of the computational process, the parameters λ_1, λ_2, λ_3 and λ_4 are tabulated in [69, 39] as functions of μ_3 and μ_4 for standardized data where $\mu_1 = 0$ and $\mu_2 = 1$. When the EGLD is fitted to the data, the choice between the GLD and the GBD is made based on the values of the kurtosis and the skewness as explained in Publication II.

4.3.2 Parameter estimation via L-moments

Other statistics can be utilized in the estimation of the parameters instead of the sample moments. Well-known drawbacks of the higher order sample moments are the high variance of estimators and the lack of robustness. The concept of L-moments [52] can be seen as a solution to these problem. The L-moments are analogous to the conventional moments but they can be estimated by linear combinations of order statistics i.e. by L-statistics. The first four theoretical L-moments are defined as

$$L_1 = \int_0^1 F^{-1}(p)\mathrm{d}p \tag{4.17}$$

$$L_2 = \int_0^1 F^{-1}(p)(2p-1)\mathrm{d}p \tag{4.18}$$

$$L_3 = \int_0^1 F^{-1}(p)(6p^2-6p+1)\mathrm{d}p \tag{4.19}$$

$$L_4 = \int_0^1 F^{-1}(p)(20p^3-30p^2+12p-1)\mathrm{d}p. \tag{4.20}$$

The L-moments exist if and only if the distribution has a finite mean. Furthermore, a distribution with a finite mean is characterized by its L-moments [52]. Analogously to the conventional moments, L_1 measures the location, L_2 measures the scaling, L_3 measures the skewness and L_4 measures the kurtosis. Scaling invariant measures are obtained by using L-moment ratios defined as

$$\tau_r \triangleq L_r/L_2, \qquad r = 3, 4, \ldots \tag{4.21}$$

Unlike the conventional moments, the L-moments of the GLD may be expressed in a closed form

$$L_1 = \lambda_1 - \frac{1}{\lambda_2}\left(\frac{1}{1+\lambda_4} - \frac{1}{1+\lambda_3}\right) \tag{4.22}$$

$$L_2\lambda_2 = -\frac{1}{1+\lambda_3} + \frac{2}{2+\lambda_3} - \frac{1}{1+\lambda_4} + \frac{2}{2+\lambda_4} \tag{4.23}$$

$$L_3\lambda_2 = \frac{1}{1+\lambda_3} - \frac{6}{2+\lambda_3} + \frac{6}{3+\lambda_3} - \frac{1}{1+\lambda_4} + \frac{6}{2+\lambda_4} - \frac{6}{3+\lambda_4} \tag{4.24}$$

$$\begin{aligned} L_4\lambda_2 = &-\frac{1}{1+\lambda_3} + \frac{12}{2+\lambda_3} - \frac{30}{3+\lambda_3} + \frac{20}{4+\lambda_3} - \\ &\frac{1}{1+\lambda_4} + \frac{12}{2+\lambda_4} - \frac{30}{3+\lambda_4} + \frac{20}{4+\lambda_4} \end{aligned} \tag{4.25}$$

The details for the parameter estimation are presented in the Publication VI.

Since the L-moments are *linear* combinations of order statistics, the variances of the sample L-moments are usually smaller than the variances of the conventional sample moments. This implies that the models fitted using the sample L-moments are more reliable than the models fitted using the conventional sample moments, especially when the sample size is small. Additionally, the L-moments are more robust against outliers.

4.3.3 Other estimation techniques

In addition to method of moments and method of L-moments, some other techniques are recently proposed for the estimation of the GLD parameters. Karian and Dudewicz [37] proposed the use of percentiles. The percentiles have similar desirable properties as the L-moments but the difference is that in the percentile method, only certain order statistics are used, whereas in the method of L-moments all order statistics are employed. This suggests that the L-moments based estimators are more efficient than the percentile based estimators.

Purely computational methods, such as, least square fit (Öztürk and Dale method) [92] and the starship method [75, 91] are also applicable. The starship method has the following three steps [75]

1. For a set of data and a range of λ_1, λ_2, λ_3 and λ_4 values, apply the reverse transformation, i.e. a data value x is transformed to $F(x)$. (Note that as F does not exist in closed form for the GLD, numerical methods are needed.)

2. Calculate the value of a suitable goodness-of-fit measure for the closeness of the resulting values to the uniform(0,1) distribution.

3. Choose the λ_1, λ_2, λ_3 and λ_4 values that minimize the chosen goodness-of-fit measure to the uniform, as the fitted values.

According to the simulation results in [75] Öztürk and Dale method and the starship method give good estimates. The computational cost, however, is higher than in the method of moments or in the method of L-moments.

4.4 Adaptive Estimating Functions

Adaptive estimating functions proposed in Publication VII can be presented as a weighted sum of two estimating functions

$$\varphi(s_i) = \omega_1\varphi_1(s_i) + \omega_2\varphi_2(s_i), \tag{4.26}$$

where $\varphi_1(s_i)$ and $\varphi_2(s_i)$ are two fixed estimating functions and ω_1 and ω_2 are the weighting parameters. The corresponding objective function may be presented as

$$\Phi(y_i;\omega_1,\omega_2) = \omega_1|\Phi_1(y_i)| + \omega_2|\Phi_2(y_i)|. \tag{4.27}$$

The idea is iteratively update the weighting parameters in optimal manner. The optimal weighting is solved maximizing an efficacy measure based on the performance analysis [17, 18, 73, 16] of contrast functions. It is usually assumed in the analysis that all the sources are identically distributed. Local stability is found to depend on the following nonlinear moments

$$\vartheta_i = E\{\varphi^{'}(s_i)\} - E\{s_i\varphi(s_i)\} \tag{4.28}$$

and the variance of the separation solution is found to depend on

$$\xi_i = E\{\varphi(s_i)^2\} - E\{s_i\varphi(s_i)\}^2. \tag{4.29}$$

In [73] it is proposed that the following measure can be used as a performance criterion

$$\Upsilon = \frac{\vartheta_i^2}{\xi_i}. \tag{4.30}$$

This measure is called BSS efficacy and it is independent of the scaling of estimating function φ. The BSS efficacy gives us an analytical way to compare contrast functions. The solution maximizing BSS efficacy is given in [72] and Publication VII.

4.4.1 Estimating functions based on cumulants and absolute moments

The simplest choice for the symmetric and the asymmetric objective function is to use the cumulant based kurtosis (2.25) and skewness (2.24). In Publication VII the cumulant based objective functions are modified to the absolute moments based objective functions that possess more complicated theoretical properties but may in some cases have better performance in practice. The absolute moment [106] of the order q is defined by

$$\nu_q(y_i) = E\left\{|y_i - \mu|^q\right\}, \tag{4.31}$$

where μ is the expected value of the distribution. The even absolute moments are equal to the conventional central moments of the same order but the odd absolute moments cannot be directly written in the terms of the central moments. In addition, we may define the skewed absolute moments by

$$\begin{aligned}\nu_q^*(y_i) =& E\left\{(y_i - \mu)|y_i - \mu|^{q-1}\right\} = \\ & E\left\{\mathrm{sign}(y_i - \mu)|y_i - \mu|^q\right\}.\end{aligned} \tag{4.32}$$

Analogously to the absolute moments, the odd skewed absolute moments are equal to the conventional central moments of the same order but the even skewed absolute moments cannot be directly written in the terms of the central moments.

The kurtosis of a distribution with unit variance can be measured by the third absolute moment

$$\nu_3(y_i) = E\left\{|y_i - \mu|^3\right\}. \tag{4.33}$$

As a measure for skewness we can use the second skewed absolute moment

$$\nu_2^*(y_i) = E\left\{|y_i - \mu|(y_i - \mu)\right\}. \tag{4.34}$$

Exploiting ν_3 and ν_2^* we may construct an ICA objective function. First, we find that for a Gaussian random variable y_i with $\mu = 0$ and $\sigma^2 = 1$

$$\nu_3(y_i) = \int_{-\infty}^{\infty} |y_i|^3 \frac{1}{\sqrt{2\pi}} e^{-y_i^2/2} dy_i = 2\sqrt{\frac{2}{\pi}} \approx 1.59577 \tag{4.35}$$

and $\nu_2^*(y_i) = 0$. Furthermore, we define measures resembling the cumulant based kurtosis and skewness

$$\nu_3^\circ(y_i) = \nu_3\left(\frac{y_i - \mu}{\sigma}\right) - 2\sqrt{\frac{2}{\pi}} \tag{4.36}$$

$$\nu_2^\circ(y_i) = \nu_2^*\left(\frac{y_i - \mu}{\sigma}\right). \tag{4.37}$$

Based on these measures the following objective function is proposed in Publication VII

$$\Phi_\nu(y_i) = \omega_{\nu,1}|\nu_3^\circ(y_i)| + \omega_{\nu,2}|\nu_2^\circ(y_i)|. \tag{4.38}$$

The expressions for the optimal weighting parameters, $\omega_{\nu,1}$ and $\omega_{\nu,2}$ and other details are provided in Publication VII.

4.4.2 Gaussian moments based estimating functions

The cumulant based approach can be generalized to other suitable nonlinearities [72]. The basic idea is that the objective function is a sum of the absolute values of symmetric and asymmetric functions. The theoretical results for an arbitrary nonlinearities are difficult to obtain and thus the validity of the objective functions must be checked in simulations. We propose using the Gaussian moments as symmetric and asymmetric objective functions. The Gaussian moments of order zero to three are defined by

$$\begin{aligned} G_0(y_i;b) &= e^{-by_i^2/2} - \frac{1}{\sqrt{b+1}} & (4.39)\\ G_1(y_i;b) &= -by_i e^{-by_i^2/2} & (4.40)\\ G_2(y_i;b) &= (by_i^2 - b)e^{-by_i^2/2} & (4.41)\\ G_3(y_i;b) &= (3b^2 y_i - b^3 y_i^3)e^{-by_i^2/2}, & (4.42)\end{aligned}$$

where b is a positive constant. The Gaussian moments form the basis of Gram-Charlier and Edgeworth series [106]. Usually (4.39) is given in the form

$$G_0^*(y_i;b) = e^{-by_i^2/2}. \tag{4.43}$$

The rationale behind the constant $-\frac{1}{\sqrt{b+1}}$ becomes obvious when we consider the expected value of $G_0(y_i)$ in the case where the distribution of y_i is Gaussian ($\mu = 0$, $\sigma^2 = 1$)

$$E\left\{G_0(y_i)\right\} = \int_{-\infty}^{\infty} e^{-by_i^2/2} \frac{1}{\sqrt{2\pi}} e^{-y_i^2/2} dy_i - \frac{1}{\sqrt{b+1}} = 0. \tag{4.44}$$

In addition, we notice that the expected value of the asymmetric part equals zero $E\left\{G_1(y_i)\right\} = 0$ because of the symmetry of Gaussian distribution.

The nonlinearity in expression (4.43) may be employed as a robust ICA objective function as proposed in [57, 56]. We propose the use of G_0 and G_1 as measures of the kurtosis and the skewness in the ICA framework. We now define theoretical measures for the kurtosis and the skewness as follows

$$\gamma_0^\circ(y_i; b) = E\left\{G_0(\frac{y_i - \mu}{\sigma}; b)\right\} \tag{4.45}$$

$$\gamma_1^\circ(y_i; b) = E\left\{G_1(\frac{y_i - \mu}{\sigma}; b)\right\}, \tag{4.46}$$

where μ is the expected value of y_i and σ is the standard deviation. The measures γ_0° and γ_1° are analogous to κ_4 and κ_3 in sense that they are zero for Gaussian distribution and in general at least one of them is nonzero for other distributions. However, γ_0° and γ_1° do not measure the kurtosis and the skewness in the same sense as κ_4 and κ_3 or ν_3° and ν_2°. For instance, the signs of γ_1° and κ_3 may differ.

Now, the objective function based on Gaussian moments with $b = 1$ can be expressed as

$$\Phi_G(y_i) = \omega_{G,1}|G_0(y_i)| + \omega_{G,2}|G_1(y_i)|. \tag{4.47}$$

The estimating function related to the objective function (4.47) and the derivative of the estimating function are

$$\varphi_G(y_i) = \omega_{G,1}\text{sign}(\gamma_0^\circ)G_1(y_i) + \omega_{G,2}\,\text{sign}(\gamma_1^\circ)G_2(y_i) \tag{4.48}$$

$$\varphi_G^{'}(y_i) = \omega_{G,1}\text{sign}(\gamma_0^\circ)G_2(y_i) + \omega_{G,2}\,\text{sign}(\gamma_1^\circ)G_3(y_i). \tag{4.49}$$

The statistic sign(γ_0°) has a similar role as the sign of the kurtosis has in many algorithms. The sign of γ_0° is either known in advance, or more practically, estimated from the data for each source.

4.5 Performance

Several simulations presented in the original publications demonstrate the reliable performance of the proposed methods. Special attention is paid on the separation of asymmetric source distributions. There are three types of design in the simulations examples: First, simple examples illustrate that the proposed methods can separate both sub- and super-Gaussian sources. Second, it is shown that some widely applied algorithms fail in the separation of asymmetric zero kurtosis sources but the proposed methods separate them reliably. Third, it is demonstrated that the proposed methods may be highly beneficial also in the cases were the symmetrical properties are theoretically sufficient for the separation. The performance is measured quantitatively comparing the source signals and the separated signals or comparing the inverse of mixing matrix and the estimated separating matrix. Signal to Interference Ratio (SIR(dB)= $-10\log_{10}$(MSE), where MSE stands for Mean Square Error MSE= $E\left\{(s(t)-y(t))^2\right\}$) is calculated between the source signals and the scale, sign and permutation adjusted separated signals. The matrices are compared using Performance Index [5].

$$E_1 = \sum_{i=1}^{m}(\sum_{j=1}^{m}\frac{|p_{ij}|}{\max_k |p_{ik}|} - 1) + \sum_{j=1}^{m}(\sum_{i=1}^{m}\frac{|p_{ij}|}{\max_k |p_{kj}|} - 1), \tag{4.50}$$

where $P = (p_{ij}) = WA$.

Figure 4.3 summarizes the performance in a simulation with six Rayleigh distributed sources. The pdf of Rayleigh distribution is

$$f(x) = \frac{x}{\sigma^2}e^{-\frac{x^2}{2\sigma^2}}, \quad x \geq 0, \tag{4.51}$$

where σ is a scaling parameter. The results are similar to the results in the original publications. The score adaptive methods outperform the methods in comparison. The order between the score adaptive methods depends on the particular source distributions. Usually, the methods using parametric models perform slightly better than adaptive estimating functions.

4.6 Discussion

Methods for the ICA score function adaptation are proposed. The properties of these methods can be now summarized with respect to the design guidelines presented in the beginning of this chapter. The Pearson system includes both symmetric and asymmetric distributions. The extended Pearson system can also model multimodal distributions. The parameters of

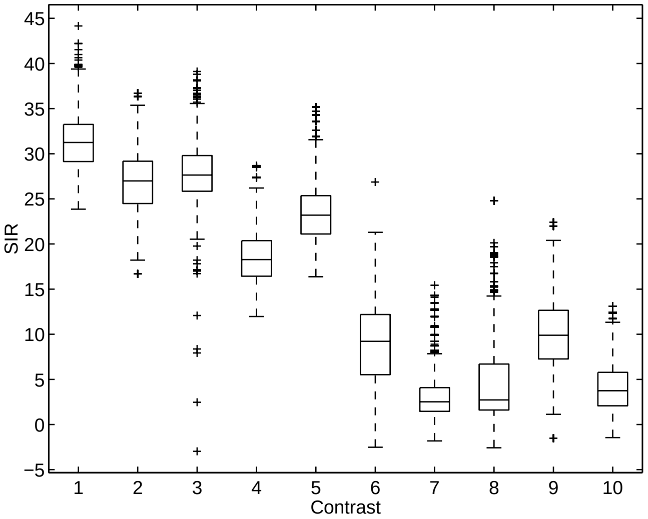


Figure 4.3: Boxplot of the SIR values of Pearson-ICA (boxplot number 1), EGLD-ICA with moments (2), EGLD-ICA with L-moments (3), Absolute moments (4), Gaussian moments (5) Original FastICA (contrasts 'Pow3' (6), 'Tanh' (7) and 'Gauss' (8)) , JADE (9) and Extended infomax (10) algorithm. Six Rayleigh distributed source signals of length 5000 were mixed. The number of realizations was 1001. The presented SIR-values are the SIR-values between the first source signal and its estimate; the SIR-values of the other source signals are similar. The Pearson-ICA and the EGLD-ICA, which exploit also skewness, perform very well: median SIR-values are 31.25 dB and 26.99 dB (27.63 dB with L-moments), respectively. For Absolute moments and Gaussian moments median SIR-values are 18.28 dB and 23.19 dB, respectively. The median values for FastICA(Pow3) and JADE are 9.22 dB and 9.89 dB. The median SIR-values of FastICA(Tanh), FastICA(Gauss) and Extended Infomax are under 4 dB.

the Pearson system can be estimated using method of moments. The EGLD also includes symmetric and asymmetric distributions. The parameters of the EGLD can be estimated using method of moments or method of L-moments. Because the cdf and the pdf of the EGLD are not available in the closed form, numerical methods are needed in the parameter estimation. Numerical methods are also needed in solving the EGLD score function. The properties of the adaptive estimating functions depend on the chosen pair of estimating functions. The estimates for the weighting parameters can be obtained as functions of sample statistics. The functional form of the estimating function is easy to compute and the robustness depends on the chosen functions.

The limitations of the proposed methods are related to the chosen parametric model. It is assumed that the source and the mixture distributions belong to the chosen parametric family. If this is not true, the methods are still supposed to work if the estimated score functions are close to the true score functions. The situation is similar to the case of the fixed estimating functions that are supposed to perform the separation even if they do not correspond to the true score functions. However, due to the adaptive score estimation it is possible to separate a much wider class of source distribution than with any fixed estimating function.

Simulation comparisons between the methods are easy to perform but it is sometimes difficult to generalize the results. Many practical algorithms have tuning parameters that make the comparison problematic. For instance, we have not used kernel density estimation in the simulations because one can always argue that better results could be obtained with a better choice of the kernel and other tuning parameters. The simulation results indicate that the source adaptive concept is highly useful. Further, the simulations indicate that Pearson-ICA, EGLD-ICA and the adaptive estimating functions are reliable implementations of the concept.

Chapter 5

Summary

This thesis considers developing source adaptive methods for ICA and BSS. In BSS, blindness means that neither the mixing system nor the source distributions is known. This contradicts with the result that the score functions of the sources are needed for the optimal maximum likelihood solution. In many widely used ICA methods, fixed estimating functions are employed, which implicitly corresponds to the direct modeling of the source distributions. More flexible methods can be derived starting from the minimization of mutual information. The usage of mutual information as a measure of independence leads to iterative estimation of the score functions of the mixtures. The goal of this thesis is to develop widely applicable adaptive ICA methods that can be implemented in a computationally efficient way.

Three adaptive approaches based on the Pearson system, the EGLD and adaptive estimating functions are proposed for ICA. The Pearson system and the EGLD are parametric families of distributions and they are used to model the distributions of the mixtures. Both families have four parameters that can be estimated from the data using e.g. method of moments or method of L-moments. The strength of these parametric families is that they contain a wide class of distributions, including asymmetric distributions with positive and negative kurtosis, while the estimation of the parameters is still a relatively simple procedure.

Adaptive estimating functions modeling the score function directly as a weighted sum of two estimating functions are developed. The weighting parameters are iteratively updated based on the data. The optimal weighting is solved using the concept of BSS efficacy.

The reliable performance of the proposed methods was demonstrated in extensive simulations. In addition to symmetric source distribution, asymmetric distributions, such as Rayleigh and lognormal distribution, were studied in simulations. The reliability of the proposed methods was also demonstrated when the number of sources is large. The score adaptive methods outperformed the methods in comparison due to their ability to adapt to asymmetric distributions.

Future directions to continue the work of this thesis include applying the proposed score models in recursive online algorithms and in nonlinear ICA. Especially, in post-nonlinear ICA the estimation of the sources is even more essential than in linear ICA model.

Bibliography

[1] L. B. Almeida. ICA of linear and nonlinear mixtures based on mutual information. In *Proc. of International Joint Conference on Neural Networks*, pages 122–127, 2001.

[2] S.-I. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10:251–276, 1998.

[3] S.-I. Amari and J.-F. Cardoso. Blind source separation – semiparametric statistical approach. *IEEE Transactions on Signal Processing*, 45(11):2692–2700, 1997.

[4] S.-I. Amari and A. Cichocki. Adaptive blind signal processing – neural network approaches. *Proceedings of the IEEE*, 86(10):2026–2048, 1998.

[5] S.-I. Amari, A. Cichocki, and H. Yang. A new learning algorithm for blind signal separation. In *Advances in Neural Information Processing Systems*, volume 8, pages 757–763. MIT Press, Cambridge MA, 1996.

[6] S.-I. Amari, A. Cichocki, and H. H. Yang. Blind signal separation and extraction: Neural and information-theoretic approaches. In Haykin [50], pages 63–138.

[7] H. Attias. Independent factor analysis. *Neural Computation*, 11(4):803–851, 1999.

[8] A. Back and A. Weigend. A first application of independent component analysis to extracting structure from stock returns. *International Journal on Neural Systems*, 8(4):473–484, 1997.

[9] A. Bell and T. Sejnowski. The ”independent components” of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338, 1997.

[10] A. Belouchrani and M. Amin. Jammer mitigation in spread spectrum communications using blind sources separation. *Signal Processing*, 80(4):723–729, 2000.

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

[12] R. Boscolo, H. Pan, and V. P. Roychowdhury. Non-parametric ICA. In *Proc. of the Third International Conference on Independent Component Analysis and Signal Separation, ICA2001*, pages 13–18, 2001.

[13] K. O. Bowman and L. R. Shenton. Notes on the distribution of $\sqrt{b_1}$ in sampling from pearson distributions. *Biometrika*, 60(1):155–167, 1973.

[14] J. Cao and N. Murata. A stable and robust ICA algorithm based on t-distribution and generalized gaussian distribution models. *Neural Networks for Signal Processing IX, 1999. The IEEE 1999 Proceedings.*, pages 283–292, 1999.

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

[16] J. F. Cardoso. On the performance of orthogonal source separation algorithms. In *Proc. EUPISCO*, pages 776–779, 1994.

[17] J. F. Cardoso. Blind signal separation: Statistical principles. *Proceedings of the IEEE*, 86(10):2009–2025, 1998.

[18] J.-F. Cardoso and B. H. Laheld. Equivariant adaptive source separation. *IEEE Transactions on Signal Processing*, 44(12):3017–3030, Dec. 1996.

[19] L. Castedo, C. Escudero, and A. Dapena. A blind signal separation method for multiuser communications. *IEEE Transactions on Signal Processing*, 45(5):1343–1348, 1997.

[20] B. Chen, A. P. Petropulu, and L. De Lathauwer. Blind identification of convolutive MIMO system with 3 sources and 2 sensors. In *Proc. of The Thirty-Fifth Annual Conference on Information Sciences and Systems, Volume II*, pages 697–701, 2001.

[21] S. Choi, A. Cichocki, and S. Amari. Flexible independent component analysis. In *Neural Networks for Signal Processing VIII, 1998. Proceedings of the 1998 IEEE Signal Processing Society Workshop*, pages 83–92, 1998.

[22] A. Cichocki, J. Karhunen, W. Kasprzak, and R. Vigario. Neural networks for blind separation with unknown number of sources. *Neurocomputing*, 24:55–93, 1999.

[23] A. Cichocki, W. Kasprzak, and S. Amari. Neural network approach to blind separation and enhancement of images. In *Signal Processing VIII (EUSIPCO)*, volume 1, pages 579–582, 1996.

[24] A. Cichocki and S. Vorobyov. Application of ICA for automatic noise and interference cancellation in multisensory biomedical signals. In *Proceedings of the Second International Workshop on Independent Component Analysis and Blind Signal Separation, ICA2000*, pages 621–626, 2000.

[25] L. Cobb, P. Koppstein, and N. H. Chen. Estimation and moment recursion relations for multimodal distributions of the exponential family. *Journal of the American Statistical Association*, 78(381):124–130, 1983.

[26] A. C. Cohen. Estimation of parameters in truncated pearson frequency distributions. *Annals of Mathematical Statistics*, 22(2):256–265, 1951.

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

[28] P. Comon and O. Grellier. Nonlinear inversion of underdetermined mixtures. In *Proceedings of the First International Workshop on Independent Component Analysis and Blind Signal Separation, ICA'99*, pages 461–465, 1999.

[29] R. Cristescu, T. Ristaniemi, J. Joutsensalo, and J. Karhunen. Delay estimation in CDMA communications using a Fast ICA algorithm. In *Proceedings of the Second International Workshop on Independent Component Analysis and Blind Signal Separation, ICA2000*, pages 585–590, 2000.

[30] S. Cruces, L. Castedo, and A. Cichocki. Novel blind source separation algorithms using cumulants. In *Proc. of ICASSP 2000*, pages 3152–3155, 2000.

[31] L. De Lathauwer, B. De Moor, and J. Vandewalle. Independent component analysis based on higher-order statistics only. In *Proc. IEEE Signal Processing Workshop on Statistical Signal Array Processing*, pages 356–359, 1996.

[32] L. De Lathauwer, B. De Moor, and J. Vandewalle. ICA techniques for 3 sources and 2 sensors. In *Proc. of Sixth IEEE Signal Processing Workshop on Higher Order Statistics*, pages 116–120, 1999.

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

[34] A. P. Dempster, N. M. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of Royal Statistical Society, Series B*, 39:1–38, 1977.

[35] S. C. Douglas, A. Cichocki, and S.-I. Amari. Multichannel blind separation and deconvolution of sources with arbitrary distributions. In *Proc. IEEE Workshop on Neural Networks for Signal Processing, NNSP'97*, pages 436–445, 1997.

[36] S. C. Douglas, A. Cichocki, and S. I. Amari. A bias removal technique for blind source separation with noisy measurements. *Electronic Letters*, 34:1379–1380, 1998.

[37] E. J. Dudewicz and Z. A. Karian. Fitting the generalized lambda distribution (GLD) system by a method of percentiles, II: Tables. *American Journal of Mathematical and Management Sciences*, 19(1), 1999.

[38] J. Eriksson, A. Kankainen, and V. Koivunen. Novel characteristic function based criteria for ICA. In *Proc. of the Third International Conference on Independent Component Analysis and Signal Separation, ICA2001*, pages 108–113, 2001.

[39] J. Eriksson, J. Karvanen, and V. Koivunen. *EGLD-ICA Matlab code available at http://wooster.hut.fi/statsp/publications.html*, 2000.

[40] J. Eriksson and V. Koivunen. Blind identifiability of a class of nonlinear instantaneous ICA models. to appear in EUSIPCO 2002, Toulouse, France, 2002.

[41] R. M. Everson and S. J. Roberts. Independent component analysis: a flexible nonlinearity and decorrelating manifold approach. *Neural Computation*, 11(8):1957–1983, 1999.

[42] M. Feng and K.-D. Kammayer. Application of source separation algorithms for mobile communications enviroment. In *Proc. of the First International Workshop on Independent Component Analysis and Signal Separation, ICA'99*, pages 431–436, 1999.

[43] S. Fiori. Blind signal processing by the adaptive activation function neurons. *Neural Networks*, 13(6):597–611, 2000.

[44] M. Freimer, G. Mudholkar, G. Kollia, and C. Lin. A study of the generalized Tukey lambda family. *Communications in Statistics - Theory and Methods*, 17:3547–3567, 1988.

[45] G. Giannakis, Y. Hua, P. Stoica, and L. Tong. *Trends in Channel Estimation and Equalization*, volume 1 of *Signal Processing Advances in Wireless and Mobile Communications*. Prentice Hall, 2001.

[46] G. Giannakis, Y. Hua, P. Stoica, and L. Tong. *Trends in Single- and Multi-User Systems*, volume 2 of *Signal Processing Advances in Wireless and Mobile Communications*. Prentice Hall, 2001.

[47] J. D. Gibbons. *Nonparametric Statistical Inference*. Statistics: textbooks and monographs. Marcel Dekker, second edition, 1985.

[48] M. Girolami. An alternative perspective on adaptive independent component analysis algorithms. *Neural Computation*, 10(8):2103–2114, 1998.

[49] M. Girolami. *Independent Component Analysis and Blind Source Separation*. Self-Organising Neural Networks. Springer-Verlag, London, 1999.

[50] S. Haykin, editor. *Blind Source Separation*, volume 1 of *Unsupervised Adaptive Filtering*. Wiley, 2000.

[51] A. B. Hoadley. Use of the pearson densities for approximating a skew density of whose left terminal and first three moments are known. *Biometrika*, 55(3):559–563, 1968.

[52] J. Hosking. L-moments: Analysis and estimation of distributions using linear combinations of order statistics. *Journal of Royal Statistical Society B*, 52(1):105–124, 1990.

[53] P. J. Huber. Projection pursuit. *The Annals of Statistics*, 13(2):435–475, 1985.

[54] A. Hyvärinen. One-unit contrast functions for independent component analysis: A statistical analysis. In *Neural Networks for Signal Processing VII. Proc. of IEEE Workshop on Neural Networks for Signal Processing*, pages 388–397, 1997.

[55] A. Hyvärinen. *FastICA Matlab code with references available at http://www.cis.hut.fi/projects/ica/fastica/*, 1998.

[56] A. Hyvärinen. New approximations of differential entropy for independent component analysis and projection pursuit. In *Advances in Neural Information Processing Systems*, volume 10, pages 273–279. MIT Press, 1998.

[57] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999.

[58] A. Hyvärinen. The fixed-point algorithm and maximum likelihood estimation for independent component analysis. *Neural Processing Letters*, 10(1):1–5, 1999.

[59] A. Hyvärinen and P. Hoyer. Emergence of phase and shift invariant features by decomposion of natural images into independent feature subspaces. *Neural Computation*, 12(7):1705–1720, 2000.

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

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

[62] S. Ikeda and N. Murata. A method of ICA in time-frequency domain. In *Proc. of the First International Workshop on Independent Component Analysis and Signal Separation, ICA'99*, pages 365–370, 1999.

[63] M. Inki and A. Hyvärinen. Two methods for estimating overcomplete independent component bases. In *Proc. of the Third International Conference on Independent Component Analysis and Signal Separation, ICA2001*, pages 343–348, 2001.

[64] J. O. Irwin. The generalized waring distribution. part III. *Journal of the Royal Statistical Society. Series A*, 138(3):374–384, 1975.

[65] M. C. Jones and R. Sibson. What is projection pursuit? *Journal of the Royal Statistical Society. Series A (General)*, 150(1):1–37, 1987.

[66] T. Jung, S. Makeig, M. McKeown, A. Bell, T. Lee, and T. Sejnowski. Imaging brain dynamics using independent component analysis. *Proceedings of the IEEE*, 89(7):1107–1122, 2001.

[67] C. Jutten and C. Herault. Blind separation of sources. Part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24(1):1–10, 1991.

[68] A. M. Kagan, Y. V. Linnik, and C. R. Rao. *Characterization Problems in Mathematical Statistics.* Wiley, New York, 1973.

[69] Z. A. Karian and E. J. Dudewicz. The extended generalized lambda distribution (EGLD) system for fitting distributions to data with moments, II: Tables. *American Journal of Mathematical and Management Sciences*, 1996.

[70] 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. *Communications in Statistics: Simulation and Computation*, 25(3):611–642, 1996.

[71] J. Karvanen, J. Eriksson, and V. Koivunen. *Pearson-ICA Matlab code available at http://wooster.hut.fi/statsp/publications.html*, 2000.

[72] J. Karvanen and V. Koivunen. Blind source separation exploiting measures of skewness and kurtosis. Manuscript.

[73] S. A. Kassam, Y. Zhang, and G. V. Moustakides. Some results on a BSS algorithm under non-standard conditions. In *Proc. of the 33rd Annual Conference on Information Sciences and Systems*, 1999.

[74] A. Khalique and I. H. Tajuddin. On the choice of using four moments or three moments and the left boundary for fitting a pearsonian distribution. *Statistician*, 36(4):393–395, 1987.

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

[76] K. Kiviluoto and E. Oja. Independent component analysis for parallel financial time series. In *Proc. of International Conference on Neural Information Processing, ICONIP'98*, volume 2, pages 895–898, 1998.

[77] V. Koivunen, M. Enescu, and E. Oja. Adaptive algorithm for blind separation from noisy time-varying mixtures. *Neural Computation*, 13(10):2339–2358, 2001.

[78] H. Lappalainen. Ensemble learning for independent component analysis. In *Proceedings of the Second International Workshop on Independent Component Analysis and Blind Signal Separation, ICA'99*, pages 7–12, 1999.

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

[80] T.-W. Lee. *Independent Component Analysis: Theory and applications*. Kluwer Academic Publishers, Boston, 1998.

[81] T.-W. Lee, M. Girolami, and T. J. Sejnowski. Independent component analysis using an extended infomax algorithm for mixed sub-gaussian and super-gaussian sources. *Neural Computation*, 11(2):417–441, 1999.

[82] S. Makeig, A. Bell, T.-P. Jung, and T. Sejnowski. Independent component analysis of electroencephalographic data. In *Advances in Neural Information Processing Systems*, volume 8. MIT Press, 1996.

[83] A. Mansour and C. Jutten. What should we say about the kurtosis. *IEEE Signal Processing Letters*, 6(12), Dec. 1999.

[84] H. Mathis, T. P. von Hoff, and M. Joho. Blind separation of signals with mixed kurtosis signs using threshold activation functions. *IEEE Transactions on Neural Networks*, 12(3):618–624, 2001.

[85] M. McKeown, S. Makeig, S. Brown, T.-P. Jung, S. Kindermann, A. J. Bell, and V. I. andT. Sejnowski. Blind separation of functional magnetic resonance imaging (fMRI) data. *Human Brain Mapping*, 6(5–6):368–372, 1998.

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

[87] P.-H. Muller and H. Vahl. Pearson's system of frequency curves whose left boundary and first three moments are known. *Biometrika*, 63(1):191–194, 1976.

[88] C. J. Nachtsheim and M. E. Johnson. A new family of multivariate distributions with applications to monte carlo studies. *Journal of the American Statistical Association*, 83(404):984–989, 1988.

[89] J. K. Ord. On a system of discrete distributions. *Biometrika*, 54(3/4):649–656, 1967.

[90] J. K. Ord. *Families of Frequency Distributions.* Griffin, London, 1972.

[91] D. Owen. The starship. *Communications in Statistics - Simulation and Computation*, 17:315–323, 1988.

[92] A. Öztürk and R. Dale. Least squares estimation of parameters of the generalized lambda distribution. *Technometrics*, 27:81–84, 1985.

[93] E. M. P. Comon and L. Rota. Blind separation of convolutive mixtures: A contrast-based diagonalization approach. In *Proc. of the Third International Conference on Independent Component Analysis and Signal Separation, ICA2001*, pages 686–691, 2001.

[94] E. S. Pearson. Karl Pearson - An appreciation of some aspects of his life and work. *Biometrika*, 28:193–257, 1936.

[95] E. S. Pearson. Karl Pearson - An appreciation of some aspects of his life and work. *Biometrika*, 29(3/4):161–248, 1938.

[96] E. S. Pearson. Some problems arising in approximating to probability distributions, using moments. *Biometrika*, 50(1/2):95–112, 1963.

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

[98] D.-T. Pham and P. Garat. Separation of a mixture of independent sources through a quasimaximum likelihood approach. *IEEE Transactions on Signal Processing*, 45:1712–1725, 1997.

[99] J. S. Ramberg, E. J. Dudewicz, P. R. Tadikamalla, and E. F. Mykytka. A probability distribution and its uses in fitting data. *Technometrics*, 21:201–204, 1979.

[100] J. S. Ramberg and B. W. Schmeiser. An approximate method for generating asymmetric random variables. *Communications of Association for Computing Machinery*, 15:987–990, 1972.

[101] J. S. Ramberg and B. W. Schmeiser. An approximate method for generating asymmetric random variables. *Communications of Association for Computing Machinery*, 17:78–82, 1974.

[102] T. Rutkowski, A. Cichocki, and A. K. Barros. Speech enhancement from interfering sounds using CASA techniques and blind source separation. In *Proc. of the Third International Conference on Independent Component Analysis and Signal Separation, ICA2001*, pages 728–733, 2001.

[103] *SAS/STAT User's Guide, Version 8*, 2000.

[104] B. W. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London/New York, 1986.

[105] H. Solomon and M. A. Stephens. Approximations to density functions using pearson curves. *Journal of American Statistical Association*, 73(361):153–160, 1978.

[106] A. Stuart and J. K. Ord. *Kendall's Advanced Theory of Statistics: Distribution Theory*, volume 1. Edward Arnold, sixth edition, 1994.

[107] Student. The probable error of a mean. *Biometrika*, 6(1):1–25, 1908.

[108] 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, ISSPA*, pages 5–8, 2001.

[109] A. Taleb. Source separation in structured nonlinear models. In *Proc. of 26th Int. Conf. on Acoustics, Speech and Signal Processing*, pages 3513–3516, 2001.

[110] A. Taleb and C. Jutten. Batch algorithm for source separation in postnonlinear mixtures. In *Proceedings of the First International Workshop on Independent Component Analysis and Blind Signal Separation (ICA'99)*, pages 155–160, 1999.

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

[112] F. I. Toranzos. An asymmetric bell-shaped frequency curve. *Annals of Mathematical Statistics*, 23(3):467–469, 1952.

[113] K. Torkkola. Blind separation of audio signals - are we there yet? In *Proc. of the First International Workshop on Independent Component Analysis and Signal Separation, ICA'99*, pages 239–244, 1999.

[114] K. Torkkola. Blind separation of delayed and convolved sources. In Haykin [50], pages 321–375.

[115] 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.

[116] M. Valkama, M. Renfors, and V. Koivunen. Advanced methods for i/q imbalance compensation in communication receivers. *IEEE Transactions on Signal Processing*, 49:2335–2344, 2001.

[117] H. Valpola. Nonlinear independent component analysis using ensemble learning: Theory. In *ICA2000. Proceedings of the Second International Workshop on Independent Component Analysis and Blind Signal Separation*, pages 251–256, 2000.

[118] R. Vigario, J. Särelä, V. Jousmäki, M. Hämäläinen, and E. Oja. Independent component approach to the analysis of EEG and MEG recordings. *IEEE Transactions on Biomedical Engineering*, 47(5):589–593, 2000.

[119] N. Vlassis and Y. Motomura. Efficient source adaptivity in independent component analysis. *IEEE Transactions on Neural Networks*, 12(3):559–566, 2001.

[120] L. Xu, C. Cheung, and S.-I. Amari. Learned parameter mixture based ICA algorithm. *Neurocomputing*, 22:69–80, 1998.

[121] H. H. Yang and S.-I. Amari. Adaptive on-line learning algorithms for blind separation - maximum entropy and minimum mutual information. *Neural Computation*, 9(7):1457–1482, 1997.

[122] A. Yeredor. Blind source separation vie the second characteristic function. *Signal Processing*, 80:897–902, 2000.

[123] Y. Zhang and S. Kassam. Blind separation and equalization using fractional sampling of communication signals. *Signal Processing*, 81(12):2591–2608, 2001.