Fast and Accurate Methods for Independent Component Analysis

Ph.D. Thesis

Ing. Zbyněk Koldovský

Supervisor: Ing. Petr Tichavský, CSc.
Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Prague

August 2005
Abstract

The thesis deals with several problems in blind separation of linear mixture of unknown sources using independent component analysis. Among other things, it focuses on a key question: how accurate the separation can be done, and how to achieve the best possible separation in practice.

First, the problem with indeterminacy of order and signs of original sources is addressed. The indeterminacies need to be retrieved in several situations, therefore, a general method for optimum assignment of separated sources to the original or some desired ones is proposed.

In order to find out some limit of separation accuracy, Cramér-Rao Lower Bound for linear independent component analysis is derived, which is an algorithm independent bound. It is shown that the bound depends on the distribution of the original sources only. Next, performance analysis of both original versions of a well-known algorithm FastICA is done. It is shown that the bound can be approached in certain scenarios.

Based on a simple idea of generalization of a symmetric version of algorithm FastICA, an improved algorithm, called EFICA, is proposed. The novel method is shown to be efficient, i.e. its accuracy attains the Cramér-Rao bound, provided that score functions of the original sources are known. The algorithm is tuned to be efficient for signals with Generalized Gaussian distributions. Computer simulations validate the efficiency and show that the method outperforms other competitive methods in different scenarios. Computational complexity of the algorithm is shown to be reasonably low.
Acknowledgements

This thesis would not have been written without many people, who helped me with my Ph.D. studies. My thanks are due to my parents, my future Zuzana, friends and colleagues.

I am most grateful to my supervisor Petr Tichavský for his collaboration, advice, insightful reviews and fellowship. I greatly appreciate the time he spent with me.

I am grateful to Professor Erkki Oja from Helsinki University of Technology for the pleasant and worthwhile collaboration I enjoyed.

I am grateful as well to the Department of Mathematics of the Faculty of Nuclear Sciences and Physical Engineering and to the Institute of Information Theory and Automation for invaluable support.

Financial support of this work was provided by the Czech Technical University in Prague through the project CTU0508214.
Contents

1 Introduction 1
  1.1 Motivation ................................. 1
  1.2 Blind Techniques ........................... 2

2 Linear ICA 9
  2.1 Definition ................................... 9
    2.1.1 Preprocessing ............................ 10
  2.2 Methods ..................................... 12
  2.3 Algorithm FastICA ........................... 15
    2.3.1 Preprocessing ............................ 16
    2.3.2 Choice of the Contrast Function ......... 17
    2.3.3 Optimization ............................. 18

3 Objectives and Results 25
  3.1 Thesis Objectives ............................ 25
  3.2 Results ..................................... 28
    3.2.1 Optimal Pairing of Signal Components .... 28
    3.2.2 Analysis of the FastICA Algorithm and Cramér-Rao
         Bounds for Linear ICA ....................... 29
    3.2.3 Efficient Variant of Algorithm FastICA ..... 30

Bibliography 31

Reprints 39
List of Symbols

\( A \) Regular \( d \times d \) mixing matrix
\( W \) The demixing matrix equal to \( A^{-1} \)
\( S \) \( d \times N \) matrix of the original signals
\( X \) \( d \times N \) matrix of the mixed signals
\( N \) Number of processed data
\( d \) Number of signals
\( \mathbb{R} \) Set of real numbers
\((S)_{k\ell}\) \( k\ell \)-th element of matrix \( S \)
\((S)_k\) A random variable with the same distribution function like \((S)_k\)
\( I \) The identity matrix
\( 1_N \) \( N \times 1 \) vector of ones
\( F_S \) Joint distribution function of variables \((S)_1, \ldots, (S)_d\)
\( f_S \) Joint probability density function of variables \((S)_1, \ldots, (S)_d\)
\( F_{s_k}, F_k, F_{(S)_k} \) Distribution function of variable \((S)_k\)
\( f_{s_k}, f_k, f_{(S)_k} \) Probability density function of variable \((S)_k\)
\( E_f[\cdot] \) The expectation value with respect to the density \( f \)
\( \text{diag}[x] \) A diagonal matrix with the elements of the vector \( x \) on the diagonal
Chapter 1

Introduction

1.1 Motivation

The meaning of what everything does the word “information” involve is, with all the awakening, very wide. For instance, for human, this covers the whole contact with his/her environment. In general, it can be said that there is an all the time repetitious process, in which we receive some information, usually by means of some “sensor”. Then, we process the information so that an output of this processing is another information which is more or less beneficial, for example in sense that we are able to understand it. It would be beyond the scope to deeply dispute about how this process can be described in more general way; what everything does the word information mean. Therefore, let us focus on the thing which is very important for us: how to process the information effectively. Eventually, remark that there has always been an effort to develop devices which can do the processing instead of human.

If the processing should be effective so that it is able to abstract “the important part” of the information, it is necessary to have some knowledge about how the information comes into existence, i.e., some prior knowledge about the process. Actually, hereby we are approaching the fundamentals of the whole science, i.e. the studies of natural processes. If we know the core of the mechanism of some process we are able to understand the information which it is produced by, and so to gain the useful information. Then, based on such knowledge, it is possible to develop a device which can do the processing itself. Henceforth, instead of the word “information” the word “signal” will be used since, in this general sense, their meanings are very close.

Our limited understanding of some mechanisms often constrains us to
regard the generated signals as random even when they are not so. Then the nature can be cognised by processing the random signals so that the correctness of our result can be related to the likelihood, i.e. the probability that the result is correct. This is useful in most physical sciences, as well as in many social sciences, due to their own needs for signals analysis, where the processed signals seem to be random. This is closely related to statistical signal processing. Blind Source Separation, which is the main topic of this thesis, falls into this discipline, as well.

The general principles known from the probability theory [18, 47, 49], e.g. the Central Limit Theorem, provide very powerful tools in signal processing. For non-experts, the stochastic modeling may even seem to enable us to puzzle out the fundamentals of the modeled process, which is, indeed, untrue. However, this sensation-desire motivate each of us, because the approach seems to give promising results without paying much effort when describing the process mechanisms, especially, if this is almost impossible due to its chaotic behavior. Finally, this is what makes blind methods so much popular: they provide a chance to retrieve lightsome information from often very complex and unintelligible data.

1.2 Blind Techniques

Blind Source Separation (BSS) consists in recovering unobserved signals or sources without any prior information given only the sensor observations that are unknown transformations (mixtures) of the source signals. Typically the observations are obtained at the output of a set of sensors, where each sensor receives a different transformation of the source signals. There are two important features which characterize the task and explain the adjective “blind”. First, the source signals are not observed and second, no information is available about the mixing transformation.

![Diagram of the mixing process](image-url)

Figure 1.1: Sketch of the mixing process.

The formulation of the problem cannot be completely general since it would not be solvable. Therefore, there are several models of the mixing
transformation which somehow restrict the generality of the task in order to be used for a specific application. Various models of signals are considered, as well.

In nature, the mixing of signals often proceeds as their weighted superposition at each sensor. The basic model of the mixing transformation, therefore, is an instantaneous linear mixing model, whose parametrization can be simply provided by an unknown $m \times n$ mixing matrix $A$, whose $k\ell$-th element will be denoted by $a_{k\ell}$. The model is useful in situations when the conditions for the propagation of signals remains unchanged through the time, and the sampling frequency is small in comparison with the speed of the propagation. This is, indeed, too restrictive for many real situations, for instance, due to multi-path propagation and reverberation in real environment (e.g., speech processing applications) discussed below. On the other hand, it is sufficient for many applications such as in biomedical engineering. Moreover, it is an underlying task for any blind models and the solutions of other studied problems coming out of it.

In the whole thesis, $S$ and $X$ are, respectively, $n \times N$ and $m \times N$ matrices, whose each row contains the observations of each original or mixed signal. Then, the instantaneous linear mixing model can be written in form

$$X = AS,$$  \hspace{1cm} (1.1)

where

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1N} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mN} \end{pmatrix} \quad \text{and} \quad S = \begin{pmatrix} s_{11} & \cdots & s_{1N} \\ \vdots & \ddots & \vdots \\ s_{m1} & \cdots & s_{mN} \end{pmatrix}$$

Different number of observed signals than the original $m \neq n$ can be considered, but, from the theoretical point of view, the case when $m = n = d$ is mainly of importance. As a motivation example see an artificial mixing of two speech signals with a random regular mixing matrix (Fig. 1.2).

A direct application of the linear BSS can be found, in particular, in biomedical engineering [64, 42] since the signals measured by electroencephalogram (EEG) or magneto-encephalogram (MEG) [56] on human head fit the linear model (1.1) well. Another application area are telecommunications [66], economic time series analysis [36], noise reduction, etc.

An often referred example of blind source separation is the so-called cocktail party problem [40]. Imagine that several people are speaking simultaneously in a room. The aim is to separate their voices using recordings of several microphones. In this situation, indeed, the linear model is
not sufficient, because the recordings involve miscellaneously delayed original sources together with their repetitions. This is typical feature of real environment, where the signals impinged on an array of sensors are convolutive mixtures of the original sources. Thus, a convolutional mixing model should be considered, which is called blind equalization/deconvolution [4, 5, 24, 26, 58]; see also [14] and references therein. Each received sample at \( k \)-th sensor can be written in form

\[
x_{k\ell} = \sum_{i=1}^{n} \sum_{t=0}^{+\infty} h_{it} s_{i,\ell-t},
\]

where \( h_{it} \) are parameters of the mixing system; in practice, indeed, finite number of the parameters is considered. Even though this thesis deals only with the linear model, note that applying the Fourier transformation on (1.2) the problem can be settle as a set of instantaneous linear mixtures for each frequency channel, consequently, the tasks are very closely related [22].

Naturally, the convolutional model can be applied in more signal processing areas; examples include communications (see Fig. 1.5), radar, sonar, audio, image processing, or control systems [14]. In the latter branches, which are currently growing, and this trend is likely to continue for several years, blind methods received considerable attention in last two decades. For instance, in wireless communication systems, blind channel identification is an attractive alternative to its training sequence based counterparts as it does not squander channel capacity with the pe-
1.2. BLIND TECHNIQUES

Figure 1.3: Example of a mobil phone signal transmission. The received signal is a convolutive mixture of the original signal due to multi-path propagation in real environment.

periodic transmission of training sequences [4, 20, 27, 34, 54]. Another example of application is speaker voice enhancement for speech recognition [40].

An important issue are noisy models, which, however, will not be studied in this thesis. Note that both models, the linear and the convolutive one, approach the reality more if an additive noise is considered, i.e., for example (1.1) takes form

\[ X = AS + N, \]  

(1.3)

where \( N \) denotes a matrix of white Gaussian noise. In that case, the original signals are no more exactly retrievable, consequently, identification of the mixing model is not the same task as the estimation of the original sources.

The second important question in BSS is how to model the original signals. It is not possible to simply assume that each original signal is just an arbitrary stochastic process, otherwise, convergence of any separating method cannot be assured. On the other hand, facile models may be too restrictive for real applications, e.g. the data from electro-encephalogram are very complex and any valid model is still unknown. It can be said, in general, that the more restrictive the BSS model is the more powerful method for its solution can be developed. Consequently, semi-blind methods [44, 38], which assumptions about the signals are often based on some more specific a priori information, are considered, also.

The basic assumption about the signal is that it is just a sequence of i.i.d. random variables [10, 59]. The advantage is that the methods based on such model can be easily analyzed and are sufficient for many real appli-
cations. Nevertheless, the drawback is that the model does not reflect any time-structure of the modeled signal and cannot separate Gaussian signals. Therefore, there are various methods which regard the signals as random stationary Gaussian processes [21, 70], non-stationary Gaussian processes [50, 72], etc.

In the literature, various BSS models have been studied; nonlinear or time-dependent models included. There are still many questions both theoretical or practical, and due to the universality of BSS still many problems arise in various fields. Nowadays, blind methods are very popular in the community, and their further applications in real practice are very feasible.
Figure 1.4: Various random signals: (a) non-stationary audio signal (b) i.i.d. signal equal to $\pm 1$ with the same probability (c) non-stationary EEG signal (d) stationary Gaussian process, e.g. produced by some industrial process.
Figure 1.5: Example of how a binary signal from a mobile phone can be distorted due to multi-way propagation and reconstructed using blind deconvolution. It is convolved with a FIR filter via (1.2). The blind deconvolution is done by means of algorithm FastICA described in the following chapter, where FIR inversion (four taps) of the convolving filter is considered.
Chapter 2

Linear Independent Component Analysis

In this chapter, the most popular method for BSS, Independent Component Analysis (ICA), will be introduced, and the underlying task, solution of the linear model, will be stated. A brief survey of existing methods for ICA and short state-of-the-art will be provided. Then, a very famous algorithm FastICA, which is an important issue of the thesis, will be described and demonstrated.

2.1 Definition

In order to retrieve the original sources from their mixture it is necessary to define them properly, to hit the correct property which characterize them, in general, if possible. Such a good property, on which ICA is based, is the stochastic independence [10]. The goal of ICA can be formulated as finding a transformation of the observed signals so that the transformed signals are independent, or at least as independent as is possible. In case of the linear model (1.1) this means to estimate the demixing matrix \( W = A^{-1} \) (the regular case \( m = n = d \) is considered) using the fact that the estimated signals

\[
\hat{S} = \hat{W}X
\]  

(2.1)

should be independent. Here, \( \hat{W} \) denotes the estimate of \( W \).

In general, the independence of random processes is defined so that the joint distribution function of any finite subset of elements of the processes is product of the marginal distributions for all the finite subsets. Nevertheless, this thesis is focused on the i.i.d. model of the signal, i.e., the signal is regarded as a sequence of identically and independently distributed random
variables. Let \( F_{s_k}(x) \) be the distribution function of \( s_k \), where \( s_k \) denotes a random variable with the same distribution as the distribution of \( s_{k1} \). Then, from the independence assumption follows that

\[
F_{s_1, \ldots, s_d}(y_1, \ldots, y_d) = \prod_{i=1}^{d} F_{s_i}(y_i)
\]

(2.2)

This is the condition that the distribution of the estimated signals (2.1) should fulfil.

There are several approaches how to perform the separation using the independence assumption either based on some criterion or maximum likelihood principle [16, 57] etc. Now, the common features of ICA will be focused, while the separating methods will be introduced in the next section. The solution of ICA problem has several indeterminacies [29] caused by invariance of criterion (2.2):

1. Obviously, the original order of signals cannot be retrieved without any a priori information as it is irrelevant in (2.2).

2. Multiplying each original signal with a positive number does not affect their independence, also. Therefore, it is often, for simplicity, assumed that the signals have unit variance, which may be restrictive in some cases [37] since finite variance assumption is required. However, this is more or less problem in theory, because in practice finite number of samples is processed and can be normalized.

3. Similarly, signs of the signals cannot be recovered. In case of complex-valued signals their original phases remain unknown.

4. Any orthogonal transformation of two independent random variables with Gaussian distribution results in two independent Gaussian variables again. Therefore, only one of the original signals \( S \) may have Gaussian distribution.

Note that the mean values of the signals is irrelevant, as well, therefore, zero mean value of each signal is a frequently claimed simplifying assumption.

### 2.1.1 Preprocessing

Due to the indeterminacies described above it is very common to define the separating criterion for signals with zero mean and unit variance. Therefore, the mean value of the observed data denoted by \( \mathbf{X} \) is removed, at first,
and the data $\mathbf{X} - \overline{\mathbf{X}}$ are substituted by any orthogonal system of signals $\mathbf{Z}$ which span the same space (with the same norm equal to $\sqrt{N}$). From statistical point of view it means that signals $\mathbf{Z}$ are uncorrelated. This is a useful transformation of the data due to the fact that the unit variance of any signal $w^T \mathbf{Z}$ is automatically ensured providing that $\|w\| = 1$.

The transformation can be viewed as factorizing the estimate of the demixing matrix $\hat{\mathbf{W}}$, which separates the signals $\mathbf{X} - \overline{\mathbf{X}}$, as

$$\hat{\mathbf{W}} = \mathbf{W}(\mathbf{Z}) \cdot \mathbf{R},$$

where $\mathbf{Z} = \mathbf{R}(\mathbf{X} - \overline{\mathbf{X}})$, and the matrix $\mathbf{W}(\mathbf{Z})$ is the to-be-estimated transformation of signals $\mathbf{Z}$ which separates them. The matrix $\mathbf{R}$ is the orthogonalizing transformation, usually called whitening, that obeys

$$\mathbf{Z} \mathbf{Z}^T = \mathbf{R}(\mathbf{X} - \overline{\mathbf{X}})(\mathbf{X} - \overline{\mathbf{X}})^T \mathbf{R}^T = N \cdot \mathbf{I}.$$ 

Note that the matrix $\mathbf{R}$ is not uniquely defined, because any matrix $\mathbf{Q} \cdot \mathbf{R}$, where $\mathbf{Q}$ is orthogonal, fulfils the latter condition, also.

Note that the choice of the matrix $\mathbf{R}$ is based only on the second-order statistics since the orthogonality of signals $\mathbf{Z}$ is equivalent with their uncorrelatedness. However, this is only the necessary condition of the independence, thus, from the ICA point of view none of the orthogonal systems $\mathbf{Z}$ is preferable. Therefore, Cardoso et al. [13] introduced a class of so-called equivariant algorithms whose estimate of the original signals is independent of any regular transformation of the observed signals $\mathbf{X}$. Hence, such equivariant algorithm is independent of the choice of the orthogonal system $\mathbf{Z}$. Nevertheless, it is often suggested to choose the normalized principal components of $\mathbf{X} - \overline{\mathbf{X}}$, because the directions with the highest variance may be (usually based on some prevalent a priori information) somehow interesting. As an example, see part C of the summarizing Figure 2.1.

Now, the goal is to estimate the matrix $\mathbf{W}(\mathbf{Z})$ so that the signals $\mathbf{W}(\mathbf{Z}) \cdot \mathbf{Z}$ are independent. The so-called one-unit approaches estimate only one row of $\mathbf{W}(\mathbf{Z})$ denoted by $w^T$ usually via optimizing some criterion subject to the constraint $\|w\| = 1$. In general, it is not known in advance which row of $\mathbf{W}(\mathbf{Z})$ is being estimated: it largely depends on some initialization of the optimization. Then, the question is how to estimate the other rows.

Based on the necessary condition that the signals should be uncorrelated it is often claimed that the matrix $\mathbf{W}(\mathbf{Z})$ is orthogonal [9, 30], i.e., its rows form an orthonormal system. Then, the other rows can be found under the condition that the currently estimated row should be orthogonal to the previous ones, or there are several different approaches how to estimate all the rows simultaneously under the constraint. In any case, the
orthogonality condition reduces the problem from finding $d(d - 1)$ parameters to $d(d - 1)/2$. This simplification is worthwhile, but note that the orthogonality is not based on true covariances but on their sample-based counterparts. **Thus, the methods using the orthogonality condition cannot achieve the best possible separation [11].**

### 2.2 Methods

Many methods for linear ICA have been proposed in the literature since the topic became very popular in the signal processing community. Therefore, only brief description of the leading features of the separation principles will be provided, and it will be shown that different approaches are very closely related, in general. Comprehensive surveys of the methods are provided by several articles and books [10, 29, 40].

Considering the given model of the original signals and the mixing transformation, the log-likelihood function can be derived. First, assume, for simplicity, that the probability density function (pdf) $f_i(\cdot)$ of each distribution function $F_i(\cdot)$ $i = 1, \ldots, d$ exists. The density of the mixed data $X$ is given by

$$f_X(X) = |\det A^{-1}|^N \prod_{k=1}^{d} \prod_{\ell=1}^{N} f_k ((A^{-1}X)_{k\ell}),$$

thus, the log-likelihood function divided by $N$ of the transformed signals $Y = B^{-1}X$ is

$$\frac{1}{N} \ln f_{B^{-1}X}(Y|B) = \frac{1}{N} \sum_{k=1}^{d} \sum_{\ell=1}^{N} \ln f_k ((A^{-1}BY)_{k\ell}) + \ln |\det A^{-1}B|. \quad (2.5)$$

The maximum likelihood estimate (MLE) of the matrix $B$ is worth considering since it gives the best asymptotically normal estimate of $A$ for large scale of distributions. It is closely related to the Kullback-Leibler divergence [17] between the density of $S$ and $B^{-1}X$. The divergence between two random matrices $Y_1$ and $Y_2$ with joint probability densities, respectively, $f_{Y_1}$ and $f_{Y_2}$ is defined as

$$K(Y_1|Y_2) \overset{\text{def}}{=} \int_{\mathbb{R}^d} f_{Y_1}(t) \ln \frac{f_{Y_1}(t)}{f_{Y_2}(t)} dt = \int_{\mathbb{R}^d} f_{Y_1}(t) \ln f_{Y_1}(t) dt - \int_{\mathbb{R}^d} f_{Y_1}(t) \ln f_{Y_2}(t) dt = E_{f_{Y_1}}[\ln f_{Y_1}] - E_{f_{Y_1}}[\ln f_{Y_2}] \quad (2.6)$$
whenever the integrals (expectations) exist. Now, for \( N \to +\infty \) the strong
law of large numbers gives

\[
\frac{1}{N} \ln f_{B^{-1}X}(Y|B) \xrightarrow{a.s.} \mathbb{E}_f s[\ln f_{B^{-1}X}] = -K(S|B^{-1}X) + \mathbb{E}_f s[\ln f_s]\text{ const.} \tag{2.7}
\]

Hence, MLE minimizes the divergence between the distribution of the original signals \( S \) and the estimated signals \( B^{-1}X \). The divergence can be regarded as a contrast function, which is a function that should be minimized by adjusting the model parameters (elements of \( B \)).

Assume that \( \tilde{Y} \) is a random vector with independent entries whose marginal distributions are the same as the corresponding marginal distributions of \( B^{-1}X \). A simple calculus shows that

\[
K(S|B^{-1}X) = K(S|\tilde{Y}) + \int_{\mathbb{R}^d} f_S(t) \ln \frac{f_{\tilde{Y}}(t)}{f_{B^{-1}X}(t)} dt \tag{2.8}
\]

It is obvious that (2.8) is zero if \( f_{\tilde{Y}} = f_{B^{-1}X} \) a.e. and \( f_{\tilde{Y}} = f_S \) a.e. at the same time. In fact, the former equality requires mutual independence of variables \( (B^{-1}X)_1, \ldots, (B^{-1}X)_d \) since joint distribution of \( \tilde{Y} \) is product of its marginal distributions. Thus, a criterion for minimizing \( K(S|B^{-1}X) \) can be the Kullback-Leibler divergence between \( B^{-1}X \) and \( \tilde{Y} \)

\[
K(B^{-1}X|\tilde{Y}) = \int_{\mathbb{R}^d} f_{B^{-1}X}(x) \ln \frac{f_{B^{-1}X}(x)}{\prod_{k=1}^d f_{(B^{-1}X)_k}(x_k)} dx. \tag{2.9}
\]

This is usually called mutual information [17] or multiinformation [61] of entries of \( B^{-1}X \) denoted by

\[
I((B^{-1}X)_1, \ldots, (B^{-1}X)_d) .
\]

In general, the Kullback-Leibler divergence (2.6) is zero if and only if the distributions \( f_{X_1} \) and \( f_{X_2} \) are the same. Then, it follows that the mutual information (2.9) is zero if and only if the entries of \( B^{-1}X \) are independent, which gives us a direct criterion of independence matching the basic formulation of the task of ICA.

Under the orthogonality constraint discussed in the previous section, i.e. minimizing the mutual information of \( W(Z) \cdot Z \)

\[
I((W(Z) \cdot Z)_1, \ldots, (W(Z) \cdot Z)_d) \tag{2.10}
\]
provided that \( W(Z) \) is an orthogonal matrix, a simple rearrangement of the definition (2.9), using the fact that \( |\det W(Z)| = 1 \), gives

\[
I((W(Z) \cdot Z)_1, \ldots, (W(Z) \cdot Z)_d) =
\]

\[
= \int_{R^d} f_{W(Z) \cdot Z}(z) \ln f_{W(Z) \cdot Z}(z) dz - \int_{R^d} f_{W(Z) \cdot Z}(z) \ln \prod_{k=1}^{d} f_{((W(Z) \cdot Z)_k)(z_k)}(z_k) dz =
\]

\[
= -\sum_{k=1}^{d} \int_{R} f_{(W(Z) \cdot Z)_k}(z_k) \ln f_{(W(Z) \cdot Z)_k}(z_k) dz_k + \text{const.}
\]

From this follows that (under the orthogonality constraint) equivalent criterion to (2.10) is

\[
\sum_{k=1}^{d} H((W(Z) \cdot Z)_k),
\]  

(2.11)

where \( H(X) \) denotes the Shannon entropy [17] of a random variable \( X \) with density \( p(x) \)

\[
H(X) = -\int_{R} p(x) \ln p(x) dx.
\]  

(2.12)

This also suggest a criterion for one-unit approach: marginal entropy of the estimated signal \( H((W(Z) \cdot Z)_k) \) since (2.11) is minimized whenever each term of the sum is minimized. However, note that in practice finite number of samples is processed, and then the criteria (2.9), (2.11), and the marginal entropy minimization are no more equivalent.

In case of blind methods, of course, the distributions of the original signals are unknown. Therefore, as a contrast function various approximations/surrogates of (2.7-2.11) are used. Basically, they can be divided into three groups:

1. the true distribution is modeled by some hypothetical distribution [2, 5, 31],

2. a nonparametric estimate of the distribution is used [7, 39], or

3. higher-order-statistics based approximation is used, e.g. the cumulants-based expansions [9].

The next important part of any separating algorithm is the optimization approach for the selected criterion. As soon as the criterion (the contrast function) is defined, any gradient-based algorithm can be used (either
2.3. ALGORITHM FASTICA

batch-oriented or recursive) in order to adjust the elements of the demixing matrix provided that the criterion is differentiable subject to the parameters. It is also very common that the optimization step lead up to the joint diagonalization problem [9, 72, 71]. Several particular tasks may be solved analytically, as well [62].

There are several aspects which should be regarded when qualifying a given algorithm for ICA: computational complexity (convergence speed), reliability (stability), versatility (applicability for wide range of distributions), demandingness of implementation (in case we want to implement the algorithm in some special device), accuracy (theoretical and practical), etc. In the large number of algorithms that have been proposed it is difficult, if not impossible, to choose the best one. This is not only due to the fact that some aspects, e.g. reliability and versatility, cannot be optimized simultaneously, but there are only few algorithms whose performance has been studied in more details. This gives rise to the question: “What is the limitation for the performance and can we reach that?”

Analysis of performance of the given algorithm plays important role both in theory and in practice since it provides insight into the method and reveals its possible optimization. Such a combination: algorithm proposal, analysis, and optimization originate a method which can be widely used for being a reliable alternative in particular case. Despite the fact that the linear ICA is nowadays well-studied problem, and several algorithms which have been sufficiently investigated provide good background for BSS tasks solutions, the whole problem still cannot be regard as closed. Missing sophisticated analysis of algorithm FastICA, which will be introduced in the following section, is one of the still many unresolved problems in ICA.

2.3 Algorithm FastICA

Algorithm FastICA was first proposed by Hyvärinen and Oja in 1997 [30]. It is one of the most popular and widely used ICA algorithm thanks to its speed, accuracy and selectivity, mainly. In addition, this all is supported by an user-friendly implementation published on website [33]. The algorithm is using a fixed-point iteration scheme for finding the local extrema of a contrast function, which might be considered for an extension of the algorithm by Shalvi and Weinstein [57] or of any kurtosis-based or fourth-order cumulants-based algorithms.

The contrast function derivation in [30] is based on marginal entropy minimization principle described in previous section. For a signal with zero
mean \( Y \) it is defined as
\[
J(Y) = (E[G(Y)] - E[G(\zeta)])^2, \tag{2.13}
\]
where \( G(\cdot) \) is a suitable non-linear non-quadratic function, and \( \zeta \) is a Gaussian variable with zero mean and the same variance like \( Y \). It is obvious that the contrast function can be considered as a surrogate (approximation) of negentropy of \( Y \) defined by
\[
N(Y) = H(\zeta) - H(Y). \tag{2.14}
\]
Taking \( G(x) = -\ln p(x) \) where \( p(x) \) is pdf of \( Y \) (2.13) becomes just a quadratic function of \( H(Y) \). The generality of the criterion admits as a special choice \( G(x) = x^4 \) so that then \( E[\zeta^4] = 3 \) and, consequently, (2.13) becomes square of kurtosis defined by (for \( Y \) having zero mean and variance one)
\[
\]
Kurtosis is zero for Gaussian signals, thus, it is commonly used as a measure of Gaussianity (see Figure 2.1 part D). In summary, the goal is to minimize the entropy of \( Y \) while the negentropy (2.14) is to-be maximized.

### 2.3.1 Preprocessing

The first step of the algorithm is removing mean from the input data \( X \) and whitening them. As it was noted in section 2.1.1, the whitening is not uniquely defined. In the original version of FastICA published on web [33] principle component analysis (PCA) is used, but other whitening transformation can be used instead. PCA can be used for reduction of the data dimension, which is required when there are more sensors than the original signals \((m > n)\). However, note that this reduction is efficient only in case that the original data have Gaussian distribution.

The whitening can be simply done by multiplying the data with a **whitening matrix** \( C^{-\frac{1}{2}} \) (thus the matrix \( R \) from section 2.1.1 is \( C^{-\frac{1}{2}} \))
\[
Z = C^{-\frac{1}{2}}(X - \bar{X}), \tag{2.15}
\]
where \( C \) is their sample covariance matrix, i.e.
\[
C = (X - \bar{X})(X - \bar{X})^T/N.
\]
The matrix \( C^{-\frac{1}{2}} \) is defined using eigenvalue decomposition \( C = U\Lambda U^T \), where \( U \) is a unitary matrix and \( \Lambda \) is a diagonal matrix, as
\[
C^{-\frac{1}{2}} = U\Lambda^{-\frac{1}{2}}U^T.
\]
Figure 2.1: Example of the preprocessing and kurtosis-based identification of two uniformly distributed signals within $[-\sqrt{3}, \sqrt{3}]$. In Figure B the variance of each direction is shown (magenta line), where red arrows indicate directions with the highest variances (principal components). In Figure D the kurtosis-based contrast function (i.e. $E[G(x^4)]$) of each direction computed from the whitened data is painted. The minimum-kurtosis directions correspond to independent components, i.e. original signals.

Now, it follows that

$$\frac{ZZ^T}{N} = C^{-\frac{1}{2}}(X - \bar{X})(X - \bar{X})^T C^{-\frac{1}{2}} = C^{-\frac{1}{2}}CC^{-\frac{1}{2}} = I,$$

thus, the transformed signals $Z$ are orthogonal. In agreement with the section 2.1.1, the goal is to estimate the rows of matrix $W(Z)$ by optimizing the criterion (2.13).

### 2.3.2 Choice of the Contrast Function

The value of the contrast function (2.13) should involve the information content of the signal (variable). The second order information, i.e. variance, is irrelevant in this matter, thus, it is natural to consider the value for signals with a fixed variance, e.g. one. The transformation (2.15) simplifies evaluation of the contrast function for the unit variance signal (2.13) since we do not need to normalize the input variable (signal). An estimated signal in an iteration $w^T Z$ has zero sample mean and unit sample variance simply provided that $\|w\| = 1$.

The drawback of this approach is that the accuracy of the contrast function estimate depends on accuracy of the estimation of the mean value and the variance via sample moments. For instance, if the estimated variance $\hat{\sigma}^2$ of the signal is too inaccurate, the signal is normalized (divided by $\hat{\sigma}$),
then the contrast function estimate may be violated by the fact that it is, actually, evaluated for signal with variance much different from one.

Despite the assumption of existence of the moments the estimation via sample moments may be poor, which occurs, in particular, in case when the signals have heavy tailed distribution. This often causes stability, numerical, or convergence problems. Fortunately, the effect of variance-estimation error is negligible otherwise. Moreover, suitably chosen function $G(\cdot)$ in (2.13) can assure robustness of the contrast function estimate against this feature for sufficiently large class of distributions.

Finally, note that analogous problems may be caused by badly estimated mean values of the signals.

### 2.3.3 Optimization

After defining the contrast function the goal is to find a unit-norm vector which maximizes $J(w^T Z)$. Note that this approach allows us to estimate only one of the original signals (one row of the demixing matrix) since the criterion function is derived from marginal entropy principle. This is called one-unit approach. Methods for estimating all the original signals will be derived later.

The choice of an optimization method of the contrast function has an influence on important properties of the final algorithm, e.g., speed of convergence or reliability. There is number of methods which can be used to this end with the steepest descent being the most used. A crucial issue when using this gradient-based optimization is an iteration step-size parameter. Hence, the idea of authors of the algorithm FastICA was to employ Newton optimization iteration method. In this way the problem with the step-size choice is avoided, moreover, the convergence speed can be hopefully increased. (A generalization via including a step-size parameter into the Newton iteration is possible though.)

Nevertheless, the Newton method has several flaws. First, the method turns out to be nonstable for ICA problem and second, each iteration is computationally demanding since computation of Hessian and its inversion is necessary. Therefore, authors used an ad hoc approximation of Hessian instead, which results, after some rearrangement, in a fixed-point algorithm. The derived method cannot be regarded as the exactly Newton one, but it has some common promising qualities such as the convergence rate (order) in a neighborhood of the solution [46].
2.3. ALGORITHM FASTICA

Figure 2.2: Comparison of iteration steps of (A) gradient-based optimization with a step size taken from [54] $\mu = 1/(N \cdot E[G(w^T Z)])$ (B) FastICA iteration. The experiment was performed with two uniformly distributed signals within $[-\sqrt{3}, \sqrt{3}]$. Black arrows indicate the result of the optimization; the red arrows indicate the initialization. The black dotted lines denote values of the kurtosis-based contrast function like in Figure 2.1.

One-unit approach

Instead of finding a maximum of (2.13), i.e.

$$J(w^T Z) \quad \text{subject to} \quad ||w|| = 1$$

the algorithm consists in finding a stationary point of

$$E[G(w^T Z)] \quad \text{subject to} \quad ||w|| = 1,$$

(2.16)

where $E[\cdot]$ is the sample average operator and function $G(\cdot)$ is applied elementwise. The reason is that the latter contrast function has straightforward derivative, and those stationary points that correspond to a minimum of $J(w^T Z)$ are nonstable, which was shown in [46] for infinite length of data. On the other hand, some troubles may occur in a real situation when processing finite amount of data. Solution to this problem is a part of this thesis.

The derivation of the algorithm FastICA is as follows. The stationary point of (2.16) obeys

$$E[Zg(w^T Z)] - \beta w = 0,$$

(2.17)

where $g(\cdot)$ denotes the derivative of $G(\cdot)$ and $\beta$ is a Lagrange multiplicator. A Jacobian of the left-hand side of (2.17) is

$$\mathcal{J} = E[ZZ^T g'(w^T Z)] - \beta I,$$

(2.18)
where \( I \) denotes the identity matrix. The Newton iteration is defined by

\[
    w^+ = w - J^{-1}E[Zg(w^T Z)] \\
    w^\text{new} = w^+/\|w^+\| (2.19)
\]

Note that \( \beta = E[w_0^T Zg(w_0^T Z)], \) where \( w_0^+ \) is the to-be found solution of (2.17). Following ad hoc approximations of \( \beta \) and (2.18) are used

\[
    \beta \approx E[w_0^T Zg(w_0^T Z)] (2.21) \\
    J \approx E[ZZ^T] E[g'(w^T Z)] - \beta I = (E[g'(w^T Z)] - \beta) I, (2.22)
\]

where \( w \) stands for an up-to-date unit-norm vector in the iteration process. Now, the steps (2.19,2.20) are, after few simplifications, replaced by

\[
    w^+ = E[Zg(w^T Z)] - E[g'(w^T Z)]w \\
    w^\text{new} = w^+/\|w^+\| (2.23)
\]

The algorithm iterates until convergence is achieved. Several features of the one-unit approach are as follows.

- It is not known in advance which of the original signals is being estimated. This largely depends on the initialization.
- It can be easily predicted that the (asymptotic) accuracy of estimation of the estimated signal is independent of the other signals. This can be seen when replacing sample averages with the expectation operator in (2.23) assuming that \( w \) is the correct solution.
- A Gaussian signal (only one of the original is possible) cannot be retrieved this way.

### Estimating other signals

There are several ways how to estimate several or all the original signals. It is possible to run the one-unit algorithm using different initialization, but we need to prevent it from converging to the same solutions. A common approach is to enforce the estimated signals to be orthogonal.

A deflation scheme consists of estimating signals in sequence where each signal is orthogonal to the previously estimated signals. Consequently, the first signal is estimated via ordinary one-unit approach, and the order of how the signals are being estimated depends on the initialization. A serious drawback of this approach is that the estimation of each signal is biased due to error in estimation of the previous ones.
A symmetric approach estimates all the original signals (or equivalently, all rows of $\mathbf{W}$) in parallel. The parallel one-unit iterations are followed by a symmetric orthogonalization of the estimated rows of $\mathbf{W}$. The whole algorithm can be written in matrix form
\begin{align*}
\mathbf{W}^+ &= g(\mathbf{WZ})\mathbf{Z}^T - \text{diag}[g'(\mathbf{WZ})\mathbf{1}_N] \mathbf{W} \\
\mathbf{W} &= (\mathbf{W}^+\mathbf{W}^{+T})^{-1/2}\mathbf{W}^+,
\end{align*}
(2.25)
(2.26)
where $\mathbf{1}_N$ stands for an $N \times 1$ vector of 1’s.

The orthogonality condition allows us to estimate all the signals, i.e., even the Gaussian one, if included. Similarly, the estimation of badly estimable signals can be improved thanks to the well estimated ones (the symmetric approach). On the contrary, the signals which can be well estimated by one-unit approach may be degraded this way. Moreover, as it was noted above, the orthogonality condition forbids the algorithm to achieve the best possible separation.

This suggests combining advantages of all the approaches together. For instance, well estimable signals can be improved via one-unit iterations after a primary run of the symmetric approach since the primary estimate can be taken as the initialization, because we can assume that it lies in the right domain of attraction. In any case, this all cannot be accomplished without theoretical knowledge about performance of the approaches.

To conclude this subsection a brief survey on resolved questions concerning the algorithm FastICA will be provided. An experimental comparison of the performance with another ICA methods has been done in [25], where promising results of the symmetric version were demonstrated. Rigorous theoretical analysis of the performance is missing, however, the performance of the one-unit approach was shown [32] to depend on the same characteristic like other gradient-based one-unit methods [19], which is
\begin{equation}
\frac{\mathbb{E}[g^2(s_k)] - (\mathbb{E}[s_kg(s_k)])^2}{(\mathbb{E}[g'(s_k)] - \mathbb{E}[s_kg(s_k)])^2}
\end{equation}
(2.27)
Here, $s_k$ denotes the original signal estimated by the one-unit method and $\mathbb{E}[\cdot]$ denotes the expectation operator.

The quantity characterizes asymptotic variance of estimation of elements of the demixing vector $\mathbf{w}$. The smaller the value (2.27) is, the smaller the variance. It is worth to note that (2.27) depends on the distribution of $s_k$ and on the choice of nonlinearity $g(\cdot)$ only. In addition, for Gaussian signal the denominator is zero regardless the choice of $g(\cdot)$, thus the variance is infinite. This corresponds to the statement that the Gaussian signal
cannot be retrieved by means of the one-unit approach. The characteristic can be shown to be minimized if and only if the function $g(\cdot)$ equals to the score function of the distribution of $s_k$, i.e.

$$g(x) = \psi(x) \overset{\text{def}}{=} -\frac{f_k'(x)}{f_k(x)}.$$

The speed of convergence was shown to be quadratic [46, 31] at least (depends on the choice of $g(\cdot)$); see also [29]. The monotonic convergence for the related gradient algorithm was considered in [53]. In [6] a version for complex-valued signals is derived.

The algorithm FastICA received a general popularity due to its simplicity, speed, accuracy, and last but not least good publication too, but it suffers from missing satisfactory theoretical examination. Therefore, this thesis focuses on this problem in order to place the algorithm into well-deserved pantheon of standard ICA methods.
Blind source separation example

Figure 2.3: Example of separation of three different signals by means of algorithm FastICA using nonlinearity \( g(x) = \tanh(x) \), equivalently \( G(x) = \ln \cosh(x) \). The separated signals correspond to the original ones up to the original order and signs. The basic functionality of the method can be easily seen, but the accuracy cannot be measured from the picture. Next, one can see that mere preprocessing (whitening) is insufficient for the separation.
Chapter 3

Objectives and Results

The previous chapter suggested general problems concerning ICA methods and the whole blind separation task at all. In this chapter, the problems in frame of the previous matter are described, the objectives of the thesis are formulated, and the results are summarized.

3.1 Thesis Objectives

In spite of the general effort to effectively apply BSS in multifarious scale of fields nowadays many important issues regarding fundamental theoretical background remain open. As it was noted above, this often consists in unknown theoretical behavior of a given method. If the method is derived to go right in some particular situation, it may fail even if the conditions are only slightly changed. Therefore, the theoretical knowledge of the method is so important, because, in addition, it clearly defines limits for its usability. The same is valid for algorithm FastICA, which is the main subject of the thesis.

When doing a performance analysis of some ICA method it is important to clearly define an unified criterion of accuracy of the separation. This is very simple in case of noiseless linear ICA model since the estimation of system parameters (the (de)mixing matrix elements) and the estimation of the original signals are directly proportional. A presence of the $\ell$-th signal in the $k$-th estimated signal is characterized by $k, \ell$-th element of so-called gain matrix $G$ defined by

$$G = \hat{W}A,$$  \hspace{1cm} (3.1)

where $\hat{W}$ is estimate of the demixing matrix.
If the separation was successful, the gain matrix should be closed to a matrix which has one and only one non-zero element per column and row, which is equal to $\pm 1$. This is due to the indeterminacy of signs and order of the original signals (see section 2.1). If the signs and the order is known, the gain matrix is closed to the identity matrix. In contrast to the theoretical analysis, where this can be simply assumed, the problem cannot be avoided, for instance, when measuring quality of separation in an experimental setup. The first part of this thesis (subsection 3.2.1) deals with this simply called permutation problem, where an optimal assignment of the estimated signals to the original ones is proposed.

The criterion (elements of the gain matrix $G$) measures separation quality of a particular case. However, as the model of the original signals is stochastic, the value of the criterion is stochastic as well. The ideal case would be if we were able to derive the final distribution law of the criterion depending on the distributions of the original signals and system parameters. Unfortunately, this is possible in only very special cases. Thus, it is common to aim at the variance of the distribution assuming that it exists, because it characterizes amplitude of fluctuations of the criterion around the ideal value.

Nevertheless, still the exact derivation of the variance can be intractable, in particular, in the case of iteration methods for parameter estimation, i.e., when a closed form expression for the estimate does not exist. Then, asymptotic methods [52] provide a powerful tool since knowledge of the asymptotic variance of the estimate is sufficient for most practical purposes.

The FastICA algorithm is the case when asymptotic approach is worth to use. As it was noted, the one-unit version of FastICA was partially analyzed in [32], thus the result of the exact analysis can be foreseen. On the other hand, it is necessary to do the analysis exactly subject to the criterion (3.1). Accuracy of the symmetric approach is one of subjects of this thesis.

A natural question is, whether there exists some theoretical limit for separation performance, and whether it is possible to reach it, or how close to the limit the accuracy of algorithm FastICA is. In statistics, such a limit provides the Cramér-Rao lower bound (CRB) [47, 18]. Considering any unbiased estimate of some parameter, this bound, if exists, is the lowest attainable variance of the estimate. It is worth to note that the asymptotic variance of the maximum likelihood estimate (MLE) is, under mild conditions, equal to the CRB. This property of MLE was used to derive the asymptotic CRB for ICA in [10, 69, 59, 51].

The aim of the second part (subsection 3.2.2) of the thesis is to carry out the asymptotic analysis of both versions of algorithm Fas-
3.1. THESIS OBJECTIVES

tICA in terms of asymptotic variance of elements of the gain matrix $G$, to derive the exact Cramér-Rao lower bound in terms of the same criterion, and to compare the bound with the asymptotic performance both in theory and in practice.

The general properties of one-unit approaches and of those approaches which uses the orthogonality constraint, whereto the symmetric approach belongs, suggest that both the versions of algorithm FastICA cannot be efficient in general. Of course, the results in subsection 3.2.2 will confirm this reflection. Thus, the goal of the third part (subsection 3.2.3) is to derive an improved version of algorithm FastICA, which relaxes the orthogonality constraint and is as effective as is possible.
3.2 Results

The thesis consists of three papers which will be briefly described in following subsections.

3.2.1 Optimal Pairing of Signal Components Separated by Blind Techniques


The indeterminacies of the original sources is essential feature of blind methods. In linear ICA model, the sources can be estimated up to the original order, signs, and scales. The separation, which is the primary task, can be done, but then the frequent problem is how to interpret the results.

As it was noted, we can assume, to avoid the indeterminacy of the original scales, that the original signals have unit variance. Then, the signs and the order remain unknown, which is called a permutation problem. Sometimes the original order of the signals needs to be retrieved, for example, in order to compare each estimated signal with its original counterpart, which occurs when testing accuracy of some algorithm for separation. Then, usually a definition of some pairwise matching criterion is stated (correlation), and, based on this, the signals are reordered so that their mutual similarity is maximized.

The problem, whereto the reordering leads up, can be formulated in term of graph theory as the one to find optimal matching in a complete bipartite graph. In this paper, it is pointed out that the optimal matching can be achieved by the Kuhn-Munkres algorithm, in general, in contrast to the commonly used method known in graph theory as the greedy algorithm.

The proposed method is demonstrated on data from electroencephalogram. The data are separated via algorithm FastICA in a sliding time-window. There being a different order of unknown original sources in each time-window, therefore, the reordering was done so that the estimated sources correspond to the estimated in the very first time-window as much as is possible. It is shown that the method results in smoother changes of the mixing system than that obtained by the greedy method.
3.2.2 Performance Analysis of the FastICA Algorithm and Cramér-Rao Bounds for Linear Independent Component Analysis


This very comprehensive article can be divided into two parts. In the first part, the asymptotic analysis of both versions of algorithm FastICA was done. The result is summarized in a proposition. It validates the general statements concerning both the versions of the algorithm (subsection 2.3.2).

The original versions of the algorithm suffer from random convergence failures. This is caused by the fact that finite amount of data is processed, or finite stop criterion for optimization is used. Then the iteration process may stop in a stationary point of the contrast function that corresponds to a minimum of (2.13). Therefore, a simple check of saddle points of the contrast function is proposed to settle this problem. This is an important improvement of the algorithm since it allows us to validate the analysis by computer simulations.


The second part of this paper involve a straightforward derivation of the Cramér-Rao lower bound for linear ICA. The result coincides with the asymptotic bound in literature [10, 69, 59, 51, 65]. The bound is compared with the theoretical performance of the algorithm FastICA, which gives an important conclusion: FastICA can approach the CRB (the estimate can be nearly efficient) in two situations: (1) when the distribution of the original signals is not too much different from Gaussian (the symmetric version), and (2) when the distribution of the sources is very different from Gaussian (the one-unit version). Both methods achieve the best performance when the nonlinear function \( g(\cdot) \) equals the score function of each independent component \( \psi_k(\cdot) \).

3.2.3 Efficient Variant of Algorithm FastICA for Independent Component Analysis Attaining the Cramér-Rao Lower Bound


It is very important that the analysis of FastICA provides a way how to estimate the accuracy of separation from the acquired signals. A simple utilization, called for brevity Smart FastICA, was also presented in the previous work. It starts with the symmetric approach, and, based on the estimated accuracy of each signal separation, it decides whether to use the one-unit approach or not. However, the best possible separation cannot be achieved this way.

In this paper, an improvement of the algorithm FastICA is proposed. It comes from generalization of the symmetric approach via weighted step \( (2.26) \). Performance analysis of this modification is straightforward, and it provides a criterion for choice of the optimum weighting. Based on this, an optimum refinement procedure is proposed, and it is shown that the method is efficient, i.e. its accuracy attains the CRB, provided that the nonlinear function \( g(\cdot) \) in the algorithm equals to the score function of each independent component.

However, the distributions of the original signals are unknown, therefore, a parametric choice of the nonlinearity \( g_k(\cdot) \) is done for each estimated signal. The choice was optimized for signals with Generalized Gaussian Distribution [67], because this class covers a wide variety of typical unimodal distributions.

The proposed improvement is compared with several competitive ICA methods in terms of accuracy and speed. It gives superior results provided that the signals have Generalized Gaussian distribution. The complexity is only slightly (about three times) higher than that of the original FastICA.

Bibliography


[67] K. Waheed, F. M. Salem, “Blind source recovery using an adaptive
generalized Gaussian score function”, The 2002 45th Midwest Sympos-
sium on Circuits and Systems, MWSCAS-2002., 4-7 Aug. 2002 Vol.3,
pp. 656-659.

to the Permutation Problem of Frequency Domain Blind Source Sep-
aration”, Proc. Fifth Intern. Conf. on ICA and BSS, Granada, Spain,

[69] D. Yellin, B. Friedlander, “Multichannel system identification and de-
convolution: performance bounds”, IEEE Tr. Signal Processing, Vol. 47,

[70] A. Yeredor, “Blind Separation of Gaussian Sources via Second-Order
Statistics with Asymptotically Optimal Weighting”, IEEE Signal Pro-

[71] A. Yeredor, “Non-Orthogonal Joint Diagonalization in the Least-
Squares Sense with Application in Blind Source Separation”, IEEE

1998.
Reprints