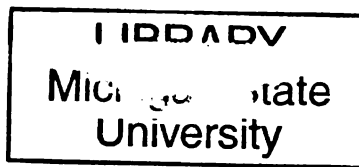




141
933
THS



This is to certify that the
dissertation entitled

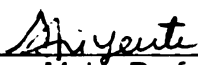
MULTICHANNEL SIGNAL DECOMPOSITION AND
SEPARATION IN THE TIME-FREQUENCY DOMAIN

presented by

Zeyong Shan

has been accepted towards fulfillment
of the requirements for the

Ph.D. degree in Electrical Engineering


Major Professor's Signature

08/25/09

Date

PLACE IN RETURN BOX to remove this checkout from your record.
TO AVOID FINES return on or before date due.
MAY BE RECALLED with earlier due date if requested.

DATE DUE	DATE DUE	DATE DUE

**MULTICHANNEL SIGNAL DECOMPOSITION AND
SEPARATION IN THE TIME-FREQUENCY DOMAIN**

By

Zeyong Shan

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Electrical Engineering

2009

ABSTRACT

MULTICHANNEL SIGNAL DECOMPOSITION AND SEPARATION IN THE TIME-FREQUENCY DOMAIN

By

Zeyong Shan

The extraction of signals or components from observed data is a fundamental and challenging problem in many signal processing applications. In many practical situations, observations may be modeled as linear mixtures of a number of source signals, i.e. a linear multi-input multi-output system. A typical example is speech recordings made in an acoustic environment in the presence of background noise and/or competing speakers. Other examples include multichannel biological signal recordings such as the electroencephalogram, passive sonar applications and cross-talk in data communications. The well-known approaches to the signal decomposition and separation problems include second or higher order statistics based methods, principal component analysis, and independent component analysis. Most of these methods are developed in the time domain, and thus inherently assume the stationarity of the underlying signals. However, most real world signals are non-stationary and have highly complex time-varying characteristics. For non-stationary signals, common signal analysis techniques such as the standard Fourier transform are not useful since the transient part of the signal such as spikes and high frequency bursts cannot be easily detected from the Fourier transform. These problems could be overcome by using non-stationary signal analysis tools such as the quadratic time-frequency distributions (TFDs). TFDs provide a two-dimensional representation of the time-varying energy information in the signal, and are suitable for tracking the non-stationary behavior of signals. Hence, there have been efforts to perform the signal decomposition

and separation in the time-frequency domain.

In this dissertation, the multichannel signal decomposition problem in the time-frequency domain is first considered. A new adaptive signal component extraction method is proposed based on the minimum entropy criterion. This method decomposes the signals into the components that are well-concentrated on the time-frequency plane. Unlike the traditional Gabor decomposition, the signal is expressed as a finite sum of the components extracted by the proposed algorithm whose time and frequency centers are determined by the signal and not by a pre-determined dictionary. Next, the overdetermined blind source separation problem is addressed in the time-frequency domain. We present a novel approach to achieve source separation using an information-theoretic cost function. Jensen-Rényi divergence, as adapted to time-frequency distributions, is introduced as an effective cost function to extract sources that are disjoint on the time-frequency plane. The sources are extracted through a series of Givens rotations and the optimal rotation angle is found using the steepest descent algorithm. The proposed method is applied to several example signals to illustrate its effectiveness and the performance is quantified through simulations. After that, the underdetermined blind source separation problem is discussed. The proposed approach takes advantage of the high resolution of time-frequency distributions for obtaining a sparse representation, and separates the sources by a simple clustering algorithm followed by a convex optimization problem. Compared to other time-frequency based separation methods, the approach presented is characterized by simplicity and ease of implementation. Finally, the proposed approach for the case of underdetermined blind source separation is applied to real signals such as electroencephalogram signals to further evaluate its performance. The experimental results show that the proposed method is more effective at extracting well-localized neuronal sources in time and frequency than ICA.

To my beloved parents and wife

ACKNOWLEDGMENTS

I would like to acknowledge many people for helping me during my doctoral work. I would especially like to thank my advisor, Prof. Selin Aviyente, for her generous time and commitment. Throughout my doctoral work she encouraged me to develop independent thinking and research skills. She continually stimulated my analytical thinking and greatly assisted me with scientific writing. Her knowledge, kindness, and patience have provided me with lifetime benefits.

I am also very grateful for having an exceptional doctoral committee and wish to thank Profs. John Deller, Guowei Wei, and Hayder Radha for their continual support and encouragement, and valuable comments and suggestions on my dissertation draft. I would also like to thank many faculty members in ECE department of MSU who were the instructors for the courses I ever took. These course works have enriched my knowledge and provided the background and foundations for my doctoral research.

I extend many thanks to my fellow graduate students at MSU. In particular, I would like to thank Jacob Swary for helping me finish a part of the experiment.

Finally, I would like to thank my family. My parents extended their passion for plants and nature to me. I'm especially grateful to my wife, Yongmei, for her support, encouragement, and unwavering love, and helping me keep my life in proper perspective and balance.

TABLE OF CONTENTS

List of Tables	ix
List of Figures	x
1 Introduction	1
1.1 Overview of Contributions	5
2 Background on Time-Frequency Analysis and Information-Theoretic Measures	7
2.1 Introduction to Time-Frequency Analysis	7
2.1.1 Short-Time Fourier Transform	8
2.1.2 The Wigner Distribution	9
2.1.3 Cohen's General Class of Time-Frequency Distributions	10
2.1.4 Reduced Interference Distributions	11
2.2 Introduction to Information-Theoretic Measures	12
2.2.1 Entropy	12
2.2.2 Rényi Entropy	13
2.2.3 Divergence Measures for Time-Frequency Distributions	15
Kullback-Leibler Divergence	15
Rényi Divergence	16
Jensen-Shannon Divergence	16
Jensen-Rényi Divergence	17
3 Review of Signal Decomposition and Source Separation Methods	18
3.1 Principal Component Analysis	18
3.1.1 Principal Components	19
3.1.2 PCA By Variance Maximization	19
3.2 Independent Component Analysis	22
3.2.1 Definition of ICA	23
3.2.2 ICA by Maximum Likelihood Estimation	24
3.2.3 ICA by Minimization of Mutual Information	25
3.3 Review of Time-Frequency Signal Decomposition and Separation Approaches	27
3.3.1 Matching Pursuit with Time-Frequency Dictionaries	27
3.3.2 Spatial Time-Frequency Distribution (STFD) Method	30
3.3.3 Blind Separation via Time-Frequency Masking	32
4 Adaptive Signal Decomposition on the Time-Frequency Plane	35
4.1 Introduction	35
4.2 Background on Gabor Decomposition and Information Measures	36

4.2.1	Gabor Signal Expansion	36
4.2.2	Chirplet Transform	37
4.2.3	Entropy Measure on the Time-Frequency Plane	38
4.3	Component Extraction Method	39
4.3.1	Problem Statement	39
4.3.2	The Proposed Approach	40
4.3.3	Convergence Analysis of the Algorithm	42
4.4	Experimental Results and Analysis	44
4.5	Summary	47
5	Overdetermined Blind Source Separation in the Time-Frequency Domain	49
5.1	Introduction	49
5.2	Information Measures in the Time-Frequency Domain	51
5.3	Problem Formulation and Method	53
5.3.1	Signal Model and Assumptions	53
5.3.2	Problem Statement in the Time-Frequency Domain	55
5.3.3	Cost Function	57
5.3.4	Rotation	58
5.3.5	Proposed Algorithm	60
5.4	Experimental Results and Analysis	61
5.5	Summary	71
6	Underdetermined Blind Source Separation in the Time-Frequency Domain	74
6.1	Introduction	74
6.2	Sparse Factorization Approach for UBSS in the Time-Frequency Domain	75
6.2.1	Linear Mixture Model and Assumptions	76
6.2.2	K -means Clustering	77
6.2.3	Determination of the Mixing Matrix	78
6.2.4	Estimation of the Source Signals for a Given Mixing Matrix	79
6.3	Experimental Results and Analysis	80
6.4	Summary	85
7	Applications of UBSS Algorithm on Electroencephalogram Signals	86
7.1	Introduction to Electroencephalogram and Event-Related Potential	87
7.1.1	Electroencephalogram	87
7.1.2	Event-Related Potential	88
7.2	Experimental Results and Performance Analysis	90
7.2.1	EEG/ERP Data Set	90
7.2.2	Single-Trial EEG	90
7.2.3	Data Reduction	91
7.2.4	Performance Evaluation	95
7.3	Summary	99
8	Conclusions and Future Work	101

Bibliography 108

LIST OF TABLES

4.1	Entropy Comparison	46
7.1	Mean measure of l_1 norm to show sparsity	98
7.2	Mean measure of entropy to show time-frequency localization	99
7.3	Measure of disjointness by correlation between components	99
7.4	Average component projection to electrodes	100

LIST OF FIGURES

4.1	The average time-frequency distribution of 128 trials and the 5 extracted components of the proposed method	46
4.2	The average time-frequency distribution of 128 trials and the 6 elements of OMP	48
5.1	Comparison of Kullback-Leibler and Jensen-Rényi divergence measures under an additive signal perturbation model	53
5.2	The mixture and the separation of a chirp and two Gabor logons: (a) the mixture; (b) and (d) the two extracted Gabor logons; (c) the extracted chirp	64
5.3	The cost function versus the number of iterations for Example 1	65
5.4	The mixture and the separation of two crossing chirp signals: (a) the mixture; (b) and (c) the separated signals	66
5.5	TFDs of the two individual speech signals: (a) TFD of a female speaker; (b) TFD of a male speaker	67
5.6	TFDs of the observed signals: (a) TFD of the first mixture; (b) TFD of the second mixture	68
5.7	TFDs of the extracted signals: (a) TFD of estimate of the female speaker; (b) TFD of estimate of the male speaker	69
5.8	Output SIR versus input SNR for speech signals	70
5.9	Comparison of output SIR versus input SNR for three different source separation methods	71

5.10	Comparison with PCA for two Gabor logon extraction: (a) the mixture, (b) and (c) the components extracted by the proposed method, (d) and (e) the components extracted by PCA	72
5.11	Separation of a chirp and a Gabor logon from their three mixtures: (a) the mixture, (b) the extracted Gabor logon, (c) the extracted chirp	73
6.1	Scatter plot of two mixtures of four Gabor logons in the time-frequency domain	81
6.2	The mixtures and the separation of four Gabor logons: (a) and (b) two mixtures; (c), (d), (e), and (f) four extracted Gabor logons	82
6.3	Scatter plot of two mixtures of a chirp and two Gabor logons in the time-frequency domain	83
6.4	The mixtures and the separation of a chirp and two Gabor logons: (a) and (b) two mixtures; (c) extracted chirp; (d) and (e) two extracted Gabor logons	84
6.5	Comparison of MSE versus SNR for extracted sources with TFD and WP	85
7.1	Single-trial results using 32 frequency bins: 6 extracted sources from ICA	92
7.2	Single-trial results using 32 frequency bins: 8 extracted sources from the proposed UBSS method	93
7.3	Results of component clustering over all single-trial results for stimulus position $u = 1$ using ICA	96
7.4	Results of component clustering over all single-trial results for stimulus position $u = 1$ using the proposed UBSS method	97

CHAPTER 1

INTRODUCTION

Signal decomposition and separation are two important and fundamental problems in signal processing with a broad range of applications including communications, speech signal processing, biomedical signal processing, and sensor networks [1–4]. The research in this dissertation focuses on multichannel signal decomposition and source separation from the perspective of time-frequency distributions taking into account the non-stationarity of real life signals.

The purpose of signal decomposition is to extract a set of features characterizing the signal of interest. Often this is realized by decomposing the signal on a set of elementary functions. An example of such a decomposition is the Fourier transform, which decomposes a given signal using harmonic functions. However, in the case of non-stationary signals, i.e., signals whose characteristics change with time, the Fourier transform does not yield a useful characterization of the signal. Such signals can be adequately decomposed on a set of locally supported elementary functions, giving rise to a so-called time-frequency decomposition. In a general time-frequency decomposition, the signal is decomposed using a set of elementary functions, characterized by their time and frequency centers. Such functions are called time-frequency atoms (*tf*-atoms). The majority of linear decomposition methods, including matching pursuit [5], basis pursuit [6], and the chirplet decomposition [7], decompose the signal on a set of *tf*-atoms, selected appropriately among a large and redundant dictionary. A problem with these decomposition methods is that the representations are not satisfactory unless all signal components are at least reasonably well approximated by dictionary elements.

The first part of this dissertation focuses on the decomposition of the observed

multichannel signals into a few number of components in the time-frequency domain. The major objective is to obtain a compact set of signal components that can represent the observed/measured signals. A new adaptive signal component extraction method is proposed based on the minimum entropy criterion. This method decomposes the signals into the components that are well-concentrated on the time-frequency plane. The concentration of the components are quantified through an entropy measure in the time-frequency domain. Extracting “minimum” entropy components orthogonal to each other produces compact components that are similar to Gabor logons and describe the given data set in a minimum mean square sense. Unlike the traditional Gabor decomposition, where the signal is expressed as an infinite sum of time and frequency shifted Gabor logons, the components extracted by this algorithm have time and frequency centers determined by the signal, and not by a pre-determined dictionary. Moreover, the components extracted in this approach have chirp rates and local spread adapted to the given set of signals. The results show that the proposed approach is effective in determining a few number of components that can be used to represent a large set of data.

In many signal processing applications, one has only access to measurements of mixed, i.e. superimposed, signals and the question is how to construct suitable projections that allow to demix and thus find the underlying (unmixed) signals of interest. Blind source separation (BSS) techniques aim at answering this question to reveal unknown sources using two ingredients (a) a model of the mixing process (typically a linear superposition) and (b) the assumption of statistical independence. As opposed to other signal processing techniques like beamforming [8], BSS uses no geometrical information about the sensor array of the underlying sources, therefore BSS is called “blind”.

In the last several years there has been much work on the problem of blind source separation, which has resulted in many diverse approaches. Most of these approaches

use higher-order statistics, minimum mutual information, and maximum entropy in their solutions. The concept of independent component analysis (ICA) is defined in [9] which measures the degree of independence among outputs using contrast functions approximated by the expansion of the Kullback-Leibler divergence. The higher order statistics is approximated by cummulants up to fourth order and requires intensive computation. Researchers in neural computation have developed adaptive learning algorithms which are simpler and biologically more plausible [10–12]. An information-theoretic approach has been proposed for the blind source separation and blind deconvolution problem [13]. The ICA has been reformulated in a maximum likelihood (ML) framework where the underlying density is estimated in a context sensitive manner [14].

Most of these methods are developed in the time domain, and thus inherently assume the stationarity of the underlying signals. However, most real world signals are non-stationary and have highly complex time-varying characteristics. Since the quadratic time-frequency distributions (TFDs) provide a two-dimensional representation of the time-varying energy information in the signal and thus are suitable for tracking the non-stationary behavior of signals, there have been efforts to perform the blind source separation in the time-frequency domain.

The second part of the proposed research focuses on the blind separation of the source signals from their mixtures in the time-frequency domain when the number of mixtures is greater than or equal to the number of sources, i.e. the overdetermined case. A new approach is introduced combining time-frequency representations with information-theoretic measures. An information-theoretic criterion, Jensen-Rényi divergence as adapted to time-frequency distributions, is used as the objective function for source separation thanks to its robustness against perturbations and noise. It is shown that this cost function achieves its maximum when the source signals are disjoint with each other. The proposed approach performs signal separation through a

multidimensional Givens rotation transformation using a steepest descent algorithm under the assumption of the approximate disjointness of the underlying source signals in the time-frequency domain. Issues regarding the convergence rate and robustness under noise of the proposed algorithm are investigated.

In the third part of the dissertation, an underdetermined blind source separation problem, i.e. the number of the mixtures is less than the number of the sources, is considered in the time-frequency domain. Compared with the (over)determined case, the underdetermined source separation is more challenging due to the noninvertibility of the mixing matrix. A two-stage sparse factorization approach is proposed to achieve source separation. The first stage of the algorithm is to determine the mixing matrix. It is shown that the mixing matrix can be estimated using K -means clustering algorithm under the condition that the source signals are sparse in the time-frequency domain. The column vectors of the mixing matrix are cluster centers of normalized mixture vectors. The second stage of the algorithm is to estimate the sources. For a given mixing matrix, although there exists an infinite number of solutions in general, the sparse solution with minimum l_1 -norm is proven to be unique, which can be obtained by using linear programming methods.

In the fourth part of the dissertation, we apply the proposed underdetermined source separation approach to the real life electroencephalogram (EEG) signals using the time-frequency distributions so as to evaluate its effectiveness. The proposed approach is capable of extracting more sources than sensors. This is important since the number of sources is unknown, and since many EEG setups do not have large electrode arrays. This approach is compared to the popular ICA algorithm when applied to the same multiple trial EEG/ERP data set. Data reduction by clustering is performed over all single-trial results to extract components that represent the results. The components are consistently more sparse compared to ICA, showing that ICA probably tends to extract components that are sums of sources. The technique

presented provides components that are more localized in the time-frequency domain and that are more distinct from each other than does ICA.

1.1 Overview of Contributions

The contributions of the dissertation consist of four parts: signal decomposition based on an information-theoretic criterion, overdetermined source separation by combining time-frequency representations with information-theoretic measures, underdetermined blind source separation achieved by a two-stage sparse factorization approach, and the applications of the proposed separation methods to biological signals.

In signal decomposition, a new adaptive component extraction method is proposed based on the minimum entropy criterion. The main contributions of this part of research work can be summarized as follows:

1. Time-frequency data reduction is accomplished by producing a few meaningful components on the time-frequency plane that explain most of the signal's energy.
2. This time-frequency domain decomposition can extract activity that overlaps in time and frequency domains, which is not possible using either time domain or frequency domain decomposition approaches.
3. The proposed approach has the ability to separate and extract parts of chirped signals, which cannot be achieved using the conventional Gabor expansion.

For the overdetermined source separation problem, a novel separation approach is presented with the following contributions:

1. Maximizing the information-theoretic divergence can effectively separate disjoint sources in the time-frequency domain.
2. The proposed method is superior to typical time domain or frequency domain

separation methods like PCA and ICA for extracting the source signals overlapping with each other in both the time and frequency domains.

3. The proposed approach also outperforms some time-frequency methods in the literature for high noise levels since it assumes the cross-terms between sources are negligible which effectively denoises the observed time-frequency matrix.

In underdetermined blind source separation, a new extraction algorithm is introduced combining the K -means clustering and linear programming. The main contributions of this part of the dissertation are:

1. The source signals are assumed to be sparse in the time-frequency domain, and do not necessarily have to be orthogonal or independent to each other unlike PCA or ICA.
2. The algorithm for determining the mixing matrix is simple and effective.
3. The proposed two-stage approach is more robust than wavelet packets under noisy environments.

In the fourth part of the research, the proposed underdetermined separation approach is applied to the EEG signals using the time-frequency distributions with the following contributions:

1. Single-trial source separation can detect any changes of state in the subject which is not possible with averaging of multiple trials, since it ignores trial-to-trial variability.
2. Components extracted by the proposed approach are more sparse, localized, and distinct in the time-frequency domain than those extracted by ICA.
3. The presented method can also be used as an effective data reduction method.

CHAPTER 2

BACKGROUND ON TIME-FREQUENCY ANALYSIS AND INFORMATION-THEORETIC MEASURES

In this chapter, we briefly introduce the theory of time-frequency analysis and relevant information-theoretic measures.

2.1 Introduction to Time-Frequency Analysis

The most common methods for representing a signal are its time and frequency domain representations. Although frequency domain representations such as the power spectrum of a signal often give information about the frequency content of a signal, the representations do not show how the frequency content evolves over time. For the majority of signals encountered in everyday life, the frequency content of the signals varies over time. Since the basis functions used in the classical Fourier analysis do not associate with any particular time instant, the resulting measurement, the Fourier transform, does not explicitly reflect the signal's time-varying nature. Thus, it is difficult to establish the point-to-point relationship between the time domain and the frequency domain based on the conventional Fourier analysis.

The fundamental idea of time-frequency analysis is to understand and describe situations where the frequency content of a signal is changing with time. There are numerous applications in both research and industry for time-frequency analysis. Examples include speech analysis [15], telecommunications [16], bioacoustics [17], geophysics [18], and structural analysis [19]. There are a number of different transforms available for time-frequency analysis. In the following sections, we will introduce some of the main time-frequency transforms, including the short-time Fourier transform (STFT), Wigner distribution (WD), Cohen's general class of transforms, and the reduced interference distributions (RID).

2.1.1 Short-Time Fourier Transform

A simple way to overcome the deficiency possessed by the regular Fourier transform is to combine the signal with elementary functions that are localized in time and frequency domains simultaneously,

$$S_h(t, \omega) = \int s(\tau) h^*(\tau - t) e^{-j\omega\tau} d\tau, \quad (2.1)$$

which is a regular inner product and reflects the similarity between the signal $s(t)$ and the elementary function $h(\tau - t)\exp\{j\omega\tau\}$. The function $h(t)$ usually has a short time duration and thereby it is named the window function. Equation (2.1) is called the short-time Fourier transform (STFT) or windowed Fourier transform. The spectrogram, which is the energy density spectrum at time t , is defined as:

$$P(t, \omega) = |S_h(t, \omega)|^2. \quad (2.2)$$

To obtain a good time resolution, a narrow window, $h(t)$, in the time domain has to be picked. Similarly, to get a good frequency resolution, a narrow window, $H(\omega)$, in the frequency domain has to be picked. Since both $h(t)$ and $H(\omega)$ can not be made arbitrarily narrow, there is an inherent trade-off between time and frequency resolution in the spectrogram for a particular window. This is the reason why the spectrogram is not preferred for high resolution time-frequency analysis.

2.1.2 The Wigner Distribution

The Wigner distribution is defined mathematically in terms of the signal, $s(t)$, as [20–22]:

$$W(t, \omega) = \int s(t + \frac{\tau}{2}) s^*(t - \frac{\tau}{2}) e^{-j\tau\omega} d\tau. \quad (2.3)$$

The Wigner distribution is said to be bilinear in the signal since the signal enters twice in its calculation.

The Wigner distribution has many desired properties. For example, it satisfies the marginals requirement, and therefore preserves the energy. It is always real, even if the signal is complex. In addition, it is time and frequency shift invariant, and satisfies finite support property in time and frequency. One of the major shortcomings of the Wigner distribution is the existence of negative energy terms. The Wigner distribution of multicomponent signals also exhibits the disturbing tendency of generating interference or cross-terms.

Despite these shortcomings, the Wigner distribution still shows some remarkable advantages over the spectrogram: the conditional averages are exactly the instantaneous frequency and the group delay, whereas the spectrogram fails to achieve this result, no matter what window is chosen; the spectrogram can not often provide the resolution required to distinguish the components in multicomponent signals as the Wigner distribution. Thus, there is a need to develop more general distributions which preserve the advantages of the Wigner distribution and address most of its drawbacks. This leads to the Cohen's class of generalized distributions.

2.1.3 Cohen's General Class of Time-Frequency Distributions

There is a considerable advantage to having a simple method to generate different time-frequency distributions. This allows one to pick and choose those with desirable properties. The most direct way is to generate the distributions from [23]:

$$C(t, \omega) = \int \int \int \phi(\theta, \tau) s(u + \frac{\tau}{2}) s^*(u - \frac{\tau}{2}) e^{j(\theta u - \theta t - \tau \omega)} du d\theta d\tau, \quad (2.4)$$

where $\phi(\theta, \tau)$ is a two dimensional function called the kernel function, a term coined by Claasen and Mecklenbrauker [24] and whom, with Janssen [25], made many important contributions to the general understanding of the general class, particularly in the signal analysis context. The kernel function determines the distribution and its properties. For the Wigner distribution, the kernel function is one.

There are three main reasons why the kernel idea is particularly useful for the study of time-frequency distributions. First of all it is easy to generate them: just choose a kernel function. The second reason is that one can design the distributions with certain properties by putting constraints on the kernel function. For example, for a distribution to satisfy the marginals

$$\int C(t, \omega) d\omega = |s(t)|^2, \quad \int C(t, \omega) dt = |S(\omega)|^2, \quad (2.5)$$

it has been shown that the kernel function must have the property

$$\phi(0, \tau) = \phi(\theta, 0) = 1. \quad (2.6)$$

An extensive discussion of the properties of a distribution and the corresponding constraints on the kernel function can be found in [26–29]. The third reason is that when a new distribution is considered, its properties can readily be ascertained by examining its kernel. For example, if the kernel does not satisfy equation (2.6), then we know the distribution can not satisfy the marginals.

2.1.4 Reduced Interference Distributions

It is known that both the spectrogram and the Wigner distribution are the members of Cohen’s class of distributions. Although the spectrogram has many useful properties, it often presents serious difficulties when used to analyze rapidly varying signals. If the analysis window is made short enough to capture rapid changes in the signal, it becomes impossible to resolve frequency components of the signal which are close in frequency. The Wigner distribution has been employed as an alternative to overcome this shortcoming. It provides a high resolution representation in time and frequency for a non-stationary signal such as a chirp. However, its energy distribution is non-positive and it often suffers from severe cross-terms between components in different

time-frequency regions, potentially leading to confusion and misinterpretation. An excellent discussion on the geometry of interferences has been provided in [30–32].

Since the Wigner distribution sometimes gives artificial and undesirable values in the time-frequency domain particularly when the signal is multicomponent, the conditions on the kernel that minimize these spurious values in some sense are developed in [33–36]. These conditions are that the kernel $\phi(\theta, \tau)$ value decays as you move away from the θ and τ axes. A way to describe this region is to observe that the product $\theta\tau$ is large when we are away from either axis. Therefore, it is concluded that for cross-term minimization, $\phi(\theta, \tau)$ should satisfy

$$\phi(\theta, \tau) \ll 1 \quad \text{for } \theta\tau \gg 0. \quad (2.7)$$

These kernels produce reduced interference distributions.

2.2 Introduction to Information-Theoretic Measures

Using entropy based distance functionals is a well-known discrimination method in signal processing. These functionals are known as divergence measures and are applied directly on statistical models describing the signals. Measures of divergence between two probability distributions are used to associate, cluster, classify, compress, and restore signals, images and patterns, in many applications [37, 38]. Many different measures of divergence have been constructed and characterized [39, 40].

Recent research in the application of information and entropy functionals on time-frequency distributions (TFDs) has proven the usefulness of distance measures for non-stationary signal analysis [41, 42]. Entropy when applied to a TFD measures the number of components in a given signal, i.e. the complexity. Similarly, divergence measures computed between two time-frequency distributions can indicate the difference in complexity between the two signals. These measures could prove useful as

time-frequency detection statistics in applications comparing reference and data distributions. In this section, we review some well-known information-theoretic distance measures for time-frequency distributions.

2.2.1 Entropy

Before introducing divergence measures, we first give a brief review of entropy, a basic concept in information theory. Entropy H , also called Shannon entropy, is defined for a discrete-valued random variable X as

$$H(X) = - \sum_i P(X = a_i) \log(P(X = a_i)), \quad (2.8)$$

where $P(\cdot)$ is the probability mass function of X , and the a_i are the possible values of X . Depending on what the base of the logarithm is, different units of entropy are obtained. Usually the logarithm with base 2 is used, in which case the unit is called a bit.

According to the definition, the entropy of a random variable can be interpreted as the degree of information that the observation of the variable gives. The more “random”, i.e., unpredictable and unstructured the variable is, the larger its entropy. Assume that the probabilities are all close to 0, except for one that is close to 1 (the probabilities must sum up to one). In that case, there is little randomness in the variable, since it almost always takes the same value and this is reflected by a small entropy. On the other hand, if all the probabilities are equal, then they are relatively far from 0 and 1. This means that the entropy is large, which reflects the fact that the variable is really random; we can not predict which value it takes.

The definition of entropy for a discrete-valued random variable can be generalized for a continuous-valued random variable, in which case it is often called differential entropy. The differential entropy H of a random variable x with density $p_x(\cdot)$ is

defined as

$$H(x) = - \int p_x(\xi) \log(p_x(\xi)) d\xi. \quad (2.9)$$

Differential entropy can be interpreted as a measure of randomness in the same way as entropy. Note that differential entropy can be negative since probability densities can be larger than 1.

2.2.2 Rényi Entropy

Rényi entropy, a generalization of Shannon entropy, is one of a family of functionals for quantifying the diversity, uncertainty or randomness of a system. Rényi entropy of order α , where $\alpha \geq 0$, is defined as [43]

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \left(\sum_i P^\alpha(X = a_i) \right). \quad (2.10)$$

If the probabilities are all the same, then all Rényi entropies of the distribution are equal with $H_\alpha(X) = \log n$, where n is the number of a_i . Otherwise, the entropies are weakly decreasing as a function of α .

Some particular cases are:

1. For $\alpha = 0$,

$$H_0(X) = \log n = \log |X|, \quad (2.11)$$

which is the logarithm of the cardinality of X .

2. In the limit that α approaches 1, it can be shown that H_α converges to

$$H(X) = - \sum_i P(X = a_i) \log(P(X = a_i)), \quad (2.12)$$

which is Shannon entropy.

3. Rényi entropy refers to the case $\alpha = 2$,

$$H_2(X) = -\log \left(\sum_i P^2(X = a_i) \right). \quad (2.13)$$

4. As $\alpha \rightarrow \infty$, the limit exists as

$$H_\infty(X) = -\log (\sup_i P(X = a_i)). \quad (2.14)$$

This is called Min-entropy, because it is the smallest value of H_α .

The two latter cases are related by $H_\infty < H_2 < 2H_\infty$, while on the other hand Shannon entropy can be arbitrarily high for a random variable X with fixed min-entropy.

2.2.3 Divergence Measures for Time-Frequency Distributions

The most general class of distance measures is known as Csiszar's f -divergence which includes some well-known measures like Hellinger distance, Kullback-Leibler divergence and Rényi divergence [40]. The divergence between two probability density functions, p_1 and p_2 for this class of distance measures can be expressed as:

$$d(p_1, p_2) = g \left\{ E_1 \left[f \left(\frac{p_2}{p_1} \right) \right] \right\}, \quad (2.15)$$

where f is a continuous convex function, g is an increasing function and E_1 is the expectation operator with respect to p_1 . The distance measures and their properties for time-frequency distributions are given below.

2.2.3.1 Kullback-Leibler Divergence

The most common distance measure used for probability distributions is the Kullback-Leibler divergence measure. This measure can be adapted to the time-frequency

distributions as follows:

$$K(C_1, C_2) = \int \int C_1(t, f) \log \frac{C_1(t, f)}{C_2(t, f)} dt df, \quad (2.16)$$

where C_1, C_2 represent two different normalized time-frequency distributions defined in equation (2.4). This measure belongs to the class of Csiszar's f -divergence with $f(x) = -\log x$, and $g(x) = x$. $0 \leq K(C_1, C_2) \leq \infty$, the first equality holds if and only if $C_1 = C_2$ and the second equality holds if and only if $\text{Supp } C_1 \cap \text{Supp } C_2 = \emptyset$. This is not a symmetric distance measure but can easily be symmetrized by taking the average of $K(C_1, C_2)$ and $K(C_2, C_1)$. The main disadvantage of this measure is that it can only be applied to positive TFDs.

2.2.3.2 Rényi Divergence

Rényi divergence is a generalized formulation of Kullback-Leibler divergence and can be expressed as:

$$D_\alpha(C_1, C_2) = \frac{1}{\alpha - 1} \log \int \int C_1^\alpha(t, f) C_2^{1-\alpha}(t, f) dt df, \quad (2.17)$$

where $\alpha \in [0, 1]$ is the order of Rényi divergence. This measure converges to Kullback-Leibler distance as $\alpha \rightarrow 1$. It is also a member of Csiszars f -divergence with $f(x) = x^{1-\alpha}$, and $g(x) = \frac{1}{\alpha-1} \log(x)$. $0 \leq D_\alpha(C_1, C_2) \leq \infty$, the first equality holds if and only if $C_1 = C_2$ and the second equality holds if and only if $\text{Supp } C_1 \cap \text{Supp } C_2 = \emptyset$.

2.2.3.3 Jensen-Shannon Divergence

One common approach for constructing divergence measures is to apply Jensen inequality on the entropy functional. For time-frequency distributions, Jensen-Shannon divergence can be defined as:

$$J(C_1, C_2) = H\left(\frac{C_1 + C_2}{2}\right) - \frac{H(C_1) + H(C_2)}{2}. \quad (2.18)$$

This distance measure is always positive since

$$H\left(\frac{C_1 + C_2}{2}\right) \geq \frac{H(C_1)}{2} + \frac{H(C_2)}{2} \quad (2.19)$$

by concavity of H . It is equal to zero when $C_1 = C_2$ and is a symmetric divergence measure. Unlike the Kullback-Leibler divergence, Jensen-Shannon distance does not diverge when the two distributions are disjoint.

2.2.3.4 Jensen-Rényi Divergence

The Rényi entropy is derived from the same set of axioms as the Shannon entropy, the only difference being the employment of a more general exponential mean instead of the arithmetic mean in the derivation. This realization inspires the modification of Jensen-Shannon divergence from an arithmetic to a geometric mean, and the following quantity is obtained for two positive TFDs C_1 and C_2 .

$$J_1(C_1, C_2) = H_\alpha(\sqrt{C_1 C_2}) - \frac{H_\alpha(C_1) + H_\alpha(C_2)}{2}, \quad (2.20)$$

where $\sqrt{C_1 C_2}(t, f) = \sqrt{C_1(t, f) C_2(t, f)}$. This quantity is obviously null when $C_1 = C_2$. The positivity of this quantity can be proven using the Cauchy-Schwartz inequality.

CHAPTER 3

REVIEW OF SIGNAL DECOMPOSITION AND SOURCE SEPARATION METHODS

Blind signal processing is one of the important topics in the fields of neural computation, advanced statistics, and signal processing with solid theoretical foundations and many potential applications. In this chapter, we will review the basic approaches and techniques for signal decomposition and source separation, especially principal component analysis, independent component analysis, and several time-frequency based methods.

3.1 Principal Component Analysis

Principal component analysis (PCA) is a classic technique in statistical data analysis, feature extraction, and data compression, stemming from the early work of Pearson [44]. Given a set of multivariate measurements, the purpose of PCA is to find a smaller set of variables with less redundancy, that would give as good a representation as possible. The redundancy is measured by correlations between data elements. Using the correlations as in PCA has the advantage that the analysis can be based on the second-order statistics only.

3.1.1 Principal Components

The starting point of PCA is a n -dimensional random vector \mathbf{x} . There is an available sample $\mathbf{x}(1), \dots, \mathbf{x}(T)$ from this random vector. No explicit assumptions on the probability density of the vectors are made in PCA, as long as the first- and second-order statistics are known or can be estimated from the sample. Also, no generative model is assumed for vector \mathbf{x} . Typically the elements of \mathbf{x} are measurements like pixel gray levels or values of a signal at different time instants. It is essential in

PCA that the elements are mutually correlated, and there is thus some redundancy in \mathbf{x} , making compression possible. If the elements are independent, the resulting components are exactly the same as the original signal measurements.

In the PCA transform, the vector \mathbf{x} is first centered by subtracting its mean $E\{\mathbf{x}\}$. The mean is in practice estimated from the available sample. Let us assume in the following that the centering has been done and thus $E\{\mathbf{x}\} = 0$. Next, \mathbf{x} is linearly transformed to another vector \mathbf{y} with m elements, $m < n$, so that the redundancy induced by the correlations is removed. This is done by finding a rotated orthogonal coordinate system such that the elements of \mathbf{x} in the new coordinates become uncorrelated. At the same time, the variance of projections of \mathbf{x} on the new coordinate axes are maximized so that the first axis corresponds to the maximal variance, the second axis corresponds to the maximal variance in the direction orthogonal to the first axis, and so on.

3.1.2 PCA By Variance Maximization

In mathematical terms, consider a linear combination

$$y_1 = \sum_{k=1}^n w_{k1} x_k = \mathbf{w}_1^T \mathbf{x} \quad (3.1)$$

of the elements x_1, \dots, x_n of the vector \mathbf{x} . The w_{11}, \dots, w_{n1} are scalar coefficients or weights, elements of an n -dimensional vector \mathbf{w}_1 , and \mathbf{w}_1^T denotes the transpose of \mathbf{w}_1 .

The factor y_1 is called the first principal component of \mathbf{x} , if the variance of y_1 is maximally large. Because the variance depends on both the norm and orientation of the weight vector \mathbf{w}_1 and grows without limits as the norm grows, we impose the constraint that the norm of \mathbf{w}_1 is constant, in practice equal to 1. Thus we look for

a weight vector \mathbf{w}_1 maximizing the PCA criterion

$$J_{PCA}^1 = E\{y_1^2\} = E\{(\mathbf{w}_1^T \mathbf{x})^2\} = \mathbf{w}_1^T E\{\mathbf{x}\mathbf{x}^T\} \mathbf{w}_1 = \mathbf{w}_1^T \mathbf{C}_\mathbf{x} \mathbf{w}_1 \quad \text{such that } \|\mathbf{w}_1\| = 1, \quad (3.2)$$

where the norm of \mathbf{w}_1 is the usual Euclidean norm defined as

$$\|\mathbf{w}_1\| = (\mathbf{w}_1^T \mathbf{w}_1)^{1/2} = \left(\sum_{k=1}^n w_{k1}^2 \right)^{1/2}, \quad (3.3)$$

and the matrix $\mathbf{C}_\mathbf{x} = E\{\mathbf{x}\mathbf{x}^T\}$ is the $n \times n$ covariance matrix of the zero-mean vector \mathbf{x} . It is well known from basic linear algebra [45,46] that the solution to PCA problem is given in terms of the unit-length eigenvectors $\mathbf{e}_1, \dots, \mathbf{e}_n$ of the matrix $\mathbf{C}_\mathbf{x}$. The ordering of the eigenvectors is such that the corresponding eigenvalues d_1, \dots, d_n satisfy $d_1 \geq d_2 \geq \dots \geq d_n$. The solution maximizing equation (3.2) is given by $\mathbf{w}_1 = \mathbf{e}_1$. Thus the first principal component of \mathbf{x} is

$$y_1 = \mathbf{e}_1^T \mathbf{x}. \quad (3.4)$$

The criterion J_{PCA}^1 in equation (3.2) can be generalized to m principal components, with m any number between 1 and n . Denoting the m -th ($1 \leq m \leq n$) principal component by $y_m = \mathbf{w}_m^T \mathbf{x}$, with \mathbf{w}_m the corresponding unit norm weight vector, the variance of y_m is now maximized under the constraint that y_m is uncorrelated with all the previously found principal components:

$$E\{y_m y_k\} = 0, \quad k < m. \quad (3.5)$$

Note that the principal components y_m have zero means because

$$E\{y_m\} = \mathbf{w}_m^T E\{\mathbf{x}\} = 0. \quad (3.6)$$

The condition (3.5) yields:

$$E\{y_m y_k\} = E\{(\mathbf{w}_m^T \mathbf{x})(\mathbf{w}_k^T \mathbf{x})\} = \mathbf{w}_m^T \mathbf{C}_\mathbf{x} \mathbf{w}_k = 0. \quad (3.7)$$

For the second principal component, we have the condition that

$$\mathbf{w}_2^T \mathbf{C}_\mathbf{x} \mathbf{w}_1 = d_1 \mathbf{w}_2^T \mathbf{e}_1 = 0, \quad (3.8)$$

because we already know that $\mathbf{w}_1 = \mathbf{e}_1$. We are thus looking for maximal variance $E\{y_2^2\} = E\{(\mathbf{w}_2^T \mathbf{x})^2\}$ in the subspace orthogonal to the first eigenvector of $\mathbf{C}_\mathbf{x}$. The solution is given by $\mathbf{w}_2 = \mathbf{e}_2$. Likewise, recursively it follows that $\mathbf{w}_k = \mathbf{e}_k$. Thus, the k th principal component is

$$y_k = \mathbf{e}_k^T \mathbf{x}. \quad (3.9)$$

From the above result, it follows that

$$E\{y_m^2\} = E\{\mathbf{e}_m^T \mathbf{x} \mathbf{x}^T \mathbf{e}_m\} = \mathbf{e}_m^T \mathbf{C}_\mathbf{x} \mathbf{e}_m = d_m, \quad (3.10)$$

which shows that the variances of the principal components are directly given by the eigenvalues of $\mathbf{C}_\mathbf{x}$. The vectors \mathbf{x} in the original data set can be approximated by the truncated PCA expansion

$$\hat{\mathbf{x}} = \sum_{i=1}^m y_i \mathbf{e}_i. \quad (3.11)$$

Then we have that the mean-squared error $E\{\|\mathbf{x} - \hat{\mathbf{x}}\|^2\}$ is equal to $\sum_{i=m+1}^n d_i$.

As the eigenvalues are all positive, the error decreases when more and more terms

are included in equation (3.11), until the error becomes zero when $m = n$ or all the principal components are included. A very important practical problem is how to choose m in equation (3.11); this is a trade-off between error and the amount of data needed for the expansion. Sometimes a rather small number of principal components are sufficient. The disciplined approaches to this problem are given by [47, 48].

3.2 Independent Component Analysis

Independent component analysis (ICA), introduced by J. Héroult, C. Jutten, and B. Ans [49] in the early 1980s, is a statistical and computational technique for revealing hidden factors that underlie sets of random variables, measurements, or signals. ICA defines a generative model for the observed multivariate data, which is typically given as a large database of samples. In the model, the data variables are assumed to be linear mixtures of some unknown latent variables, and the mixing system is also unknown. The latent variables are assumed nongaussian and mutually independent, and they are called the independent components of the observed data. These independent components, also called sources or factors, can be found by ICA.

ICA is a much more powerful technique and capable of finding the underlying factors or sources when the classic methods like PCA fail completely. The data analyzed by ICA could originate from many different kinds of application fields, including digital images and document databases, as well as economic indicators and psychometric measurements. In many cases, the measurements are given as a set of parallel signals or time series; the term blind source separation is used to characterize this problem. Typical examples are mixtures of simultaneous speech signals that have been picked up by several microphones [50], brain waves recorded by multiple sensors [51], interfering radio signals arriving at a mobile phone [52], or parallel time series obtained from some industrial process [53].

3.2.1 Definition of ICA

There are n observed random variables z_1, \dots, z_n , which are modeled as linear combinations of n random variables s_1, \dots, s_n :

$$z_i = a_{i1}s_1 + a_{i2}s_2 + \dots + a_{in}s_n, \quad \text{for all } i = 1, \dots, n \quad (3.12)$$

where the s_i are unknown and statistically mutually independent, the $a_{ij}, i, j = 1, \dots, n$ are some unknown real coefficients. This is the basic ICA model. All what are observed are the random variables z_i , and both the mixing coefficients a_{ij} and the independent components s_i must be estimated using the z_i .

It is usually more convenient to use vector-matrix notation instead of the sums as in the previous equation. Let us denote by \mathbf{z} the random vector whose elements are the mixtures z_1, \dots, z_n , and likewise by \mathbf{s} the random vector whose elements are the source signals s_1, \dots, s_n . Let us denote by \mathbf{A} the matrix with elements a_{ij} . All vectors are assumed to be column vectors. Using this vector-matrix notation, the mixing model is written as

$$\mathbf{z} = \mathbf{A}\mathbf{s}. \quad (3.13)$$

Sometimes the columns of matrix \mathbf{A} , denoted by \mathbf{a}_i , are needed, and the model can also be written as

$$\mathbf{z} = \sum_{i=1}^n \mathbf{a}_i s_i. \quad (3.14)$$

Compared with PCA, it is easy to see that in the ICA model the following ambiguities or indeterminacies will hold:

1. The variance of the independent components can not be determined.

The reason is that, both \mathbf{s} and \mathbf{A} being unknown, any scalar multiplier in one of the sources s_i could always be cancelled by dividing the corresponding column

\mathbf{a}_i of \mathbf{A} by the same scalar, say α_i :

$$\mathbf{z} = \sum_i \left(\frac{1}{\alpha_i} \mathbf{a}_i \right) (\alpha_i s_i). \quad (3.15)$$

As a consequence, the magnitudes of the independent components may be fixed as well. Since they are random variables, the most natural way to do this is to assume that each source has unit variance, $E\{s_i^2\} = 1$. Then the matrix \mathbf{A} will be adapted in the ICA solution methods to take this restriction into account.

2. The order of the independent components can not be determined.

The reason is that, again both \mathbf{s} and \mathbf{A} being unknown, the order of the terms in the sum in equation (3.14) can be freely changed, and any of the independent components can be called the first one. Formally, a permutation matrix \mathbf{P} and its inverse can be substituted in the model to give $\mathbf{z} = \mathbf{A}\mathbf{P}^{-1}\mathbf{P}\mathbf{s}$. The elements of $\mathbf{P}\mathbf{s}$ are the original independent variables s_i , but in another order. The matrix $\mathbf{A}\mathbf{P}^{-1}$ is just a new unknown mixing matrix, to be solved by the ICA algorithms.

3.2.2 ICA by Maximum Likelihood Estimation

A very popular approach for estimating the ICA model is maximum likelihood (ML) estimation. Maximum likelihood estimation is a fundamental method of statistical estimation. One interpretation of ML estimation is that those parameter values, which give the highest probability for the observations, are taken as estimates.

According to the properties of the density of a linear transform, the density p_z of the mixture vector $\mathbf{z} = \mathbf{A}\mathbf{s}$ can be formulated as

$$p_z(\mathbf{z}) = |\det \mathbf{B}| p_s(\mathbf{s}) = |\det \mathbf{B}| \prod_i p_i(s_i), \quad (3.16)$$

where $\mathbf{B} = \mathbf{A}^{-1}$, and the p_i denote the densities of the independent components. This can be expressed as a function of $\mathbf{B} = (\mathbf{b}_1, \dots, \mathbf{b}_n)^T$ and \mathbf{z} , giving

$$p_z(\mathbf{z}) = |\det \mathbf{B}| \prod_i p_i(\mathbf{b}_i^T \mathbf{z}). \quad (3.17)$$

Assume that we have K observations of \mathbf{z} , denoted by $\mathbf{z}(1), \dots, \mathbf{z}(K)$. Then the likelihood can be obtained as the product of this density evaluated at the K points. This is denoted by L and considered as a function of \mathbf{B} :

$$L(\mathbf{B}) = \prod_{t=1}^K \prod_{i=1}^n p_i(\mathbf{b}_i^T \mathbf{z}(t)) |\det \mathbf{B}|. \quad (3.18)$$

Very often it is more practical to use the logarithm of the likelihood, since it is algebraically simpler. This does not make any difference here since the maximum of the logarithm is obtained at the same point as the maximum of the likelihood. The log-likelihood is given by

$$\log L(\mathbf{B}) = \sum_{t=1}^K \sum_{i=1}^n \log p_i(\mathbf{b}_i^T \mathbf{z}(t)) + K \log |\det \mathbf{B}|. \quad (3.19)$$

Divide the likelihood by K to obtain

$$\frac{1}{K} \log L(\mathbf{B}) = \mathbb{E} \left\{ \sum_{i=1}^n \log p_i(\mathbf{b}_i^T \mathbf{z}) \right\} + \log |\det \mathbf{B}|. \quad (3.20)$$

To perform ML estimation in practice, an algorithm is needed to perform the numerical maximization of likelihood. In fact, there are many different methods, among which the simplest algorithms for maximizing likelihood are obtained by gradient methods [13].

3.2.3 ICA by Minimization of Mutual Information

An important approach for ICA estimation, inspired by information theory, is minimization of mutual information. The motivation of this approach is that it may not be very realistic in many cases to assume that the data follows the ICA model. Therefore, an approach that does not assume anything about the data needs to be developed. The goal is to have a general-purpose measure of the dependence of the components of a random vector. With such a measure, ICA could be defined as a linear decomposition that minimizes that dependence measure. Such an approach can be developed using mutual information, which is an information-theoretic measure of statistical dependence. One of the main utilities of mutual information is that it serves as a unifying framework for many estimation principles, in particular ML estimation.

Mutual information I between n random variables y_i , $i = 1, \dots, n$ is defined as follows

$$I(y_1, \dots, y_n) = \sum_{i=1}^n H(y_i) - H(\mathbf{y}), \quad (3.21)$$

where $H(y_i)$ and $H(\mathbf{y})$ are y_i 's entropy and joint entropy, respectively. Mutual information is a natural measure of the dependence between random variables. It is always nonnegative, and zero if and only if the variables are statistically independent. Mutual information takes into account the whole dependence structure of the variables, and not just the covariance, like PCA and related methods. Therefore, mutual information can be used as the criterion for finding the ICA representation. This approach is an alternative to the model estimation approach. The ICA of a random vector \mathbf{z} is defined as an invertible transformation:

$$\mathbf{s} = \mathbf{B}\mathbf{z}, \quad (3.22)$$

where the matrix \mathbf{B} is determined so that the mutual information of the transformed component s_i is minimized. If the data follows the ICA model, this allows estimation of the data model. On the other hand, in this definition, it is not needed to assume that the data follows the model. In any case, minimization of mutual information can be interpreted as giving the maximally independent components.

Mutual information and likelihood are intimately connected. A detailed analysis of the connection between mutual information and maximum likelihood can be seen in [10]. The same gradient algorithm can be used to optimize mutual information due to its connection with likelihood. In addition, a nonparametric algorithm for minimization of mutual information is proposed in [54], and an approach based on order statistics is proposed in [55].

3.3 Review of Time-Frequency Signal Decomposition and Separation Approaches

The most common methods for component extraction including PCA and ICA are effective at extracting orthogonal or independent components and assume the stationarity of the underlying signals. Since most real life signals are not stationary and thus do not obey this underlying assumption, recent research has focused on source/component extraction in the joint time-frequency domain. In this section, we review signal decomposition and source separation approaches based on the time-frequency distributions.

3.3.1 Matching Pursuit with Time-Frequency Dictionaries

Matching pursuit [5] is a method to decompose a signal into a linear expansion of waveforms which belong to a redundant dictionary of functions, and whose time-frequency properties are adapted to the local structures of the signal. These waveforms are called time-frequency atoms. This algorithm offers a decomposition particularly important for representing signal components whose localizations in time and frequency vary

widely.

A general family of time-frequency atoms can be generated by scaling, translating and modulating a single window function $g(t) \in \mathbf{L}^2(\mathbf{R})$, and is defined as

$$g_\gamma(t) = \frac{1}{\sqrt{s}} g\left(\frac{t-u}{s}\right) e^{j\xi t}, \quad (3.23)$$

where $s > 0, u, \xi$ are the parameters of the scale, translation, and frequency modulating, respectively, and $\gamma = (s, u, \xi) \in \mathbf{\Gamma} = \mathbf{R}^+ \times \mathbf{R}^2$. The factor $\frac{1}{\sqrt{s}}$ normalizes the norm of $g_\gamma(t)$ to 1. The family $D = (g_\gamma(t))_{\gamma \in \mathbf{\Gamma}}$ is extremely redundant, and its properties have been studied in [56]. A linear expansion of a signal $f(t)$ over a set of vectors selected from D can be done by successive approximations of $f(t)$ with orthogonal projections on elements of D , in order to best match its inner structures. Let $g_{\gamma_0} \in D$. The signal f is decomposed into

$$f = \langle f, g_{\gamma_0} \rangle g_{\gamma_0} + R^1 f, \quad (3.24)$$

where $\langle \cdot, \cdot \rangle$ represents the inner product of two functions, and $R^1 f$ is the residue after approximating f in the direction of g_{γ_0} . Clearly, g_{γ_0} is orthogonal to $R^1 f$, hence

$$\|f\|^2 = |\langle f, g_{\gamma_0} \rangle|^2 + \|R^1 f\|^2. \quad (3.25)$$

To minimize $\|R^1 f\|$, g_{γ_0} is chosen from D such that $|\langle f, g_{\gamma_0} \rangle|$ is maximum. After m iterations, the signal f is decomposed into

$$f = \sum_{n=0}^{m-1} \langle R^n f, g_{\gamma_n} \rangle g_{\gamma_n} + R^m f, \quad (3.26)$$

and an energy conservation equation is yielded as

$$\|f\|^2 = \sum_{n=0}^{m-1} |\langle R^n f, g_{\gamma_n} \rangle|^2 + \|R^m f\|^2. \quad (3.27)$$

It is proven that the matching pursuit algorithm is convergent with respect to the iteration number m , so as $m \rightarrow \infty$,

$$f = \sum_{n=0}^{\infty} \langle R^n f, g_{\gamma_n} \rangle g_{\gamma_n}, \quad (3.28)$$

and

$$\|f\|^2 = \sum_{n=0}^{\infty} |\langle R^n f, g_{\gamma_n} \rangle|^2. \quad (3.29)$$

It is thus shown that any signal $f(t) \in \mathbf{L}^2(\mathbf{R})$ can be decomposed into a sum of complex time-frequency atoms that best match its residues by matching pursuit.

Although matching pursuit gives a flexible signal decomposition, a problem with this method is the restricted number of waveforms in the dictionary. While dictionaries containing a wide variety of elements can be employed at the expense of high computational cost, the representations are not satisfactory unless all signal components are at least reasonably well approximated by dictionary elements.

A modified matching pursuit algorithm called Orthogonal Matching Pursuit (OMP) is developed in [57]. For nonorthogonal dictionaries, OMP in general converges faster than matching pursuit. Furthermore for any finite size dictionary of N elements, OMP converges to the projection onto the span of the dictionary elements in no more than N steps. OMP is simply described as follows: assume that after m iterations, the signal f is decomposed into

$$f = \sum_{n=1}^m a_n^m g_{\gamma_n} + R^m f, \quad \text{with } \langle R^m f, g_{\gamma_n} \rangle = 0, \quad n = 1, 2, \dots, m. \quad (3.30)$$

It is desired that for the $(m + 1)$ th iteration, the signal f can be represented as

$$f = \sum_{n=1}^{m+1} a_n^{m+1} g_{\gamma_n} + R^{m+1} f, \quad \text{with } \langle R^{m+1} f, g_{\gamma_n} \rangle = 0, \quad n = 1, 2, \dots, m+1. \quad (3.31)$$

Since elements of the dictionary D are not required to be orthogonal, to perform such an iteration, an auxiliary model for the dependence of $g_{\gamma_{m+1}}$ on the previous g_{γ_n} 's ($n = 1, 2, \dots, m$) is required. Let

$$g_{\gamma_{m+1}} = \sum_{n=1}^m b_n^m g_{\gamma_n} + p_m, \quad \text{with } \langle p_m, g_{\gamma_n} \rangle = 0, \quad n = 1, 2, \dots, m. \quad (3.32)$$

Using the above auxiliary model, it may be shown that the correct update from the m th iteration to the $(m + 1)$ th iteration is given by

$$\begin{aligned} a_n^{m+1} &= a_n^m - \beta_m b_n^m, \quad n = 1, 2, \dots, m \\ \text{and } a_{m+1}^{m+1} &= \beta_m, \\ \text{where } \beta_m &= \frac{\langle R^m f, g_{\gamma_{m+1}} \rangle}{\langle p_m, g_{\gamma_{m+1}} \rangle} = \frac{\langle R^m f, g_{\gamma_{m+1}} \rangle}{\|p_m\|^2} \\ &= \frac{\langle R^m f, g_{\gamma_{m+1}} \rangle}{\|g_{\gamma_{m+1}}\|^2 - \sum_{n=1}^m b_n^m \langle g_{\gamma_n}, g_{\gamma_{m+1}} \rangle}. \end{aligned} \quad (3.33)$$

It also follows that the residual $R^{m+1} f$ satisfies

$$R^m f = R^{m+1} f + \beta_m p_m, \quad (3.34)$$

and

$$\|R^m f\|^2 = \|R^{m+1} f\|^2 + \frac{|\langle R^m f, g_{\gamma_{m+1}} \rangle|^2}{\|p_m\|^2}. \quad (3.35)$$

3.3.2 Spatial Time-Frequency Distribution (STFD) Method

For non-stationary signals, a blind source separation method using spatial time-frequency distributions is introduced in [58].

The multidimensional data model is

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t), \quad (3.36)$$

where $\mathbf{x}(t) = [x_1(t), \dots, x_m(t)]^T$ is a noisy instantaneous linear mixture of source signals $\mathbf{s}(t) = [s_1(t), \dots, s_n(t)]^T$, \mathbf{A} is the mixing matrix, and $\mathbf{n}(t)$ is the additive noise. The discrete-time form of the Cohen's class of TFD for signal $x_1(t)$ is given by [23]

$$D_{x_1 x_1}(t, \omega) = \sum_{l=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \psi(m, l) x_1(t + m + l) x_1^*(t + m - l) e^{-j2\omega l}, \quad (3.37)$$

where t and ω represent the time index and the frequency index, respectively. The cross-TFD of two signals $x_1(t)$ and $x_2(t)$ is defined by

$$D_{x_1 x_2}(t, \omega) = \sum_{l=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \psi(m, l) x_1(t + m + l) x_2^*(t + m - l) e^{-j2\omega l}. \quad (3.38)$$

The two equations given above are used to define the spatial time-frequency distribution (STFD) matrix as follows

$$\mathbf{D}_{\mathbf{xx}}(t, \omega) = \sum_{l=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \psi(m, l) \mathbf{x}(t + m + l) \mathbf{x}^*(t + m - l) e^{-j2\omega l}, \quad (3.39)$$

where $[\mathbf{D}_{\mathbf{xx}}(t, \omega)]_{ij} = D_{x_i x_j}(t, \omega)$, for $i, j = 1, \dots, n$.

The blind identification method is presented based on a two-step process: the first step consists of whitening the data in order to transform the mixing matrix \mathbf{A} into a unitary matrix \mathbf{U} ; the second step consists of retrieving this unitary matrix \mathbf{U} by

jointly diagonalizing a set of whitened data STFD matrices. Under the assumption that the source signals $s_i(t), 1 \leq i \leq n$ are mutually uncorrelated, the whitening matrix \mathbf{W} can be determined from the array output autocorrelation \mathbf{R}

$$\mathbf{W}(\mathbf{R} - \sigma^2 \mathbf{I})\mathbf{W}^T = \mathbf{W}\mathbf{A}\mathbf{A}^T\mathbf{W}^T = \mathbf{I}, \quad (3.40)$$

where $\mathbf{R} = \frac{1}{T} \sum_{t=1}^T \mathbf{x}(t)\mathbf{x}^*(t)$ as $T \rightarrow \infty$, \mathbf{I} is the $n \times n$ identity matrix, and σ^2 is the noise variance. In the second step, the whitened STFD matrix is obtained as

$$\mathbf{D}_{\mathbf{z}\mathbf{z}}(t, \omega) = \mathbf{W}\mathbf{D}_{\mathbf{x}\mathbf{x}}(t, \omega)\mathbf{W}^T = \mathbf{U}\mathbf{D}_{\mathbf{s}\mathbf{s}}(t, \omega)\mathbf{U}^T, \quad (3.41)$$

where $\mathbf{z}(t) = \mathbf{W}\mathbf{x}(t)$ is the whitened data vector, and $\mathbf{U} = \mathbf{W}\mathbf{A}$ is a unitary matrix. Since $\mathbf{D}_{\mathbf{s}\mathbf{s}}(t, \omega)$ is diagonal, \mathbf{U} may be obtained as a unitary diagonalizing matrix of the whitened STFD matrices $\mathbf{D}_{\mathbf{z}\mathbf{z}}(t, \omega)$ for time-frequency points corresponding to signal autoterms. In the end, the source signals are estimated as $\mathbf{s}(t) = \mathbf{U}^T\mathbf{W}\mathbf{x}(t)$.

In contrast to blind source separation methods using second-order and/or high-order statistics, the proposed approach allows the separation of Gaussian sources with identical spectral shapes but with different time-frequency localization properties. However, due to the joint diagonalization of the STFD matrix, it has a higher computational complexity.

In [59], an underdetermined separation algorithm for nondisjoint sources is proposed based on the STFD method. Source separation is achieved by combining the STFD matrix with a clustering approach.

3.3.3 Blind Separation via Time-Frequency Masking

In [60], binary time-frequency masks are created to achieve demixing provided the time-frequency representations of the sources do not overlap.

Without loss of generality, suppose $x_1(t)$ and $x_2(t)$ are two mixtures of source

signals $s_1(t), \dots, s_N(t)$ such that

$$\begin{aligned} x_1(t) &= \sum_{i=1}^N s_i(t), \\ x_2(t) &= \sum_{i=1}^N a_i s_i(t - \delta_i), \end{aligned} \tag{3.42}$$

where a_i and δ_i are the attenuation coefficients and the time delays. Using the shift-invariance of the short-time Fourier transform (STFT), the time-frequency representation of the mixing model (3.42) is

$$\begin{bmatrix} X_1(t, \omega) \\ X_2(t, \omega) \end{bmatrix} = \begin{bmatrix} 1 & \dots & 1 \\ a_1 e^{-j\omega\delta_1} & \dots & a_N e^{-j\omega\delta_N} \end{bmatrix} \begin{bmatrix} S_1(t, \omega) \\ \vdots \\ S_N(t, \omega) \end{bmatrix}. \tag{3.43}$$

It is assumed that the STFTs, $S_i(t, \omega)$ and $S_k(t, \omega)$, of any two source signals $s_i(t)$ and $s_k(t)$ are disjoint

$$S_i(t, \omega) S_k(t, \omega) = 0, \quad \forall t, \omega \quad \forall i \neq k. \tag{3.44}$$

To demix, one creates the time-frequency mask corresponding to each source and applies each mask to the mixture to produce the original source time-frequency representations. For example, defining

$$M_i(t, \omega) = \begin{cases} 1, & S_i(t, \omega) \neq 0 \\ 0, & \text{otherwise} \end{cases}, \tag{3.45}$$

one obtains the time-frequency representation of $s_i(t)$ from the mixture $X_1(t, \omega)$ via

$$S_i(t, \omega) = M_i(t, \omega) X_1(t, \omega), \quad \forall t, \omega. \tag{3.46}$$

Let $\Omega_i = \{(t, \omega) : M_i(t, \omega) = 1\}$ for any $i \in (1, \dots, N)$ so that $M_i = 1_{\Omega_i}$. Consider

$$R(t, \omega) = \frac{X_2(t, \omega)}{X_1(t, \omega)}. \quad (3.47)$$

Clearly, on Ω_i

$$R(t, \omega) = a_i e^{-j\delta_i \omega}. \quad (3.48)$$

In this case, $|R(t, \omega)| = a_i$ and $-\frac{1}{\omega} \angle R(t, \omega) = \delta_i$, where $\angle z$ denotes the phase of the complex number z taken between $-\pi$ and π . Hence, one simply labels each time-frequency point (t, ω) with the pair $(|R(t, \omega)|, -\frac{1}{\omega} \angle R(t, \omega))$. Since the sources are disjoint, there will be N distinct labels. By grouping the time-frequency points (t, ω) with the same label, the sets Ω_i are constructed, and then the masks $M_i = 1_{\Omega_i}$. Therefore, from equation (3.46), the time-frequency representations $S_i(t, \omega)$ of the original sources $s_i(t)$ can be obtained. Although this approach may separate any number of sources from their two mixtures, the problem with more than two mixtures is not addressed. In addition, it is hard to separate the source signals which have the same parameters a_i and δ_i in their mixtures.

CHAPTER 4

ADAPTIVE SIGNAL DECOMPOSITION ON THE TIME-FREQUENCY PLANE

4.1 Introduction

Signal decomposition aims to extract the components comprising the observed signals. The majority of methods for performing linear signal decomposition involve over-complete waveform dictionaries. By selecting the optimum set of available waveforms from the dictionary based on some criterion, a sparse model of the signal can be obtained. Such decomposition schemes include matching pursuit [5], basis pursuit [6], and the chirplet decomposition [7]. A problem with these decomposition methods is the restricted number of waveforms in the dictionary. While dictionaries containing a wide variety of elements can be employed at the expense of high computational cost, the representations are not satisfactory unless all signal components are at least reasonably well approximated by dictionary elements.

For this reason, in this chapter we introduce an adaptive component extraction approach on the time-frequency plane. This approach relies on extracting components that are well-concentrated on the time-frequency plane. The concentration of the components are quantified through an entropy measure on the time-frequency plane. Since it has been shown in the literature that signals that achieve a small entropy value on the time-frequency plane are Gabor logons, our component extraction algorithm reduces to extracting the Gabor logons that best describe the given data set in a minimum mean square sense. Unlike the traditional Gabor decomposition [61], where the signal is expressed as an infinite sum of the time and frequency shifted Gabor logons, we do not have to create a dictionary beforehand, and the components extracted by the proposed method have time and frequency centers determined by

the signal. Moreover, these extracted components have chirp rates and local spread adapted to the given set of signals. The goal is to represent the given data set with a few number of the chirped Gabor logons.

4.2 Background on Gabor Decomposition and Information Measures

4.2.1 Gabor Signal Expansion

In 1946, Gabor presented an approach to characterize a time function in time and frequency simultaneously, which later became known as the Gabor signal expansion [62]. He showed that any signal in L_2 could be represented as the weighted sum of modulated and shifted Gaussian functions (logons) centered on a rectangular lattice in time and frequency under the constraint that $T\Omega \leq 2\pi$ where T is the time sampling interval and Ω is the frequency sampling interval. That is, for signal $s(t)$, the Gabor expansion is defined as

$$s(t) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} a_{mn} g_{mn}(t) \quad (4.1)$$

with

$$g(t) = \sqrt[4]{2} e^{-\pi(t/T)^2} \quad (4.2)$$

$$g_{mn}(t) = g(t - mT) e^{jn\Omega t}. \quad (4.3)$$

The Gabor expansion coefficients a_{mn} are computed by the usual inner product rule for projecting $s(t)$ onto an auxiliary function $\gamma(t)$, i.e.,

$$a_{mn} = \int_{-\infty}^{\infty} s(t) \gamma_{mn}^*(t) dt \quad (4.4)$$

$$\gamma_{mn}(t) = \gamma(t - mT) e^{jn\Omega t}, \quad (4.5)$$

where $*$ denotes the complex conjugate operation. Equations (4.4) and (4.5) are in fact a sampled version of the windowed Fourier transform of the signal $s(t)$ with the analysis window $\gamma(t)$, which is known as the Gabor transform. The analysis window and the synthesis window satisfy the following biorthogonality relationship [63]:

$$\frac{T_0\Omega_0}{2\pi} \int_{-\infty}^{\infty} g(t)\gamma^*(t - mT_0)e^{-jn\Omega_0 t} dt = \delta(m)\delta(n), \quad (4.6)$$

where $T_0 = 2\pi/\Omega$, and $\Omega_0 = 2\pi/T$.

4.2.2 Chirplet Transform

The Gabor transform essentially provides expansions of signals as linear combinations of time-frequency atoms with fixed time and frequency “concentration” properties. However, it fails to represent the chirp-like components in a compact and precise way. In other words, more atoms are needed to approximate the chirp-like components with frequency modulation, which results in the reduction of the effectiveness and compactness of the time-frequency representation.

For these reasons, the chirplet transform, a generalized Gabor transform, is developed [64]. The time-frequency atoms for a Gaussian chirplet transform, the so-called Gaussian chirplets, are derived from a single Gaussian function through the operations of scaling, chirping, time- and frequency-shifting, which leads to a family of wave packets with four adjustable parameters:

$$g_k(t) = \sqrt[4]{\frac{\alpha_k}{\pi}} \exp \left\{ -\frac{\alpha_k}{2}(t - t_k)^2 + j \left[\omega_k + \frac{\beta_k}{2}(t - t_k) \right] (t - t_k) \right\}, \quad (4.7)$$

where the parameters (t_k, ω_k) determine the time and frequency centers of the linear Gaussian chirplets; the variance $\alpha_k (> 0)$ controls the time duration of the chirplet; β_k is the frequency modulation rate (chirp rate) that characterizes the “quickness” of frequency changes. Compared with the Gabor logon used for the Gabor expansion,

the Gaussian chirplet has more freedom and thereby can better match the signal under consideration.

4.2.3 Entropy Measure on the Time-Frequency Plane

Since a TFD, $C(t, \omega)$, from Cohen's general class has many desired properties such as the energy preservation and the marginals, it is analogous to the probability density function (pdf) of a two-dimensional random variable. This analogy has inspired the adaptation of information-theoretic measures such as entropy and mutual information to the time-frequency plane. The adaptation of classical Shannon entropy to the time-frequency plane yields

$$H(C) = - \iint C(t, \omega) \log_2 C(t, \omega) dt d\omega. \quad (4.8)$$

This measure is only defined when $C(t, \omega) > 0, \forall t, \omega$. Therefore, it is valid for positive distributions such as the spectrogram, but yet invalid for non-positive distributions. For this reason, a more generalized class of entropy measures known as Rényi entropy has been adapted to the time-frequency plane. In [42], Rényi entropy was introduced as an alternative way of measuring the complexity of TFDs and the properties of this measure were proved extensively in [65]:

$$H_\alpha(C) = \frac{1}{1-\alpha} \log_2 \iint \left(\frac{C(t, \omega)}{\iint C(u, v) du dv} \right)^\alpha dt d\omega. \quad (4.9)$$

where $\alpha > 0$. This measure is well-defined as long as $\iint C(t, \omega) dt d\omega > 0$ and has been shown to be finite for a large class of signals and distributions [65]. It is important to note that as $\alpha \rightarrow 1$, Rényi entropy becomes Shannon entropy.

It has been shown that the minimum value of entropy on the time-frequency plane is achieved for a Gabor logon [65]. This is also consistent with the fact that

the Gabor logon is the signal that achieves the lower bound on the uncertainty on the time-frequency plane [23]. For this reason, our signal decomposition algorithm is based on extracting a set of well-concentrated components, that best describe the given data set.

4.3 Component Extraction Method

4.3.1 Problem Statement

Given M measurements of a signal, $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M\}$, we want to extract the first L components, $L < M$, that minimize entropy on the time-frequency plane. Each measurement, \mathbf{x}_i , is transformed to the time-frequency plane as:

$$C_i(n, \omega; \psi) = \sum_m \sum_l \psi(n-l, m) x_i \left(l + \frac{m}{2}\right) x_i^* \left(l - \frac{m}{2}\right) e^{-j\omega m}. \quad (4.10)$$

The time-frequency distribution corresponding to each trial is vectorized and a matrix of time-frequency distributions is formed:

$$\mathbf{C} = \begin{bmatrix} C_1 \\ C_2 \\ \vdots \\ C_M \end{bmatrix}, \quad (4.11)$$

where C_i is a vector of length $N \times K$ points, N and K being the number of time and frequency points, respectively. The components on the time-frequency plane are found based on this time-frequency data matrix. Each component S_i is a linear combination of the rows of this matrix, i.e.

$$S_i(n, k) = \sum_{j=1}^M a_j C_j(n, k), \quad (4.12)$$

where $\sqrt{\sum_j a_j^2} = 1$ and a_j 's are chosen such that $H_\alpha(S_i)$ is minimized on the time-frequency plane.

4.3.2 The Proposed Approach

Since Gabor logon signals have minimum entropy in the time-frequency domain, the cost function is chosen as $e = H_\alpha(S_i) - H_\alpha^*$, where $H_\alpha(S_i)$ is Rényi entropy of i th component with order α , and H_α^* represents Rényi entropy of the corresponding desired Gabor logon signal. The weight vector $\mathbf{a} = [a_1, a_2, \dots, a_M]^T$ is updated using the method of Steepest Descent [66], which is

$$\hat{\mathbf{a}} = \mathbf{a} - \mu \frac{\partial e}{\partial \mathbf{a}}, \quad (4.13)$$

where μ is the step size parameter. In the discrete case, Rényi entropy of the component S_i is

$$H_\alpha(S_i) = H_\alpha(\mathbf{a}^T C) = \frac{1}{1-\alpha} \log_2 \sum_n \sum_k \left(\sum_{j=1}^N a_j C_j(n, k) \right)^\alpha, \quad (4.14)$$

where S_i and C_j are normalized. The gradient of the cost function e with respect to the l th weight coefficient a_l is derived as:

$$\frac{\partial e}{\partial a_l} = \frac{\alpha}{1-\alpha} \frac{\sum_n \sum_k (S_i(n, k))^{\alpha-1} C_l(n, k)}{\sum_n \sum_k (S_i(n, k))^\alpha}, \quad (4.15)$$

where $l = 1, \dots, M$. For the special case of $\alpha = 2$,

$$\frac{\partial e}{\partial a_l} = -2 \frac{\sum_n \sum_k S_i(n, k) C_l(n, k)}{\sum_n \sum_k (S_i(n, k))^2}. \quad (4.16)$$

Substituting the results in equation (4.16) into equation (4.13) yields the update equation for \mathbf{a} as:

$$\hat{a}_l = a_l + 2\mu \frac{\sum_n \sum_k S_i(n, k) C_l(n, k)}{\sum_n \sum_k (S_i(n, k))^2}. \quad (4.17)$$

The algorithm can be summarized as follows:

1. Find the Gabor logon that best describes the average of all trials, $C_{av} = \frac{1}{M} \sum_j C_j$. This first Gabor logon is found by finding the average time duration, average frequency, the spread, and the chirp rate of C_{av} . A logon with these estimated parameters is constructed and chosen as the first desired signal, $G(n, k; n_0, k_0, \sigma, \beta)$.
2. Set the initial value for $a_j = \frac{1}{\sqrt{N}}$, and use the adaptive filtering algorithm to update the weights until the error converges. Here, when the absolute value of the difference of two neighboring weights is less than a given error value, the update is stopped. The first component is then determined as, $S_1 = \mathbf{a}_*^T \mathbf{C}$, where \mathbf{a}_* is the optimal weighting vector.
3. Project all the trials on S_1 and compute the residue.

$$\hat{C}_i = C_i - \langle S_1, C_i \rangle C_i, \quad i = 1, 2, \dots, M \quad (4.18)$$

4. Repeat the same algorithm on this residue matrix $\hat{\mathbf{C}}$, and extract the next component.
5. Stop when the average energy of the residues drops below a pre-determined threshold value.

4.3.3 Convergence Analysis of the Algorithm

An important issue in adaptation is the convergence of the algorithm. We investigate the convergence of the proposed entropy adaptation algorithm, whose weight update

is given in equation (4.16), for the special case of entropy error minimization with order $\alpha = 2$ in the linear filter $S = \mathbf{a}^T \mathbf{C}$.

Rényi entropy of the extracted component S is written in the matrix-vector format for $\alpha = 2$ as follows

$$H_2(S) = H_2(\mathbf{a}^T \mathbf{C}) = -\log_2(SS^T) = -\log_2(\mathbf{a}^T \mathbf{C} \mathbf{C}^T \mathbf{a}). \quad (4.19)$$

The weight vector \mathbf{a} at step $(k + 1)$ is updated by

$$\mathbf{a}_{k+1} = \mathbf{a}_k - \mu \Delta \mathbf{a}_k, \quad (4.20)$$

where

$$\Delta \mathbf{a}_k = \frac{\partial e}{\partial \mathbf{a}_k} = \frac{\partial H_2(S)}{\partial \mathbf{a}_k} = -\frac{2(\mathbf{C} \mathbf{C}^T) \mathbf{a}_k}{\mathbf{a}_k^T (\mathbf{C} \mathbf{C}^T) \mathbf{a}_k}. \quad (4.21)$$

Assume that the desired Gabor logon is $G = \mathbf{a}_*^T \mathbf{C}$ with the optimal weight vector \mathbf{a}_* . Consider the weight error vector $\varepsilon_k = \mathbf{a}_k - \mathbf{a}_*$. Subtracting \mathbf{a}_* from both sides of equation (4.20), we get

$$\varepsilon_{k+1} = \varepsilon_k - \mu \Delta \mathbf{a}_k. \quad (4.22)$$

Multiplying both sides of above equation with its transpose to get the norm of the weight error yields

$$\|\varepsilon_{k+1}\|^2 = \|\varepsilon_k\|^2 - 2\mu \varepsilon_k^T \Delta \mathbf{a}_k + \mu^2 \|\Delta \mathbf{a}_k\|^2. \quad (4.23)$$

In order for the weights to converge to the true weights, we require

$$\|\varepsilon_{k+1}\|^2 < \|\varepsilon_k\|^2, \quad (4.24)$$

which is guaranteed when the step size satisfies the inequality

$$0 < \mu < \frac{2\varepsilon_k^T \Delta \mathbf{a}_k}{\|\Delta \mathbf{a}_k\|^2}. \quad (4.25)$$

Since

$$\begin{aligned} \varepsilon_k^T \Delta \mathbf{a}_k &= (\mathbf{a}_* - \mathbf{a}_k)^T \left(-\frac{2(\mathbf{C}\mathbf{C}^T)\mathbf{a}_k}{\mathbf{a}_k^T(\mathbf{C}\mathbf{C}^T)\mathbf{a}_k} \right) \\ &= 2 \left(1 - \frac{\mathbf{a}_*^T(\mathbf{C}\mathbf{C}^T)\mathbf{a}_k}{\mathbf{a}_k^T(\mathbf{C}\mathbf{C}^T)\mathbf{a}_k} \right) \\ &= 2 \left(1 - \frac{(\mathbf{a}_*^T \mathbf{C})(\mathbf{a}_k^T \mathbf{C})^T}{(\mathbf{a}_k^T \mathbf{C})(\mathbf{a}_k^T \mathbf{C})^T} \right) \\ &= 2 \left(1 - \frac{\langle G, S_k \rangle}{\langle S_k, S_k \rangle} \right), \end{aligned} \quad (4.26)$$

where \langle, \rangle represents the inner product of two vectors, S_k is the extracted component at step k , and G is the corresponding desired Gabor logon, the upper bound on the positive step size becomes

$$0 < \mu < \frac{4}{\|\Delta \mathbf{a}_k\|^2} \left(1 - \frac{\langle G, S_k \rangle}{\langle S_k, S_k \rangle} \right). \quad (4.27)$$

Notice that since $\langle G, S_k \rangle$ is less than $\langle S_k, S_k \rangle$ for the normalized TFDs S_k and G , the upper bound on the step size is positive and valid. It can be concluded that the proposed adaptation algorithm is convergent.

4.4 Experimental Results and Analysis

In order to evaluate the effectiveness of our method, we consider the following example. The set of observed signals are linear combinations of two Gabor logons and a chirp signal, i.e. $x_i = w_{i1}s_1 + w_{i2}s_2 + w_{i3}s_3$, where w_{i1}, w_{i2}, w_{i3} are the weights for each signal and are distributed as $N(0, 1)$. The first Gabor logon is centered at

the time sample point 50 and normalized frequency of 0.7, the second Gabor logon is centered at time sample point 150 and normalized frequency of -0.7. The linear chirp signal has an initial normalized frequency of -0.2 and its instantaneous frequency increases to a normalized frequency of 0.2. Rényi entropy with $\alpha = 2$ is used as the cost function to ensure that entropy is well-defined. The data set consists of $M = 128$ linear combinations of these three signals. Each signal is transformed to the time-frequency domain with $N = 50$ time samples and $K = 64$ frequency samples. Each TFD is then vectorized to form a TFD matrix of size 128×3200 .

First, the average of M TFDs corresponding to each trial is computed. Then, the time-frequency location of the peak energy on the time-frequency plane is found as n_0 and k_0 . A window centered at (n_0, k_0) is constructed to determine a local region around this peak. The size of the window is determined based on the energy distribution of the signal, i.e. the window is expanded until the energy value drops below 10% of the peak value. This windowing approach around the peak helps us extract local features. The same window is applied to all trials to extract the corresponding regions in each trial. The standard deviation σ and gradient (the chirp rate), β , of this local TFD are estimated. Based on the parameters $(n_0, k_0, \sigma, \beta)$, a Gabor logon is constructed and chosen as the first desired signal. Using the steepest descent algorithm, the weight coefficients a_j 's are updated to minimize the difference of Rényi entropy between the linear combination of the M local TFDs and the TFD of the first desired logon to obtain the first time-frequency component, S_1 . This first component is projected onto all of the M trials and the residue is found. This same algorithm is repeated for the residue on the time-frequency plane, i.e. pick the peak, construct a window, determine the desired Gabor logon, and adaptively filter the signals to get close to the desired Gabor logon. This process is repeated until the energy of the residue is below a certain threshold. In this example, 11 components were enough to represent 90% of the total energy of the signal.

Table 4.1 gives the entropy values for the first 3 of the 11 extracted components, the corresponding desired logon signals, and the first 3 components obtained using PCA. It is shown in Table 4.1 that the entropy of the extracted components are closer to the entropy of the Gabor logons. Since the entropy differences between these extracted components and the desired logon signals are small, we can infer that the extracted signals are quite close to the actual logons. It is also seen that the entropy of components extracted by our method is less than the entropy of PCA components. This indicates that we obtain components that are more compact than the ones obtained by PCA.

The time-frequency surfaces in Figure 4.1 indicate that the 5 extracted components include both the logon signals and the first three chirped logons that represent the linear chirp signal. The topographical plots of the extracted components make it clear that each component was appropriately isolated in terms of the topographical region of origin.

The results of this example show that the decomposition of time-frequency energy using our approach can extract meaningful time-frequency components for analysis of large sets of data. This decomposition algorithm achieves several goals. First, time-frequency data reduction is accomplished by producing a few meaningful components on the time-frequency plane that explain most of the signal's energy. A second benefit of this time-frequency domain decomposition is that it can extract activity that overlaps in time, but not in frequency, which is not possible using time domain decomposition approaches. Finally, another benefit of our method is the ability to separate and extract parts of chirped signals, which cannot be achieved using the conventional Gabor expansion.

Next, the performance of the proposed approach is compared with that of Orthogonal Matching Pursuit (OMP) introduced in Chapter 3. In this example, in order to represent the same 90% of the total energy of the signal as in the proposed method,

Table 4.1. Entropy Comparison

Entropy	Decomp Comps	PCA Comps	Desired Logons
1	2.8319	4.5011	2.7809
2	2.7825	3.0461	2.7413
3	2.7517	2.9724	2.7252

24 dictionary elements for OMP are needed, among which the six ones are shown in Figure 4.2. It is indicated that the number of components required by OMP is much larger than that of the proposed approach which is only eleven. Moreover, the computation of OMP is much more complex, about four times of the proposed method. One reason why the performance of matching pursuit is not well is that the decomposition is not satisfactory with matching pursuit unless all signal components are well approximated by dictionary elements; on the other hand, although the dictionary contains a wide variety of elements, this kind of redundancy leads to the elements to be employed at the expense of high computational cost.

4.5 Summary

In this chapter, a new signal decomposition method on the time-frequency plane is proposed based on the minimum entropy criterion. The major difference of the proposed approach from conventional component extraction or decomposition methods is the cost function. The cost function which is minimized is entropy on the time-frequency plane, thus producing compact components that are similar to Gabor logons. Using entropy as the cost function and adopting an adaptive filtering method to update the weights corresponding to each trial, we extract “minimum” entropy components orthogonal to each other. Experimental results show that the presented approach is effective in determining a few number of components that can be used to represent a large set of data.



Figure 4.1. The average time-frequency distribution of 128 trials and the 5 extracted components of the proposed method



Figure 4.2. The average time-frequency distribution of 128 trials and the 6 elements of OMP

CHAPTER 5

OVERDETERMINED BLIND SOURCE SEPARATION IN THE TIME-FREQUENCY DOMAIN

5.1 Introduction

Blind source separation (BSS) is an important and fundamental problem in signal processing with a broad range of applications. Several unobservable source signals first pass through an intermediate media, and then arrive at an array of sensors. The observed output of each sensor is a mixture of all the source signals. The goal in BSS is to recover the original source signals from the observed mixtures. Typical BSS applications include communications [1], speech signal processing [2], and biomedical signal processing applications [3]. A number of BSS algorithms have been proposed based on the instantaneous mixture model, in which the observed signals are linear combinations of the source signals and no time delays are involved in the mixtures. Among these methods, the most common ones are second order statistics based methods [67], and information-theoretic approaches which utilize cost functions such as mutual information or divergence measures, e.g. independent component analysis (ICA) [9, 10, 68–70], sparse component analysis (SCA) [71], and nonnegative matrix factorization (NMF) [72]. These methods in general assume a certain structure for the underlying source signals. Some examples include higher-order statistics based methods which assume non-Gaussian and i.i.d source signals, and ICA which assumes the independence of the source signals.

Most real life signals are non-stationary, and thus do not obey the underlying assumption of stationarity that is embedded in the current methods. For this reason, recently various methods have been introduced to exploit the non-stationarity property of the source signals to solve the separation problem, including frequency

domain [73], [74] and time domain [75], [76] approaches. In general the frequency-domain estimation algorithms have a simpler implementation, less computational time, and better convergence properties over the time-domain ones. However, the disadvantages of using frequency-domain methods are the arbitrary permutation and scaling ambiguities of the estimated frequency response of the un-mixing system at each frequency bin.

Motivated by these problems, researchers have resorted to the powerful tool of time-frequency signal representations. For non-stationary signals, a blind separation approach using a spatial time-frequency distribution is proposed in [58] to achieve the separation by joint diagonalization of the auto-terms in the spatial time-frequency distributions. This approach has been modified and improved as discussed in [77, 78]. Another time-frequency based method described in [60] uses binary time-frequency masks to separate more than two speech sources from two mixtures using the sparsity of the time-frequency representations of speech signals.

In this chapter, we introduce a new approach to the source separation problem combining time-frequency representations with information-theoretic measures. An information-theoretic criterion, Jensen-Rényi divergence as adapted to the time-frequency distributions, is used as the objective function to separate the sources. The underlying sources are assumed to be disjoint on the time-frequency plane and it is shown that this new cost function achieves its maximum when the signals are disjoint. With the assumption that the source signals are disjoint on the time-frequency plane, signal separation is performed through a rotation transformation using a steepest descent algorithm.

5.2 Information Measures in the Time-Frequency Domain

The information-theoretic measures such as entropy have been successfully applied to the time-frequency plane [42, 65], due to an analogy between a TFD and the probabil-

ity density function (pdf) of a two-dimensional random variable. Although entropy measures have proven to be useful in quantifying the complexity of individual signals, they cannot be used directly to quantify the difference between signals. For this reason, well-known divergence measures from information theory have been adapted to the time-frequency plane [79, 80]. The most common divergence measures used for probability distributions belong to Csiszar's f-divergence such as Kullback-Leibler divergence based on Shannon entropy and α -divergence based on Rényi entropy [39]. Another common class of divergence measures is based on the Jensen difference such as the Jensen-Shannon divergence constructed by applying Jensen inequality to the entropy functional. Jensen-Rényi divergence is the modification of Jensen-Shannon divergence from an arithmetic to a geometric mean introduced by Michel [79]. For time-frequency distributions, Jensen-Rényi divergence can be defined as:

$$\bar{J}_{12}^{\alpha}(C_1, C_2) = H_{\alpha}(\sqrt{C_1 C_2}) - \frac{H_{\alpha}(C_1) + H_{\alpha}(C_2)}{2}, \quad (5.1)$$

where C_1 and C_2 are the general TFDs of two different signals defined in equation (2.4) respectively, and H_{α} represents Rényi entropy defined in equation (4.9). Jensen-Rényi divergence is equal to zero when $C_1 = C_2$, and its positivity can be proven using the Cauchy-Schwartz inequality. This measure has some desired properties such as being symmetric and monotonically increasing as the overlap between the two distributions decreases, i.e. $C_1(t, \omega)C_2(t, \omega) \rightarrow 0$. Therefore, maximizing this measure corresponds to obtaining disjoint time-frequency representations.

When Jensen-Rényi divergence is compared to other symmetric and monotonically increasing information-theoretic measures on the time-frequency plane, several advantages emerge. First of all, Jensen-Rényi divergence is defined based on the Rényi entropy which is well-defined for a larger class of time-frequency distributions compared to Shannon entropy which is defined for only positive distributions. Therefore,

Shannon entropy based divergence measures are limited in their applicability. Second, recent work in the analysis of sensitivity or robustness of divergence measures on the time-frequency plane reveals that Jensen-Rényi divergence is more robust against perturbations and noise, which makes it more suitable for source detection and separation applications [79, 80]. The following simulation example compares the robustness of two different distance measures under an additive signal perturbation model. The original signal is a Gabor logon, $s_1(t) = \exp(-(t - t_0)^2) \exp(-j\omega_0 t)$, centered at time $t_0 = 32$, normalized frequency $\omega_0 = 0.2$, and the perturbation signal is another Gabor logon, $s_2(t) = \exp(-(t - t_1)^2) \exp(-j\omega_0 t)$, centered at $t_1 = 64, \omega_0 = 0.2$. The perturbed signal is $z(t) = (1 - \epsilon)s_1(t) + \epsilon s_2(t)$, where $\epsilon \in [0, 1]$. The distance between the time-frequency distributions of the perturbed signal and the original one is computed as ϵ goes from 0 to 1. Figure 5.1 shows the comparison between the symmetric Kullback-Leibler and the Jensen-Rényi divergences for different values of ϵ . When ϵ is small, the Jensen-Rényi divergence is smaller than the Kullback-Leibler divergence showing robustness against small perturbation. However, as ϵ increases, the Jensen-Rényi divergence reacts faster to the change and detects the second signal component.

5.3 Problem Formulation and Method

5.3.1 Signal Model and Assumptions

In this chapter, we consider the problem of determining the source signals when the number of observed mixtures is equal to or greater than the number of source signals. Assume that the N -dimensional vector $\mathbf{s}(t) = [s_1(t), s_2(t), \dots, s_N(t)]^T$ corresponds to the N non-stationary complex source signals. The source signals are transmitted through a medium and the M sensors pick up a set of mixed signals represented by $\mathbf{z}(t) = [z_1(t), z_2(t), \dots, z_M(t)]^T$, where $M \geq N$.

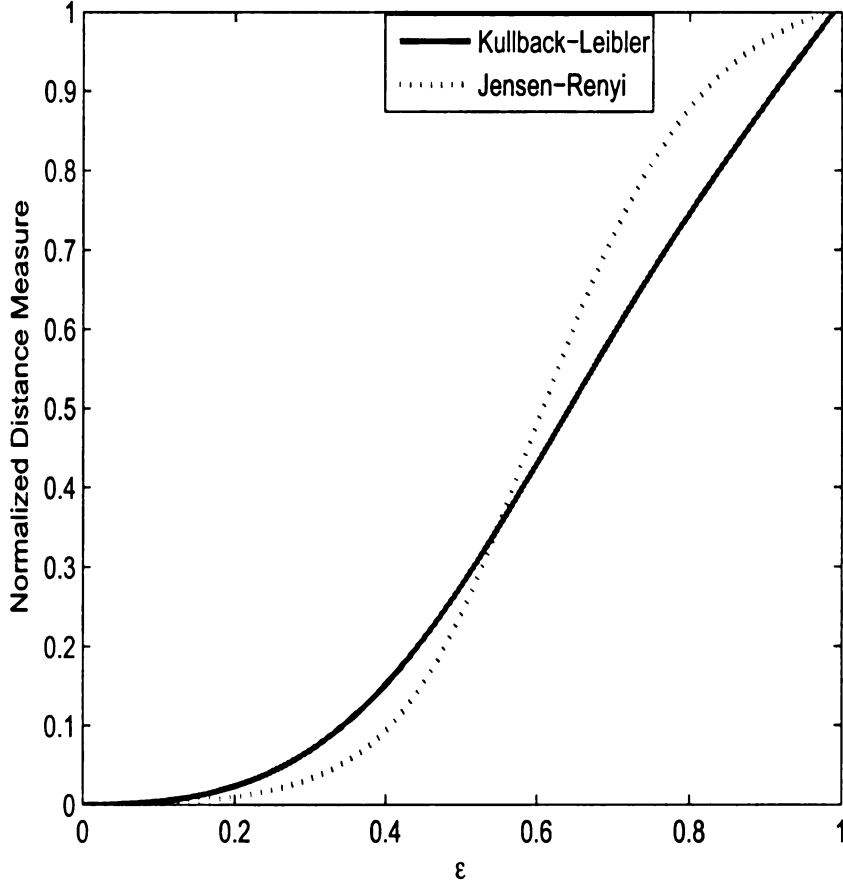


Figure 5.1. Comparison of Kullback-Leibler and Jensen-Rényi divergence measures under an additive signal perturbation model

Given M observations or mixtures, $\mathbf{z}(t) = [z_1(t), z_2(t), \dots, z_M(t)]^T$, with

$$\mathbf{z}(t) = \mathbf{A}(t)\mathbf{s}(t), \quad (5.2)$$

where $\mathbf{A}(t)$ is the mixing matrix, we want to extract the underlying sources $\mathbf{s}(t)$. In this chapter, we assume an instantaneous mixture of the sources, i.e., $\mathbf{A}(t) = \mathbf{A}$, where \mathbf{A} is a $M \times N$ matrix. The following assumption is made about the underlying sources: the sources are assumed to have different structures and localization properties on the

time-frequency plane, i.e. the sources are disjoint on the time-frequency plane. This implies that $C_{s_i}(t, \omega)C_{s_j}(t, \omega) = 0, \forall t, \omega$ for $i \neq j$. In practice, this condition is never satisfied exactly. However, as long as the inner product is small, source separation can be achieved. This condition on the disjointness of the sources on the time-frequency plane has already been used by several authors for separation of speech and music signals [60], [81].

Before proceeding further, it is important to specify the notion of blind identification. In the blind context, a full identification of the mixing matrix \mathbf{A} is impossible since the exchange of a fixed scalar factor between a given source signal and the corresponding column of \mathbf{A} does not affect the observations, as is shown by the following relation:

$$\mathbf{z}(t) = \mathbf{A}s(t) = \sum_{i=1}^N \frac{\mathbf{a}_i}{\beta_i} \beta_i s_i(t), \quad (5.3)$$

where β_i is an arbitrary complex factor, and \mathbf{a}_i denotes the i th column of \mathbf{A} . Advantage can be taken of this indeterminacy by assuming that the source signals have unit variance so that the dynamic range of the sources is accounted for by the magnitude of the corresponding columns of \mathbf{A} . This normalization convention turns out to be convenient in the sequel; it does not affect the performance results. For the proposed algorithm, the sources are extracted on the time-frequency plane up to a scalar factor, and permutation.

5.3.2 Problem Statement in the Time-Frequency Domain

This section will briefly outline the overall structure of the source separation. The different components of the algorithm such as the cost function, and the optimization method will be discussed subsequently.

Each observation, $z_i(t)$, is first transformed to the time-frequency plane as:

$$X_i(n, \omega; \psi) = \sum_m \sum_l \psi(n-l, m) z_i\left(l + \frac{m}{2}\right) z_i^*\left(l - \frac{m}{2}\right) e^{-j\omega m}. \quad (5.4)$$

Since $z_i(t) = \sum_{k=1}^N a_{ik} s_k(t)$ from equation (5.2), where a_{ik} is the element of the mixing matrix \mathbf{A} located at the i th row and k th column, we have:

$$\begin{aligned}
X_i(n, \omega; \psi) &= \sum_m \sum_l \psi(n-l, m) \left[\sum_{k=1}^N a_{ik} s_k \left(l + \frac{m}{2} \right) \right] \left[\sum_{r=1}^N a_{ir}^* s_r^* \left(l - \frac{m}{2} \right) \right] e^{-j\omega m} \\
&= \sum_m \sum_l \psi(n-l, m) \left[\sum_{k=1}^N |a_{ik}|^2 s_k \left(l + \frac{m}{2} \right) s_k^* \left(l - \frac{m}{2} \right) \right] e^{-j\omega m} + \\
&\quad \sum_m \sum_l \psi(n-l, m) \left[\sum_{k=1}^N \sum_{r=1(r \neq k)}^N a_{ik} a_{ir}^* s_k \left(l + \frac{m}{2} \right) s_r^* \left(l - \frac{m}{2} \right) \right] e^{-j\omega m}, \\
&\quad i = 1, 2, \dots, M.
\end{aligned} \tag{5.5}$$

In the right hand side of the above equation, the first term represents the auto-terms, and the second term represents the cross-terms. We are assuming that the kernel function $\psi(\cdot, \cdot)$ used in this chapter is a reduced interference distribution (RID), so that the cross-terms are negligible. Thus,

$$\begin{aligned}
X_i(n, \omega; \psi) &\approx \sum_m \sum_l \psi(n-l, m) \left[\sum_{k=1}^N |a_{ik}|^2 s_k \left(l + \frac{m}{2} \right) s_k^* \left(l - \frac{m}{2} \right) \right] e^{-j\omega m} \\
&= \sum_{k=1}^N |a_{ik}|^2 \left[\sum_m \sum_l \psi(n-l, m) s_k \left(l + \frac{m}{2} \right) s_k^* \left(l - \frac{m}{2} \right) e^{-j\omega m} \right] \\
&= \sum_{k=1}^N |a_{ik}|^2 C_k(n, \omega; \psi),
\end{aligned} \tag{5.6}$$

where $C_k(n, \omega; \psi)$ is the discrete time-frequency distribution of the source signal $s_k(t)$. This shows that the instantaneous mixtures of the source signals in the time domain transforms into the instantaneous mixtures of TFDs. This is an important underlying assumption that makes the proposed approach easier to implement compared to other

TFD based BSS methods [58].

The time-frequency distribution $X_i(n, \omega; \psi)$ corresponding to each observation $z_i(t)$ is vectorized and a matrix of time-frequency distributions is formed:

$$\begin{aligned} \mathbf{X} &= \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_M \end{bmatrix} = \begin{bmatrix} X_1(1) & \cdots & X_1(P) \\ X_2(1) & \cdots & X_2(P) \\ & \vdots & \\ X_M(1) & \cdots & X_M(P) \end{bmatrix} \\ &= \mathbf{A}^2 \mathbf{C} = \begin{bmatrix} |a_{11}|^2 & \cdots & |a_{1N}|^2 \\ |a_{21}|^2 & \cdots & |a_{2N}|^2 \\ & \vdots & \\ |a_{M1}|^2 & \cdots & |a_{MN}|^2 \end{bmatrix} \begin{bmatrix} \mathbf{C}_1 \\ \mathbf{C}_2 \\ \vdots \\ \mathbf{C}_N \end{bmatrix}, \end{aligned} \quad (5.7)$$

where \mathbf{X}_i and \mathbf{C}_i are vectors of length $P = K \times L$ points, K and L are the numbers of time and frequency points respectively, and \mathbf{A}^2 is the element-by-element square of the mixing matrix \mathbf{A} . The extracted sources on the time-frequency plane are defined as:

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_N \end{bmatrix} = \begin{bmatrix} Y_1(1) & \cdots & Y_1(P) \\ Y_2(1) & \cdots & Y_2(P) \\ & \vdots & \\ Y_N(1) & \cdots & Y_N(P) \end{bmatrix}. \quad (5.8)$$

In order to make the following discussions simpler, we concentrate on the case where $M = N$. The discussion can be easily generalized for $M > N$ as illustrated through an example in Section 5.4.

5.3.3 Cost Function

The cost function used in this chapter is the total pairwise Jensen-Rényi divergence defined as:

$$\bar{J}_\alpha \triangleq \sum_{i=1}^{N-1} \sum_{j=i+1}^N \left[H_\alpha(\sqrt{\mathbf{Y}_i \mathbf{Y}_j}) - \frac{H_\alpha(\mathbf{Y}_i) + H_\alpha(\mathbf{Y}_j)}{2} \right]. \quad (5.9)$$

Maximizing this cost function will ensure that the extracted components do not overlap with each other on the time-frequency plane.

The pairwise Jensen-Rényi divergence between two time-frequency distributions is defined as:

$$\bar{J}_{ij}^\alpha = H_\alpha(\sqrt{\mathbf{Y}_i \mathbf{Y}_j}) - \frac{H_\alpha(\mathbf{Y}_i) + H_\alpha(\mathbf{Y}_j)}{2}. \quad (5.10)$$

This expression can be further simplified as:

$$\begin{aligned} \bar{J}_{ij}^\alpha &= H_\alpha(\sqrt{\mathbf{Y}_i \mathbf{Y}_j}) - \frac{H_\alpha(\mathbf{Y}_i) + H_\alpha(\mathbf{Y}_j)}{2} \\ &= \frac{1}{1-\alpha} \log \left(\sum_{k=1}^P \left(\sqrt{Y_i(k) Y_j(k)} \right)^\alpha \right) \\ &\quad - \frac{1}{2(1-\alpha)} \left[\log \left(\sum_{k=1}^P Y_i^\alpha(k) \right) + \log \left(\sum_{k=1}^P Y_j^\alpha(k) \right) \right] \\ &= \frac{1}{1-\alpha} \log \left(\frac{\sum_{k=1}^P \left(\sqrt{Y_i(k) Y_j(k)} \right)^\alpha}{\sqrt{\left(\sum_{k=1}^P Y_i^\alpha(k) \right) \left(\sum_{k=1}^P Y_j^\alpha(k) \right)}} \right), \end{aligned} \quad (5.11)$$

which represents the ratio of the energy of the overlap between the two TFDs to the product of the energy of the individual TFDs. Let

$$J_{ij}^\alpha = \frac{\sum_{k=1}^P \left(\sqrt{Y_i(k) Y_j(k)} \right)^\alpha}{\sqrt{\left(\sum_{k=1}^P Y_i^\alpha(k) \right) \left(\sum_{k=1}^P Y_j^\alpha(k) \right)}}, \quad (5.12)$$

and

$$J_\alpha = \sum_{i=1}^{N-1} \sum_{j=i+1}^N J_{ij}^\alpha. \quad (5.13)$$

Since $\log(\cdot)$ is a monotonic function, maximizing \bar{J}_α is equivalent to minimizing J_α for $\alpha > 1$, or maximizing J_α for $\alpha < 1$. This means that we can equivalently use J_α as our cost function. In this chapter, we will consider orders of $\alpha > 1$. The results are similar for $\alpha < 1$. One special case of $\alpha > 1$ is the quadratic one when $\alpha = 2$. When $\alpha = 2$, the cost function J_α simplifies to:

$$J_2 = \sum_{i=1}^{N-1} \sum_{j=i+1}^N \left[\frac{\sum_{k=1}^P Y_i(k) Y_j(k)}{\sqrt{\left(\sum_{k=1}^P Y_i^2(k)\right) \left(\sum_{k=1}^P Y_j^2(k)\right)}} \right]. \quad (5.14)$$

In this chapter, we will use $\alpha = 2$ since the Rényi entropy will be well-defined for this order even when the distributions are non-positive. Minimizing J_2 is equivalent to minimizing the sum of pairwise normalized inner products between the extracted sources, and ensures disjoint source extraction.

5.3.4 Rotation

In our source separation problem, the observed time-frequency distributions, \mathbf{X} , can be written as a linear combination of the original sources' TFDs, \mathbf{C} , assuming negligible cross-terms between the sources:

$$\mathbf{X} = \mathbf{A}^2 \mathbf{C} = \mathbf{B} \mathbf{C}, \quad (5.15)$$

where \mathbf{A}^2 is the square of the mixing matrix in the time domain, and $\mathbf{B} = \mathbf{A}^2$. The goal of source separation in this chapter is to find a linear transform \mathbf{Q} of the observed signals, \mathbf{X} , such that the extracted signals, \mathbf{Y} , are as disjoint as possible from each other. The cost function in Section 5.3.3 quantifies the disjointness of the extracted sources using divergence. In this section, we will show how to obtain this

linear transform \mathbf{Q} . \mathbf{Q} should be chosen such that the elements of $\mathbf{Y} = \mathbf{Q}\mathbf{X} = \mathbf{Q}\mathbf{B}\mathbf{C}$ are disjoint. If the elements of \mathbf{Y} are exactly disjoint, then $\mathbf{Y}\mathbf{Y}^T$ will be a diagonal matrix, which means that $\mathbf{Q}\mathbf{B}\mathbf{C}\mathbf{C}^T\mathbf{B}^T\mathbf{Q}^T$ will also be diagonal. Since the original sources' TFDs, \mathbf{C} , are disjoint and normalized, $\mathbf{C}\mathbf{C}^T = \mathbf{I}$. Therefore, finding a linear transform \mathbf{Q} for unmixing the observations reduces to finding an unitary matrix that will diagonalize $\mathbf{B}\mathbf{B}^T$. Since \mathbf{B} is not known a priori, we try to estimate the unitary transform \mathbf{Q} iteratively. Any unitary matrix \mathbf{Q} can be written as a product of Givens rotation matrices and this formulation allows us to parameterize the estimation of \mathbf{Q} in terms of the rotation angles θ .

It is well-known that any unitary matrix \mathbf{Q} can be written as the product of $N(N-1)/2$ Givens rotation matrices, $\mathbf{Q} = \mathbf{G}_1\mathbf{G}_2\cdots\mathbf{G}_{N(N-1)/2}$. In N -dimensional space, the simplest rotation is in the two-dimensional plane. If a rotation is through an angle θ_{ab} in the $a - b$ plane, then the Givens rotation matrix $\mathbf{G}_{ab}(\theta_{ab})$ is:

$$\mathbf{G}_{ab}(\theta_{ab}) = \begin{bmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & \cos(\theta_{ab}) & \cdots & \sin(\theta_{ab}) & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & -\sin(\theta_{ab}) & \cdots & \cos(\theta_{ab}) & \cdots & 0 \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{bmatrix}, \quad (5.16)$$

where $\mathbf{G}_{ab}(\theta_{ab})$ equals the $N \times N$ identity matrix \mathbf{I}_N except that the elements $I_N(a, a)$, $I_N(a, b)$, $I_N(b, a)$, and $I_N(b, b)$ are replaced by $\cos(\theta_{ab})$, $\sin(\theta_{ab})$, $-\sin(\theta_{ab})$, and $\cos(\theta_{ab})$, respectively, where $I_N(a, b)$ is the element of \mathbf{I}_N located at the a th row and b th column. From [82], we know that any N -dimensional rotation matrix can be written as the product of $N(N-1)/2$ two-dimensional-plane N -dimensional rotation

matrices, which is:

$$\mathbf{G}(\theta) = \mathbf{G}_{12}(\theta_{12}) \cdots \mathbf{G}_{ab}(\theta_{ab}) \cdots \mathbf{G}_{(N-1)N}(\theta_{(N-1)N}), \quad (5.17)$$

where $\theta = [\theta_{12}, \dots, \theta_{ab}, \dots, \theta_{(N-1)N}]^T$, and $a < b$.

In order to have exact source separation in this formulation, there should exist an unitary matrix \mathbf{Q} that will diagonalize the mixing matrix $\mathbf{B} = \mathbf{A}^2$ on the time-frequency plane. Since \mathbf{B} has all positive entries, there is no such \mathbf{Q} unless \mathbf{B} is already diagonal which corresponds to the observations that are scalar multiples of the sources.

5.3.5 Proposed Algorithm

The objective of the proposed algorithm is to determine the optimal rotation transform such that the total pairwise divergence measure is maximized to achieve signal separation. We use the gradient adaptation algorithm also known as the steepest descent [66] to update the rotation angles. Gradient adaptation is not the only choice, but it is preferred in many practical paradigms due to its simplicity and efficient convergence [83].

The overall update equation for stochastic gradient descent is:

$$\theta(n+1) = \theta(n) - \mu \frac{\partial J_2}{\partial \theta}, \quad (5.18)$$

where μ is the step size parameter. The gradient of the cost function J_2 with respect to the rotation angle θ_{ab} is derived as:

$$\frac{\partial J_2}{\partial \theta_{ab}} = \sum_{i=1}^{N-1} \sum_{j=i+1}^N \frac{\partial J_{ij}^2}{\partial \theta_{ab}}, \quad (5.19)$$

where

$$\begin{aligned}
\frac{\partial J_{ij}^2}{\partial \theta_{ab}} = & \frac{\sum_{k=1}^P \left(\frac{\partial \mathbf{G}_i}{\partial \theta_{ab}} \mathbf{X}(k) Y_j(k) + Y_i(k) \frac{\partial \mathbf{G}_j}{\partial \theta_{ab}} \mathbf{X}(k) \right)}{\sqrt{\left(\sum_{k=1}^P Y_i^2(k) \right) \left(\sum_{k=1}^P Y_j^2(k) \right)}} \\
& - \frac{\sum_{k=1}^P Y_i(k) Y_j(k)}{\left(\sqrt{\left(\sum_{k=1}^P Y_i^2(k) \right) \left(\sum_{k=1}^P Y_j^2(k) \right)} \right)^3} \\
& \times \left(\sum_{k=1}^P Y_i(k) \frac{\partial \mathbf{G}_i}{\partial \theta_{ab}} \mathbf{X}(k) \right) \left(\sum_{k=1}^P Y_j^2(k) \right) \\
& - \frac{\sum_{k=1}^P Y_i(k) Y_j(k)}{\left(\sqrt{\left(\sum_{k=1}^P Y_i^2(k) \right) \left(\sum_{k=1}^P Y_j^2(k) \right)} \right)^3} \\
& \times \left(\sum_{k=1}^P Y_i^2(k) \right) \left(\sum_{k=1}^P Y_j(k) \frac{\partial \mathbf{G}_j}{\partial \theta_{ab}} \mathbf{X}(k) \right), \tag{5.20}
\end{aligned}$$

where \mathbf{G}_i is the i th row of $\mathbf{G}(\theta)$, and $\mathbf{X}(k)$ is the k th column of \mathbf{X} . The explicit gradient expression for $N = 3$ is given in the Appendix.

5.4 Experimental Results and Analysis

In order to evaluate the effectiveness of the proposed method, we consider various source separation examples. In all of the examples with the synthesized signals, the sources are assumed to be approximately disjoint on the time-frequency plane. Each observation is transformed to the time-frequency domain with $K = 50$ time samples and $L = 64$ frequency samples. Each TFD is vectorized to form a TFD observation matrix of size $M \times 3200$ as in equation (5.7), where M is the number of observations. Jensen-Rényi divergence with order $\alpha = 2$ is used as the cost function to ensure that the divergence is well-defined. The binomial kernel [23] is used for computing the TFD since it belongs to the class of reduced interference distributions (RIDs) and thus will have negligible cross-terms. This property of the distributions will improve

the performance of our source separation algorithm. The performance of the proposed method is quantified in terms of the accuracy of the extracted sources, convergence rate, and robustness to noise. All of the experimental results will be evaluated using the signal to interference ratio (SIR) defined as:

$$\begin{aligned}
\text{SIR} &= \frac{1}{N} \sum_{i=1}^N \text{SIR}_i, \\
\text{SIR}_i &= \text{SIR}_{O_i} - \text{SIR}_{I_i}, \\
\text{SIR}_{O_i} &= 10 \log \left(\frac{\sum_{k=1}^P Y_{is_i}^2(k)}{\sum_{k=1}^P \left(\sum_{j \neq i} Y_{is_j}(k) \right)^2} \right), \\
\text{SIR}_{I_i} &= 10 \log \left(\frac{\sum_{k=1}^P X_{is_i}^2(k)}{\sum_{k=1}^P \left(\sum_{j \neq i} X_{is_j}(k) \right)^2} \right),
\end{aligned} \tag{5.21}$$

where Y_{is_j} and X_{is_j} are the outputs and inputs of the system when only the signal s_j is active, respectively, and N is the number of sources.

Example 1: Separation of a chirp signal and two Gabor logon signals

In this example, the set of observed signals are the three linear combinations of a chirp signal and two Gabor logon signals, i.e. $z_i(t) = a_{i1}s_1(t) + a_{i2}s_2(t) + a_{i3}s_3(t)$, where a_{i1}, a_{i2}, a_{i3} are the weights for each signal distributed as $N(0, 1)$, $i = 1, 2, 3$, and $s_1(t), s_2(t), s_3(t)$ correspond to the two Gabor logon signals and the chirp signal, respectively. A Gabor logon is a modulated Gaussian expressed as:

$$s_i(t) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(t-t_{i0})^2}{2\sigma^2}} e^{j\omega_{i0}t} \quad (i = 1, 2). \tag{5.22}$$

The Gabor logon has virtually compact support in both time and frequency centered in time at $t = t_{i0}$ and frequency at $\omega = \omega_{i0}$. In this example, the first Gabor logon is centered at the time sample point $t_{10} = 50$ and normalized frequency of $\omega_{10} = 0.7$,

and the second Gabor logon is centered at the time sample point $t_{20} = 150$ and normalized frequency of $\omega_{20} = -0.7$. A linear Gaussian chirp is expressed as:

$$s_3(t) = \sqrt[4]{\frac{\sigma}{\pi}} e^{\{-\frac{\sigma}{2}(t-t_0)^2 + j[\omega_0 + \frac{\beta}{2}(t-t_0)](t-t_0)\}}, \quad (5.23)$$

where t_0, ω_0 are the time and frequency centers, σ, β are the time spread and frequency modulation rates of the chirp, respectively. In this example, the linear chirp signal has an initial normalized frequency of -0.2 and its instantaneous frequency increases to a normalized frequency of 0.2 with $t_0 = \omega_0 = 0$. It is known that the chirp signal overlaps with these two Gabor logons in the time domain, so it is not possible to separate them using time domain decomposition approaches. However, it is illustrated in Figure 5.2 that these three signals can be effectively extracted on the time-frequency plane using the proposed method through an optimal rotation under the divergence criterion with an average SIR of 37.5169 dB. Moreover, the convergence rate is high as shown in Figure 5.3.

Example 2: Separation of two crossing chirp signals

In this example, we consider the separation of two signals overlapping in the time-frequency domain. A mixture of two linear chirp signals is used for source separation. One of the chirp signals has an initial normalized frequency of -0.8 and its instantaneous frequency increases to a normalized frequency of 0.8. The other one has an initial normalized frequency of 0.8 and its instantaneous frequency decreases to a normalized frequency of -0.8. Obviously, these two chirp signals overlap with each other in both the time and frequency domains. Typical time domain or frequency domain separation methods can not be used to perfectly recover them. Figure 5.4 shows that using the proposed approach, we can successfully separate these two chirp signals from their mixtures with an average SIR of 32.1164 dB. It can be seen from the figures that most of the error occurs in the time-frequency region where the two

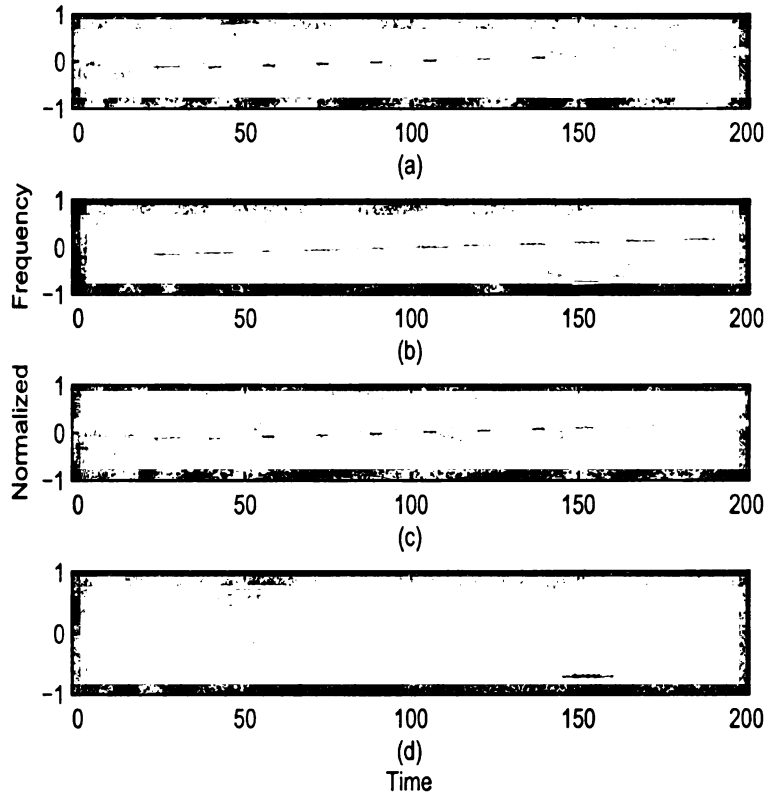


Figure 5.2. The mixture and the separation of a chirp and two Gabor logons: (a) the mixture; (b) and (d) the two extracted Gabor logons; (c) the extracted chirp

signals overlap. This is due to the fact that the crossing chirps do not exactly satisfy our underlying assumption of disjoint sources.

Example 3: Separation of two speech sources

In this example, we consider the mixtures of two speech signals from two speakers, one female and one male. The two speakers' voices are recorded by two microphones 3m directly in front of the speakers in an anechoic chamber. Due to time delay, these two signals partly overlap with each other in the time domain. The TFDs of the original speech signals and their mixtures are shown in Figure 5.5 and Figure 5.6, respectively. Figure 5.7 shows the TFDs of the speech signals extracted by the proposed method. The SIR is 26.0482 dB in this case, since the two speech signals

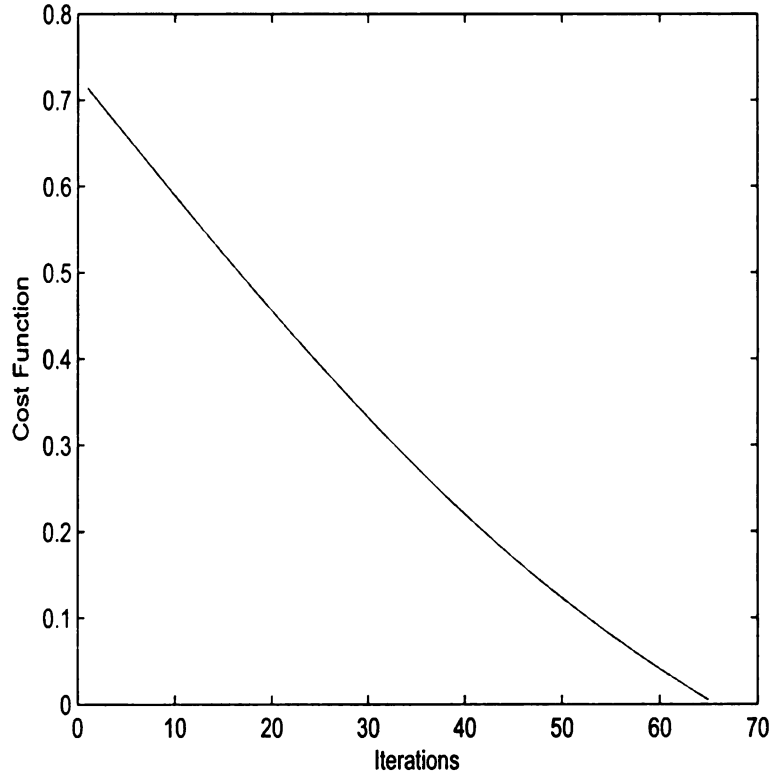


Figure 5.3. The cost function versus the number of iterations for Example 1

have partial overlap on the time-frequency plane.

In order to further investigate the robustness of the proposed algorithm for real-life signals, we add white Gaussian noise into the two speech signals over a SNR range of $[-8 - 8 \text{ dB}]$. It is shown in Figure 5.8 that the proposed method is robust against noise and results in the separation of the speech signals even at low input SNRs.

Example 4: Performance comparison with the STFD and FastICA methods

In order to further evaluate the performance of the proposed approach, we compare it with two different methods, one of which is a time-frequency based source separation method, the spatial time-frequency distribution (STFD) [58], and the other one is an information-theoretic method, FastICA [84] adapted to the time-frequency domain.

The STFD method is based on the joint diagonalization of a combined set of spa-

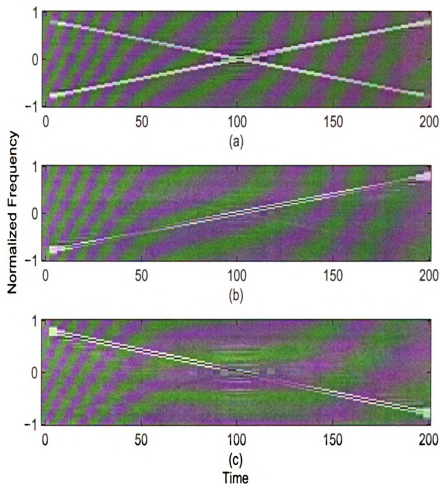


Figure 5.4. The mixture and the separation of two crossing chirp signals: (a) the mixture; (b) and (c) the separated signals

tial time-frequency distribution matrices. STFD matrices are made up of the auto- and cross-TFDs of the data snapshots across the multisensor array, and they are expressed in terms of the TFD matrices of the sources. The diagonal structure of the TFD matrix of the sources is essential for the STFD method and is enforced by using only the information in the time-frequency points corresponding to the signal auto-terms. The benefit of using STFDs in a non-stationary signal environment is the direct exploitation of the information brought by the non-stationarity of the signals.

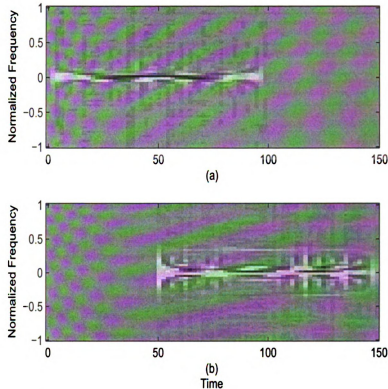


Figure 5.5. TFDs of the two individual speech signals: (a) TFD of a female speaker; (b) TFD of a male speaker

FastICA combines Comon's information-theoretic approach [9] and the projection pursuit approach [85]. A family of contrast (objective) functions for ICA are introduced using maximum entropy approximations of differential entropy. These contrast functions enable both the estimation of the whole decomposition by minimizing mutual information, and estimation of individual independent components as projection pursuit directions. FastICA algorithm has a fast convergence rate and is robust under noise. In this example, the TFD observation matrix, \mathbf{X} in equation (5.7), is considered as the input to the FastICA algorithm to achieve source separation.

SIR is used as the performance criterion for the two crossing chirp signals discussed in Example 2 by adding white Gaussian noise over a SNR range of $[-8 - 8 \text{ dB}]$. We use

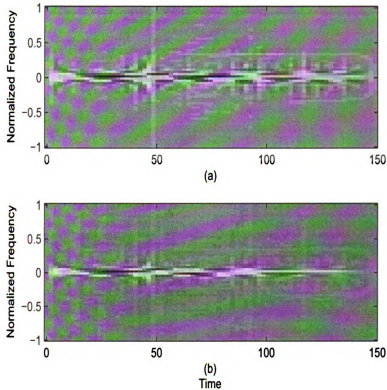


Figure 5.6. TFDs of the observed signals: (a) TFD of the first mixture; (b) TFD of the second mixture

100 Monte Carlo simulations for each noise level. Figure 5.9 compares the robustness of the the proposed approach in noise to the STFD and FastICA methods. It is evident that both the proposed approach and the STFD method are superior to FastICA under noise. This is due to the fact that the assumption that the underlying source signals are independent does not necessarily apply to the time-frequency distributions. We can also see that the proposed approach performs better than the STFD method as the noise level increases. The reason is that as the noise level increases, the energy of the cross-terms will increase, thus making it harder to differentiate between the auto- and the cross-terms in the STFD method. This results in errors in the estimation of the sources, consequently reducing the SIR. On the other hand, the

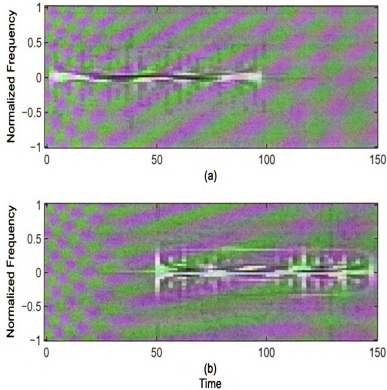


Figure 5.7. TFDs of the extracted signals: (a) TFD of estimate of the female speaker; (b) TFD of estimate of the male speaker

proposed method assumes that the cross-terms are negligible. When the signal is very noisy, the cross-terms between the signal and the noise become significant, and neglecting these cross-terms amounts to denoising of the observed TFDs, resulting in higher SIRs. The STFD method also has a higher computational complexity than the proposed method, since it computes both the auto- and cross-terms of the observed signals whereas our method uses only the auto-terms by using a RID kernel.

Example 5: Performance comparison with PCA

In this example, we compare the proposed source separation method with PCA for the mixture of the two Gabor logon signals in Example 1. PCA is an orthogonal decomposition of the observed data matrix just like the proposed method. However,

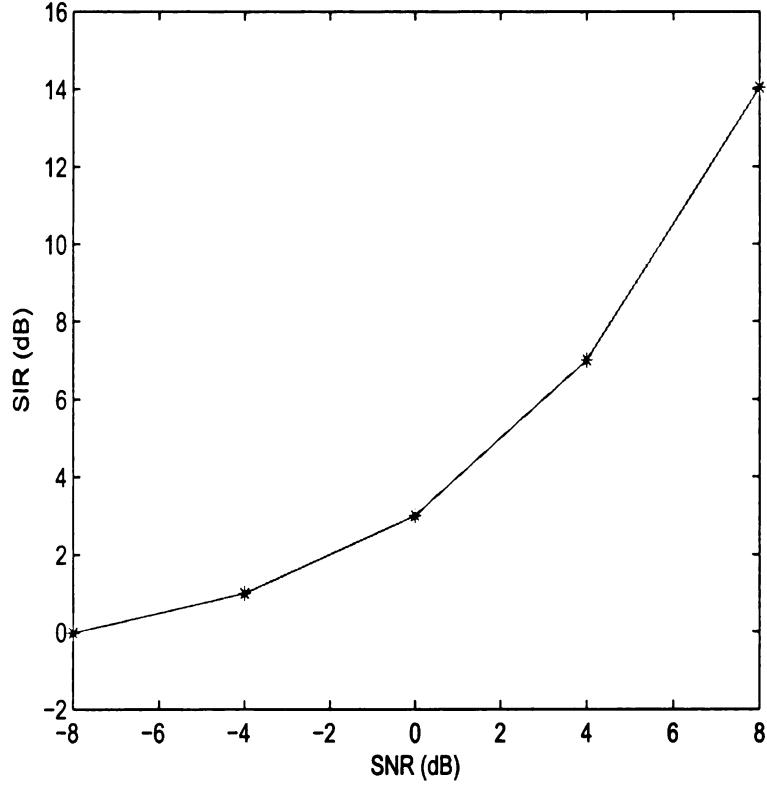


Figure 5.8. Output SIR versus input SNR for speech signals

the cost functions used are different and results in the difference seen in the extracted components in Figure 5.10. With PCA the variance explained by each component is maximized whereas the proposed method maximizes the divergence between the components resulting in better separated sources.

Example 6: Number of mixtures greater than the number of sources

In this example, we consider a more general situation where the number of mixtures is larger than the number of sources. For M mixtures and N sources ($M > N$), we construct a new $N \times M$ rotation matrix as follows:

$$\mathbf{G}_{NM}(\theta) = \mathbf{I}_{NM} \mathbf{G}_M(\theta), \quad (5.24)$$

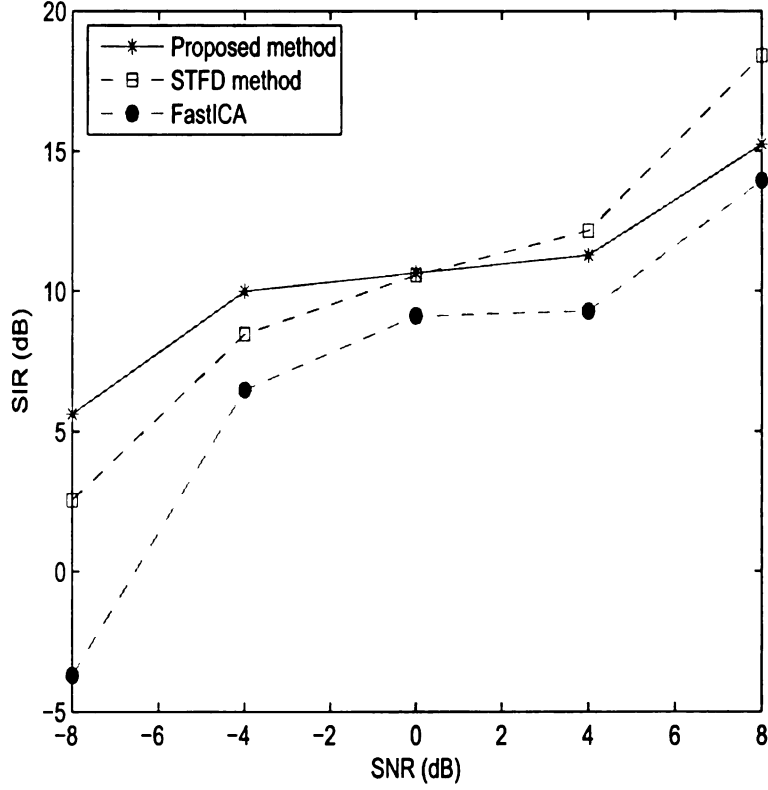


Figure 5.9. Comparison of output SIR versus input SNR for three different source separation methods

where $\mathbf{G}_M(\theta)$ is an $M \times M$ rotation matrix given by equation (5.17), and \mathbf{I}_{NM} is an $N \times M$ matrix with elements equal to 1 if $i = j$, 0 otherwise, where i, j represent the row and column indices, respectively. The source signals are the chirp signal and one of the two Gabor logons in Example 2. We use the proposed approach with this new rotation matrix to extract these two signals from their three mixtures. It is shown in Figure 5.11 that the source signals can be effectively extracted when the number of mixtures is greater than the number of sources with an average SIR of 39.8827 dB.

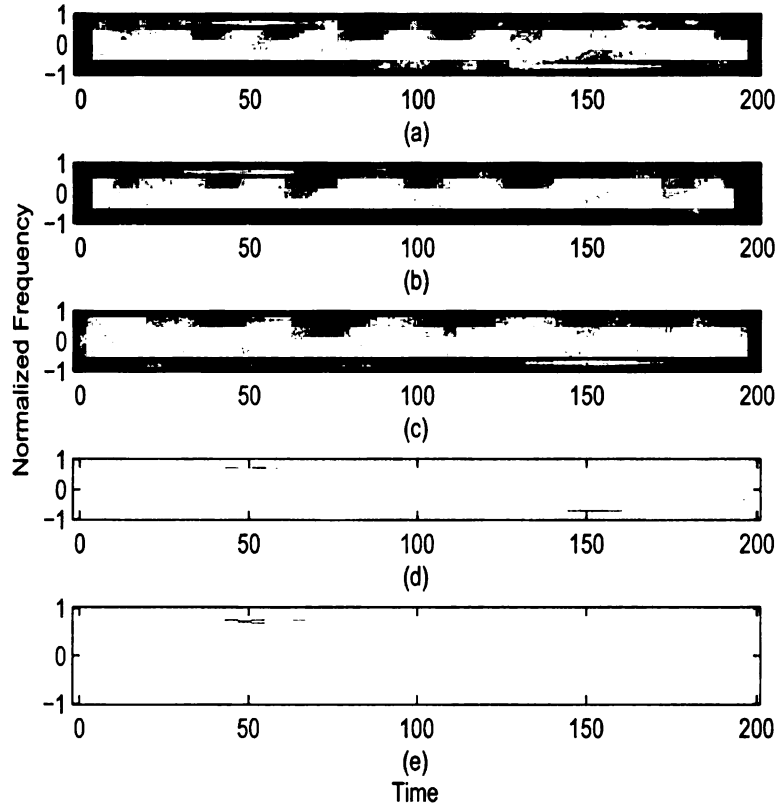


Figure 5.10. Comparison with PCA for two Gabor logon extraction: (a) the mixture, (b) and (c) the components extracted by the proposed method, (d) and (e) the components extracted by PCA

5.5 Summary

In this chapter, a new approach is presented for the separation of non-stationary signals on the time-frequency plane using an information-theoretic cost function. The proposed algorithm assumes the disjointness of the underlying signals on the time-frequency plane. This assumption allows us to extract the sources through a N -dimensional Givens rotation. Using Jensen- Rényi divergence as the cost function, a steepest descent algorithm is implemented to update the rotation angles. Several examples are given to illustrate the performance of the proposed algorithm for synthesized and real life signals. Issues regarding convergence rate and robustness under

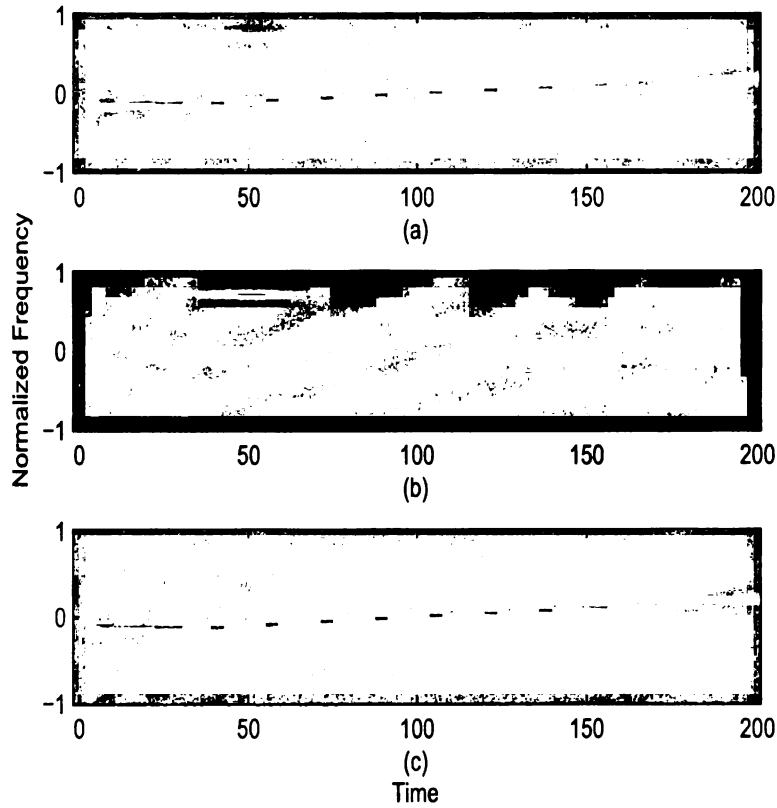


Figure 5.11. Separation of a chirp and a Gabor logon from their three mixtures: (a) the mixture, (b) the extracted Gabor logon, (c) the extracted chirp

noise are investigated. The performance of the algorithm is illustrated under noise and is compared to PCA and ICA as adapted to the time-frequency plane, and STFD. The results illustrate that maximizing the divergence on the time-frequency plane can separate sources that are disjoint in the time-frequency domain, and is better than the mutual information cost function used in ICA in terms of fidelity to the original sources. The proposed method also outperforms STFD for high noise levels since it assumes the cross-terms between sources are negligible which effectively denoises the observed time-frequency matrix, and is apparently superior to PCA.

CHAPTER 6

UNDERDETERMINED BLIND SOURCE SEPARATION IN THE TIME-FREQUENCY DOMAIN

6.1 Introduction

Underdetermined blind source separation (UBSS) is a more challenging problem compared to the (over)determined case, because contrary to the (over)determined case, estimating the mixing system is not sufficient for reconstruction of the sources, since the mixing matrix is not invertible. Therefore, we need additional *a priori* information about the sources to allow for reconstruction. One increasingly popular and powerful assumption is the sparsity of the sources on a given basis [5,86,87]. A signal is said to be sparse when it is zero or nearly zero more than might be expected from its variance. Such a signal has a probability density function or distribution of values with a sharp peak at zero and fat tails. The advantage of a sparse signal representation is that the probability of two or more sources being simultaneously active is low. Thus, sparse representations lend themselves to good separability because most of the energy on a basis coefficient at any time instant belongs to a single source. This statistical property of the sources results in a nicely defined structure being imposed by the mixing process on the resultant mixtures, which can be exploited to make estimating the mixing process much easier.

Sparse representation of the signals, which is modelled by matrix factorization, has been receiving a great deal of interest and has been applied to blind source separation in recent years. In several references, the mixing matrix and sources are estimated using the maximum posterior approach, the maximum likelihood approach, and the expectation maximization algorithm [50,88–92]. However, these algorithms may stick at a local minima and have poor convergence property. In [93], a blind

source separation is developed via multi-node sparse representation. Based on several subsets of wavelet packet coefficients, the mixing matrix is estimated by using Fuzzy C -means clustering algorithm, and the sources are recovered using the inverse of the estimated mixing matrix. However, the case of less sensors than sources is not discussed.

In this chapter, we introduce a sparse factorization approach to the UBSS problem in the time-frequency domain, in which the mixing matrix is estimated using the K -means clustering method, while the sources are estimated using a linear programming method. In [94], the equivalence results of the l_0 -norm solution and the l_1 -norm solution are obtained using a probabilistic approach. These results show that if the sources are sufficiently sparse in the analyzed domain, they are more likely to be equal to the l_1 -norm solution, which can be obtained using a linear programming method.

6.2 Sparse Factorization Approach for UBSS in the Time-Frequency Domain

In this section, a sparse factorization approach including two stages for the UBSS problem in the time-frequency domain are presented, in which the first stage is for determining the mixing matrix, and then the second stage is for estimating the source signals.

6.2.1 Linear Mixture Model and Assumptions

We first give out the system model and assumptions. The observed M mixtures, $\mathbf{z}(t) = [z_1(t), z_2(t), \dots, z_M(t)]^T$, of the N non-stationary complex source signals, $\mathbf{s}(t) = [s_1(t), s_2(t), \dots, s_N(t)]^T$, may be modeled linearly in the time domain as

$$\mathbf{z}(t) = \mathbf{B}\mathbf{s}(t), \quad (6.1)$$

where \mathbf{B} is the $M \times N$ instantaneous mixing matrix.

Each mixture, $z_i(t)$, is transformed to the time-frequency plane, and then the corresponding time-frequency distribution is vectorized to form a matrix of time-frequency distributions, \mathbf{X} . From Chapter 5, it is known that the time-frequency distributions of the mixtures, \mathbf{X} , can be written approximately as a linear combination of the original sources' TFDs, \mathbf{S} , assuming the cross-terms between the sources are negligible by using a RID:

$$\mathbf{X} \approx \mathbf{B}^2 \mathbf{S} = \mathbf{A} \mathbf{S}, \quad (6.2)$$

where $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_P] \in R^{M \times P}$, $\mathbf{S} = [\mathbf{s}_1, \dots, \mathbf{s}_P] \in R^{N \times P}$, $P = I \times L$, I and L are the numbers of time and frequency points respectively, $\mathbf{A} = \mathbf{B}^2 = [\mathbf{a}_1, \dots, \mathbf{a}_N] \in R^{M \times N}$ with normalized columns, i.e., $\|\mathbf{a}_1\| = \dots = \|\mathbf{a}_N\| = 1$, and \mathbf{B}^2 is the element-by-element square of the mixing matrix in the time domain. The task of BSS is to recover the sources \mathbf{S} only using the mixture matrix \mathbf{X} . Here, we assume $M < N$, which indicates that BSS is underdetermined, and the source signals are sparse in the time-frequency domain. The sparsity of the sources plays a key role in this chapter.

It is well known that in general there exist many possible solutions for the model (6.2). For a given mixing matrix, under the sparsity measure of l_1 -norm, the uniqueness result of sparse solution is obtained. And the number of nonzero entries of the sparse solution can not be reduced. It is also found that the mixing matrix of which the column vectors are composed by cluster centers of the mixtures \mathbf{X} is a sub-optimal mixing matrix, which can be obtained using K -means clustering algorithm [94].

6.2.2 K -means Clustering

K -means clustering is an iterative algorithm that seeks to minimize a squared-error criterion function in order to separate a completely unknown set of data into K different groupings [95]. Suppose that $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ are the vector observations in a data set and make up realizations of K different distributions of random variables.

Then $\mu_1, \mu_2, \dots, \mu_K$ are the mean vectors of these distributions, and K -means clustering seeks to categorize the observations, \mathbf{x}_i , into one of the K distributions such that the squared Euclidean distance, $\|\mathbf{x}_i - \mu_j\|^2$, is minimized. However, since the properties of the data set are unknown, $\mu_1, \mu_2, \dots, \mu_K$ must be estimated first as $\hat{\mu}_1, \hat{\mu}_2, \dots, \hat{\mu}_K$.

As a starting point, K random samples of the data are chosen as the initial mean estimates, $\hat{\mu}_j$. The distributions are then estimated by classifying all points, \mathbf{x}_i , into the group whose estimated mean it is closest to in the squared Euclidean sense, so that $\mathbf{x}_i \in \hat{\mu}_j$ when j is subject to

$$\min_j \|\mathbf{x}_i - \hat{\mu}_j\|^2. \quad (6.3)$$

Once all data points are classified, the mean of each group is recalculated. Suppose m_j is the number of data points in the j th distribution, and $\mathbf{x}_{1j}, \mathbf{x}_{2j}, \dots, \mathbf{x}_{m_jj}$ are all data points. The new mean is then calculated as

$$\hat{\mu}_j = \frac{1}{m_j} \sum_{i=1}^{m_j} \mathbf{x}_{ij}. \quad (6.4)$$

This process is repeated until convergence, when the estimated means do not change upon further iterations.

6.2.3 Determination of the Mixing Matrix

Due to the sparsity of the source signals in the time-frequency domain, there exists many columns of \mathbf{S} with only one nonzero entry. For instance, suppose that $\mathbf{s}_{i_1}, \dots, \mathbf{s}_{i_K}$ are K columns of \mathbf{S} , where only the first entry of each of these columns is nonzero, then we have

$$\mathbf{A}\mathbf{s}_{i_j} = \mathbf{a}_1 s_{1i_j} \quad j = 1, \dots, K, \quad (6.5)$$

and

$$[\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_K}] = \mathbf{A}[\mathbf{s}_{i_1}, \dots, \mathbf{s}_{i_K}] = [\mathbf{a}_1 s_{1i_1}, \dots, \mathbf{a}_1 s_{1i_K}], \quad (6.6)$$

where, \mathbf{x}_{i_j} is the i_j th column of \mathbf{X} , \mathbf{a}_1 is the first column of \mathbf{A} , and s_{1i_j} is the first entry of \mathbf{s}_{i_j} . From equation (6.6), we see that each \mathbf{x}_{i_j} is equal to \mathbf{a}_1 multiplied by a scalar s_{1i_j} , which means that these K column vectors of \mathbf{X} , $\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_K}$, are distributed along the direction of \mathbf{a}_1 . Thus, ideally after normalization, $\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_K}$ are mapped to a unique vector on the multidimensional unit circle which is equal to \mathbf{a}_1 . However, in practice, the sources are likely not completely sparse in the time-frequency domain. That is, $\mathbf{s}_{i_1}, \dots, \mathbf{s}_{i_K}$, have the dominant first entries so that $s_{1i_j} \gg s_{ri_j}$ for $r \neq 1$ and $j = 1, \dots, K$. When more than one source are nonzero, $\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_K}$ are not exactly in the same direction as \mathbf{a}_1 , but rather in the neighborhood of \mathbf{a}_1 . This means that \mathbf{a}_1 lies at the center of $\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_K}$.

Therefore, we use the K -means clustering method to cluster the column vectors of the mixture matrix \mathbf{X} into different clusters, where the center of each cluster corresponds to one column vector of the mixing matrix \mathbf{A} . By doing so, we can obtain an estimate of the mixing matrix \mathbf{A} . The algorithm is summarized as follows:

Algorithm 1: Determining the mixing matrix

1. Normalize the column vectors of the mixture matrix \mathbf{X} .
2. Take a sufficiently large positive integer K as the number of clusters. Choose the initial points of iteration and the distance measure criterion. In this part of the proposal, we choose the squared Euclidean distance as the criterion.
3. Do K -means clustering iteration followed by normalization to estimate the sub-optimal mixing matrix. Note that if two column vectors have opposite directions, only one is taken.

6.2.4 Estimation of the Source Signals for a Given Mixing Matrix

After obtaining the estimated mixing matrix, the next stage is to estimate the source signals. For a given mixing matrix \mathbf{A} in model (6.2), the source signals can be estimated by maximizing posterior distribution $P(\mathbf{S}|\mathbf{X}, \mathbf{A})$ of \mathbf{S} [96]. Under the assumption that the prior is Laplacian, maximizing posterior distribution can be implemented by solving the following optimization problem [6]:

$$\min \sum_{i=1}^N \sum_{j=1}^P |s_{ij}|, \quad \text{subject to } \mathbf{AS} = \mathbf{X}. \quad (6.7)$$

Hence, the l_1 -norm

$$J_1(\mathbf{S}) = \sum_{i=1}^N \sum_{j=1}^P |s_{ij}| \quad (6.8)$$

can be used as the sparsity measure.

It is not difficult to prove that the optimization problem (6.7) is equivalent to the following set of P smaller scale linear programming (LP) problems:

$$\min \sum_{i=1}^N |s_{ij}|, \quad \text{subject to } \mathbf{As}_j = \mathbf{x}_j \quad \text{for } j = 1, \dots, P. \quad (6.9)$$

By setting $\mathbf{S} = \mathbf{U} - \mathbf{V}$, where $\mathbf{U} = [u_{ij}] \in R^{N \times P} \geq 0$ and $\mathbf{V} = [v_{ij}] \in R^{N \times P} \geq 0$, equation (6.9) can be converted to the following standard LP problems with non-negative constraints:

$$\begin{aligned} & \min \sum_{i=1}^N (u_{ij} + v_{ij}), \\ & \text{subject to } [\mathbf{A}, -\mathbf{A}][\mathbf{u}_j^T, \mathbf{v}_j^T]^T = \mathbf{x}_j, \quad \mathbf{u}_j \geq 0, \mathbf{v}_j \geq 0 \quad \text{for } j = 1, \dots, P. \end{aligned} \quad (6.10)$$

Combining the discussion of this section and the previous sections, we have the algorithm for estimating the source signals:

Algorithm 2: Blindly separating the sparse source signals

1. Prethreshold the mixture matrix \mathbf{X} to obtain a sparser data matrix $\hat{\mathbf{X}}$.
2. Estimate the mixing matrix \mathbf{A} using *Algorithm 1* and $\hat{\mathbf{X}}$.
3. Using the estimated mixing matrix \mathbf{A} and the mixtures \mathbf{X} , estimate the time-frequency representations \mathbf{S} by solving the set of LP problems (6.10).

6.3 Experimental Results and Analysis

In this section, several examples will be used to illustrate the effectiveness of the proposed approach to separate the sparse source signals from their fewer mixtures in the time-frequency domain. The binomial kernel [23] is used for computing the TFD since it belongs to the class of reduced interference distributions (RIDs).

Example 1: The set of observed signals are two linear combinations of four Gabor logons. These four Gabor logons are centered at the time sample point and the normalized frequency (30,0.7), (160,-0.7), (70,-0.4), and (120,0.1), respectively. Each observed signal is first transformed to the time-frequency domain with $I = 50$ time samples and $L = 64$ frequency samples. Each TFD is then vectorized to form a TFD mixture matrix $\mathbf{X} = [\mathbf{X}_1; \mathbf{X}_2]$ of size 2×3200 .

Figure 6.1 presents a scatter plot of the mixtures \mathbf{X} (\mathbf{X}_2 vs. \mathbf{X}_1) in the time-frequency domain. It can be seen from this plot that almost all significant data points are distributed along four different directions, thus providing very good separability. The separation results using the proposed approach are illustrated in Figure 6.2. Figure 6.2 (a) and (b) represent the two mixtures. The four extracted Gabor logon signals are shown in Figure 6.2 (c), (d), (e), and (f). The results indicate that these four Gabor logons can be successfully separated from their two mixtures using the proposed approach based on their sparsity with an average signal to interference ratio (SIR) of 36.1251 dB.

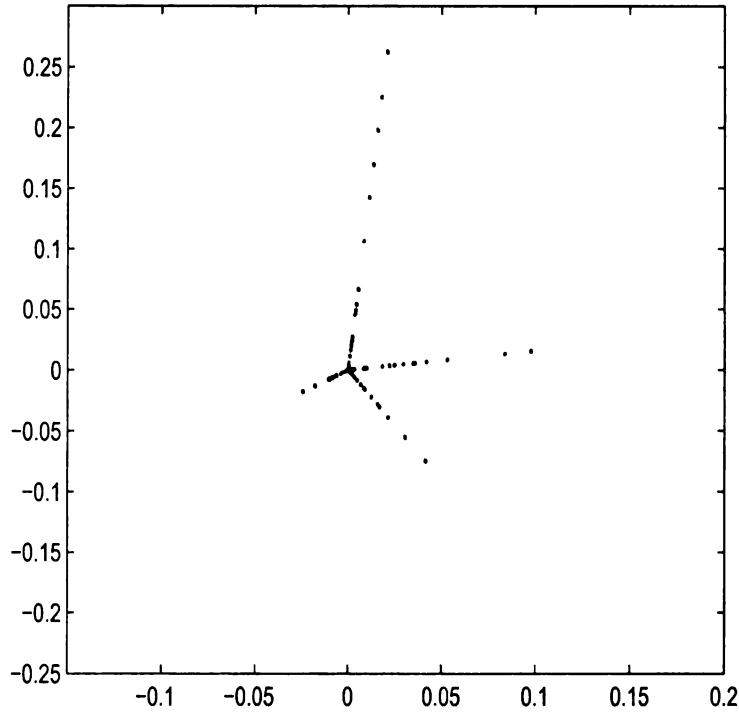


Figure 6.1. Scatter plot of two mixtures of four Gabor logons in the time-frequency domain

Example 2: Two mixtures of a chirp signal and two Gabor logons are given. The chirp signal has a linear frequency increasing from an initial normalized frequency of -0.2 to a normalized frequency of 0.2. The Gabor logons are the first two Gabor logons given in Example 1. A scatter plot of the two mixtures in Figure 6.3 shows that it is similar to the first example in that the distributions of data points belonging to different sources are along three different directions. Since the chirp signal overlaps with the two Gabor logons in the time domain, typical time domain separation methods can not be used to perfectly recover them. However, it is illustrated in Figure 6.4 that these three signals can be effectively extracted in the time-frequency domain using the proposed method with an average SIR of 32.7634 dB.

Example 3: In this example, the same two mixtures of four Gabor logons given in Example 1 are used. The effectiveness of the presented approach is compared for TFDs and wavelet packets (WP) in the presence of noise. Haar wavelet with five levels is used for the wavelet packet decomposition.

To show the effect of increased sparsity of TFDs, the mixtures at different levels of white Gaussian noise are considered. 100 Monte Carlo simulations are used for each noise level. The average mean squared error (MSE) between the extracted sources and the original sources is calculated for both the TFD and WP. The TFD and WP provide similar results when there is no noise. However, as the noise level increases, the performance of the WP rapidly degrades compared to that of the TFD. The MSE versus the signal-to-noise ratio (SNR) is shown in Figure 6.5 for both the TFD and WP. This result shows that the RID results in a more sparse time-frequency surface compared to the WP, which improves the robustness of BSS under noise.

6.4 Summary

In this chapter, a two-stage approach is introduced for underdetermined blind separation of sparse and non-stationary sources using TFDs. The mixing matrix is estimated using K -means clustering algorithm based on the sparsity of the sources; for a given mixing matrix, the sources are extracted using a linear programming method. The performance of the proposed approach is compared with wavelet packets under different noise levels. The results show that the presented method is simple and effective at separating the sources from their mixtures, and is more robust than wavelet packets under noisy environments.

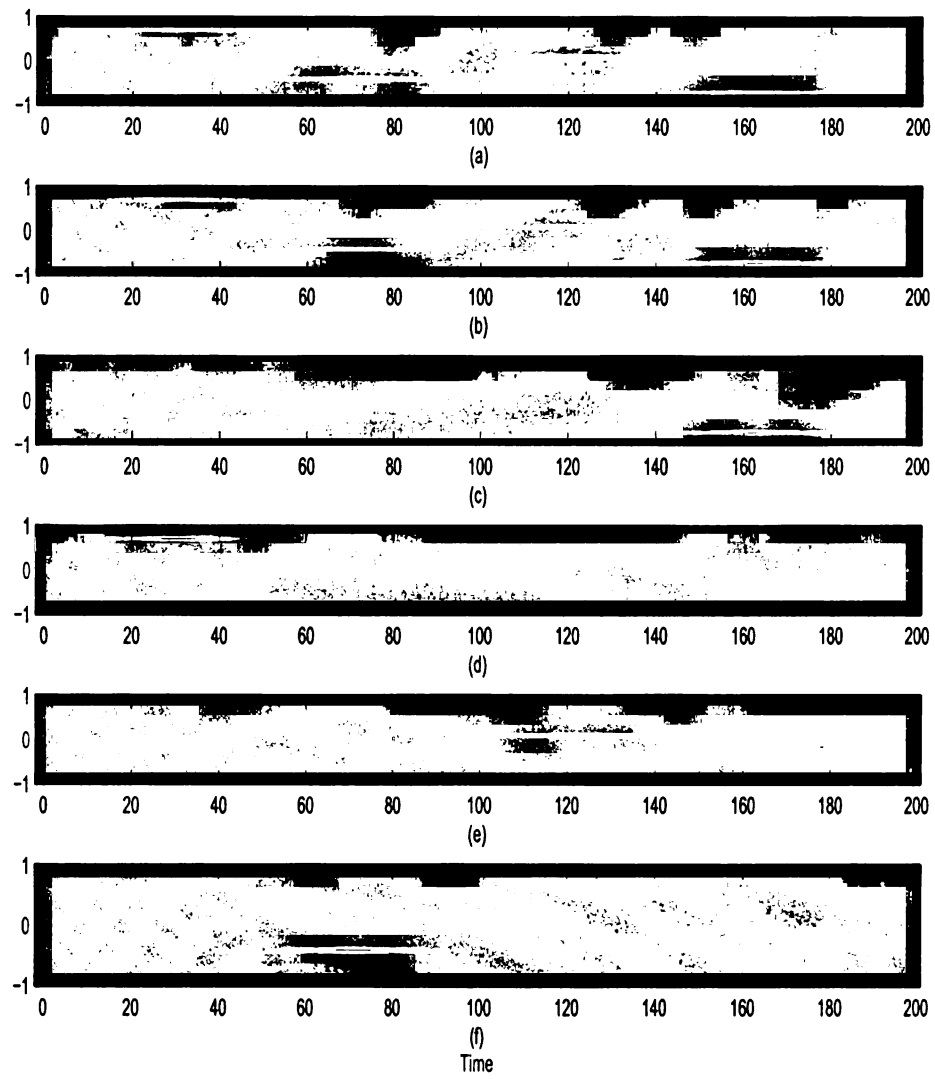


Figure 6.2. The mixtures and the separation of four Gabor logons: (a) and (b) two mixtures; (c), (d), (e), and (f) four extracted Gabor logons

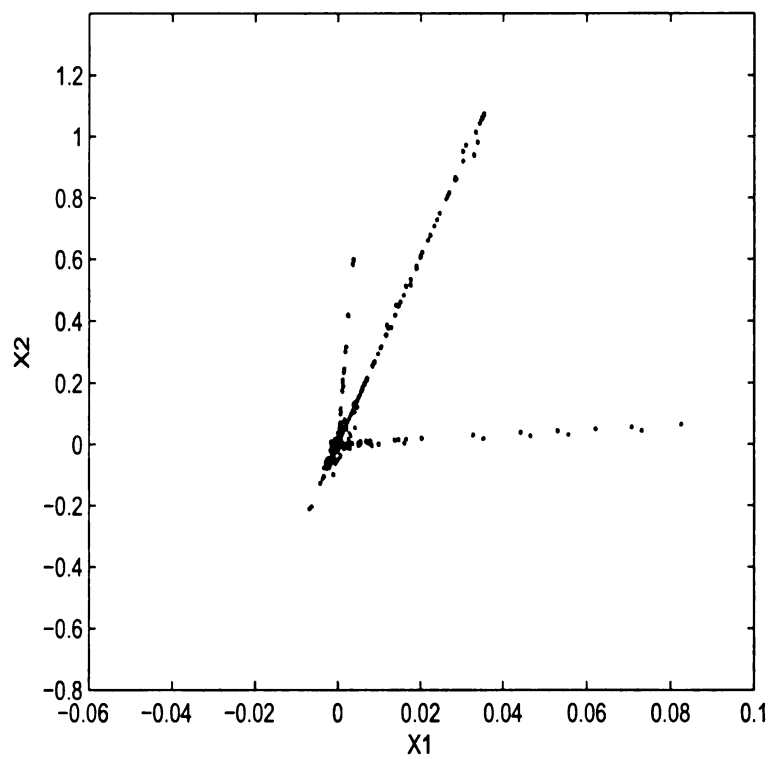


Figure 6.3. Scatter plot of two mixtures of a chirp and two Gabor logons in the time-frequency domain

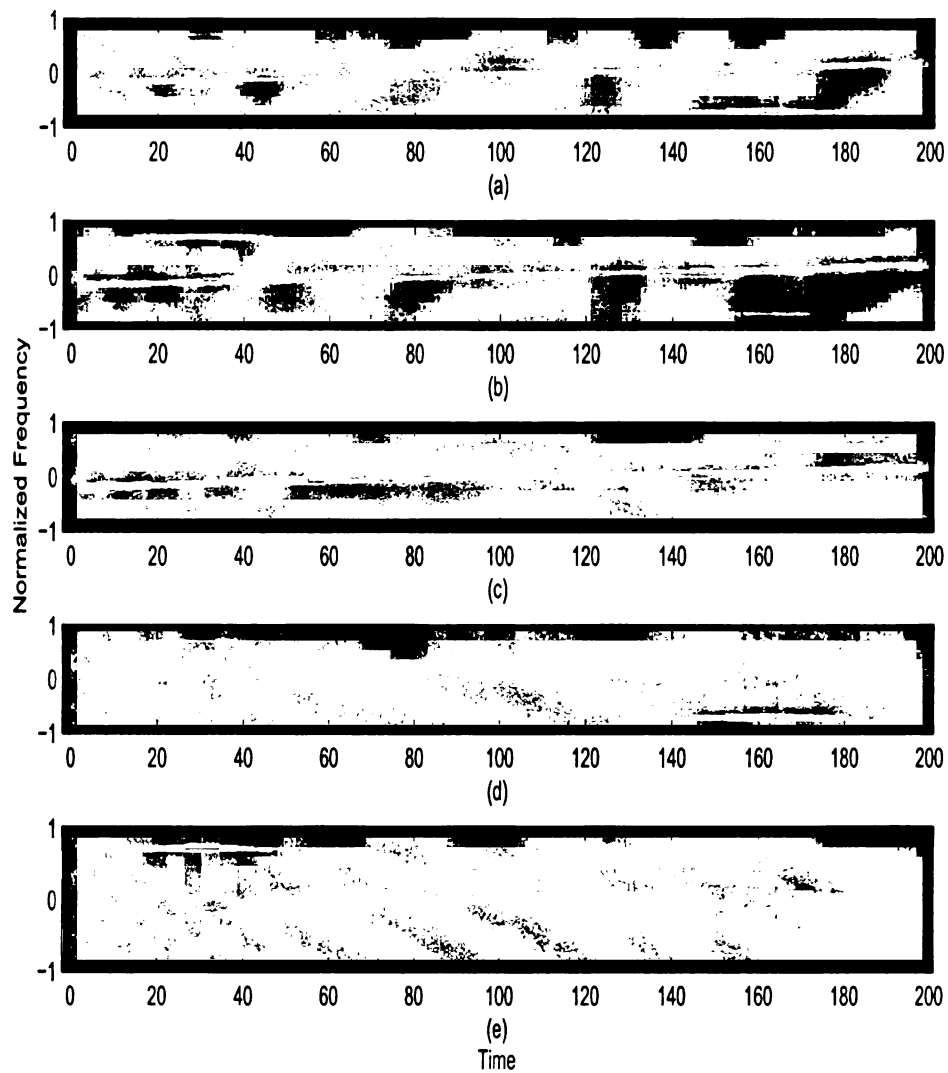


Figure 6.4. The mixtures and the separation of a chirp and two Gabor logons: (a) and (b) two mixtures; (c) extracted chirp; (d) and (e) two extracted Gabor logons

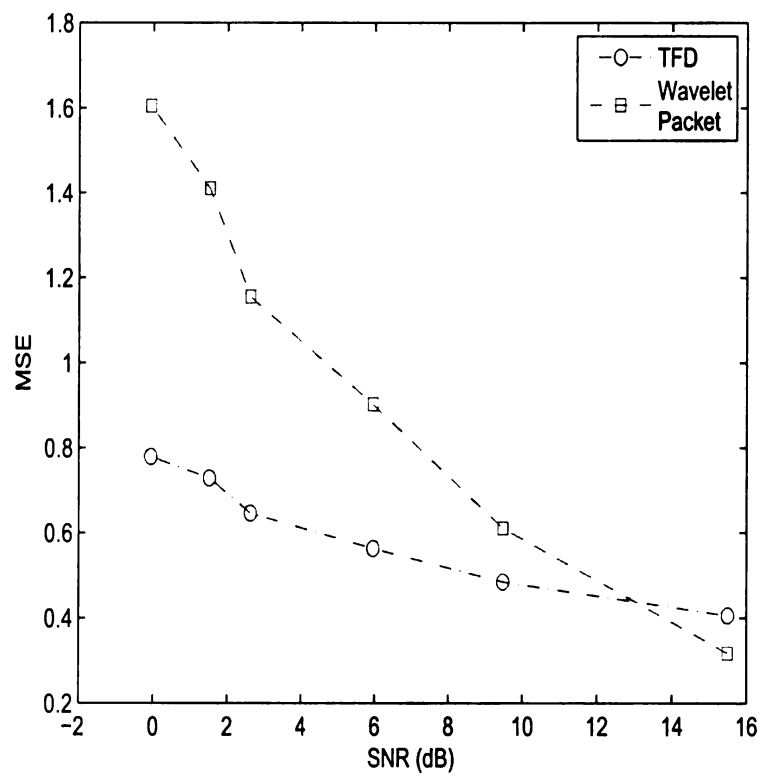


Figure 6.5. Comparison of MSE versus SNR for extracted sources with TFD and WP

CHAPTER 7

APPLICATIONS OF UBSS ALGORITHM ON ELECTROENCEPHALOGRAPH SIGNALS

It is well-known that there is a broad range of applications for blind source separation. A classical example is the cocktail party problem. A more practical application is noise reduction. Another application area is economic time series [97]. Recently, telecommunications applications have also been published [1]. Besides these applications, one popular application of source separation is biomedical signal processing [3], such as separation of electroencephalogram (EEG) signals which consist of recordings of brain activity obtained using electrodes attached to the scalp. Decomposition of evoked field potentials measured by EEG [98] is an application of considerable interest in the neurosciences.

In this chapter, we will apply the UBSS approach proposed in Chapter 6 to the EEG signals so as to evaluate its effectiveness on real life signals.

7.1 Introduction to Electroencephalogram and Event-Related Potential

7.1.1 Electroencephalogram

Electroencephalography is the neurophysiologic measurement of the electrical activity of the brain recorded by electrodes placed on the scalp or, in special cases, subdurally or in the cerebral cortex. The resulting traces are known as an electroencephalogram (EEG) and are reflections of temporal and spatial summation of synchronized post-synaptic cortical potentials [99]. Specifically, EEG data represents the synchronous activity of large cortical groups of neurons, measured as integrated electrical signals on the scalp.

In conventional scalp EEG, the recording is obtained by placing the electrodes

on the scalp in special positions with a conductive gel. Some EEG systems use a plastic cap into which the electrodes are embedded. The electrode positions on the scalp are identified by the recordist who measures the head using the International 10-20 System. This relies on taking measurements between certain fixed points on the head. The electrodes are then placed at points that are 10% and 20% of these distances. Each electrode site is labeled with a letter and a number. The letter refers to the area of brain underlying the electrode e.g. F - Frontal lobe and T - Temporal lobe. Even numbers denote the right side of the head and odd numbers the left side of the head.

EEG activity can be subdivided into various types of frequency rhythms or bands. Research has indicated that different EEG frequency bands are associated with different mental states [100]. In general, EEG activity is broken down into 4 distinct frequency bands:

1. Beta activity 13 Hz–30 Hz. Beta activity is a normal activity present when the eyes are open or closed. It tends to be seen in the channels recorded from the center or front of the head.
2. Alpha activity 8 Hz–13 Hz. Alpha activity is also a normal activity when present in waking adults. It is mainly seen in the channels recorded from the back of the head. It is fairly symmetrical and has an amplitude of 40 μV to 100 μV . It is only seen when the eyes are closed and should disappear or reduce in amplitude when the eyes are open.
3. Theta activity 4 Hz–7 Hz. Theta activity can be classed as both a normal and abnormal activity depending on the age and state of the patient. In adults it is normal if the patient is drowsy. However it can also indicate brain dysfunction if it is seen in a patient who is alert and awake. In younger patients, theta activity may be the main activity seen in channels recorded from the back and

central areas of the head.

4. Delta activity < 4 Hz. Delta activity is only normal in an adult patient if they are in a moderate to deep sleep. If it is seen at any other time it would indicate brain dysfunction.

EEG is preferred in many applications for exploring the brain activity thanks to its high time resolution. Other methods for studying brain activity have time resolution between seconds and minutes, while the EEG has a temporal resolution down to sub-millisecond [101]. As the brain is thought to work through its electric activity, EEG is the only method to measure it directly. Other methods for exploring functions in the brain rely on blood flow or metabolism which may be decoupled from the brain electric activity.

7.1.2 Event-Related Potential

An event-related potential (ERP) is an electrophysiological response to an internal or external stimulus. More simply, it is a measured brain response that is directly the result of a thought or perception. ERPs can be reliably measured using EEG. As the EEG reflects thousands of simultaneously ongoing brain processes, the brain response to a certain stimulus or event of interest is usually not visible in the EEG. One of the most robust features of the ERP response is a response to unpredictable stimuli. This response, known as the P300 (or simply “P3”), manifests as a positive deflection in voltage approximately 300 milliseconds after the stimulus is presented.

In actual recording situations, it is difficult to see an ERP after the presentation of a single stimulus. Rather the most robust ERPs are seen after many dozens or hundreds of individual presentations are averaged together. This technique cancels out noise in the data allowing only the voltage response to the stimulus to stand out clearly. While evoked potentials reflect the processing of the physical stimulus, ERPs are caused by the “higher” processes, that might involve memory, expectation,

attention, or changes in the mental state, among others.

ERPs have found numerous applications in clinical neurophysiology and psychiatry [102,103]. This is because their recording is noninvasive and accurate, and they are consistently shown to be an indicator of brain functions and its abnormalities. The clinical applications of ERPs could be significantly extended if they could be interpreted more accurately and effectively. This requires the development of novel signal processing methods. In recent years, there has been a tremendous growth in applying signal processing techniques such as independent component analysis, wavelet and time-frequency methods for separating the source signals and extracting useful information about the underlying brain activity [13,104].

Event-related potentials like most other real life signals are non-stationary, and thus can be best tackled by using non-stationary signal analysis techniques such as time-frequency distributions (TFDs) and wavelet analysis. In the next section, we will apply the UBSS algorithm presented in Chapter 6 to ERP data set in the time-frequency domain and compare its performance with ICA which is one of the main methods used in the extraction of EEG/ERP sources in both aspects of research and practice.

7.2 Experimental Results and Performance Analysis

7.2.1 EEG/ERP Data Set

The EEG/ERP data analyzed in this chapter is released by Swartz Center for Computational Neuroscience at the University of California, San Diego [105]. The study consisted of one subject and 32 electrodes. In the selective visual attention experiment, there were two types of events “square” and “rt”, “square” events corresponding to the appearance of a green colored square in the display and “rt” to the reaction time of the subject. The square could be presented at five locations on the screen distributed along the horizontal axis. Here we only consider presentation on the left,

i.e. positions 1 and 2 as indicated by the “position” field (at about 3 degree and 1.5 degree of visual angle respectively). In this experiment, the subject covertly attended to the selected location on the computer screen responded with a quick thumb button press only when a square was presented at this location. The subject was to ignore circles presented either at the attended location or at an unattended location. To reduce the amount of data required to process, the data set used in our analysis contains only target (i.e., “square”) stimuli presented at the two left-visual-field attended locations. And 6 electrodes are chosen from 32 electrodes, which are F3, F4, Cz, P3, P4, and Oz in the International 10-20 System. The stimulus is repeated 40 times resulting in a total of 80 trials per electrode.

7.2.2 Single-Trial EEG

The goal in single-trial EEG analysis is to be able to extract individual underlying sources in the brain which are generated in a localized area. With successful source extraction, analysis of individual responses of the brain can be performed, and the dynamic variability of the EEG responses can be compared on a trial to trial basis. In this way, observations can be made on all factors affecting subject’s performance. A comparison is made between the algorithm outlined in Chapter 6 and ICA applied to the same data. Both these BSS techniques are applied to all 80 trials of data available.

In the application of the proposed UBSS approach in Chapter 6, first the number of sources to extract, k , must be chosen. This value is empirically chosen such that it is greater than the number of electrodes, 6. In our analysis, the experiment is done using 32 frequency bins for which k is 8. This number is chosen since higher number of sources resulted in sources that were too sparse and did not correspond to actual neuronal activity. ICA is then applied to the same data. Since only 6 mixtures are used, ICA can only extract 6 components per trial. The results for ICA are in the time domain, so they are converted to the time-frequency plane at the frequency

resolution level using 32 frequency bins for the purposes of comparison.

Figure 7.1 and Figure 7.2 illustrate the results for one trial in the time-frequency domain. Similar results are obtained over all 80 trials. The sources from the proposed technique show in general less activity, i.e. more sparsity, on the time-frequency plane than the sources from ICA. It is, however, difficult to compare results on the single-trial level since the underlying source generators are actually not known, and since a different number of components are extracted from each technique. It is also difficult because there are 80 individual trials. An attempt must be made to generalize the results.

7.2.3 Data Reduction

In order to evaluate the performance of ICA and the proposed UBSS method, the single-trial results are put together in their respective groups depending on stimulus type. K -means clustering is carried out over all extracted components from the subject and the extracted cluster centers represent similar components across all trials. These components are then representative of the most prevalent sources extracted throughout the trials for each stimulus. Evaluation of these cluster centers is then carried out in an attempt to quantify the general results of ICA to those of the proposed method.

The results of one trial are represented by the matrix, \mathbf{S}_v , which is of size $k \times P$. Each component in the time-frequency domain is first vectorized to form a vector of length P , which in our case is equal to 2112. These vectors are then put into a matrix. This represents k extracted components from trial v , each over P time-frequency points. For the data reduction of all results for a particular stimulus, the extracted matrices over all trials are each appended to form a new matrix, $\tilde{\mathbf{S}}_u$, such

that

$$\tilde{\mathbf{S}}_u = \begin{bmatrix} \mathbf{S}_1 \\ \mathbf{S}_2 \\ \vdots \\ \mathbf{S}_{40} \end{bmatrix} = \begin{bmatrix} s_1^1(1) & \cdots & s_1^1(P) \\ \vdots & \vdots & \vdots \\ s_k^1(1) & \cdots & s_k^1(P) \\ s_1^2(1) & \cdots & s_1^2(P) \\ \vdots & \vdots & \vdots \\ s_k^{40}(1) & \cdots & s_k^{40}(P) \end{bmatrix}, \quad (7.1)$$

where $u = \{1, 2\}$ represents the stimulus position number, and $\tilde{\mathbf{S}}_u$ is of size $40k \times P$. Each element, $s_i^v(j)$, is one time-frequency point of source i from trial v .

K -means clustering is then carried out on each $\tilde{\mathbf{S}}_u$ where each of its rows is grouped into one of K clusters based on its squared Euclidean distance to that cluster center as described in Section 6.2.2. The clustering algorithm is run 10 times to avoid randomness in the final cluster results. We run K at 8, 12, and 16 to get an idea of how the different number of components may affect the outcome. The rows of $\tilde{\mathbf{S}}_u$ are then grouped by a hierarchical clustering method based on the results of the 10 K -means runs. A matrix, \mathbf{R} , of zeros of size $40k \times 40k$ is created. Each entry is updated iteratively. The entry, r_{ij} , represents how many times out of 10 row i of $\tilde{\mathbf{S}}_u$ was grouped into the same cluster as row j of $\tilde{\mathbf{S}}_u$. This matrix then serves as a similarity measure, the more times two sources were grouped together by K -means, the more similar they are. All diagonal entries, r_{ii} , represent how many times each source was grouped with itself. These entries are ignored because they are all 10 and are meaningless.

A hierarchical clustering is then carried out using the similarity matrix, \mathbf{R} , as its distance metric. In the first step, each row of $\tilde{\mathbf{S}}_u$ is in its own cluster. The second step then groups all rows together with a similarity value of 10 in the matrix, \mathbf{R} . Next, all rows with similarity of 9 are grouped. If a group already exists, then the average similarity between one row and all rows already in the cluster is used. The

next step will then group together a cluster with another cluster or individual row if it has the highest similarity value. If not, then all rows with similarity value of 8 are grouped together. This is repeated until the number of clusters is reduced to K . Cluster centers are then calculated by the mean of the time-frequency components in each cluster, and these are the components that categorize all single-trial EEG results. For example, a set of extracted components for $K = 8$ is shown in Figure 7.3 and Figure 7.4 for ICA and the proposed UBSS method, respectively.

7.2.4 Performance Evaluation

The levels to which the extracted components are sparse, disjoint, and localized in the time-frequency domain all speak to how close they may be to an actual underlying source in the brain. The components obtained from the clustering method described in the previous section are evaluated based on these factors. Sparsity will be measured using the l_1 -norm, disjointness using the total inner product between the components, and localization using a measure of entropy on the time-frequency plane.

Since a sparse component must have most of its values close to zero, the l_1 -norm is a good measurement of how sparse a component is and a smaller l_1 -norm means a sparser component. The extracted clusters are represented by the $K \times P$ matrix $\mathbf{C}_u, u = \{1, 2\}$. Each row of \mathbf{C}_u represents one extracted component. Thus each component's sparsity is measured with

$$\sum_{m=1}^P |c_i^u(m)|, \quad (7.2)$$

where u refers to stimulus position, i represents component number between 1 and K , and P is the number of time-frequency points.

Disjointness between two components is measured by using the inner product. A summation of all the pairwise inner products between K components represents a

total level of disjointness over all extracted components. This is computed as

$$\sum_{i \neq j} \sum_{m=1}^P |c_i^u(m) c_j^u(m)|. \quad (7.3)$$

Time-frequency localization of each component is computed using a measurement of entropy. This is calculated as

$$- \sum_{m=1}^P |c_i^u(m)| \log_2 |c_i^u(m)|. \quad (7.4)$$

A smaller entropy corresponds to a more localized component.

The results calculated for these parameters are shown in Table 7.1, Table 7.2, and Table 7.3. This shows that under the proposed UBSS approach, the extracted components are typically more sparse, localized, and disjoint than the extracted components under ICA. This means that under the proposed approach, the components are more likely a closer representation of a true source.

Finally, the extracted components from both methods are projected back to the electrodes to show the amount of variance of the original data explained by the sources. From Table 7.4, it is seen that in general a little bit more amount of variance is explained by the components extracted from ICA than by the presented UBSS method. This is because the presented UBSS method seeks to find the sparsest sources, while ICA seeks to find maximally independent sources. The components with less sparse representations (from ICA) project better back to the original measurements, but it is likely that they are linear sums of further reducible sources.

Table 7.1. Mean measure of l_1 norm to show sparsity

Running Conditions	Position 1 (u=1)		Position 2 (u=2)	
	UBSS	ICA	UBSS	ICA
$K=8$	23.03	29.06	21.92	27.63
$K=12$	22.36	28.31	22.54	28.15
$K=16$	23.29	28.18	22.43	27.27

Table 7.2. Mean measure of entropy to show time-frequency localization

Running Conditions	Position 1 (u=1)		Position 2 (u=2)	
	UBSS	ICA	UBSS	ICA
$K=8$	9.80	10.29	9.73	10.25
$K=12$	9.79	10.25	9.80	10.26
$K=16$	9.85	10.24	9.81	10.20

7.3 Summary

This chapter discusses the applications of the proposed UBSS approach in Chapter 6 on the study of ERPs using EEG measurements to help understand mental processes. Since EEG signals have been shown to be non-stationary, the proposed method is applied to ERP data using TFDs and is compared to the popular ICA algorithm when applied to the same multiple trial ERP data set. Data reduction by clustering is performed over all single-trial results to extract components that represent the results. The components were consistently more sparse using the proposed UBSS technique than with ICA, showing that ICA probably tends to extract components that are sums of sources, and can help explain the higher correlation value to the original data. The UBSS technique provided components that are more localized in

Table 7.3. Measure of disjointness by correlation between components

Running Conditions	Position 1 (u=1)		Position 2 (u=2)	
	UBSS	ICA	UBSS	ICA
$K=8$	3.12	3.63	3.84	4.35
$K=12$	8.92	9.38	8.30	8.90
$K=16$	14.05	14.51	15.87	16.35

Table 7.4. Average component projection to electrodes

Running Conditions	Position 1 (u=1)		Position 2 (u=2)	
	UBSS	ICA	UBSS	ICA
$K=8$	0.026	0.039	0.029	0.045
$K=12$	0.058	0.091	0.065	0.103
$K=16$	0.129	0.217	0.147	0.246

the time-frequency domain and that are more distinct from each other than ICA.

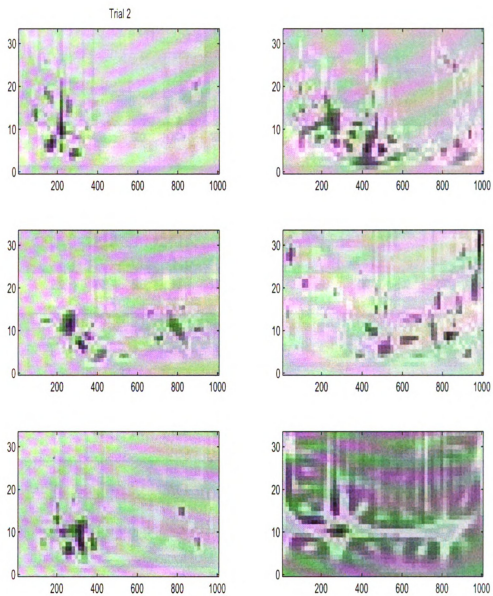


Figure 7.1. Single-trial results using 32 frequency bins: 6 extracted sources from ICA

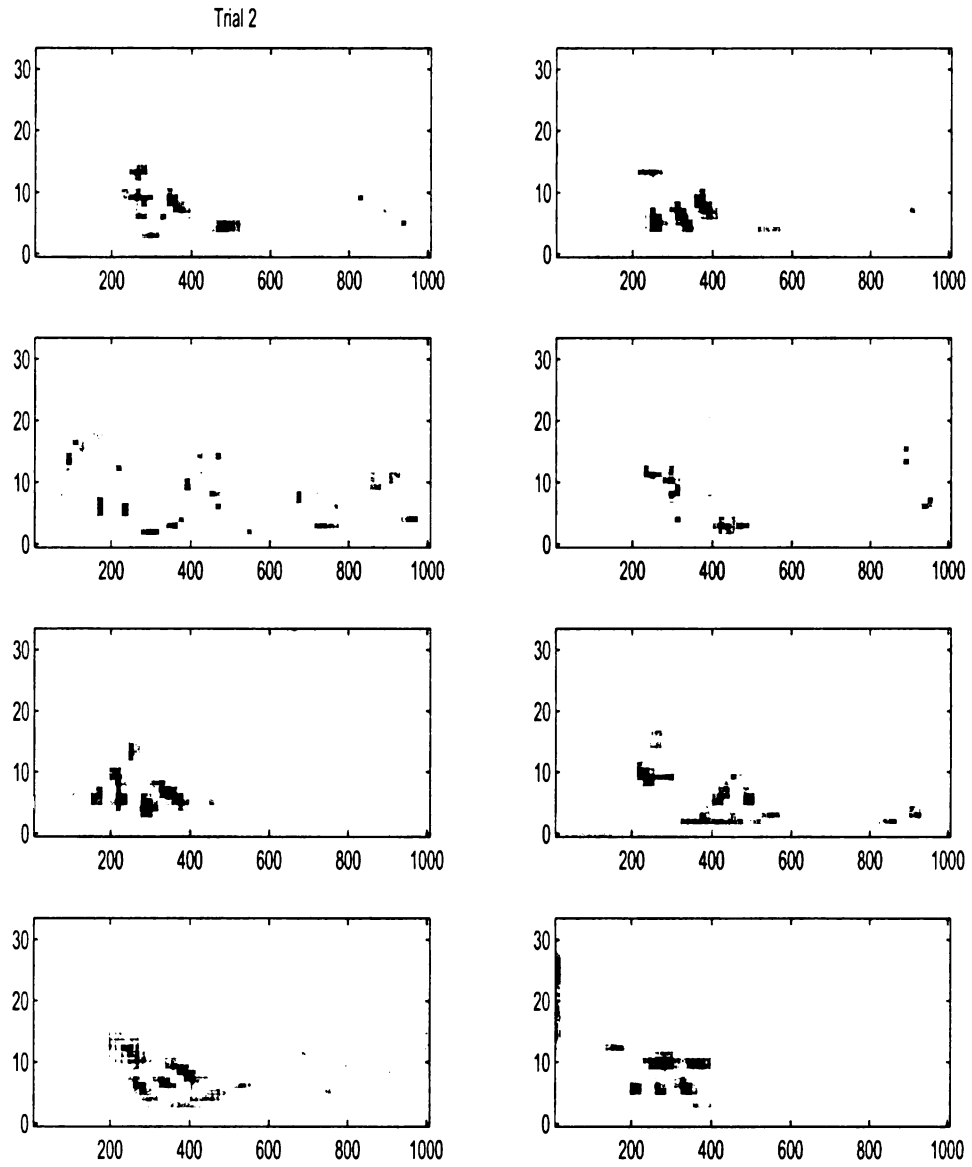


Figure 7.2. Single-trial results using 32 frequency bins: 8 extracted sources from the proposed UBSS method

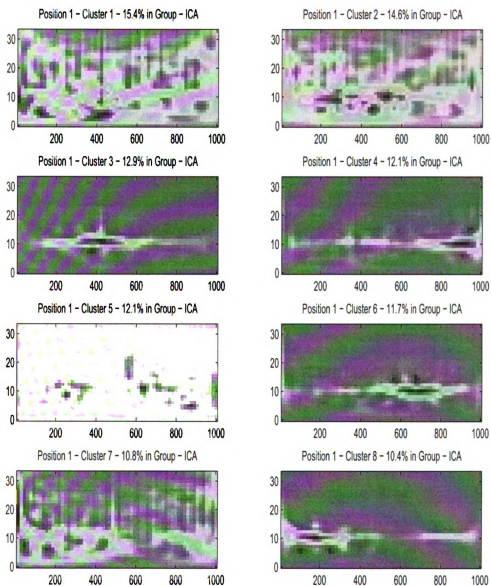


Figure 7.3. Results of component clustering over all single-trial results for stimulus position $u = 1$ using ICA

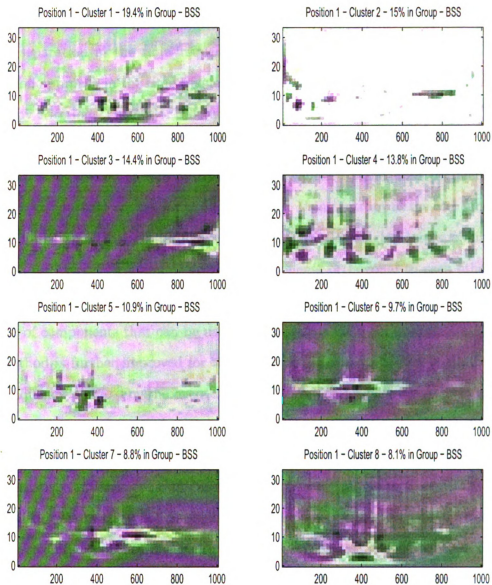


Figure 7.4. Results of component clustering over all single-trial results for stimulus position $u = 1$ using the proposed UBSS method

CHAPTER 8

CONCLUSIONS AND FUTURE WORK

In this dissertation, several problems regarding multichannel signal decomposition and source separation in the time-frequency domain are addressed. A new signal decomposition approach in the time-frequency domain is proposed based on the minimum entropy criterion. The major difference of the proposed approach from conventional component extraction or decomposition methods is the cost function. The cost function that is minimized is entropy on the time-frequency plane, thus producing compact components that are similar to Gabor logons. Using entropy as the cost function and adopting an adaptive filtering method to update the weights corresponding to each trial, we extract “minimum” entropy components orthogonal to each other. Experimental results show that the proposed approach is effective in determining a few number of components that can be used to represent a large set of data.

This method is further improved for the separation of non-stationary signals on the time-frequency plane. For the overdetermined case, the proposed algorithm assumes the disjointness of the underlying signals on the time-frequency plane. This assumption allows us to extract the sources through a N -dimensional Givens rotation. Using Jensen-Rényi divergence as the cost function, a steepest descent algorithm is implemented to update the rotation angles. Several examples are given to illustrate the performance of the proposed algorithm for synthesized and real life signals. Issues regarding convergence rate and robustness under noise are investigated. The performance of the algorithm is illustrated under noise and is compared to PCA and ICA as adapted to the time-frequency plane, and STFD. The results illustrate that maximizing the divergence on the time-frequency plane can separate sources that are disjoint in the time-frequency domain, and is better than the mutual information

cost function used in ICA in terms of fidelity to the original sources. The proposed method also outperforms STFD for high noise levels since it assumes the cross-terms between sources are negligible which effectively denoises the observed time-frequency matrix, and is apparently superior to PCA.

In the next part of this dissertation, the BSS problem is extended for the under-determined case. Most BSS algorithms are not suitable to be applied in this case. In this part, a two-stage sparse factorization approach is proposed for UBSS. The first stage is to determine the mixing matrix. The mixing matrix can be estimated using K -means clustering algorithm. The column vectors of the mixing matrix are cluster centers of normalized mixture vectors. The second stage is to estimate the sources. For a given mixing matrix, although there exist an infinite number of solutions in general, the sparse solution with minimum l_1 -norm is proven to be unique, which can be obtained by using linear programming methods. The results show that if the sources are sufficiently sparse in the time-frequency domain, the proposed approach can separate them effectively from their mixtures. Compared to PCA and ICA, the proposed method does not require the assumption that the sources have to be orthogonal or mutually independent.

The final part of the work focuses on the applications of the proposed UBSS algorithm on multichannel EEG/ERP recordings. Under the assumption that sources are sparse in the time-frequency domain, all single-trial components are extracted in the condition of the number of sources selected in advance. Then data reduction by clustering is performed over all single-trial results to extract components that represent the results. The performance of the proposed approach is compared with that of ICA. It is concluded that the proposed method is more effective at extracting well localized neuronal sources in time and frequency than ICA. These sources are shown to be more sparse, and distinct from each other.

Future work includes determining the convergence rates of the proposed algorithms

and investigating the effect of order α in the information-theoretic cost functions on the performance of the proposed algorithms. One problem for K -means clustering is the arbitrary selection of how many sources to extract. This is still an open question. If the number of extracted sources is less than the number of actual sources, some of actual sources can not be obtained; on the other hand, if the number of separated components is more than that of the actual sources, that means some components belonging to the same source are thought to be different sources. Thus, it would be more efficient to have the algorithm automatically select the number of sources. In addition, the requirement that the sources must be approximately disjoint limits the algorithms. If this assumption could be relaxed, results could be more reliable since the real sources may not be disjoint. Another area of future work is using signal synthesis methods to transform the extracted sources from the time-frequency domain to the time domain.

APPENDICES

In this appendix, the derivation for the steepest descent algorithm for $N = M = 3$ in Chapter 5 will be given explicitly. In this case, the matrix of mixture is

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \mathbf{X}_3 \end{bmatrix} = \begin{bmatrix} X_1(1) & \cdots & X_1(P) \\ X_2(1) & \cdots & X_2(P) \\ X_3(1) & \cdots & X_3(P) \end{bmatrix}, \quad (1)$$

and the extracted sources in the time-frequency domain are

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \mathbf{Y}_3 \end{bmatrix} = \begin{bmatrix} Y_1(1) & \cdots & Y_1(P) \\ Y_2(1) & \cdots & Y_2(P) \\ Y_3(1) & \cdots & Y_3(P) \end{bmatrix}. \quad (2)$$

We aim to find the optimal rotation transform $\mathbf{R}(\theta)$ under the Jensen-Rényi divergence criterion to obtain the signals $\mathbf{Y} = \mathbf{R}(\theta)\mathbf{X}$ that are disjoint on the time-frequency plane. From [82], we know that any 3-D rotation matrix can be written in the following form:

$$\mathbf{R}(\theta) = \mathbf{R}_1(\theta_1)\mathbf{R}_2(\theta_2)\mathbf{R}_3(\theta_3), \quad (3)$$

where

$$\mathbf{R}_1(\theta_1) = \begin{bmatrix} \cos(\theta_1) & \sin(\theta_1) & 0 \\ -\sin(\theta_1) & \cos(\theta_1) & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad (4)$$

$$\mathbf{R}_2(\theta_2) = \begin{bmatrix} \cos(\theta_2) & 0 & \sin(\theta_2) \\ 0 & 1 & 0 \\ -\sin(\theta_2) & 0 & \cos(\theta_2) \end{bmatrix}, \quad (5)$$

$$\mathbf{R}_3(\theta_3) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\theta_3) & \sin(\theta_3) \\ 0 & -\sin(\theta_3) & \cos(\theta_3) \end{bmatrix}, \quad (6)$$

and $\theta = [\theta_1, \theta_2, \theta_3]^T$. Hence the entries of \mathbf{Y} are

$$\begin{aligned} Y_1(i) = & (\sin(\theta_1) \cos(\theta_3) - \cos(\theta_1) \sin(\theta_2) \sin(\theta_3)) X_2(i) \\ & + (\sin(\theta_1) \sin(\theta_3) + \cos(\theta_1) \sin(\theta_2) \cos(\theta_3)) X_3(i) \\ & + \cos(\theta_1) \cos(\theta_2) X_1(i), \end{aligned} \quad (7)$$

$$\begin{aligned} Y_2(i) = & (\cos(\theta_1) \cos(\theta_3) + \sin(\theta_1) \sin(\theta_2) \sin(\theta_3)) X_2(i) \\ & + (\cos(\theta_1) \sin(\theta_3) - \sin(\theta_1) \sin(\theta_2) \cos(\theta_3)) X_3(i) \\ & - \sin(\theta_1) \cos(\theta_2) X_1(i), \end{aligned} \quad (8)$$

$$\begin{aligned} Y_3(i) = & -\cos(\theta_2) \sin(\theta_3) X_2(i) + \cos(\theta_2) \cos(\theta_3) X_3(i) \\ & - \sin(\theta_2) X_1(i), \end{aligned} \quad (9)$$

where $i = 1, 2, \dots, P$, and P is the length of the time-frequency vector. The gradients of \mathbf{Y} with respect to θ_1 are derived as follows:

$$\begin{aligned} \frac{\partial Y_1(i)}{\partial \theta_1} = & (\cos(\theta_1) \cos(\theta_3) + \sin(\theta_1) \sin(\theta_2) \sin(\theta_3)) X_2(i) \\ & + (\cos(\theta_1) \sin(\theta_3) - \sin(\theta_1) \sin(\theta_2) \cos(\theta_3)) X_3(i) \\ & - \sin(\theta_1) \cos(\theta_2) X_1(i), \end{aligned} \quad (10)$$

$$\begin{aligned} \frac{\partial Y_2(i)}{\partial \theta_1} = & (-\sin(\theta_1) \cos(\theta_3) + \cos(\theta_1) \sin(\theta_2) \sin(\theta_3)) X_2(i) \\ & - (\sin(\theta_1) \sin(\theta_3) + \cos(\theta_1) \sin(\theta_2) \cos(\theta_3)) X_3(i) \\ & - \cos(\theta_1) \cos(\theta_2) X_1(i), \end{aligned} \quad (11)$$

$$\frac{\partial Y_3(i)}{\partial \theta_1} = 0. \quad (12)$$

Similarly, we can derive the gradients of \mathbf{Y} with respect to θ_2 and θ_3 , respectively.

The cost function with the order $\alpha = 2$ is

$$J = \sum_{i=1}^2 \sum_{j=i+1}^3 J_{ij} = J_{12} + J_{13} + J_{23}, \quad (13)$$

where,

$$J_{ij} = \frac{\sum_{k=1}^P Y_i(k) Y_j(k)}{\sqrt{\left(\sum_{k=1}^P Y_i^2(k)\right) \left(\sum_{k=1}^P Y_j^2(k)\right)}} \quad (i < j). \quad (14)$$

The derivatives of the numerator and denominator of J_{ij} with respect to θ_l are given

as

$$\frac{\partial \left(\sum_{k=1}^P Y_i(k) Y_j(k) \right)}{\partial \theta_l} = \sum_{k=1}^P \left(\frac{\partial Y_i(k)}{\partial \theta_l} Y_j(k) + Y_i(k) \frac{\partial Y_j(k)}{\partial \theta_l} \right), \quad (15)$$

and

$$\begin{aligned} \frac{\partial \sqrt{\left(\sum_{k=1}^P Y_i^2(k)\right) \left(\sum_{k=1}^P Y_j^2(k)\right)}}{\partial \theta_l} &= \frac{1}{\sqrt{\left(\sum_{k=1}^P Y_i^2(k)\right) \left(\sum_{k=1}^P Y_j^2(k)\right)}} \times \\ &\quad \left[\left(\sum_{k=1}^P Y_i(k) \frac{\partial Y_i(k)}{\partial \theta_l} \right) \left(\sum_{k=1}^P Y_j^2(k) \right) + \left(\sum_{k=1}^P Y_i^2(k) \right) \left(\sum_{k=1}^P Y_j(k) \frac{\partial Y_j(k)}{\partial \theta_l} \right) \right], \end{aligned} \quad (16)$$

respectively, where $l = 1, 2, 3$. With equations (15) and (16), we get the gradient of

the pairwise cost function J_{ij} with respect to θ_l as

$$\begin{aligned} \frac{\partial J_{ij}}{\partial \theta_l} = & \frac{\partial \left(\sum_{k=1}^P Y_i(k) Y_j(k) \right) / \partial \theta_l}{\sqrt{\left(\sum_{k=1}^P Y_i^2(k) \right) \left(\sum_{k=1}^P Y_j^2(k) \right)}} - \frac{\sum_{k=1}^P Y_i(k) Y_j(k)}{\left(\sum_{k=1}^P Y_i^2(k) \right) \left(\sum_{k=1}^P Y_j^2(k) \right)} \times \\ & \partial \left(\sqrt{\left(\sum_{k=1}^P Y_i^2(k) \right) \left(\sum_{k=1}^P Y_j^2(k) \right)} \right) / \partial \theta_l. \end{aligned} \quad (17)$$

By summing up of the pairwise gradients of J_{ij} , we obtain the gradient of the total cost function J with respect to any rotation angle θ_l

$$\frac{\partial J}{\partial \theta_l} = \sum_{i=1}^2 \sum_{j=i+1}^3 \frac{\partial J_{ij}}{\partial \theta_l} = \frac{\partial J_{12}}{\partial \theta_l} + \frac{\partial J_{13}}{\partial \theta_l} + \frac{\partial J_{23}}{\partial \theta_l}. \quad (18)$$

This expression is used in updating the rotation angle in the steepest descent algorithm.

BIBLIOGRAPHY

BIBLIOGRAPHY

- [1] A.-J. van der Veen, S. Talvar, and A. Paulraj, "A subspace approach to blind space-time signal processing for wireless communication systems," *IEEE Trans. on Signal Processing*, vol. 45, pp. 173–190, 1997.
- [2] C. B. Papadias and A. Paulraj, "A constant modulus algorithm for multi-user signal separation in presence of delay spread using antenna arrays," *IEEE Signal Processing Letters*, vol. 4, pp. 178–181, 1997.
- [3] A. Rouxel, D. Le Guennec, and O. Macchi, "Unsupervised adaptive separation of impulse signals applied to EEG analysis," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2000, vol. 1, pp. 420–423.
- [4] David A. Peterson, James N. Knight, Michael J. Kirby, Charles W. Anderson, and Michael H. Thaut, "Feature selection and blind source separation in an EEG-based brain-computer interface," *EURASIP Journal on Applied Signal Processing*, vol. 19, pp. 3128–3140, 2005.
- [5] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. on Signal Processing*, vol. 41, pp. 3397–3415, 1993.
- [6] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Scientific Comp.*, vol. 20, pp. 33–61, 1999.
- [7] R. G. Baraniuk and D. L. Jones, "Wigner-based formulation of the chirplet transform," *IEEE Trans. on Signal Processing*, vol. 44, pp. 3129–3135, 1996.
- [8] B. D. V. Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Magazine*, 1988.
- [9] P. Comon, "Independent component analysis: A new concept?," *Signal Processing*, vol. 36, pp. 287–314, 1994.
- [10] H. Yang and S. Amari, "Adaptive online learning algorithms for blind separation: Maximum entropy and minimum mutual information," *Neural Computation*, vol. 9, pp. 1457–1482, 1997.
- [11] A. Cichocki and R. Unbehauen, "Robust neural networks with on-line learning for blind identification and blind separation of sources," *IEEE Trans. on Circuits and Systems*, vol. 43, no. 11, pp. 894–906, 1996.

- [12] M. Girolami and C. Fyfe, "Negentropy and kurtosis as projection pursuit indices provide generalised ICA algorithms," in *Advances in Neural Information Processing Systems Workshop*, 1996.
- [13] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, pp. 1129–1159, 1995.
- [14] B.A. Pearlmutter and L.C. Parra, "A context-sensitive generalization of ICA," in *Int. Conf. Neural Information Processing*, 1996.
- [15] Ivan Magrin-Chagnolleau, Geoffrey Durou, and Frederic Bimbot, "Application of time-frequency principal component analysis to text-independent speaker identification," *IEEE Trans. Speech and Audio Processing*, vol. 10, no. 6, pp. 371–378, 2002.
- [16] Matteo Gandetto, Marco Guainazzo, and Carlo S. Regazzoni, "Use of time-frequency analysis and neural networks for mode identification in a wireless software-defined radio approach," *EURASIP Journal on Applied Signal Processing*, vol. 12, pp. 1778–1790, 2004.
- [17] James R. Berriman, David A. Hutchins, Adrian Neild, Tat Hean Gan, and Phil Purnell, "The application of time-frequency analysis to the air-coupled ultrasonic testing of concrete," *IEEE Trans. Ultrasonics, Perroelectrics, and Frequency Control*, vol. 53, no. 4, pp. 768–776, 2006.
- [18] I. J. Chant and L. M. Hastie, "Time-frequency analysis of magnetotelluric data," *Geophysical Journal International*, vol. 111, no. 2, pp. 399–413, 1992.
- [19] Massimiliano Pieraccini, Guido Luzi, Linhsia Noferini, Daniele Mecatti, and Carlo Atzeni, "Joint time-frequency analysis for investigation of layered masonry structures using penetrating radar," *IEEE Trans. Geoscience and Remote Sensing*, vol. 42, no. 2, pp. 309–317, 2004.
- [20] E. P. Wigner, "On the quantum correction for thermodynamic equilibrium," *Physical Review*, vol. 40, pp. 749–759, 1932.
- [21] J. Ville, "Theorie et applications de la notion de signal analytique," *Cables et Transmission*, vol. 2A, pp. 61–74, 1948.
- [22] J. E. Moyal, "Quantum mechanics as a statistical theory," in *Proc. Camb. Phil. Soc.*, 1949, vol. 45, pp. 99–124.

- [23] L. Cohen, *Time-Frequency Analysis*, Prentice Hall, New Jersey, 1995.
- [24] T. A. C. M. Claasen and W. F. G. Mecklenbrauker, "The Wigner distribution – a tool for time-frequency signal analysis – Part III: Relations with other time-frequency signal transformations," *Philips Journal of Research*, vol. 35, pp. 372–389, 1980.
- [25] A. J. E. M. Janssen, "On the locus and spread of pseudo-density functions in the time-frequency plane," *Philips Journal of Research*, vol. 37, pp. 79–110, 1982.
- [26] J. Jeong and W. J. Williams, "Kernel design for reduced interference distributions," *IEEE Trans. on Signal Processing*, vol. 40, pp. 402–412, 1992.
- [27] P. Loughlin, J. Pitton, and L. E. Atlas, "Bilinear time-frequency representations: new insights and properties," *IEEE Trans. on Signal Processing*, vol. 41, pp. 750–767, 1993.
- [28] S. Oh and R. J. Marks II, "Some properties of the generalized time-frequency representation with cone shaped kernel," *IEEE Trans. on Signal Processing*, vol. 40, pp. 1735–1745, 1992.
- [29] S. Oh, R. J. Marks II, L. E. Atlas, and J. A. Pitton, "Kernel synthesis for generalized time-frequency distributions using the method of projection onto convex sets," in *SPIE-Advanced Signal Processing Algorithms*, 1990, vol. 1348, pp. 197–207.
- [30] P. Flandrin, "Some features of time-frequency representations of multicomponent signals," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 1984, pp. 41B.4.1–4.4.
- [31] P. Flandrin and F. Hlawatsch, "Signal representations, geometry and catastrophes in the time-frequency plane," *Mathematics in Signal Processing*, pp. 3–14, 1987.
- [32] F. Hlawatsch, "Interference terms in the wigner distribution," *Digital Signal Processing-84*, pp. 363–367, 1984.
- [33] H.-I. Choi and W. J. Williams, "Improved time-frequency representation of multicomponent signals using exponential kernels," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 37, no. 6, pp. 862–871, 1989.

- [34] Y. Zhao, L. E. Atlas, and R. J. Marks, "The use of cone-shaped kernels for generalized time-frequency representations of nonstationary signals," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 38, pp. 1084–1091, 1990.
- [35] J. Jeong and W. J. Williams, "A new formulation of generalized discrete-time time-frequency distributions," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 1991, pp. 3189–3192.
- [36] G. S. Cunningham and W. J. Williams, "High-resolution signal synthesis for time-frequency distributions," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 1993, vol. 4, pp. 400–403.
- [37] E. GokGay and J. C. Principe, "New clustering evaluation function using Rényi's information potential," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2000, pp. 3490–3493.
- [38] M. Ben-Bassat and J. Raviv, "Rényi entropy and the probability of error," *IEEE Trans. on Info. Theory*, vol. 24, no. 3, pp. 324–330, May 1978.
- [39] M. Basseville, "Distance measures for signal processing and pattern recognition," *Signal Processing*, vol. 18, no. 4, pp. 349–369, Dec. 1989.
- [40] I. Csiszar, "Information-type distance measures and indirect observations," *Stud. Sci. Math. Hungar.*, vol. 2, pp. 299–318, 1967.
- [41] R. G. Baraniuk, P. Flandrin, A. J. E. M. Janssen, and O. Michel, "Measuring time-frequency information content using the Rényi entropies," *IEEE Trans. on Info. Theory*, vol. 47, no. 4, pp. 1391–1409, May 2001.
- [42] W. J. Williams, M. Brown, and A. Hero, "Uncertainty, information and time-frequency distributions," in *SPIE-Advanced Signal Processing Algorithms*, 1991, vol. 1556, pp. 144–156.
- [43] A. Rényi, "On measures of entropy and information," in *Proceedings 4th Berkeley Symp. Math. Stat. and Prob.*, 1961, vol. 1, pp. 547–561.
- [44] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine*, vol. 2, pp. 559–572, 1901.
- [45] E. Oja, *Subspace Methods of Pattern Recognition*, Research Studies Press, England, and Wiley, USA, 1983.

- [46] K. I. Diamantaras and S. Y. Kung, *Principal Component Neural Networks: Theory and Applications*, Wiley, 1996.
- [47] M. Wax and T. Kailath, "Detection of signals by information-theoretic criteria," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 33, pp. 387–392, 1985.
- [48] J. Karhunen, A. Cichocki, W. Kasprzak, and P. Pajunen, "On neural blind separation with noise suppression and redundancy reduction," *Int. Journal of Neural Systems*, vol. 8, no. 2, pp. 219–237, 1997.
- [49] B. Ans, J. Herault, and C. Jutten, "Adaptive neural architectures: detection of primitives," in *Proc. COGNITIVA '85*, 1985, pp. 593–597.
- [50] T. W. 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*, vol. 6(4), pp. 87–90, 1999.
- [51] L. Zhukov, D. Weinstein, and C. Johnson, "Independent component analysis for EEG source localization," *IEEE Eng. Med. Biol. Mag.*, vol. 19, pp. 87–96, 2000.
- [52] Erik Visser and Te-Won Lee, "Blind source separation in mobile environments using a priori knowledge," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2004, vol. 3, pp. 893–896.
- [53] Guillaume Gelle, Maxime Colas, and Christine Servire, "Blind source separation: A new pre-processing tool for rotating machines monitoring," *IEEE Trans. Instrumentation and Measurement*, vol. 52, no. 3, pp. 790–795, 2003.
- [54] Z. He, L. Yang, J. Liu, Z. Lu, C. He, and Y. Shi, "Blind source separation using clustering-based multivariate density estimation algorithm," *IEEE Trans. on Signal Processing*, vol. 48, pp. 575–579, 2000.
- [55] D.-T. Pham, "Blind separation of instantaneous mixture of sources based on order statistics," *IEEE Trans. on Signal Processing*, vol. 48, pp. 363–375, 2000.
- [56] B. Torresani, "Wavelets associated with representations of the affine Weyl-Heisenberg group," *Journal Math. Physics*, vol. 32, pp. 1273–1279, 1991.
- [57] Y.C. Pati, R. Rezaifar, and P.S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition,"

- in *Proc. of the 27th Annual Asilomar Conference on Signals, Systems and Computers*, 1993.
- [58] A. Belouchrani and M. G. Amin, "Blind source separation based on time-frequency signal representations," *IEEE Trans. on Signal Processing*, vol. 46, pp. 2888–2897, 1998.
 - [59] Abdeldjalil Aissa-El-Bey, Nguyen Linh-Trung, Karim Abed-Meraim, Adel Belouchrani, and Yves Grenier, "Underdetermined blind separation of nondisjoint sources in the time-frequency domain," *IEEE Trans. on Signal Processing*, vol. 55, pp. 897–907, 2007.
 - [60] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. on Signal Processing*, vol. 52, pp. 1830–1847, 2004.
 - [61] S. Qian and D. Chen, "Discrete Gabor transform," *IEEE Trans. on Signal Processing*, vol. 41, pp. 2429–2438, 1993.
 - [62] D. Gabor, "Theory of communication," *J. of IEE*, vol. 93, no. 26, pp. 429–457, 1946.
 - [63] J. Wexler and S. Raz, "Discrete Gabor expansions," *Signal Processing*, vol. 21, no. 3, pp. 207–221, 1990.
 - [64] S. Mann and S. Haykin, "The chirplet transform—physical considerations," *IEEE Trans. on Signal Processing*, vol. 43, no. 11, pp. 2745–2761, 1995.
 - [65] R. G. Baraniuk, P. Flandrin, A. J. E. M. Janssen, and O. Michel, "Measuring time-frequency information content using the Rényi entropies," *IEEE Trans. on Info. Theory*, vol. 47, no. 4, pp. 1391–1409, May 2001.
 - [66] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*, Prentice Hall, 1985.
 - [67] A. Belouchrani, K. Abed-Meraim, J-F. Cardoso, and E. Moulines, "A blind source separation technique using second-order statistics," *IEEE Trans. on Signal Processing*, vol. 45, no. 2, pp. 434–444, 1997.
 - [68] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley and Sons, 2001.
 - [69] J. F. Cardoso, "Infomax and maximum likelihood for blind source separation," *IEEE Signal Processing Letters*, vol. 4, pp. 112–114, 1997.

- [70] K.E. Hild II, D. Erdogmus, and J. C. Principe, "Blind source separation using Rényi's mutual information," *IEEE Signal Processing Letters*, vol. 8, pp. 174–176, 2001.
- [71] Y. Q. Li, A. Cichocki, and S. Amari, "Sparse component analysis for blind source separation with less sensors than sources," in *Proc. 4th Int. Symposium on Independent Component Analysis and Blind Signal Separation*, 2003, pp. 89–94.
- [72] D. D. Lee and H. S. Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.
- [73] E. Weinstein, M. Feder, and A. Oppenheim, "Multi-channel signal separation by decorrelation," *IEEE Trans. on Speech and Audio Processing*, vol. 1, pp. 405–413, 1993.
- [74] L. Parra and C. Spence, "Convolutional blind separation of non-stationary sources," *IEEE Trans. on Speech and Audio Processing*, vol. 8, pp. 320–327, 2000.
- [75] H. Sahlin and H. Broman, "MIMO signal separation for FIR channels: a criterion and performance analysis," *IEEE Trans. on Signal Processing*, vol. 48, pp. 642–649, 2000.
- [76] C. T. Ma, Z. Ding, and S. F. Yau, "A two-stage algorithm for MIMO blind deconvolution of nonstationary colored signals," *IEEE Trans. on Signal Processing*, vol. 48, pp. 1187–1192, 2000.
- [77] A. Belouchrani, K. Abed-Meraim, M. G. Amin, and A. M. Zoubir, "Blind separation of nonstationary sources," *IEEE Signal Processing Letters*, vol. 11, no. 7, 2004.
- [78] M. G. Amin and Y. Zhang, "Signal averaging of time-frequency distributions for signal recovery in uniform linear arrays," *IEEE Trans. on Signal Processing*, vol. 48, no. 10, pp. 2892–2902, 2000.
- [79] O. Michel, R. G. Baraniuk, and P. Flandrin, "Time-frequency based distance and divergence measure," in *Proc. IEEE Int. Symp. Time-Frequency and Time-Scale Analysis*, 1994, pp. 64–67.
- [80] S. Aviyente, "Information processing on the time-frequency plane," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2004, vol. 2, pp. 617–620.

- [81] L.-T. Nguyen, A. Belouchrani, K. Abed-Meraim, and B. Boashash, "Separating more sources than sensors using time-frequency distributions," in *Proc. Int. Symposium on Signal Processing and its Applications*, 2001, pp. 583–586.
- [82] F. D. Murnaghan, *The Unitary and Rotation Groups*, Spartan Books, Washington D.C., 1962.
- [83] S. Haykin, *Introduction to Adaptive Filters*, MacMillan, New York, 1984.
- [84] A. Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. on Neural Networks*, vol. 10, no. 3, pp. 626–634, May 1999.
- [85] P. J. Huber, "Projection pursuit," *Ann. Statist.*, vol. 13, no. 2, pp. 435–475, 1985.
- [86] I. F. Gorodnitsky and B. D. Rao, "Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm," *IEEE Trans. on Signal Processing*, vol. 45, pp. 600–616, 1997.
- [87] D. L. Donoho and M. Elad, "Maximal sparsity representation via l^1 minimization," in *Proc. Nat. Acad. Sci.*, 2003, pp. 2197–2202.
- [88] M. Zibulevsky and B. A. Pearlmutter, "Blind source separation by sparse decomposition," *Neural Computations*, vol. 13(4), pp. 863–882, 2001.
- [89] W. Lu and J. C. Rajapakse, "Approach and applications of constrained ICA," *IEEE Trans. Neural Netw.*, vol. 16(1), pp. 203–212, 2005.
- [90] L. Zhang, A. Cichocki, and S. Amari, "Self-adaptive blind source separation based on activation functions adaptation," *IEEE Trans. Neural Netw.*, vol. 15(2), pp. 233–244, 2004.
- [91] S. A. Cruces-Alvarez, A. Cichocki, and S. Amari, "From blind signal extraction to blind instantaneous signal separation: Criteria, algorithms, and stability," *IEEE Trans. Neural Netw.*, vol. 15(4), pp. 859–873, 2004.
- [92] M. Girolami, "A variational method for learning sparse and overcomplete representations," *Neural Computations*, vol. 13(11), pp. 2517–2532, 2001.
- [93] M. Zibulevsky, P. Kisilev, Y. Y. Zeevi, and B. A. Pearlmutter, "Blind source separation via multinode sparse representation," *NIPS-2001*.

- [94] Y. Q. Li, A. Cichocki, and S. Amari, "Sparse representation and blind source separation," *Neural Computations*, vol. 16(6), pp. 1193–1234, 2004.
- [95] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley, New York, 2001.
- [96] M. S. Lewicki and T. J. Sejnowski, "Learning overcomplete representations," *Neural Computations*, vol. 12(2), pp. 337–365, 2000.
- [97] K. Kiviluoto and E. Oja, "Independent component analysis for parallel financial time series," in *Proc. ICONIP'98*, 1998, vol. 2, pp. 895–898.
- [98] R. Vigario, J. Sarela, V. Jousmaki, and E. Oja, "Independent component analysis in decomposition of auditory and somatosensory evoked fields," in *Proc. Int. Workshop on Independent Component Analysis and Signal Separation (ICA '99)*, 1999, pp. 167–172.
- [99] S. J. Luck, *An introduction to the event-related potential technique*, The MIT Press, 2005.
- [100] J. N. Demos, *Getting started with neurofeedback*, W.W. Norton & Company, 2005.
- [101] O. Faugeras, F. Clement, R. Deriche, R. Keriven, T. Papadopoulos, and J. Roberts, "The inverse EEG and MEG problems: The adjoint state approach I: The continuous case," *INRIA Research Report 3673*, 1999.
- [102] H. Shevrin, J. A. Bond, L. A. Brakel, R. K. Hertel, and W. J. Williams, *Conscious and Unconscious Processes: Psychodynamic, Cognitive and Neurophysiological Convergences*, New York: Guilford, 1996.
- [103] H. Shevrin, W. J. Williams, R. E. Marshall, R. K. Hertel, J. A. Bond, and L. A. Brakel, "Event-related potential indicators of the dynamic unconscious," *Consciousness Cogn*, vol. 1, pp. 340–366, 1992.
- [104] W. J. Williams, H. P. Zaveri, and J. C. Sackellares, "Time-frequency analysis of electrophysiology signals in epilepsy," *IEEE Trans. Eng. Med. Biol.*, pp. 133–143, 1995.
- [105] <http://www.sccn.ucsd.edu/eeglab>.

MICHIGAN STATE UNIVERSITY LIBRARIES



3 1293 03062 9822