Adaptive Blind Signal Separation

By
Chi-chiu CHEUNG

Supervised By :
Prof. Lei XU

Submitted to the Division of Computer Science & Engineering
in partial fulfillment of the requirements for the
degree of Master of Philosophy
at the
Chinese University of Hong Kong
June 1997

Adaptive Blind Signal Separation

submitted by

Chi-chiu CHEUNG

for the degree of Master of Philosophy
at the Chinese University of Hong Kong

Abstract

The blind signal separation problem consists of recovering a set of statistically independent source signals from observed mixtures of them, while the source signals and the mixing process are unknown. The problem with linear, instantaneous mixing can be formulated as the Independent Component Analysis (ICA) problem. Recently, a general information-theoretic ICA scheme based on the Bayesian-Kullback YING-YANG learning theory (Xu 1995,96,97) is proposed by Xu & Amari (1996). This thesis aims at providing analysis for the information-theoretic ICA scheme, particularly in the relationship between nonlinearity and separation capability.

Firstly, a number of properties of the cost function used in the information-theoretic ICA scheme, including continuity, singularity, asymptotic properties, etc, are discussed. These properties are essential for constructing convergence proofs in later sections.

Secondly, the information-theoretic ICA scheme with cubic nonlinearity on two channels of signals is analyzed in details. We have proved that the information-theoretic ICA algorithms with cubic nonlinearity can separate two 'globally sub-Gaussian' sources and cannot separate two 'globally super-Gaussian' sources. A theorem on the global convergence is provided. Some partial results on the three-channel case have also been worked out. The theoretical works are accompanied by experimental verification. This investigation provides an interesting insight in the role of nonlinearity in adaptive ICA algorithms.

Then, we analyze the separation capability of several nonlinearities. We experimentally test the separation capability of the cubic root nonlinearity and proved another theorem of the convergence of the information-theoretic ICA algorithm using the cubic nonlinearity in one channel and linearity in the second channel. Using the comparison of the separation capabilities of the reversed sigmoids used by Bell & Sejnowski (1995), the cubic nonlinearity and the above two cases, we argue that a 'loose match-

ing' between the nonlinearity and distribution of the sources is needed for successful separation.

Meanwhile, an implementation technique using mixture of densities is proposed by Xu *et al* (1997) to achieve the matching of nonlinearity to source distribution. This thesis provides some experimental results and analysis. The experiment results support that the mixture of densities can adapt and separate sources of any density. We also test and suggest that mixture of two densities with only centers of the components changeable may be enough for achieving the loose matching of the nonlinearity to any source density.

In addition, this thesis provides an investigation on the possibility of constructing adaptive ICA algorithms from the Bayesian YING YANG learning scheme with non-Kullback separation functionals. An algorithm based on Positive Convex divergence is derived and experiments show that it possesses robustness against outliers in the sources.

Acknowledgments

I would like to give special thanks to my supervisor, Prof. Lei Xu, who teaches me so much in the academic aspect, gives me guidance in the research and tells us how to be a successful person. I would like to acknowledge Prof. Jiong Ruan, who takes part in the collaboration of the work. I would also like to give thanks to other members in Prof. Xu's research group, including Dr. Xin-sheng Zhang, Dr. Lei, Dr. Yu-ping Wang, Dr. Tai-jun Wang, Mr. Yiu-ming Cheung, Mr. Wai-man Leung, Mr. Wing-kai Lam, and Mr. Zhi-bin Lai who participate in helpful discussions with me. At a personal level, I would like to thank numerous colleagues and friends of mine and my family, who give me so much support and helps during my M. Phil. study.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The main objective of this thesis is to discuss the relationship between nonlinearity and separation capability on the information-theoretical ICA scheme [74] for the linear, instantaneous blind signal separation problem. Both theoretical analysis and computer simulation are included in the work.

In this chapter, we will introduce the blind signal separation problem and the contribution of the thesis. Section 1.1 gives brief introduction to the blind signal separation problem and the present research development in this topic. Section 1.2 lists out the contribution of this thesis for an easy glance. Section 1.3 discusses the applications of the problem. The organization of the thesis is sketched in Section 1.4.

1.1 The Blind Signal Separation Problem

The blind signal separation problem generally consists of recovering a set of independent source signals solely from a set of mixtures of them (Figure 1.1). The mathematical definition of the problem will be presented in Chapter 2. Two sub-cases of the problem are often discussed, namely the convolutive and the instantaneous blind signal separation problem. In situations like processing mixtures of speeches in room acoustic environment, due to multi-path effect, the mixing is dynamic and have to be modeled by convolution. In situations like analysis of electro-encephalographic (EEG) signals, the mixing can be satisfactorily modeled to be instantaneous and linear. Under such condition the blind signal separation problem can be formulated as the Independent Component Analysis (ICA) problem - the extraction of independent components (source signals) from a set of multi-variant data (observed signals). In the literature, the names 'Blind Signal Separation', 'Blind Separation of Sources','Blind Source Separation', are often used to refer immediately to the problem with instantaneous mixing,

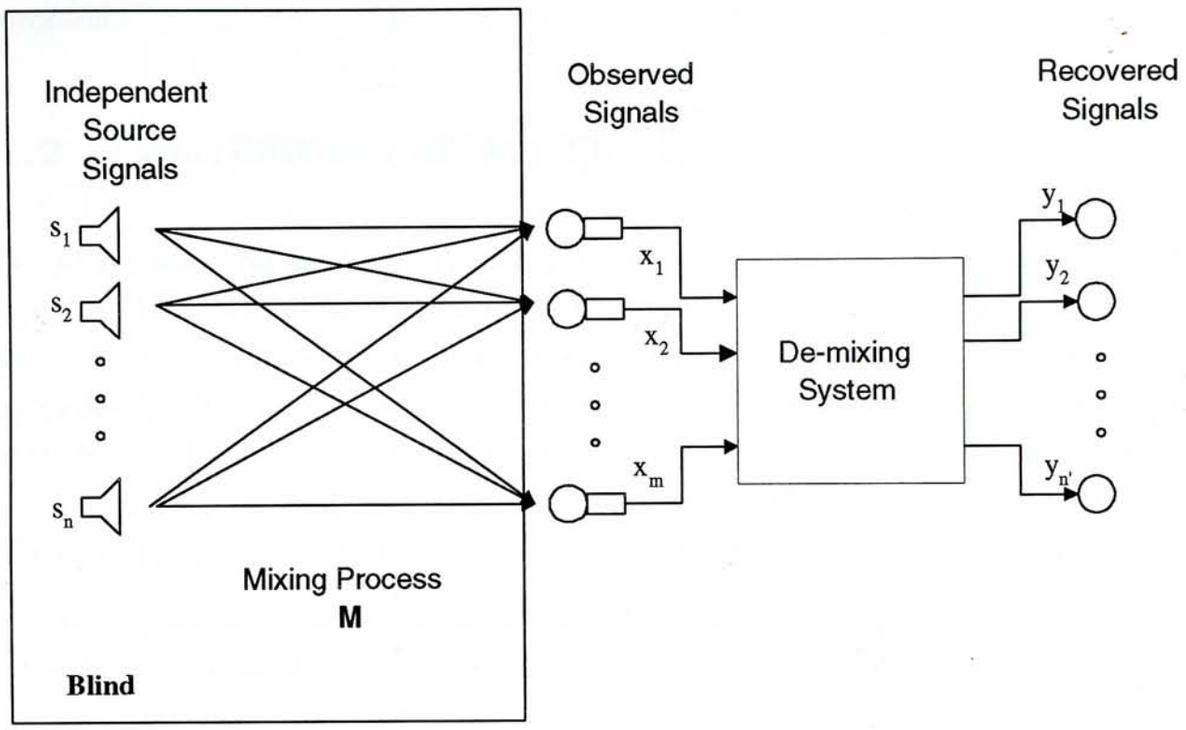


Figure 1.1: The general blind signal separation problem.

rather than the more general problem.

Strictly speaking, the word 'blind' means that we only have the mixed signals available and know nothing about the source signals and mixing process, except the statistical independence assumption of the source signals. The 'blind' problem formulation takes the advantage that no model for the source signals or mixing process is needed. Such problem formulation is very suitable for analyzing or processing observed signals that are supposed to be mixtures of some 'more basic' sources, like those situations to be discussed in the application section. However, in most literature the number of sources is assumed to be known and the problem is actually not strictly blind.

This problem has become an emerging topic in this decade, probably after C. Jutten and J. Herault published the seminal paper [33]. The blind signal separation problem becomes popular not only because it has a number of valuable applications, but also because it involves interesting theoretical problems in high order statistics and nonlinear systems. The problem is currently so hot that special sessions for this problem are held in various conferences and journals in recent years. Large proportion of the literature focused on the problem with instantaneous, linear mixing, while smaller proportion of papers have been written on convolutive mixture and other aspects of the

problem.

1.2 Contributions of this Thesis

Xu and Amari [74] suggested a general information-theoretic ICA scheme from the Bayesian-Kullback YING-YANG learning scheme [67, 69, 70, 72]. We realized that the choice of nonlinearity in the algorithm is crucial to the separation capability, or more specifically, one nonlinearity can perform separation on sources with some class of distribution, but not on sources with any distribution, and different nonlinearities can perform separation on sources with different classes of distribution. Hence we focus on the investigation in the relationship between nonlinearity and separation capability.

The contributions of this thesis is listed below for an easy glance, with the credits by individuals identified.

- I worked out a theorem on the global convergence of the information-theoretic ICA algorithms with cubic nonlinearity on two channels of sources. I proved the theorem by investigating the configuration of the cost function scalar field in the parameter space. I provide two methods in the investigation of stabilities of the equilibrium points of the algorithm, one involves the analysis of the Hessian matrix and the other one works through direct comparison of the cost function value and the simplicity of the 2-channel problem. I have also performed experiments and the results are consistent to the theorem.

- Ruan and I investigated the 3-channel case of the information-theoretic ICA algorithms with cubic nonlinearity. Some equilibrium points are determined and the condition on the stability of the correct solution is determined.

- I observed that $h_i(y_i)$ nonlinearities (see Chapter 6) of different shapes have different separation capability on sources of different distribution. Xu [76] suggested the concept of 'loose matching' between $g_i(y_i)$ nonlinearity and the source density for successful source separation. We then investigate the nonlinearity and separation capability. I obtained the experimental results on several nonlinearities that support the argument.

- Ruan and I obtained a theorem on the convergence of the 2-channel information-theoretic ICA algorithm with one channel using the cubic nonlinearity and the other channel using linearity. The theorem shows that the channel with cubic nonlinearity always recovers the source with flatter probability density function

(pdf). This result further support the intuition that the nonlinearity must have some 'loose matching' with source density for signal separation to be successful.

- I performed experiments and analysis on the implementation of the information-theoretic ICA scheme with mixture of densities, which was suggested by Xu *et al* [75]. The experiment supports that the mixture of densities can adapt sources with any distribution and perform separation on them. I seek the simplest mixture of densities and suggest that the mixture of two densities with only centers of components changeable may be sufficient to achieve the loose matching of the nonlinearity to any source distribution.

- Xu and I investigated the construction of adaptive ICA algorithms from ICA principles which is based on the use of non-Kullback separation functionals suggested by Xu [72, 73]. We constructed an adaptive algorithm from the Positive Convex divergence but we cannot construct adaptive algorithms from the L_P divergence and de-correlation index. I performed the experiments suggested by Xu and the experimental result is consistent to Xu's expectation that the ICA algorithms derived from Positive Convex divergence has robustness on large-magnitude 'outliers' in the sources [73].

1.3 Applications of the Problem

One importance of the blind signal separation problem is to account for the 'cocktail party effect'. In a noisy cocktail party, human ears actually receive mixtures of a large number of voices from different locations. However, human beings are capable of concentrating the attention on listening to one person, and occasionally switching to another person. Apart from factors from the cognitive process in the brain, it was found that there should be some mechanism in the brain to perform localization and separation of the sound source from the difference between the received signals by the two ears (see, for example, [11]). Similar effect occurs in the odor system of animals. Some animals are able to localize the source of odor so as to hunt food and flee from danger. The blind signal separation problem tries to find plausible mechanisms for the brain.

The separation of noise from speech signals has important real world applications. Such technique may be used to reduce noise in hand-free speaker phone systems, especially those in automobile and the helmet communication system in fighter aircrafts, where the noise from the engine is extremely annoying. Application of noise reduc-

tion to speech recognition system may lead to significant improvement in recognition rate [47].

In data communication, signals in different channels may have 'cross-talk' to each other, i.e. the signal in one channel physically affects, or leaks to, the channels adjacent to it. In case that the signals are statistically independent to each other, the blind signal separation technique may be applied to recover the signals.

Independent component analysis can be applied to any situation that the underlying independent components are to be found. In the investigation of brain signals, different parts of the brain emit more or less independent signals but EEG electrodes on the skin can only receive (almost instantaneous) mixtures of them. Hence ICA is used to localize and extract independent signals of the brain and some encouraging results have been found [43, 50]. Similar technique can be used to process electrocardiographic (ECG) signals, from which special ECG patterns reflecting particular heart diseases can be identified for diagnosis. The separation of source is useful in observing fetus's cardiac signal from a mixture with the mother's cardiac signal [66]. Works have been done on processing natural image by ICA and it was found that edges are independent components of natural image [7]. ICA has also been applied to natural sound to elucidate the higher-order structure of it [8].

1.4 Organization of the Thesis

In Chapter 2, the blind signal separation problem will be presented mathematically. At the beginning the general problem will be introduced. We refine the problem to convolutive mixture, then to instantaneous linear mixing and the independent component analysis problem will be formulated. The assumptions and definition of the problem used in the thesis will be presented.

Chapter 3 will present a literature survey on the blind signal separation problem. Both the development of the approaches to the ICA problem and the separation of convolutive mixture will be reviewed.

Chapter 4 reviews the information-theoretic ICA scheme proposed by Xu and Amari [74]. The construction of the information-theoretic ICA scheme from the Bayesian-Kullback YING-YANG learning scheme [67, 69, 70, 72] and the derivation of the algorithm [74] will be reviewed.

Chapter 5 to 9 provide the main contribution of this thesis. In Chapter 5, theoret-

ical analysis of the information-theoretic ICA scheme on several aspects common to a large class of nonlinearities will be presented as lemmas and corollaries.

In Chapter 6, the convergence behavior of the information-theoretic ICA algorithm with cubic nonlinearity will be theoretically analyzed in details. A theorem on global convergence will be provided, accompanied by experimental verification. This analysis serves as a case study and demonstrates a method for the investigation of nonlinearity and separation capability. Some theoretical and experimental results on the 3-channel case are also provided.

In Chapter 7, The relationship between the nonlinearity used in the algorithm and the distribution of sources that it can separate will be investigated. Several cases with different nonlinearities will be discussed with experimental findings or verification. The investigation results in the suggestion of 'loose matching' between the nonlinearity and source distribution as the requirement for successful separation.

In Chapter 8, the idea of implementation of the information-theoretic ICA scheme with mixture of densities proposed by Xu *et al* [75] will be reviewed. The algorithm will be derived and experiments will be presented. Then we present the search for the simplest mixture of densities that is sufficient to adapt any source.

In Chapter 9, the suggestion of using non-Kullback separation functionals in ICA scheme by Xu [72, 73] will be introduced. Then we investigate the possibility of deriving ICA algorithm from the new cost functions. Finally, experiments on one algorithm derived from the Positive Convex divergence will demonstrate its robustness against outliers in the sources.

At last, the conclusion of this thesis will be given in Chapter 10.

Chapter 2

The Blind Signal Separation Problem

The definition of the blind signal separation problem is given mathematically in this chapter. Section 2.1 presents the general formulation for the physical problem. Section 2.2 formulates the case that models the mixing process to be linear and convolutive. Section 2.3 formulates the case that confines the mixing process to be linear and instantaneous. Finally, we give the definition of the linear, instantaneous blind signal separation problem formulated as the Independent Component Analysis (ICA) problem in Section 2.4 and list out the assumptions used in this thesis.

2.1 The General Blind Signal Separation Problem

The name 'blind signal separation' can generally be interpreted, from its wording, as the problem of separating, or *recovering*, the source signals from mysterious mixtures of them, while the mixing process and the source signals are unknown (Figure 1.1). Suppose there are n *channels* of *source signals*, or simply called *sources*, $\mathbf{s}(t) = [s_1(t), \ldots, s_n(t)]^T, t \in \mathbb{R}^+$, where $[\ldots]^T$ denotes matrix transpose. m *sensors*, or *receivers* are used in the system to pick up the *observed signals*, or *mixed signals* $\mathbf{x}(t) = [x_1(t), \ldots, x_m(t)]^T, t \in \mathbb{R}^+$. The mixing process is represented by a vector function:

$$\mathbf{x}(t) = \mathbf{M}(\mathbf{s}(t')) + \boldsymbol{\varepsilon}(t), \qquad t' \in [0, t]) \tag{2.1}$$

or

$$x_i(t) = \mathsf{M}_i(\mathbf{s}(t')) + \varepsilon_i(t), \qquad t' \in [0, t], \qquad i = 1, ..., m \tag{2.2}$$

where $\boldsymbol{\varepsilon}(t) = [\varepsilon_1(t), ..., \varepsilon_m(t)]^T$ are noise contamination of the sensors. The physical blind signal separation problem is to obtain a set of *recovered signals*, or *output signals*, $\mathbf{y}(t) = [y_1(t), \ldots, y_{n'}(t)]^T, t \in \mathbb{R}^+$ which is as similar to the source signals as possible.

In digital processing, we pick up the observed signals at discrete time and the observed signals become $\mathbf{x}(k) = [x_1(k), \ldots, x_m(k)]^T$, $k = 1, 2, \ldots$, and $k = t/\Delta t$ where Δt is the sampling interval. The sources are assumed to be of discrete time, $\mathbf{s}(k) = [s_1(k), \ldots, s_n(k)]^T$, too. Hence, the blind signal separation problem become to obtain the recovered signals $\mathbf{y}(k) = [y_1(k), \ldots, y_{n'}(k)]^T$ which is as similar to the source signals as possible.

2.2 Convolutive Linear Mixing Process

In situations like the separation of speech signals in room acoustic environment and super-sonic radar signal processing, multi-path effect occurs because sound wave travels at relatively low speed and the signal from one source reflected by different obstacles in the environment arrive one sensor at different times. The cross-coupling effect between channels in data communication are dynamic and have similar effect.

In these cases, the mixing process is modeled as a Linear Time Invariant (LTI) system:

$$x_i(k) = \sum_{j=1}^{n} \sum_{k'=0}^{l} H_{ij}(k') s_j(k - k') + \varepsilon_i(k) \qquad i = 1, ..., m \tag{2.3}$$

where $H_{ij}(k'), k' = 0, \ldots, l$ is the (discrete) transfer function from source j to sensor i.

The whole system:

$$\mathbf{x}(k) = \sum_{k'=1}^{l} \mathbf{H}(k') \mathbf{s}(k - k') + \boldsymbol{\varepsilon}(k) \tag{2.4}$$

where $\mathbf{H}(k') = [H_{ij}(k')]$, $i = 1, \ldots, m$, $j = 1, \ldots, n$, is a Multi-Input-Multi-Output (MIMO) LTI system. The present blind signal separation problem are usually called separation of convolutive mixture, or *MIMO channel equalization problem* since it finds the inverse of the MIMO LTI system.

2.3 Instantaneous Linear Mixing Process

In cases that the mixing process can be assumed to be instantaneous, the observed signals at some time, $\mathbf{x}(k)$ will not depend on the sources at previous time. Further assuming the mixing processing is linear, we have,

$$\mathbf{x}(k) = \mathbf{A}\mathbf{s}(k) + \epsilon(k) \tag{2.5}$$

where $\mathbf{A} = [a_{ij}], i = 1, \ldots, m, j = 1, \ldots, n$ is called the *mixing matrix*. In this case, the blind signal separation problem reduce to the determination of the inverse of $\mathbf{A}$. $\mathbf{A}$ is generally assumed to be fixed in the problem. However, adaptive algorithm may also truck slowly changing $\mathbf{A}$.

2.4 Problem Definition and Assumptions in this Thesis

We make the following assumptions in our work:

(A1) The mixing process is linear and instantaneous.

(A2) The number of sources n is known and the number of sensors $m = n$.

(A3) All channels of source signals are *statistically independent* to each other. This is, we have

$$p_{\mathbf{s}}(\mathbf{s}) = \prod_{i=1}^{n} p_{s_i}(s_i) \tag{2.6}$$

(A4) The mixing matrix $\mathbf{A}$ is non-singular.

(A5) Each channel of source signal $s_i(k), k = 1, 2, \ldots$ are identically and independently distributed (iid). Hence $s_i(k), k = 1, 2, \ldots$ can be regarded as realized values of a random variable s_i [1] with probability distribution function (pdf) $p_{s_i}(s_i)$.

(A6) No more than one channel of the source signals can be Gaussian distributed.

(A7) Moments exist up to necessary order, which depends on the nonlinearity used.

(A8) Noise can be neglected.

[1] Small letters are used to refer to both the random variables and the realized values. Readers can in most case determine the nature of the symbol by its intuitive meaning or indices on it.

Under these assumptions, we can write the mixing process as:

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{2.7}$$

The blind signal separation problem is formulated as the Independent Component Analysis (ICA) problem. Each observed signal x_i consists of linearly superimposed independent components $\{s_j, j = 1, \ldots, n\}$ and source separation is to extract the independent components out.

We use a $n \times n$ feedforward de-mixing $\mathbf{W} = [w_{ij}]$ to obtain the *recovered signals* $\mathbf{y} = [y_1, \ldots, y_n]^T$:

$$\mathbf{y} = \mathbf{W}\mathbf{x} \tag{2.8}$$

We targets at obtaining $\mathbf{y} = \mathbf{s}$. However, as inspired from the matrix multiplication $\mathbf{x} = \mathbf{A}\mathbf{s}$ with both $\mathbf{A}$ and $\mathbf{s}$ unknown to us, the order of the channels and the scale of the sources $\mathbf{s}$ are indeterminable. (Many combinations of $\mathbf{A}$ and $\mathbf{s}$ with different order of channels and scale of $\mathbf{s}$ give the same $\mathbf{x}$.) Hence, source separation is said to be successful if we recover the sources up to an arbitrary scaling factor and an arbitrary permutation of channel index. That is, we target at obtaining

$$y_i = v_{ij_i} s_{j_i} \qquad i = 1, ..., n \tag{2.9}$$

where $j_k \neq j_l$ if $k \neq l$ represents the permutation of channel index and v_{ij} is the scaling factor.

In matrix form, defining

$$\mathbf{V} = [v_{ij}] = \mathbf{W}\mathbf{A} \tag{2.10}$$

source separation is said to be achieved if $\mathbf{V}$ consists of one non-zero element v_{ij_i} in each row and each column, and other elements being zero. Such $\mathbf{V}$ can be written in the form:

$$\mathbf{V} = \mathbf{P}\mathbf{D}, \tag{2.11}$$

where $\mathbf{D}$ is a diagonal matrix and $\mathbf{P}$ is a permutation matrix.

Remark 1 Explanation for the assumptions:

(a) The independence assumption (A3) is the most important assumption in the ICA problem. It is possible to recover the blind sources only because we know that they are statistically independent. Then making the output signals mutually independent ensures that the output signals recover the sources up a permutation and scalar factor. If the sources were not independent, the method of separation would break down.

(b) The iid assumption (A5) is one of the conditions for ensuring convergence of the adaptive algorithm. This assumption can be relaxed to 'time-average invariance' [49] condition which also guarantee convergence. In experiments, the adaptive algorithms can also work on real world signals like human speech signal, which is non-stationary, not iid and can only be roughly approximated by the 'time-average invariance' model. Hence, the iid assumption may not be necessary in practice.

(c) Assumptions (A6) and (A7) arise because higher order statistics are used in our approach. As Gaussian distribution has null higher-order cumulants, algorithms using higher order statistics generally cannot separate two Gaussian sources. Assumption (A7) is satisfied by many distributions, however, for example, for Cauchy distribution the variance dose not exist and in an experiment on sources with Cauchy distribution, the algorithm failed to converge.

Remark 2 It has been proved in [24] that the recovered signals being statistically independent is an equivalent criteria for source separation eq. (2.11) being achieved. It is obvious that if each recovered signal recovers one distinct source signal, they satisfy the statistical independence criteria $p_{\mathbf{y}}(\mathbf{y}) = \prod_{i=1}^{n} p_{y_i}(y_i)$ (by considering $p_{\mathbf{y}}(\mathbf{y}) = p_{\mathbf{s}}(\mathbf{s})/\det \mathbf{V}$). If source separation is not successful, different channels of recovered signals consists of common components of source signals. Then the recovered signals must be correlated and statistically dependent. Hence, ICA can be approached by considering methods that make the recovered signal statistically independent.

Remark 3 It is well-known that Higher Order Statistics (HOS) are needed to solve the ICA problem (with iid, or temporally white signals). If only second order statistics are used, we can only have $n(n-1)/2$ equations to control the correlation:

$$E[y_i y_j] = 0, \qquad i \neq j \tag{2.12}$$

and n equations that control the (self) variant $E[y_i^2]$. Hence, there are not enough equations for the $n \times n$ elements of $\mathbf{W}$ and there would be infinite number of possible

solutions. If HOS is used, we have more constraints on the HOS $E[y_i^p y_j^q]$ that reduce the number of possible solutions to finite number. In adaptive ICA algorithms, HOS are picked up by the nonlinearity in the algorithm, as seen from the Taylor expansion of the nonlinearity consisting of high order terms.

Chapter 3

Literature Review

A literature survey will be presented in this chapter. The survey is not exhaustive but most important works on the blind signal separation problem will be reviewed. Section 3.1 reviews the approaches to separation of instantaneous mixture and Section 3.2 presents the achievement in separation of convolutive mixtures.

3.1 Previous Works on Blind Signal Separation with Instantaneous Mixture

The independent component analysis problem has attracted much attention since Herault, Jutten and Ans's seminal paper [33] in French and its English journal paper version [37, 26, 59] were published. Different approaches have been proposed to solved the problem. A popular method is to optimize some suitable objective function (so-called contrast function defined by Comon [24]) that the global maximum or minimum of which ensure source separation. For some objective functions, batch-way algorithm is devised for the optimization [24]. These algorithms are often named as the *batch-way approaches*, or *algebraic approaches*. For some objective functions, simple adaptive algorithms can be devised [2, 5, 14, 27, 34, 64]. Approaches that use adaptive algorithms are sometimes named as the *neural approach* [41] since most of them are plausible for neural network implementation. Apart from the objective function optimizing method, there are also some other heuristic adaptive algorithms suggested to solve the ICA problem [37].

Some approaches need spatial pre-whitening, or spatial sphering, of the observed signals. The pre-whitening process is to remove the second-order cross moments and fix the (self) variants of the mixed signals. The observed signals are multiplied by a

whitening matrix $\mathbf{U}$ to give the *whitened signals* $\tilde{\mathbf{x}}$:

$$\tilde{\mathbf{x}} = \mathbf{U}\mathbf{x} \tag{3.1}$$

the whitened signals have identity correlation matrix $E[\tilde{\mathbf{x}}\tilde{\mathbf{x}}^T] = \mathbf{I}$. The whitened signals are then multiplied by a de-mixing matrix $\widetilde{\mathbf{W}}$ to give the recovered signals $\mathbf{y}$:

$$\mathbf{y} = \widetilde{\mathbf{W}}\tilde{\mathbf{x}} \tag{3.2}$$

The pre-whitening process can easily done by standard Principal Component Analysis (PCA) adaptively or by singular value decomposition in a batch-way manner.

In the following subsections, summaries of the several popular approaches are provided.

3.1.1 Algebraic Approaches

Algebraic approaches use batch-way algorithms, or recursive algorithms, to estimate the parameters of the de-mixing system. The whole series of signal data have to be obtained before running the algorithm and hence this class of algorithms have to be used off-line. The computation loadings are usually intensity. These undesirable properties of the algorithms hinder them from being used in real world applications.

Cardoso and Comon [15] provided a summary of some algebraic methods. Cardoso [12] suggested one algebraic method that applied diagonalization on the forth-order cumulant tensor of pre-whitened mixed signals. He also suggested another algebraic method that did not require pre-whitening and may work on cases that the number of sensors is less than the number of source [13]. Comon [24] developed a batch-way algorithm to maximize contrast functions. He used contrast function derived from the truncated Edgeworth expansion of the mutual information between the recovered signals, or some simpler contrasts that consist of the sum of the square of the cumulant of each recovered signal.

Recently, Hyvarinen [35] proposed a fixed-point iteration algorithm based on the minimization/maximization of kurtosis under a constraint on the norm of the recovered signal. The iteration algorithm works in the way that one source is extracted one time, and the algorithm can be repeated to extract another source. This algorithm has advantage of fast cubic convergence rate and having no manual adjustable parameter (like the learning rate). Convergence analysis of the algorithm has also been provided.

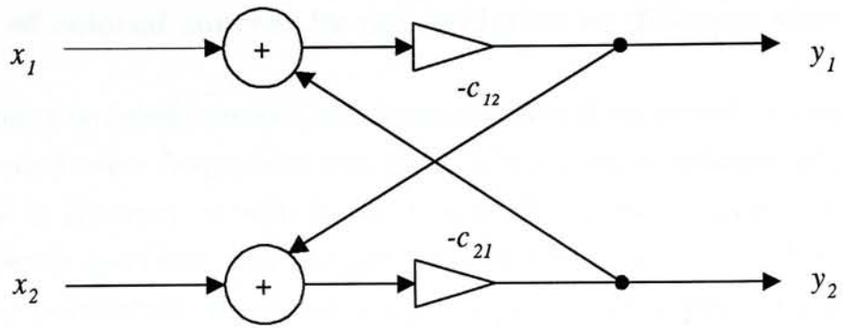


Figure 3.1: The Herault-Jutten network (2-channel).

3.1.2 Neural approaches

There have been vast number of neural approaches to the independent component analysis problem. Since adaptive algorithms can perform learning and process data online, they have high potential to be used in real application. In particular, adaptive algorithms with locality is plausible for VLSI implementation and biologically feasible. We shall introduce some popular approaches in the followings.

The Herault-Jutten network

The Herault-Jutten network [37, 26], or simply H-J network, is very famous in last decade. Successive works have been followed by other researcher and analog VLSI implementation of the H-J network has been successful [22, 23, 38, 63].

The H-J network uses a feedback network $\mathbf{C}$ with the diagonal being zero as the demixing system, such that $\mathbf{y} = [\mathbf{I}+\mathbf{C}]^{-1}\mathbf{x}$ (Figure 3.1). An heuristic adaptive algorithm is used to tune $\mathbf{C}$:

$$\Delta c_{ik} \propto f(s_i(t))g(s_k(t)) \tag{3.3}$$

The choice $f(r) = r^3$ and $g(r) = r$ works on sub-Gaussian source [59] and the choice $f(r) = r$ and $g(r) = r^3$ works on super-Gaussian signals [28]. However, the network suffers from some stability problem [30] and convergence to the correct solution is not generally guaranteed.

Separation of colored sources by decorrelation at different time lags

For narrow-band (colored) sources, the frequency spectrum consist of components concentrated around some frequencies and there is temporal correlation at different time lags. This is in contrary to wide-band (temporally white) sources that possess dispersed frequency spectrum and samples at different time point being independent. The temporal correlation of the sources provides cue for source separation. Gerven and Compernolle [31] make use of the cancelation of correlation of the recovered signals at different time lags for source separation:

$$E[y_i(k)y_j(k-\tau)] = 0 \quad \forall \tau \in \mathbb{Z} \tag{3.4}$$

It should be noticed that this class of algorithm would not work on wide band sources.

Approaches that rely on minimization/maximization of kurtosis

Comon [24] proved that the sum of squares of (self) cumunants of order $r \geq 2$ over all recovered signals is a contrast function:

$$J = \sum_{i=1}^{n} [\text{cum}^{(r)}(y_i)]^2 \tag{3.5}$$

provided that the signals are pre-whitened. The forth order cumulant, which equals to the kurtosis of the signal:

$$\text{cum}^{(4)}(y) = \text{kurt}(y) = E[y^4] - 3\{E[y^2]\}^2 \tag{3.6}$$

is of particular interest. However, Comon have only developed batch-way algorithm but not adaptive algorithm to achieve the maximization.

Moreau and Macchi [53] proved that the following function:

$$J = \sum_{i=1}^{n} |\text{kurt}(y_i)| \tag{3.7}$$

is a contrast function if all the sources have kurtosis of the same sign and devised an adaptive algorithm based on it.

The deflation approach introduced by Delfosse and Loubaton [27] uses a cascade de-mixing network (Figure 3.2) to extract the sources. The elements of the de-mixing

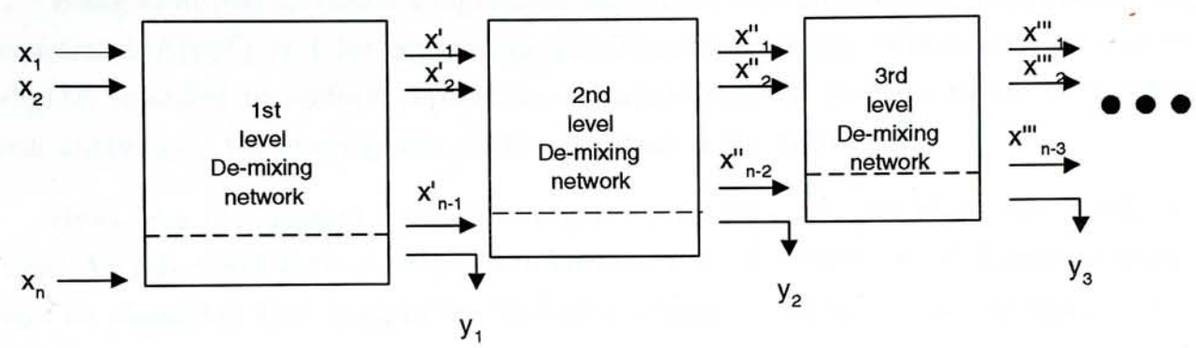


Figure 3.2: The network used in the deflation approach.

matrices are parameterized by elements $\boldsymbol{\theta}$ on the unit sphere to achieve spatial sphering. The extraction of source signal on each layer is based on the maximization of

$$J(\boldsymbol{\theta}) = [\text{kurt}(\mathbf{w}^T(\boldsymbol{\theta})\mathbf{x})]^2/4 \tag{3.8}$$

which has been proved to be able to work. The remaining channels are ensured to be a whitened mixture of the remaining sources by the algorithm. As the cascade de-mixing structure is recursive, extraction of all sources is theoretically proved. This approach is the first approach that convergence to the correct solution is guaranteed for the general case with arbitrary number of channel n. It has also been analyzed that the reconstruction errors on the last extracted source should not increase drastically with n.

Cardoso and Laheld [14] proposed an adaptive algorithm that possess the 'equivariant' property, and hence exhibit 'uniform performance' that the separation performance is independent of the conditioning of the mixing matrix in noiseless case. The EASI (Equivariant Adaptive Separation via Independence) algorithm uses a feedforward de-mixing matrix $\mathbf{W}$ like that in the problem definition and takes the following form:

$$\Delta\mathbf{W} \propto [\mathbf{y}\mathbf{y}^T - \mathbf{I} + \mathbf{g}(\mathbf{y})\mathbf{y}^T - \mathbf{y}\mathbf{g}(\mathbf{y})^T]\mathbf{W} \tag{3.9}$$

If the nonlinear function $\mathbf{g}(\mathbf{y}) = [g_1(y_1), ..., g_n(y_n)]$ is chosen to be $g_i(y_i) = y_i^3$, the algorithm is actually minimizing $\sum_i E[y_i^4]$ under the constraint $E[\mathbf{y}\mathbf{y}^T] = \mathbf{I}$. For the 2 channel case, it is proved that there is no spurious local minimum and the convergence to correct solution is ensured for two sources with sum of kurtosis being negative [14]. However, cases with more channels have not been considered.

Wang *et al* [64] proposed a bigradient algorithm that minimize $\sum_i E[y_i^4]$ under the constraint $E[\mathbf{y}\mathbf{y}^T] = \mathbf{I}$ for separating sub-Gaussian sources. The algorithm can be slightly modified to perform separation on super-Gaussian sources. However, theoretical analysis on the convergence of the algorithm is not provided.

Hyvarinen [34] devised a simple one-unit neural algorithm for blind source separation. An algorithm that minimizes the kurtosis is used to extract sub-Gaussian signal and an algorithm that maximizes the kurtosis under a constraint on the norm of the weight vector is used to extract super-Gaussian signal. One additional unit is used to determine the sign of the kurtosis of the extracted signal. The extracting units can be used in parallel to extract arbitrary ($\leq n$) number of sources simultaneously by adding some feedback. Theoretical analysis of the algorithm is provided [36].

The Maximum Entropy Approach

Bell and Sejnowski [5, 6] proposed the Information Maximization approach, or called the INFORMAX or Maximum Entropy (ME) approach, from the viewpoint of nonlinear generalization of Linsker's INFORMAX principle [45]. In this approach, a nonlinear transformation function $f_i(y_i)$ is applied to the recovered signal y_i to give the transformed output z_i:

$$\mathbf{z} = \mathbf{f}(\mathbf{y}) = [f_1(y_1), ..., f_n(y_n)]^T \tag{3.10}$$

With $\{f_i(y_i)\}$ suitably chosen to be close to the Cumulative Distribution Functions (CDF's) of the sources, the maximization of the output entropy $H(\mathbf{z})$ is proposed to be a principle for blind source separation. Bell and Sejnowski used the gradient ascent algorithm in [5, 6] to perform the maximization:

$$\Delta\mathbf{W} \propto \left\{ \left[\mathbf{W}^T\right]^{-1} + \mathbf{h}(\mathbf{y})\mathbf{x}^T \right\} \tag{3.11}$$

where $\mathbf{h}(\mathbf{y}) = [h_1(y_1), ..., h_n(y_n)]$, $h_i(y_i) = g_i'(y_i)/g_i(y_i)$ and $g_i(y_i) = f_i'(y_i)$. It has been experimentally shown that the choice $f_i(y_i) = \text{logsig}(y_i) = 1/(1 + \exp(-y_i)$ or $f_i(y_i) = \tanh(y_i)$ can perform separation on speech signals.

The gradient ascent algorithm involves an undesirable matrix inversion and the convergence is slow. At later time, the 'natural gradient' method [2] is adopted (e.g. see [7]) by multiplying the positive definite matrix $\mathbf{W}^T\mathbf{W}$ to the right of the gradient:

$$\Delta\mathbf{W} \propto [\mathbf{I} + \mathbf{h}(\mathbf{y})\mathbf{y}^T]\,\mathbf{W} \tag{3.12}$$

The new algorithm has a much faster convergence speed and have small residual fluctuation after convergence with constant learning rate.

The Minimum Mutual Information (MMI) approach

Amari *et al* [2] proposed an adaptive algorithm to minimize the mutual information between the recovered signals, which equals to the Kullback divergence between the joint density of the recovered signals and the product of marginal densities of the recovered signals:

$$J(\mathbf{W}) = \int_{\mathbf{y}} p_{\mathbf{y}}(\mathbf{y}) \log \frac{p_{\mathbf{y}}(\mathbf{y})}{\prod_{i=1}^{n} p_{y_i}(y_i)} \mathrm{d}\mathbf{y} \tag{3.13}$$

The global minima of the mutual information obviously consist only of correct solutions for source separation since they occur only when $p_{\mathbf{y}}(\mathbf{y}) = \prod_{i=1}^{n} p_{y_i}(y_i)$. However, the difficulty is to approximate the mutual information and establish the minimization algorithm appropriately. A truncated Gram-Charlier series (see [40]) is used to approximate the mutual information and the minimization is proposed to be achieved by a 'coarse implementation' that replace some expected values by their instantaneous values. The algorithm is given by

$$\begin{aligned} &\Delta\mathbf{W} \propto [\mathbf{I} + \mathbf{h}(\mathbf{y})\mathbf{y}^T]\,\mathbf{W} \\ &\mathbf{h}(\mathbf{y}) = [h_1(y_1), ..., h_n(y_n)]^T \\ &h_i(y_i) = -\frac{3}{4}y_i^{11} - \frac{25}{4}y_i^9 + \frac{14}{3}y_i^7 + \frac{47}{4}y_i^5 - \frac{29}{4}y_i^3 \qquad i = 1, ..., n \end{aligned} \tag{3.14}$$

The success of the algorithm relies on the approximation and the effectiveness of the 'coarse implementation'. It can be noticed that the algorithms used in the Maximum Entropy approach with natural gradient and the MMI approach turn out to have the same form but differ from the nonlinearity used. As to be discussed in later chapters, since $h_i(y_i)$ is nevertheless a fixed nonlinearity, the algorithm is expected to be able to separate sources with a class of distribution (that are 'appropriately approximated' by the truncated Gram-Charlier series.) but not sources with any distribution.

Other neural approaches

Apart from that approaches mentioned above, there are a large number of algorithms proposed with different features. However, most of the algorithm lack theoretical convergence analysis. Cichocki and Amari have proposed a biologically plausible recurrent

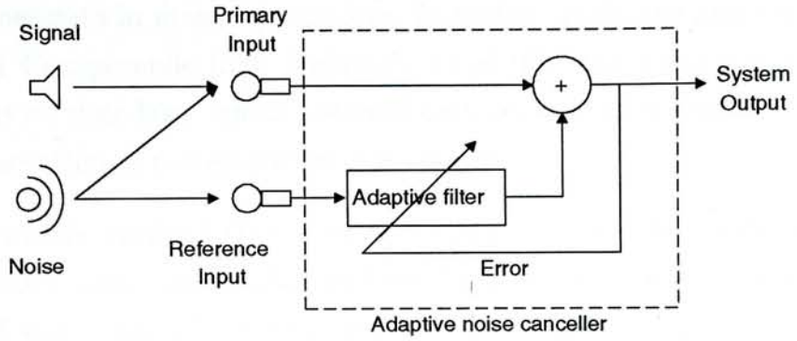


Figure 3.3: The adaptive noise cancelation problem and Widrow's network.

network [1] and a multi-layer network [21] which can handle ill-conditioned mixing matrix and badly scaled input. Cichocki and Moszczynski [20] and H. Marsman [51] tried heuristics on the control of learning rate in order to speed up convergence. Matsuoka *et al* [52] proposed an algorithm for non-stationary signals. Oja and Karhunen [42, 54] also investigate the use of nonlinear PCA learning on source separation. Self-organizing maps has been applied to source separation by Pajunen *et al* [56]. The blind signal separation has also been formulated as a maximum likelihood problem by Pham *et al* [57] who carried out an adaptive implementation, and by Belouchrani and Cardoso [9] who used batch-way or stochastic EM algorithm.

3.2 Previous Works on Blind Signal Separation with Convolutive Mixture

Reducing unwanted components from the observed signals has long been a practical problem in engineering applications. In the 1970s, Widrow *et al* [66] proposed a well-known Adaptive Noise Cancelling algorithm using power minimization principle to the problem. In this approach, the primary sensor receive both signal and noise while the reference sensor is assumed to pick up the noise only (Figure 3.3). The drawback of this 'asymmetric' approach is that the leakage of signal to the reference sensor would lead to some desired signal also cancelled together with the noise. Such drawback limited that usage of Widrow's noise canceler.

In late 1980s and 90s, researchers discarded the viewpoint that treats noise and signal distinctly, and view both noise and signal as source signals. The problem formulation of blind signal separation for convolutive mixture in the last chapter are adopted and received much attention. Feedforward or feedback causal FIR filters are used as

the de-mixing network in many approaches. In earlier trials, the approaches suggested by Gerven and Compernolle [32], Weinstein *et al* [65] and Chan *et al* [16] relied on decorrelation at all time lags, which involved only second order statistics. Convergence to the correct solution is not guaranteed generally.

It became widely realized that methods using second order statistics only is not sufficient to ensure source separation and the work moves to using higher order statistics. Yellin and Weinstein [77, 78] proved that statistical independence of the recovered signals is a sufficient condition for separation of convolutively mixed signals and devised an algorithm that involved cross-polyspectra. Thi and Jutten [60] proved that the cancelation of forth-order cross-cumulants of the recovered signals is sufficient for source separation. They proposed an algorithm based on this cancelation and another algorithm with nonlinear function from the generalization of the well-known Herault-Jutten algorithm for instantaneous mixture. Torkkola [61] and Lee *et al* [46] suggested algorithms for both causal FIR and causal IIR filter networks from the generalization of the Maximum Entropy approach [5] for instantaneous mixtures. The algorithms with causal filters succeed in separating sources from minimum-phrase mixtures but generally fail on non-minimum-phrase mixtures, that room acoustic condition often involves. This is because the inverse of a non-minimum phase system may be non-causal. Recently, Lee *et al* [47] adopted a technique to realize non-causal filters for the de-mixing network. They obtained successful result in separating real world sound signals recorded in room acoustic condition.

Besides the above contributions, Platt and Faggin [58] considered the case of delayed, but non-convolutive, mixture in early 90s. Bamford and Canagarajah [4] suggested the usage of time-delay estimation techniques to reduce the length of the de-mixing filters, which were widely adopted by other researchers. Tong *et al* [62] gave theoretical discussion on the indeterminacy and identifiability of blind identification. Comon [25] introduced 'contrasts' for convolutive mixing, as an extension to the contrasts for instantaneous mixing [24]. Lindgren *et al* [48] provided some investigation on the local convergence on a class on blind separation algorithms with FIR filter. Feder *et al* [29] suggested an approach that seek the maximum likelihood estimates of the filter parameter using EM algorithms, assuming the speech signal is auto-regressive Gaussian random process and the noise source is Gaussian. It can be seen that the development in this area is amazingly fast in the last decade.

Chapter 4

The Information-theoretic ICA Scheme

In this chapter, we review the information-theoretic ICA scheme proposed by Xu and Amari [74], which we shall analyze in the oncoming chapters. In Section 4.1, we review the Bayesian YING-YANG learning scheme [67, 69] that the information-theoretic ICA scheme based on. Section 4.2 gives the construction of the information-theoretic ICA scheme and the derivation of the information-theoretic ICA algorithm.

4.1 The Bayesian YING-YANG Learning Scheme

The Bayesian YING-YANG (BYY) learning theory is proposed by Xu [67, 68, 69, 70, 72]. In this thesis, we will only introduce the part of the BYY learning theory that the ICA framework concerns and the power of the theory. For the details of the BYY learning theory, please refer to the original papers [67, 68, 69, 70, 72].

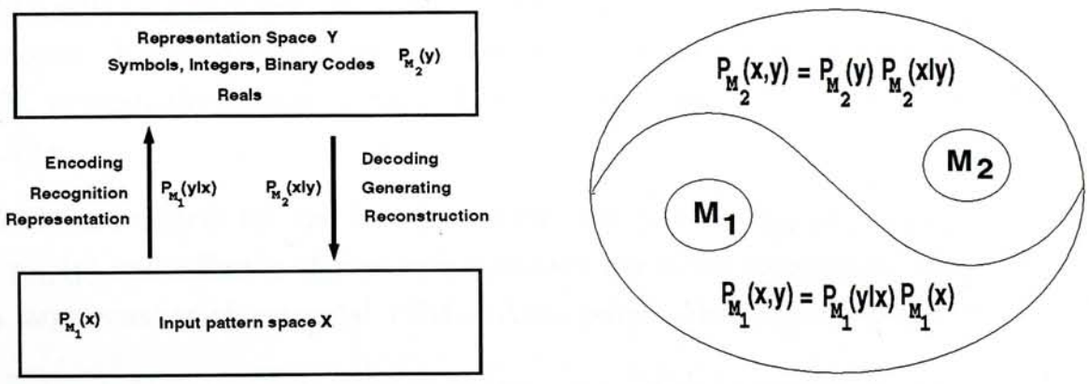


Figure 4.1: The joint spaces X, Y and the YING-YANG Machine

As shown in [67, 69, 72], learning problems can be summarized into the problem of estimating the joint density $p(x, y)$ [1] of patterns in the input space X and the representation space Y as shown in Figure 4.1. Under Bayesian framework, we have two representations for $p(x, y)$. One representation is

$$p_{M_1}(x, y) = p_{M_1}(y|x)p_{M_1}(x) \tag{4.1}$$

implemented by a model M_1 called the *YANG* (male) part, which gets its name since it performs the task of transferring a pattern (real body) into a code (seed). The other representation is

$$p_{M_2}(x, y) = p_{M_2}(x|y)p_{M_2}(y) \tag{4.2}$$

implemented by a model M_2 called the *YING* (female) part, which gets its name because it performs the task of generating a pattern (real body) from a code (seed). They are complementary to each other and together implement a cycle $x \to y \to x$. The four propabability components $p_{M_1}(x)$, $p_{M_1}(y|x)$, $p_{M_2}(x|y)$ and $p_{M_2}(y)$ are to be assigned to specific forms depending on the network architecture and the learning problem encountered.

A separation functional $F_s(M_1, M_2) = F_s(p_{M_1}(x, y), p_{M_2}(x, y))$ is used to measure the 'distance' between the YING part $p_{M_2}(x, y)$ and YANG part $p_{M_1}(x, y)$. The separation functional satisfies the requirement that $F_s(M_1, M_2) \geq 0$ and $F_s(M_1, M_2) = 0$ if and only if $p_{M_1}(x, y) = p_{M_2}(x, y)\ \forall x, y$. If the Kullback divergence

$$KL(M_1, M_2) = \int_{x,y} p_{M_1}(y|x)p_{M_1}(x) \log \frac{p_{M_1}(y|x)p_{M_1}(x)}{p_{M_2}(x|y)p_{M_2}(y)} \mathrm{d}x\mathrm{d}y \tag{4.3}$$

is used as the separation functional, the system is called the Bayesian-Kullback YING-YANG (BKYY) Machine [67, 69]. If other separation functional, such as Convex divergence, L_p-divergence or de-correlation index (which shall be introduced in Chapter 9), is used, the system is called Bayesian Non-Kullback YING-YANG (BNKYY) Machine.

Different choices for the four probability components $p_{M_1}(x)$, $p_{M_1}(y|x)$, $p_{M_2}(x|y)$ and $p_{M_2}(y)$ and different choices and manipulation of the separation functional results in a large number of potential YING-YANG pairs. Although not all of the potential

[1] The symbols in the Bayesian YING-YANG learning scheme in this chapter follow those in the original paper. Reader are advised to distinguish the symbols by their intuitive meaning and printing styles.

pairs provide sensible learning models, a number of them indeed lead to useful learning models and some of them are currently existing models.

[2] The Bayesian YING-YANG learning scheme is suggested to be a unified general statistical learning theory for *parameter learning*, *scale selection*, *structure evaluation*, *sampling design* and *regularization*. In parameter learning, the *cosupervised learning scheme* can unify unsupervised and supervised learning and let them coexist consistently [71].

For unsupervised learning, as shown in [67, 69], the special cases of the BYY learning scheme can reduce to the EM algorithm related learnings, to Amari's *Information Geometry* theory and the *em* algorithm, to Hinton & Zemel's MDL autoencoder, and to multisets modeling learning - a unified learning framework for clustering, PCA-type learnings and self-organizing map. One of its other special case reduces to Maximum Information Preservation, and another special case of it reduces to the Helmholtz Machine with new understandings. The BYY learning scheme is also used to derive ICA approaches analysed in this thesis and other new algorithms for the blind signal separation problem that the acutual number of sources is unknown and observed signals are contaminatied with noise [73].

For supervised learning, the scheme includes the mixture of expert model as special case, from which new algorithms for improving learnings for RBF networks can be derived. The scheme also includes maximum likelihood learning (least square learning in particular) as special case, and provides new algorithms as alternatives for back-propagation. Moreover, this scheme has been extended to temporal patterns with a number of new models for signal modeling [68], including the extensions of Hidden Markov Model (HMM), AMAR models, as well as the extensions of Helmholtz Machine and Maximum Information Preservation.

For scale selection, the BYY learning scheme gives new criteria on the selection of number of densities in mixture of densities learning, in particular, the scheme gives new criteria on the number of clusters in the clustering problem, which is a classical open problem [70]. The scheme also gives criteria on the dimension of subspace on the Pricipal Component Analysis (PCA) type learning [70] and number of hidden units in feed-forword nets [71].

In the following section, we review the Information-theoretic ICA Scheme derived

[2]The references for the learning theories that the BYY learning scheme unifies is not contained in the bibliography of this thesis. Please refer to the bibbliographies of the original papers of the BYY learning scheme [67, 68, 69, 70, 72].

from the BKYY learning scheme. The trial of deriving ICA algorithm from BNKYY learning is given in Chapter 9.

4.2 The Information-theoretic ICA Scheme

4.2.1 Derivation of the cost function from YING-YANG Machine

The YING-YANG learning scheme was used on the ICA problem by Xu and Amari [74]. Since the source signals $\mathbf{s}$ are hidden to us while the recovered signals $\mathbf{y}$ are the actual output we get, the source signals $\mathbf{s}$ are regarded as the 'seed' y in eq. (4.3), and the recovered signals $\mathbf{y}$ are regarded as the 'real body' x in eq.(4.3). As the hidden $\mathbf{s}$ generates $\mathbf{y}$ via the information passage

$$\mathbf{s} \rightarrow \mathbf{x} = \mathbf{A}\mathbf{s} \rightarrow \mathbf{y} = \mathbf{W}\mathbf{x} = \mathbf{W}\mathbf{A}\mathbf{s} = \mathbf{V}\mathbf{s} \tag{4.4}$$

we regard $\mathbf{s} \rightarrow \mathbf{y}$ as the YING part M_2. Similarly, the information passage $\mathbf{y} \rightarrow \mathbf{s}$ along the inverse direction in eq.(4.4) is regarded as the YANG part M_1.

Now we consider the designation of the four probability components. According to our desire that $\mathbf{y} = \mathbf{s}$ (although only $\mathbf{y} = \mathbf{PDs}$ can be achieved), we design

$$P_{M_1}(\mathbf{s}|\mathbf{y}) = \delta(\mathbf{s} - \mathbf{y}) = \delta(\mathbf{y} - \mathbf{s}) = P_{M_2}(\mathbf{y}|\mathbf{s}) \tag{4.5}$$

As the components of the $\mathbf{s}$ are independent, we have

$$p_{\mathbf{s}}(\mathbf{s}) = \prod_{i=1}^{n} p_{s_i}(s_i) \tag{4.6}$$

However, in the blind signal separation problem, $p_{s_i}(s_i)$ is blind to us and the most useful information we know is that $s_i, i = 1, \cdots, n$ are independent. Hence, we replace $p_{s_i}(s_i)$ by a manually assigned continuous probability density function (pdf) $g_i(r)$ and design $P_{M_2}(\mathbf{s})$ as:

$$P_{M_2}(\mathbf{s}) = \prod_{i=1}^{n} g_i(s_i) \tag{4.7}$$

It is found that $\{g_i(r)\}$ has a much wider choice other than the original (unknown) source marginal densities. The choice of $\{g_i(r)\}$ shall be discussed in later chapters and constitutes the main objective of this thesis.

Moreover, we let $P_{M_1}(\mathbf{y})$ be the transformed density from the density of the observed signals $\mathbf{x}$ or source signal $\mathbf{s}$ via the forward passage $\mathbf{y} = \mathbf{W}\mathbf{x} = \mathbf{V}\mathbf{s}$:

$$P_{M_1}(\mathbf{y}) = p_{\mathbf{y}}(\mathbf{y}) = \frac{p_{\mathbf{x}}(\mathbf{x})}{|\det \mathbf{W}|} = \frac{p_{\mathbf{s}}(\mathbf{s})}{|\det \mathbf{V}|} \tag{4.8}$$

The Kullback divergence of the system is written as:

$$KL(M_1, M_2) = \int_{\mathbf{s},\mathbf{y}} P_{M_1}(\mathbf{s}|\mathbf{y}) P_{M_1}(\mathbf{y}) \log \frac{P_{M_1}(\mathbf{s}|\mathbf{y}) P_{M_1}(\mathbf{y})}{P_{M_2}(\mathbf{y}|\mathbf{s}) P_{M_2}(\mathbf{s})} \mathrm{d}\mathbf{y}\mathrm{d}\mathbf{s} \tag{4.9}$$

Putting eq. (4.5) into eq. (4.9) and performing the integration for the δ-function, we get:

$$KL(M_1, M_2) = \int_{\mathbf{y}} P_{M_1}(\mathbf{y}) \log \frac{P_{M_1}(\mathbf{y})}{P_{M_2}(\mathbf{y})} \mathrm{d}\mathbf{y} \tag{4.10}$$

Putting eqs. (4.7) and (4.8) into eq. (4.10), we get the cost function:

$$J = \int_{\mathbf{y}} p(\mathbf{y}) \log \frac{p(\mathbf{y})}{\prod_{i=1}^{n} g_i(y_i)} \mathrm{d}\mathbf{y} \tag{4.11}$$

or

$$\begin{aligned} J(\mathbf{W}) &= \int_{\mathbf{x}} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{|\det \mathbf{W}| \prod_{i=1}^{n} g_i(\mathbf{w}_i^T \mathbf{x})} \mathrm{d}\mathbf{x} \\ &= \int_{\mathbf{s}} p(\mathbf{s}) \log \frac{p(\mathbf{s})}{|\det \mathbf{V}| \prod_{i=1}^{n} g_i(\mathbf{v}_i^T \mathbf{s})} \mathrm{d}\mathbf{s} \\ &= J(\mathbf{V}) \end{aligned} \tag{4.12}$$

As the two Bayesian representations $P_{M_1}(\mathbf{s}, \mathbf{y})$ and $P_{M_2}(\mathbf{s}, \mathbf{y})$ should match each other, we designate the minimization of $J(\mathbf{W})$ as the means of the learning of $\mathbf{W}$ to perform ICA. The framework above is proposed as a general information-theoretic scheme for the ICA problem [74].

4.2.2 Connections to previous information-theoretic approaches

Connection to the Minimum Mutual Information approach

In the special case that $g_i(y_i)$ is designated as $p_{y_i}(y_i)$, the marginal density of y_i, the cost function eq.(4.11) reduce to:

$$J(\mathbf{W}) = \int_{\mathbf{y}} p_{\mathbf{y}}(\mathbf{y}) \log \frac{p_{\mathbf{y}}(\mathbf{y})}{\prod_{i=1}^{n} p_{y_i}(y_i)} \mathrm{d}\mathbf{y} \tag{4.13}$$

which is exactly the mutual information used in the Minimum Mutual Information (MMI) approach proposed by Amari *et al* [2]. The MMI approach attempts to approximate $\prod_{i=1}^{n} p_{y_i}(y_i)$ to the best. Hence the MMI approach is a special case of the information-theoretic ICA scheme.

Connection to the Maximum Entropy approach

We can establish the connection to the Maximum Entropy (ME) approach proposed by Bell & Sejnowski [5, 6] using the designation of the *nonlinear transformation function* $f_i(y_i)$ as the integral of $g_i(y_i)$:

$$f_i(y_i) = \int_{-\infty}^{y_i} g_i(u)\mathrm{d}u \tag{4.14}$$

Then, from applying the nonlinear transformation function to $\mathbf{y}$, the *transformed vector* $\mathbf{z} = [z_1, \ldots, z_n]^T = \mathbf{f}(\mathbf{y}) = [f_1(y_i), \ldots, f_n(y_n)$ is obtained. By the equation

$$p_{\mathbf{y}}(\mathbf{y}) = p_{\mathbf{z}}(\mathbf{z}) \left| \det \left[\frac{\partial \mathbf{z}}{\partial \mathbf{y}^T} \right] \right| = p_{\mathbf{z}}(\mathbf{z}) \prod_{i=1}^{n} g_i(y_i) \tag{4.15}$$

$J(\mathbf{W})$ is found to be equivalent to the negative entropy of $\mathbf{z}$:

$$J(\mathbf{W}) = \int_{\mathbf{z}} p_z(\mathbf{z}) \log p_{\mathbf{z}}(\mathbf{z})\mathrm{d}\mathbf{z} = -H(\mathbf{z}) \tag{4.16}$$

Therefore, minimizing $J(\mathbf{W})$ is equivalent to maximizing the entropy $H(\mathbf{z})$ and the ME approach is reached. Bell & Sejnowski [5, 6] uses logistic sigmoid $\mathrm{logsig}(y_i) = 1/(1 + exp(-y_i))$, hyperbolic tangent $\tanh(y_i)$ and arc-tangent $\arctan(y_i)$ as $f_i(y_i)$. These nonlinearities corresponds to $h_i(y_i) = 1 - 2\,\mathrm{logsig}(y_i)$, $h_i(y_i) = -2tanh(y_i)$ and $h_i(y_i) = -2y_i/(1 + y_i^2)$ respectively. The above three $h_i(y_i)$ all have reversed sigmoid shape (see Figure 6.1) and are suggested and experimentally verified to be able to separate super-Gaussian signals. This class of reversed sigmoid $h_i(y_i)$ constitute a choice of nonlinearity for the information-theoretic ICA scheme.

4.2.3 Derivation of the Algorithms

$J(\mathbf{W})$ can be considered as a scalar field in the parameter space $\mathbf{W}$, which is better written as a column vector:

$$\mathrm{Vec}(\mathbf{W}^T) = [w_{11}, \ldots, w_{1n}, w_{12}, \ldots, w_{2n}, \ldots, w_{n1}, \ldots, w_{nn}]^T \tag{4.17}$$

where Vec(·) is an operation that cascades the columns of the matrix from the left to the right to form a column vector. Then the following form of *general gradient descent algorithm*, which is commonly used in optimization theories, can be used to perform (local) minimization on $J(\mathbf{W})$:

$$\frac{\mathrm{dVec}(\mathbf{W}^T)}{\mathrm{d}t} \propto -\mathbf{G}\frac{\partial J(\mathbf{W})}{\partial \mathrm{Vec}(\mathbf{W}^T)} = -\mathbf{G}\mathrm{Vec}\left([\nabla_W J(\mathbf{W})]^T\right) \tag{4.18}$$

where $\mathbf{G}$ is a symmetric, positive definite matrix, and $\nabla_W J(\mathbf{W})$ is a $n \times n$ matrix whose (i,j)-element is $\partial J(\mathbf{W})/\partial w_{ij}$. From [5] the gradient of $J(\mathbf{W})$ is:

$$\nabla_W J(\mathbf{W}) = -E\left\{\left[\mathbf{W}^T\right]^{-1} + \mathbf{h}(\mathbf{y})\mathbf{x}^T\right\} \tag{4.19}$$

where

$$\mathbf{h}(\mathbf{y}) = [h_1(y_1), \ldots, h_n(y_n)]^T \tag{4.20}$$

$$h_i(y_i) = \frac{g_i'(y_i)}{g_i(y_i)}, \qquad g_i(r) = f_i'(r), \qquad i = 1, ..., n \tag{4.21}$$

The descent direction is flexible up to the choice of $\mathbf{G}$. The general stochastic gradient descent algorithm with any legitimate $\mathbf{G}$ is called an information-theoretic ICA algorithm.

The gradient algorithm

If $\mathbf{G}$ is chosen as the identity matrix, the general gradient algorithm reduces to the *gradient algorithm* used in [5, 6]:

$$\frac{\mathrm{d}\mathbf{W}}{\mathrm{d}t} \propto -\nabla_W J(\mathbf{W}) \tag{4.22}$$

or in the stochastic version:

$$\Delta\mathbf{W} = \epsilon(t)\left\{\left[\mathbf{W}^T\right]^{-1} + \mathbf{h}(\mathbf{y})\mathbf{x}^T\right\} \tag{4.23}$$

where $\epsilon(t)$ is a small positive learning rate.

However, this gradient algorithm involves an undesirable matrix inversion. Experiments also show that it has a slow convergence speed, and for a fixed learning rate $\mathbf{W}$ fluctuates relatively large around the convergence point after it has converged.

The natural gradient algorithm

A better choice of descent direction is the natural gradient direction [2]. The corresponding $\mathbf{G}$ is [2]:

$$\mathbf{G} = \mathbf{W}^T\mathbf{W} \otimes \mathbf{I} \tag{4.24}$$

where $\otimes$ is an operation that for any $m \times n$ matrix $\mathbf{B}$ and $p \times q$ matrix $\mathbf{C}$,

$$\mathbf{B} \otimes \mathbf{C} = \begin{bmatrix} b_{11}\mathbf{C} & \cdots & b_{1n}\mathbf{C} \\ \vdots & \ddots & \vdots \\ b_{m1}\mathbf{C} & \cdots & b_{mn}\mathbf{C} \end{bmatrix} \tag{4.25}$$

is a $mp \times mq$ matrix. By noting the identity [2]

$$\mathrm{Vec}(\mathbf{BCD}) = (\mathbf{D}^T \otimes \mathbf{B})\mathrm{Vec}(\mathbf{C}) \tag{4.26}$$

the natural gradient algorithm can be written as

$$\frac{\mathrm{d}\mathbf{W}}{\mathrm{d}t} \propto -[\nabla_W J(\mathbf{W})]\mathbf{W}^T\mathbf{W} \tag{4.27}$$

or, in the stochastic version,

$$\Delta\mathbf{W} = \epsilon(t)[\mathbf{I} + \mathbf{h}(\mathbf{y})\mathbf{y}^T]\mathbf{W} \tag{4.28}$$

Experiments show that the natural gradient algorithm yields convergence that is orders of magnitude faster than the gradient algorithm. In addition, for a fixed learning rate $\mathbf{W}$ fluctuates less around the convergence point after it has stabilized. Hence, the quality of separation is better if it converges to a correct solution.

The learning rate $\epsilon(t)$ can be controlled to decrease according to certain scheme formulated in stochastic approximation theory to obtain (almost) exact convergence to the solution. However, to retain the ability of the system to track show changes in the mixing condition, and for convenience, we use a fixed learning rate in this thesis. The system will attain 'weak convergence' in a small volume around the solution if the fixed learning rate is suitablly small. The average magnitude of the fluctuation around the solution decreases if smaller learning rate is used, the quality of separation will be better but the convergence speed will derease.

4.2.4 Roles and Constraints on the Nonlinearities

It is important to distinguish clearly the roles and nature of the nonlinear functions used in the theory. $f_i(y_i)$ is the nonlinear transformation function acting on y_i to obtain z_i. $g_i(y_i) = f_i'(y_i)$ is the first order derivative of $f_i(y_i)$. $h_i(y_i) = [\log g_i(y_i)]/y_i = g_i'(y_i)/g_i(y_i)$ is the derivative of $\log g_i(y_i)$, which is second order derivative in $f_i(y_i)$.

The role of $\{g_i(.)\}$ is a set of manually chosen probability density functions (pdf's) we assign to the unknown $\{p_{s_i}(s_i)\}$. The role of $\{f_i(.)\}$ is a set of manually chosen Cumulative Distribution Functions (CDF's) we assign to $\{s_i\}$. The function $h_i(.)$ is the corresponding nonlinearity in the algorithm that picks up suitable higher order statistics and constrain them in the equilibrium condition of the algorithm.

It is noteworthy to mention the reason why $g_i(y_i)$ is asserted to be a pdf at the far beginning, i.e., why $g_i(y_i)$ must be a (continuous) positive function integrable to 1 [3] :

$$\int_{-\infty}^{\infty} g_i(y_i)\mathrm{d}y_i = 1 \tag{4.29}$$

If $g_i(y_i)$ has discontinuity or can drop down to negative, $h_i(y_i)$ would be undefined at the discontinuity or at the zero point of $g_i(y_i)$ and the algorithm would break down. If $g_i(y_i)$ were not integrable, $f_i(y_i)$ would be an unbounded monotonic increasing function. Then the entropy $H(\mathbf{z})$ can be unbounded above since the random variable z_i can take infinite range. Hence the maximization of $H(\mathbf{z})$ (minimization of $J(\mathbf{W})$) w.r.t. $\mathbf{W}$ would result in divergence of $\mathbf{W}$ to infinity since this would give increasing $H(\mathbf{z})$.

4.3 Direction and Motivation for the Analysis of the Nonlinearity

In the previous ME approach [5], Bell & Sejnowski suggested that the distribution of the sources should better be known in *a priori* and the nonlinear transformation function should be as close to the CDF's of the sources as possible. In the MMI approach [2], Amari *et al* suggested that the marginal densities of the recovered signals should be approximated to the best. Both approaches suggested that it was necessary

[3] As multiplying an arbitrary constant to $g_i(y_i)$ results in the same $h_i(y_i)$, and that only $h_i(y_i)$ is material to the algorithm, we can actually allow an arbitrary multiplicative constant to $g_i(y_i)$ with no effect to the algorithm. However, as $g_i(y_i)$ has the role of the pdf of y_i, we will stick to using the $g_i(y_i)$ that is normalized to give an integral of one.

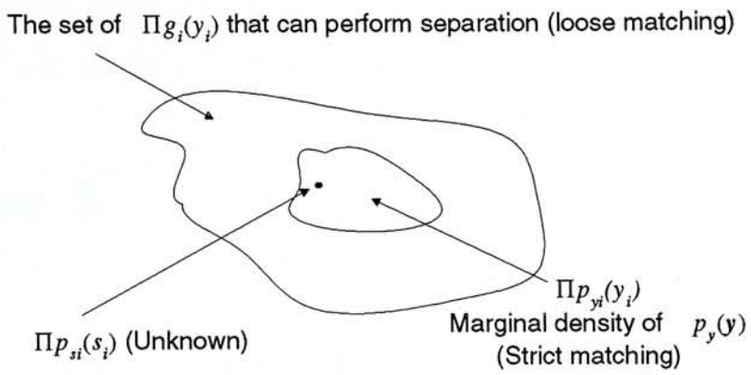


Figure 4.2: The choice of $\{g_i(y_i)\}$ that is capable of separating sources with $\{p_{s_i}(s_i)\}$.

to select $\{g_i(.)\}$ as close to $\{p_{s_i}(s_i)\}$ or $\{p_{y_i}(y_i)\}$ as possible. [4] However, the general information-theoretic ICA scheme proposed by Xu & Amari [74] suggests a new view that the approximation of $\{g_i(y_i)\}$ to the marginal densities $\{p_{y_i}(y_i)\}$ or $\{p_{s_i}(s_i)\}$ is over-necessary and a much wider class $\mathfrak{g}$ of $\{g_i(.)\}$ can also perform signal separation successfully (Figure 4.2). This is called the 'loose matching' between $\{g_i(.)\}$ and $\{p_{y_i}(y_i)\}$, but the detailed condition of the 'loose matching' is not clear yet.

Therefore, one main objective of this thesis is to investigate how large the class $\mathfrak{g}$ is, for a particular $p_{\mathbf{s}}(\mathbf{s})$, or the condition on the choice of the $\{g_i(.)\}$ for that $p_{\mathbf{s}}(\mathbf{s})$. Equivalently, we investigate the condition on the sources for them to be separated by a particular nonlinearity. Therefore, we investigate fixed nonlinearities case by case and find out which class of sources each fixed nonlinearty can separate. We perform detailed theoretical analysis and experimental verification on the simple cubic nonlinearity and empirical experiments on other nonlinearities. These results on fixed nonlinearities serve as examples of loose matching and give visualization of how loose the loose matching is.

Secondly, we try to investigate the algorithm with flexible mixture of densities proposed by Xu, *et al* [75] that tries to adapt flexible, parameterized $\{g_i(.)\}$ to $\{p_{y_i}(y_i)\}$. We try to determine the least flexible nonlinearity that can achieve the loose matching. Through these investigations, we give support to the suggestion that a much wider class of $\{g_i(.)\}$ can perform separation other than $\{p_{s_i}(s_i)\}$ and $\{p_{y_i}(y_i)\}$.

[4] It is worth noting that $p_{\mathbf{y}}(\mathbf{y})$ is changing while $\mathbf{W}$ is being tuned by the algorithm since $\mathbf{y} = \mathbf{WA}\mathbf{s}$. However, the global minima of the cost function are guaranteed to be correct solutions if (each) $g_i(.)$ is kept to be equal to the changing $p_{y_i}(y_i)$, as the mutual information eq. (3.13) is a contrast. Moreover, each y_i also equal to some s_{j_i} up to a scaling constant after successful separation.

Chapter 5

Properties of the Cost Function and the Algorithms

In this chapter, we shall present some theoretical results on the properties of the cost function $J(\mathbf{V})$ in the n^2-dimensional $\mathbf{V}$-parameter space. These results are useful for the analysis of the information-theoretic ICA algorithms and are common to many nonlinearities. We shall discuss the singularity, continuity, asymptotic behavior and other aspects of the cost function $J(\mathbf{V})$.

5.1 Lemmas and Corollaries

The analysis of the information-theoretic ICA scheme follows an idea proposed by Xu & Amari in [74] that investigates the cost function J in the $\mathbf{V}$-parameter space (v_{11}, .., v_{1n}, v_{21}, ..., v_{2n}, ..., v_{n1}, ..., v_{nn}) rather than in the $\mathbf{W}$-parameter space (w_{11}, .., w_{1n}, w_{21}, ..., w_{2n}, ..., w_{n1}, ..., w_{nn}). This is because $\mathbf{V} = \mathbf{WA}$ bears a one-to-one mapping to $\mathbf{W}$ and is a set of parameters that completely characterizes the system. The mathematical analysis using $\mathbf{V}$ is simpler since it does not explicitly involve $\mathbf{A}$ and $\mathbf{x}$.

In particular, the analysis often involves determination of equilibrium points of the cost function J. A standard method to do so is to solve the equilibrium equation for $\mathbf{W}$:

$$\nabla_{\mathbf{W}} J(\mathbf{W}) = E_{\mathbf{x}}[I + \mathbf{h}(\mathbf{Wx})(\mathbf{Wx})^T] \left[\mathbf{W}^T\right]^{-1} = \mathbf{0} \tag{5.1}$$

However, this equation is difficult to be solved directly due to the involvement of mixture $\mathbf{x}$. On the other hand, it is equivalent to solve the equilibrium equation for

V:

$$\nabla_{\mathbf{V}} J(\mathbf{V}) = E_{\mathbf{s}}[I + \mathbf{h}(\mathbf{Vs})(\mathbf{Vs}^T)] \left[\mathbf{V}^T\right]^{-1} = \mathbf{0} \tag{5.2}$$

since $\nabla_{\mathbf{W}} J(\mathbf{W}) = \nabla_{\mathbf{V}} J(\mathbf{V}) \mathbf{A}^T$ and $\mathbf{A}$ is non-singular [74]. This equation is easier to solve since sources $\mathbf{s}$ are independent. Hence, it is more appropriate to investigate the J scalar field in the $\mathbf{V}$-parameter space rather than in the $\mathbf{W}$-parameter space.

Secondly, the stability of the equilibrium points are determined by checking the Hessian matrix:

$$\nabla^2_{\mathbf{V}} J(\mathbf{V}) = \mathbf{Q} = \begin{bmatrix} \frac{\partial^2 J}{\partial v_{11} \partial v_{11}} & \cdots & \frac{\partial^2 J}{\partial v_{11} \partial v_{1n}} & \frac{\partial^2 J}{\partial v_{11} \partial v_{21}} & \cdots & \frac{\partial^2 J}{\partial v_{11} \partial v_{nn}} \\ \vdots & \ddots & \vdots & & & \vdots \\ \frac{\partial^2 J}{\partial v_{1n} \partial v_{11}} & \cdots & \frac{\partial^2 J}{\partial v_{1n} \partial v_{1n}} & & & \vdots \\ \frac{\partial^2 J}{\partial v_{21} \partial v_{11}} & & & \ddots & & \vdots \\ \vdots & & & & \ddots & \vdots \\ \frac{\partial^2 J}{\partial v_{nn} \partial v_{11}} & \cdots & \cdots & \cdots & \cdots & \frac{\partial^2 J}{\partial v_{nn} \partial v_{nn}} \end{bmatrix} \tag{5.3}$$

If the Hessian is positive definite, the equilibrium point is a local minimum; if the Hessian is negative definite, the equilibrium point is a local maximum; if the Hessian is neither positive definite nor negative definite, i.e., some of the eigenvalues are/is positive and some are/is negative, the equilibrium point is a saddle point. The stability can be checked by directly finding the signs of the eigenvalues or by checking the signs of the leading principle minors.

5.1.1 Singularity of $J(\mathbf{V})$

LEMMA 1 (Singular subspace) For the information-theoretic ICA scheme on any number of channels, as $\det \mathbf{V} \to 0$, $J(\mathbf{V}) \to +\infty$.

Proof If $\det \mathbf{V} = \det \mathbf{W} = 0$, there is a deterministic linear dependence on the recovered signals, that is, any y_i can be written as $y_i = L(y_1, ..., y_{i-1}, y_{i+1}, ..., y_n)$ where $L(.)$ is a linear function. Hence, $z_i = f_i(y_i)$ also bear a deterministic relationship with $\{z_1, ..., z_{i-1}, z_{i+1}, ..., z_n\}$ and the conditional entropy $H(z_i | z_1, ..., z_{i-1}, z_{i+1}, ..., z_n) \to -\infty$. As the joint entropy $H(\mathbf{z}) = H(z_i) + H(z_i | z_1, ..., z_{i-1}, z_{i+1}, ..., z_n)$, we have $H(\mathbf{z}) \to -\infty$. Hence, on the singular subspace, $J(\mathbf{V}) \to +\infty$. □

The (n^2-1)-dimensional subspace defined by $\det \mathbf{V} = 0$ ($\det \mathbf{W} = 0$) in the n^2-dimensional $\mathbf{V}$-parameter space ($\mathbf{W}$-parameter space) is called the 'singular subspace'.

Remark 4 For the natural gradient algorithm eq. (4.28), if $\mathbf{W}$ is initialized as a singular matrix, it will subsequently be trapped in the singular subspace $\det \mathbf{W} = 0$ because

$$\det \mathbf{W}_{t+1} = (\det [\mathbf{I} + \epsilon(t)[\mathbf{I} + \mathbf{H}]])(\det \mathbf{W}_t) = 0 \tag{5.4}$$

Surely, it cannot perform source separation.

5.1.2 Continuity of $J(\mathbf{V})$

LEMMA 2 (Continuity) For the information-theoretic ICA scheme with monotonic decreasing, odd $h_i(y_i)$ nonlinearity, including:

- the cubic nonlinearity $h_i(y_i) = -c_i y_i^3, c_i > 0$, and other $h_i(y_i) = -c_i y_i^p$, $c_i > 0$, p being a positive odd integer,
- $h_i(y_i)$ being reversed sigmoids like $h_i(y_i) = 1 - 2\,\mathrm{logsig}(y_i) = (\exp(-y_i) - 1)/(1 + \exp(-y_i))$, $h_i(y_i) = -2\tanh(y_i)$ or $h_i(y_i) = -2y_i/(1 + y_i^2)$,
- $h_i(y_i) = -c_i (y_i)^{1/p}, c_i > 0, p$ being a positive odd integer,

on any number of channels of signals, $J(\mathbf{V})$ is continuous at any non-singular $\mathbf{V}$.

Proof $J(\mathbf{V})$ is continuous at some finite point $\mathbf{V}^*$ if and only if the gradient $\nabla_V J(\mathbf{V})$ exists and is finite at $\mathbf{V}^*$. Consider

$$\nabla_V J(\mathbf{V}) = -\left\{ \frac{(\mathrm{adj}\mathbf{V})^T}{\det \mathbf{V}} + E_{\mathbf{s}}[\mathbf{h}(\mathbf{V}\mathbf{s})(\mathbf{s})^T] \right\} \tag{5.5}$$

where adj$\mathbf{V}$ denotes the adjoint of $\mathbf{V}$. It is obvious that $E_{\mathbf{s}}[h_i(\mathbf{v}_i^T \mathbf{s}) s_i]$ at a finite point $\mathbf{V}^*$ is finite for monotonic decreasing, odd $h_i(y_i)$. The magnitudes of the elements in the first term are infinitely large only when $\det \mathbf{V} = 0$, hence $\nabla_V J(\mathbf{V})$ exists and is finite for any non-singular $\mathbf{V}$. Therefore, $J(\mathbf{V})$ is continuous anywhere except on the singular subspace □.

5.1.3 Behavior of $J(\mathbf{V})$ along a radially outward line

LEMMA 3 For the information-theoretic ICA scheme with odd, monotonic decreasing $h_i(y_i)$ nonlinearity on any number of channels of signals, consider the value of $J(N\widehat{\mathbf{V}})$ along a radially outward line passing through a non-singular $\widehat{\mathbf{V}}$, where $N = ||\mathbf{V}|| = [(\mathrm{Vec}(\mathbf{V}^T))^T \cdot \mathrm{Vec}(\mathbf{V}^T)]^{1/2} = \sqrt{\sum_{i=1}^n \sum_{j=1}^n v_{ij}^2}$ is the norm of $\mathbf{V}$, $\mathbf{V} = N\widehat{\mathbf{V}}$, and $\widehat{\mathbf{V}} = \mathbf{V}/N$ is a point on the sphere of unit norm. For any non-singular $\widehat{\mathbf{V}}$, there exists a norm N_0 such that $J(N_0\widehat{\mathbf{V}})$ is the unique local minimum of $J(N\widehat{\mathbf{V}})$, $N \in [0, +\infty]$. $J(N\widehat{\mathbf{V}})$ is monotonic increasing with N (along the radially outward direction) if $N > N_0$, and is monotonic decreasing from $+\infty$ to $J(N_0\widehat{\mathbf{V}})$ if $0 < N < N_0$.

Proof The directional derivative of $J(\mathbf{V})$ along the radially outward direction $\mathrm{Vec}(\widehat{\mathbf{V}}^T)$ is:

$$\begin{aligned}
\frac{dJ(N\widehat{\mathbf{V}})}{dN} &= \frac{\partial J}{\partial \mathrm{Vec}(\mathbf{V}^T)} \cdot \mathrm{Vec}(\widehat{\mathbf{V}}^T) \\
&= -\frac{1}{N}\left\{ \mathrm{Vec}\left[\frac{(\mathrm{adj}\mathbf{V})^T}{\det \mathbf{V}} + E[\mathbf{h}(\mathbf{V}\mathbf{s})(\mathbf{s})^T] \right]^T \right\}^T \cdot \mathrm{Vec}(\mathbf{V}^T) \\
&= -\frac{1}{N}\left\{ \frac{1}{\det \mathbf{V}} \sum_{i=1}^n \sum_{j=1}^n v_{ij} \mathrm{cof} v_{ij} + \sum_{i=1}^n E_{\mathbf{s}}[h_i(\mathbf{v}_i^T\mathbf{s})(\mathbf{v}_i^T\mathbf{s})] \right\} \\
&= -\frac{1}{N}\left\{ \frac{1}{\det \mathbf{V}} (n \det \mathbf{V}) + \sum_{i=1}^n E_{\mathbf{s}}[h_i(\mathbf{v}_i^T\mathbf{s})(\mathbf{v}_i^T\mathbf{s})] \right\} \\
&= -\frac{1}{N} \sum_{i=1}^n \{1 + E_{\mathbf{s}}[h_i(\mathbf{v}_i^T\mathbf{s})(\mathbf{v}_i^T\mathbf{s})]\}
\end{aligned} \tag{5.6}$$

Denote $\widehat{\mathbf{V}} = [\widehat{\mathbf{v}_1}, \ldots, \widehat{\mathbf{v}_n}]^T$. Then

$$\frac{dJ(N\widehat{\mathbf{V}})}{dN} = -\frac{1}{N} \sum_{i=1}^n \left\{ 1 + N E_{\mathbf{s}}[h_i(N\widehat{\mathbf{v}}_i^T\mathbf{s})(\widehat{\mathbf{v}}_i^T\mathbf{s})] \right\} \tag{5.7}$$

where the factor

$$F(N) = -\sum_{i=1}^n \left\{ 1 + N E_{\mathbf{s}}[h_i(N\widehat{\mathbf{v}}_i^T\mathbf{s})(\widehat{\mathbf{v}}_i^T\mathbf{s})] \right\} \tag{5.8}$$

is a monotonic increasing function of N since $\{h_i(y_i)\}$ are odd monotonic decreasing. Noting that $F(0) = -n$ and $F(N) \to +\infty$ as $N \to +\infty$, we deduce $F(N)$ must change sign at some N_0. Since the sign of $\mathrm{d}J(N\widehat{\mathbf{V}})/\mathrm{d}N = F(N)/N$ is always the same as

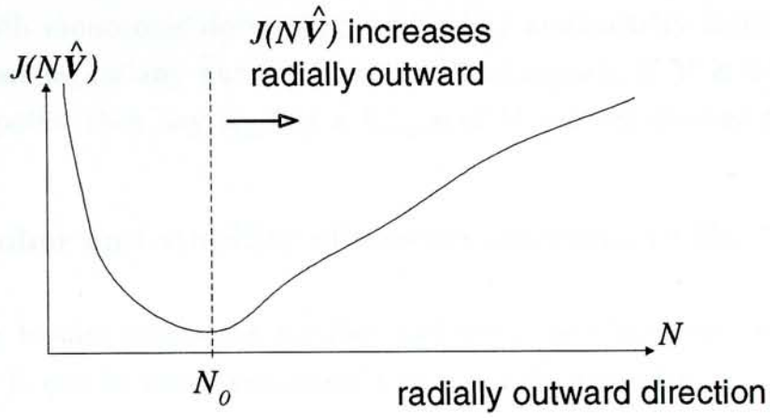


Figure 5.1: An illustration of the behavior of $J(\mathbf{V})$ along a radially outward line described by Lemma 3.

$F(N)$, we have $\mathrm{d}J(N\hat{\mathbf{V}})/\mathrm{d}N < 0$ if $0 < N < N_0$, $\mathrm{d}J(N\hat{\mathbf{V}})/\mathrm{d}N = 0$ if $N = N_0$ and $\mathrm{d}J(N\hat{\mathbf{V}})/\mathrm{d}N > 0$ if $N > N_0$. Noting also that $J(N\hat{\mathbf{V}}) \to +\infty$ as $N \to 0$, since $\mathbf{V} = \mathbf{0}$ is a singularity, the theorem is proved. □

An illustration for the behavior of $J(\mathbf{V})$ along a radially outward line described in Lemma 3 is sketched in Figure 5.1.

Remark 5 The increase of $J(\mathbf{V})$ in the outward direction of $\mathbf{V}$ can be understood intuitively by considering entropy $H(\mathbf{z})$. The sigmoidal-shaped nonlinear transformation function $f_i(y_i)$ is limiting to the lower bound as $y_i \to -\infty$ and limiting to the upper bound as $y_i \to +\infty$. Both ends of $f_i(y_i)$ tend to be more flat as $y_i \to \pm\infty$, that is, large range of y_i with large magnitude is mapped into small range of $z_i = f_i(y_i)$. As $||\mathbf{V}||$ tends to be large, the random variable y_i generally becomes large and $z_i = f_i(y_i)$ is squeezed to be more concentrated at the regions near the upper and lower bound. Hence there is less randomness in $\mathbf{z}$, the entropy $H(\mathbf{z})$ decreases or equivalently, $J(\mathbf{V})$ increases.

5.1.4 Impossibility of divergence of the information-theoretic ICA algorithms with a large class of nonlinearities

As the information-theoretic ICA algorithm performs descent in $J(\mathbf{V})$, by Lemma 3, $\mathbf{V}$ (any element of it) will not move in the outward direction if it is already in the region that $J(\mathbf{V})$ is increasing outward. Hence we have the following corollary:

COROLLARY 1 (Impossibility of divergence) For the information-theoretic ICA algorithms with monotonic decreasing, odd $h_i(y_i)$ nonlinearity including those listed in Lemma 2 acting on any number of channels of signals, if $\mathbf{V}$ is initialized at some non-singular point, then any v_{ij}, $i, j = 1, ..., n$ of $\mathbf{V}$ will not diverge to $\pm\infty$.

5.1.5 Number and stability of correct solutions in the 2-channel case

The following lemma states the number and forms of the correct solution in the 2-channel case. It can be easily generalized to the n-channel case.

LEMMA 4 For the Information-theoretic ICA scheme with monotonic decreasing, odd $h_i(y_i)$ in 2-channel case, the correct solutions have the form:

$$\textbf{Solution A1-A4} \qquad \mathbf{V} = \begin{bmatrix} \pm|v_{11}^*| & 0 \\ 0 & \pm|v_{22}^*| \end{bmatrix} \tag{5.9}$$

where $|v_{11}^*|$ is the unique magnitude of the solutions for

$$1 + E_{s_1}[h_1(v_{11}s_1)s_1v_{11}] = 0 \tag{5.10}$$

and $|v_{22}^*|$ is the unique magnitude of the solutions for

$$1 + E_{s_2}[h_2(v_{22}s_2)s_2v_{22}] = 0 \tag{5.11}$$

or

$$\textbf{Solution A5-A8} \qquad \mathbf{V} = \begin{bmatrix} 0 & \pm|v_{12}^*| \\ \pm|v_{21}^*| & 0 \end{bmatrix} \tag{5.12}$$

where $|v_{12}^*|$ is the unique magnitude of the solutions for

$$1 + E_{s_2}[h_1(v_{12}s_2)s_2v_{12}] = 0 \tag{5.13}$$

and $|v_{21}^*|$ is the unique magnitude of the solutions for

$$1 + E_{s_1}[h_2(v_{21}s_1)s_1v_{21}] = 0 \tag{5.14}$$

There are totally 8 correct solutions in the 2-channel case.

Proof Substituting $v_{12} = v_{21} = 0$ for Solutions A1-A4, the cross-coupling equilibrium equations:

$$E_{\mathbf{s}}[h_1(v_{11}s_1)s_2v_{22}] = E_{s_1}[h_1(v_{11}s_1)]E_{s_2}[s_2]v_{22} = 0 \tag{5.15}$$

$$E_{\mathbf{s}}[h_2(v_{22}s_2)s_1v_{11}] = E_{s_2}[h_2(v_{22}s_2)]E_{s_1}[s_1]v_{11} = 0 \tag{5.16}$$

are automatically satisfied. As $h_i(y_i)$ are monotonic decreasing and odd, the magnitudes $|v_{11}^*|$ and $|v_{22}^*|$ have unique solutions for the self-coupling equations eq. (5.10) and (5.11) respectively. The case for Solutions A5-A8 is similar. Counting the combination of signs of the elements, we conclude there are exactly 8 correct solutions. □

LEMMA 5 For the Information-theoretic ICA scheme with monotonic decreasing, odd $h_i(y_i)$ in 2-channel case, the sufficient and necessary condition for Solutions A1-A4 in eq. (5.9) to be stable is:

$$E[s_1^2]E[s_2^2]E_{s_1}[-h_1'(v_{11}^*s_1)]E_{s_2}[-h_2'(v_{22}^*s_2)] - \frac{1}{{v_{11}^*}^2{v_{22}^*}^2} > 0 \tag{5.17}$$

Proof The Hessian matrix for Solutions A1-A4 is:

$$\nabla_{\mathbf{V}}^2 J(\mathbf{V}) = \mathbf{Q} = \begin{bmatrix} \frac{1}{v_{11}^*} + E_{s_1}[-h_1'(v_{11}^*s_1)s_1^2] & 0 & 0 & 0 \\ 0 & E_{\mathbf{s}}[-h_1'(v_{11}^*s_1)s_2^2] & \frac{1}{v_{11}^*v_{22}^*} & 0 \\ 0 & \frac{1}{v_{11}^*v_{22}^*} & E_{\mathbf{s}}[-h_2'(v_{22}^*s_2)s_1^2] & 0 \\ 0 & 0 & 0 & \frac{1}{v_{22}^*} + E_{s_2}[-h_2'(v_{22}^*s_2)s_2^2] \end{bmatrix} \tag{5.18}$$

The sufficient and necessary condition for the Hessian to be positive definite is that all leading principal minors must be positive. The elements q_{11}, q_{44} are always positive. Hence, the condition for all leading principal minors to be positive is:

$$\det \begin{bmatrix} q_{22} & q_{23} \\ q_{32} & q_{33} \end{bmatrix} = E_{\mathbf{s}}[-h_1'(v_{11}^*s_1)s_2^2]E_{\mathbf{s}}[-h_2'(v_{22}^*s_2)s_1^2] - \frac{1}{{v_{11}^*}^2{v_{22}^*}^2} > 0 \tag{5.19}$$

By the independence assumption of the source signals, the condition eq. (5.17) is reached. □

Similarly,

LEMMA 6 For the Information-theoretic ICA scheme with monotonic decreasing, odd $h_i(y_i)$ in 2-channel case, the sufficient and necessary condition for Solutions A5-A8 in eq. (5.9) to be stable is:

$$E[s_1^2]E[s_2^2]E_{s_2}[-h_1'(v_{12}^* s_2)]E_{s_1}[-h_2'(v_{21}^* s_1)] - \frac{1}{v_{12}^{*\,2} v_{21}^{*\,2}} > 0 \tag{5.20}$$

The above conditions are actually a partial result. To apply the lemmas to a particular nonlinearity, we have to substitute the $\{h_i(.)\}$ and $\mathbf{V}^*$ in to get a condition in terms of the statistics or distribution of the sources only. However, one difficulty is to pick the elements $\{v_{ij}^*\}$ out of the expectation operation. Only in the simple cubic nonlinearity case can we expand the term in the expectation operation and pick $\{v_{ij}^*\}$ out. We cannot analyze more general nonlinearities like reversed sigmoids or cubic root nonlinearities. Moreover, investigation on other equilibrium points and stability of them are needed for global convergence analysis.

5.1.6 Scale for the equilibrium points

If a family of nonlinear function $g_i(y_i)$ can be written as $(1/\theta_i)\check{g}_i(y_i/\theta_i)$, θ_i is called the scale parameter of $g_i(y_i)$. We shall prove that this scale parameter only have the effect of controlling the scale of the equilibrium points and does not affect the separation capability of the nonlinearity.

LEMMA 7 Consider an information-theoretic ICA system A using $\tilde{g}_i(y_i)$ and another information-theoretic ICA system B using $g_i(y_i) = (1/\theta_i)\tilde{g}_i(y_i/\theta_i)$, $i = 1, ..., n$. For any $\mathbf{V}^{A*} = [\mathbf{v}_1^{A*}, ..., \mathbf{v}_n^{A*}]^T$, let $\mathbf{V}^{B*} = [\mathbf{v}_1^{B*}, ..., \mathbf{v}_n^{B*}]^T$ such that $\mathbf{v}_i^{B*} = \theta_i \mathbf{v}_i^{A*}$, $i = 1, ..., n$. Then, $\mathbf{V}^{B*}$ is an equilibrium point of system B if and only if $\mathbf{V}^{A*}$ is an equilibrium point of system A.

Proof For the $g_i(y_i)$ used in system B, we have $h_i(y_i) = (1/\theta_i)\tilde{h}_i(y_i/\theta_i)$, where $\tilde{h}_i(r) = \tilde{g}_i'(r)/\tilde{g}_i(r)$. If $\mathbf{V}^{A*}$ is an equilibrium point of system A, $\mathbf{V}^{A*}$ satisfies the equilibrium equation:

$$\nabla_{\mathbf{V}} \tilde{J}(\mathbf{V})\mathbf{V}^T|_{\mathbf{V}^{A*}} = E_{\mathbf{s}}[\mathbf{I} + \tilde{\mathbf{h}}(\mathbf{V}^{A*}\mathbf{s})(\mathbf{V}^{A*}\mathbf{s})^T] = \mathbf{0} \tag{5.21}$$

where $\tilde{\mathbf{h}}(\mathbf{y}) = [\tilde{h}_1(y_1), ..., \tilde{h}_n(y_n)]^T$. For the diagonal elements (self-coupling equations),we have, for $i = 1, ..., n$,

$$\begin{aligned} 0 &= 1 + E_{\mathbf{s}}\left[\tilde{h}_i([\mathbf{v}_i^{A*}]^T\mathbf{s})\,([\mathbf{v}_i^{A*}]^T\mathbf{s})\right] \\ &= 1 + E_{\mathbf{s}}\left[\tilde{h}_i\left(\frac{[\mathbf{v}_i^{B*}]^T}{\theta_i}\mathbf{s}\right)\left(\frac{[\mathbf{v}_i^{B*}]^T}{\theta_i}\mathbf{s}\right)\right] \\ &= 1 + E_{\mathbf{s}}\left[\frac{1}{\theta_i}\tilde{h}_i\left(\frac{[\mathbf{v}_i^{B*}]^T\mathbf{s}}{\theta_i}\right)([\mathbf{v}_i^{B*}]^T\mathbf{s})\right] \\ &= 1 + E_{\mathbf{s}}\left[h_i([\mathbf{v}_i^{B*}]^T\mathbf{s})\,([\mathbf{v}_i^{B*}]^T\mathbf{s})\right] \end{aligned} \tag{5.22}$$

For the off-diagonal elements (cross-coupling equations), we have, for $i, j = 1, ..., n$, $i \neq j$,

$$\begin{aligned} 0 &= E_{\mathbf{s}}\left[\tilde{h}_i([\mathbf{v}_i^{A*}]^T\mathbf{s})\left([\mathbf{v}_j^{A*}]^T\mathbf{s}\right)\right] \\ &= E_{\mathbf{s}}\left[\tilde{h}_i\left(\frac{[\mathbf{v}_i^{B*}]^T}{\theta_i}\mathbf{s}\right)\left(\frac{[\mathbf{v}_j^{B*}]^T}{\theta_j}\mathbf{s}\right)\right] \\ &= \frac{\theta_i}{\theta_j}E_{\mathbf{s}}\left[\frac{1}{\theta_i}\tilde{h}_i\left(\frac{[\mathbf{v}_i^{B*}]^T\mathbf{s}}{\theta_i}\right)\left([\mathbf{v}_j^{B*}]^T\mathbf{s}\right)\right] \end{aligned} \tag{5.23}$$

which implies

$$E_{\mathbf{s}}\left[h_i([\mathbf{v}_i^{B*}]^T\mathbf{s})([\mathbf{v}_j^{B*}]^T\mathbf{s})\right] = 0 \tag{5.24}$$

Combining eq. (5.22) and (5.24), we have

$$\nabla_{\mathbf{V}}J(\mathbf{V})\mathbf{V}^T|_{\mathbf{V}^{B*}} = E_{\mathbf{s}}[\mathbf{I} + \mathbf{h}(\mathbf{V}^{B*}\mathbf{s})(\mathbf{V}^{B*}\mathbf{s})^T] = \mathbf{0} \tag{5.25}$$

Hence, we have $\mathbf{V}^{B*}$ satisfies the equilibrium equation for system B, and are equilibrium points of system B. The converse can be proved using the reversed direction of the above method and the lemma holds. □

COROLLARY 2 For an information-theoretic ICA system using $g_i(y_i)$ from a scale family $(1/\theta_i)\tilde{g}_i(y_i/\theta_i)$, $i = 1, ..., n$, the magnitude of $\mathbf{v}_i^T$ of the equilibrium points is controlled by the scale parameter θ_i as $v_{ij} \propto \theta_i$, $j = 1, ..., n$, and after the system has converged to some equilibrium point, the magnitude of recovered signal $E[|y_i|]$ is proportional to θ_i.

COROLLARY 3 For an information-theoretic ICA system using $g_i(y_i)$ from a scale family $(1/\theta_i)\tilde{g}_i(y_i/\theta_i)$, $i = 1, ..., n$, the values of scale parameters $\{\theta_i\}$ do not affect the number and forms of the solutions of the equilibrium equation.

LEMMA 8 Consider an information-theoretic ICA system A using $\tilde{g}_i(y_i)$ and another information-theoretic ICA system B using $g_i(y_i) = (1/\theta_i)\tilde{g}_i(y_i/\theta_i)$ $i = 1, ..., n$. For any $\mathbf{V}^{A*} = [\mathbf{v}_1^{A*}, ..., \mathbf{v}_n^{A*}]^T$, let $\mathbf{V}^{B*} = [\mathbf{v}_1^{B*}, ..., \mathbf{v}_n^{B*}]^T$ such that $\mathbf{v}_i^{B*} = \theta_i \mathbf{v}_i^{A*}$, $i = 1, ..., n$. Then, the stability of $\mathbf{V}^{B*}$ in system B is the same as the stability of $\mathbf{V}^{A*}$ in system A.

Proof The elements of the Hessian matrix eq. (5.3) at $\mathbf{V}^{B*}$ in system B is given by:

$$\begin{aligned}\left.\frac{\partial^2 J}{\partial v_{ij}\partial v_{kl}}\right|_{\mathbf{V}^{B*}} =& \frac{\mathrm{cof} v_{ij}^{B*}\,\mathrm{cof} v_{kl}^{B*}}{(\det \mathbf{V}^{B*})^2} \\ &- (1-\delta_{ik})(1-\delta_{jl})\frac{(-1)^{i+j+k+l}\det M_{(i,j),(k,l)}^{B*}}{\det \mathbf{V}^{B*}} \\ &- \delta_{ik} E[h_i'([\mathbf{v}_i^{B*}]^T\mathbf{s})s_j s_k]\end{aligned} \tag{5.26}$$

where $\mathrm{cof} v_{ij}$ is the cofactor of v_{ij}, and $M_{(i,j),(k,l)}$ denote the matrix $\mathbf{V}$ with row i, row k, column j and column l removed. Differentiating, we have $h_i'(y_i) = (1/\theta_i^2)\tilde{h}_i(y_i/\theta_i)$.

As each row $[\mathbf{v}_i^{B*}]^T = \theta_i[\mathbf{v}_i^{A*}]^T$, we have

$$\det \mathbf{V}^{B*} = \left(\prod_{i=1}^{n}\theta_i\right)\det \mathbf{V}^{A*} \tag{5.27}$$

and similarly,

$$\mathrm{cof} v_{ij}^{B*} = \left(\prod_{\substack{m=1\\ m\neq i}}^{n}\theta_m\right)\mathrm{cof} v_{ij}^{A*} \tag{5.28}$$

$$\det M_{(i,j),(k,l)}^{B*} = \left(\prod_{\substack{m=1\\ m\neq i, m\neq k}}^{n}\theta_m\right)\det M_{(i,j),(k,l)}^{A*} \tag{5.29}$$

Hence, we can link the elements of the Hessian of system B at $\mathbf{V}^{B*}$ to the corresponding elements of the Hessian of system A at $\mathbf{V}^{A*}$:

$$\begin{aligned}\left.\frac{\partial^2 J}{\partial v_{ij}\partial v_{kl}}\right|_{\mathbf{V}^{B*}} &= \frac{\mathrm{cof}\, v_{ij}^{A*}\mathrm{cof}\, v_{kl}^{A*}}{\theta_i\theta_k(\det \mathbf{V}^{A*})^2} \\ &\quad - (1-\delta_{ik})(1-\delta_{jl})\frac{(-1)^{i+j+k+l}\det M^{A*}_{(i,j),(k,l)}}{\theta_i\theta_k \det \mathbf{V}^{A*}} \\ &\quad - \delta_{ik}\frac{1}{\theta_i^2}E[h_i'([\mathbf{v}_i^{A*}]^T\mathbf{s})s_j s_k] \\ &= \frac{1}{\theta_i\theta_k}\left.\frac{\partial^2 \tilde{J}}{\partial v_{ij}\partial v_{kl}}\right|_{\mathbf{V}^{A*}}\end{aligned} \tag{5.30}$$

The stability of the point is completely characterized by the signs of the leading principal minors $p_1,...,p_{n^2}$, which are defined by:

$$p_r = \det\begin{bmatrix} q_{11} & \cdots & q_{1r} \\ \vdots & \ddots & \vdots \\ q_{r1} & \cdots & q_{rr} \end{bmatrix} \tag{5.31}$$

(e.g. see [3]) Note from eq. (5.30), the factor $1/\theta_i$ is common to each element in the row and the factor $1/\theta_k$ is common to each element in the column of the Hessian matrix. Hence we get that the leading principal minor p_r for the Hessian matrix of system B at $\mathbf{V}^{B*}$ is proportional and has the same sign as the leading principal minor $\tilde{p}_r$ for the Hessian matrix of system A at $\mathbf{V}^{A*}$,:

$$p_r|_{\mathbf{V}^{B*}} = \Theta_r \tilde{p}_r|_{\mathbf{V}^{A*}} \qquad r = 1,...,n^2 \tag{5.32}$$

where Θ_r is the product of some corresponding $(1/\theta_i)$, $i \in \{1,...,n\}$

Therefore, the lemma is proved. □

COROLLARY 4 The scale parameter cannot affect the stability of the equilibrium points.

A nonlinearity is said to be capable of separating some sources if $\mathbf{V}$ converge to a correct solution $\mathbf{PD}$. The separation capability hence depends on the number of equilibrium points, forms (whether they are equal to some $\mathbf{PD}$) of the equilibrium points and the stability of the equilibrium points. Since the scale parameter can affect none of factors, we have the following corollary.

COROLLARY 5 The scale parameter cannot affect the separation capability and a family of information-theoretic ICA system using $g_i(y_i)$ from a scale family $(1/\theta_i)\tilde{g}_i(y_i/\theta_i)$ have the same separation capability.

5.1.7 Absence of local maximum of $J(\mathbf{V})$

LEMMA 9 For the information-theoretic ICA scheme with monotonic decreasing, odd $h_i(y_i)$ nonlinearity, including those listed in Lemma 2, on any number of channels of signals, there is no local maximum of $J(\mathbf{V})$ in the whole $\mathbf{V}$-parameter space.

Proof The diagonal elements of the Hessian matrix of $J(\mathbf{V})$ is:

$$\frac{\partial^2 J}{\partial v_{ij}\partial v_{ij}} = \frac{(\mathrm{cof}\, v_{ij})^2}{(\det \mathbf{V})^2} + E_s[-h_i'(\mathbf{v}_i^T \mathbf{s}) s_j^2] \tag{5.33}$$

For any monotonic decreasing $h_i(y_i)$, $h_i'(y_i) < 0$, $E_s[-h_i'(\mathbf{v}_i^T \mathbf{s}) s_j^2] > 0$, and hence $\partial^2 J/\partial v_{ij}\partial v_{ij} > 0$. The first leading principal minor of the Hessian matrix is a diagonal element, and hence is positive. Therefore, the Hessian matrix cannot be negative definite at any $\mathbf{V}$ and there is no local maximum in the whole $\mathbf{V}$-parameter space. □

Chapter 6

The Algorithms with Cubic Nonlinearity

In this chapter, we shall present an analysis on the information-theoretic ICA algorithms with $h_i(y_i)$ being cubic nonlinearity. We shall theoretically prove the cubic nonlinearity can perform separation on two channels of mixtures of so-called *globally sub-Gaussian* source signals. This analysis can act as a detailed case study that visualize the relationship between nonlinearity and separation capability to be discussed in the next chapter. Some results will also be given to the 3-channel case.

In Section 6.1, we will introduce the cubic nonlinearity and compare it with the reversed sigmoid used by Bell & Sejnowski [5]. Theoretical results on the 2-channel case will be presented in Section 6.2 and Section 6.3 presents some experiments to verify the theoretical results. Some theoretical results on the 3-channel case will be given in Section 6.4 and some experiments for the 3-channel case will be presented in Section 6.5.

6.1 The Cubic Nonlinearity

In this chapter, we investigate the information-theoretic ICA algorithm with $h_i(y_i)$ being cubic nonlinearity. It is because the cubic nonlinearity is the simplest polynomial and the manipulation of it in the analysis is plausible. The cubic nonlinearity takes the form (Figure 6.1):

$$h_i(y_i) = -c_i y_i^3, \qquad c_i > 0, \qquad i = 1, \ldots, n \tag{6.1}$$

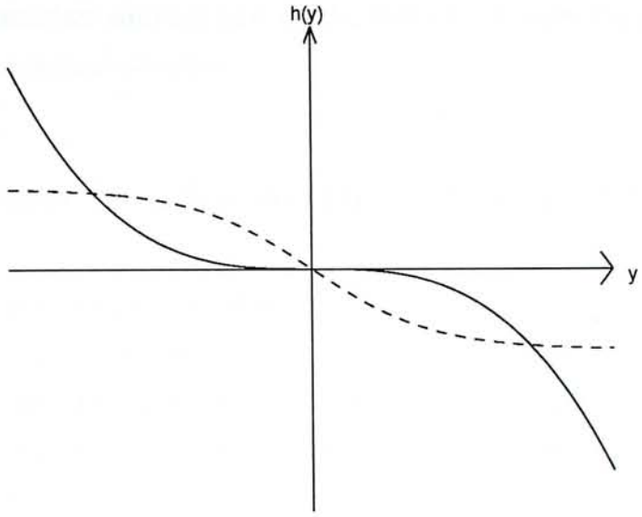


Figure 6.1: The cubic nonlinearity and reversed sigmoid for $h_i(y_i)$.
Dashed: Reversed sigmoid that works on source signals with sharply peaked pdf's.
Solid: The cubic nonlinearity that works on sub-Gaussian sources.

This $h_i(y_i)$ corresponds to

$$g_i(y_i) = C_i \exp(-\frac{c_i}{4} y_i^4), \tag{6.2}$$

and

$$f_i(y_i) = \int_{-\infty}^{y_i} g_i(r)\mathrm{d}r \tag{6.3}$$

where

$$C_i = \frac{1}{\int_{-\infty}^{\infty} \exp(-\frac{c_i}{4} y_i^4)\mathrm{d}y_i} = \frac{1}{c_i^{-1/4}} \frac{\sqrt{2}}{\gamma(1/4)}, \tag{6.4}$$

$\gamma(\cdots)$ being the gamma function, is a normalizing constant. Note that $c_i^{-1/4}$ is the scale parameter and in general c_i, $i = 1, 2, ..., n$ can be arbitrarily chosen as any positive number and can also be either equal or different.

In [5], the nonlinearity used by Bell & Sejnowski $h_i(y_i) = 1 - 2\,\mathrm{logsig}(y_i) = (\exp(-y_i) - 1)/(1 + \exp(-y_i)$, $h_i(y_i) = -2\tanh(y_i)$, and $h_i(y_i) = -2y_i/(1 + y_i^2)$ are all of reversed sigmoid shape (Figure 6.1). They are suggested to be able to perform separation on super-Gaussian sources and are experimentally verified to be able to separate human speech signals [5]. The cubic nonlinearity has different properties from the reversed sigmoids - it has opposite curvature and is unbounded. We assert that it

can separate sub-Gaussian sources and in the 2-channel case, we prove it can separate two 'globally sub-Gaussian' sources.

6.2 Theoretical Results on the 2-Channel Case

The global convergence behavior of the information-theoretic ICA algorithms with the cubic nonlinearity is investigated through the following three steps: 1) Explicitly and exhaustively determine the equilibrium points of the cost function, 2) Determine, for each equilibrium point, whether and under what condition it is a local minimum or saddle point, 3) Investigate the global configuration of the whole parameter space.

6.2.1 Equilibrium points

We start the determination of equilibrium points from the idea in [74] that investigate the **equilibrium equation** in the **V**-parameter space:

$$\nabla_V J(\mathbf{V}) = \mathbf{0}, \tag{6.5}$$

rather than $\nabla_{\mathbf{W}} J(\mathbf{W}) = \mathbf{0}$ in the **W**-parameter space for the reasons stated in Section 5.1.

Provided $\det \mathbf{V} \neq 0$, we have eq.(9) in [74] (without the constraints on that theorem), which can written as

$$E_{\mathbf{s}} \left[\mathbf{I} + \mathbf{h}(\mathbf{Vs})(\mathbf{Vs})^T\right] = \mathbf{0} \tag{6.6}$$

For the 2-channel case, we get a system of four equations with four unknowns $\{v_{11}, v_{12}, v_{21}, v_{22}\}$:

$$1 - c_1 E[y_1^4] = 1 - c_1(v_{11}^4\mu_1^4 + 2v_{11}^2 v_{12}^2 m + v_{12}^4\mu_2^4) = 0 \tag{6.7}$$

$$E[y_1^3 y_2] = v_{11}v_{21}(v_{11}^2\mu_1^4 + v_{12}^2 m) + v_{12}v_{22}(v_{12}^2\mu_2^4 + v_{11}^2 m) = 0 \tag{6.8}$$

$$E[y_2^3 y_1] = v_{11}v_{21}(v_{21}^2\mu_1^4 + v_{22}^2 m) + v_{12}v_{22}(v_{22}^2\mu_2^4 + v_{21}^2 m) = 0 \tag{6.9}$$

$$1 - c_2 E[y_2^4] = 1 - c_2(v_{21}^4\mu_1^4 + 2v_{21}^2 v_{22}^2 m + v_{22}^4\mu_2^4) = 0 \tag{6.10}$$

where $\mu_i^p = E[s_i^p]$ and $m = 3\mu_1^2\mu_2^2$. A key step to solve them is to write the cross-coupling equations (6.8) and (6.9) in the following form (suggested by Jiong Ruan):

$$\begin{bmatrix} v_{11}^2 & v_{12}^2 \\ v_{21}^2 & v_{22}^2 \end{bmatrix} \begin{bmatrix} \mu_1^4 & m \\ m & \mu_2^4 \end{bmatrix} \begin{bmatrix} v_{11}v_{21} \\ v_{12}v_{22} \end{bmatrix} = \mathbf{0} \tag{6.11}$$

Denote

$$\mathbf{M} = \begin{bmatrix} v_{11}^2 & v_{12}^2 \\ v_{21}^2 & v_{22}^2 \end{bmatrix} \begin{bmatrix} \mu_1^4 & m \\ m & \mu_2^4 \end{bmatrix} \tag{6.12}$$

Eq. (6.11) implies

$$\begin{bmatrix} v_{11}v_{21} \\ v_{12}v_{22} \end{bmatrix} = \mathbf{0} \qquad \text{or} \qquad \det \mathbf{M} = 0 \tag{6.13}$$

We treat these two exhaustive possibilities in case A and case B respectively. **Case A**

$$\begin{bmatrix} v_{11}v_{21} \\ v_{12}v_{22} \end{bmatrix} = \mathbf{0} \tag{6.14}$$

By eq. (6.7) and (6.10), we find that elements in any row of $\mathbf{V}$ cannot be all zeros simultaneously. Hence, putting $v_{12} = 0$ and $v_{21} = 0$ into the self-coupling equations (6.7) and (6.10), we get:

Solutions A1 - A4:

$$\mathbf{V} = \begin{bmatrix} \pm(c_1\mu_1^4)^{-1/4} & 0 \\ 0 & \pm(c_2\mu_2^4)^{-1/4} \end{bmatrix} \tag{6.15}$$

Putting $v_{11} = 0$ and $v_{22} = 0$ into eqs. (6.7) and (6.10), we get:

Solutions A5 - A8

$$\mathbf{V} = \begin{bmatrix} 0 & \pm(c_1\mu_2^4)^{-1/4} \\ \pm(c_2\mu_1^4)^{-1/4} & 0 \end{bmatrix} \tag{6.16}$$

Solution A1 - A8 are the eight and only eight solutions in case A. These eight solutions satisfy $\mathbf{V} = \mathbf{DP}$, and they are correct solutions that can perform source separation.

Case B

Now we consider the case

$$\det \mathbf{M} = (v_{11}^2 v_{22}^2 - v_{12}^2 v_{21}^2)(\mu_1^4 \mu_2^4 - m^2) = 0 \tag{6.17}$$

Assume that

$$\mu_1^4 \mu_2^4 - m^2 = \mu_1^4 \mu_2^4 - [3(\mu_1^2)^2][3(\mu_2^2)^2] \neq 0 \tag{6.18}$$

i.e., the two sources are not 'globally Gaussian', eq. (6.17) becomes

$$v_{11}^2 v_{22}^2 - v_{12}^2 v_{21}^2 = 0 \tag{6.19}$$

Firstly, we prove that all $v_{ij} \neq 0$ in case B. Again by eqs. (6.7) and (6.10), the elements in any row of $\mathbf{V}$ cannot be all zeros simultaneously. Hence, suppose $v_{11} = 0$, then $v_{12} \neq 0$. From eq. (6.8), $v_{22} = 0$. From eq. (6.19), $v_{21} = 0$. There is a contradiction that $v_{22} = 0$ and $v_{21} = 0$ simultaneously, so it is impossible that $v_{11} = 0$. Using the same argument on other elements of $\mathbf{V}$, the proposition all $v_{ij} \neq 0$ in case B is proved.

Secondly, we consider the possible combinations of the signs of v_{ij}. Let the sign of v_{ij} be $s_{ij} = v_{ij}/|v_{ij}|$. As all elements in the first two matrices of eq. (6.11) are positive, it is necessary that $(v_{11}v_{21})$ and $(v_{12}v_{22})$ are of different signs. Hence we have the constraint

$$s_{11}s_{12}s_{21}s_{22} = -1 \tag{6.20}$$

In other words, three of the four v_{ij} must be of the same sign, and the remaining one the opposite sign.

Coping with the constraint on the signs, eq. (6.19) implies:

$$v_{11} = -\frac{v_{12}v_{21}}{v_{22}} \tag{6.21}$$

Putting eq. (6.21) back into eqs. (6.7) to (6.10), we get:

Solutions B1 - B8

$$\mathbf{V} = \begin{bmatrix} s_{11}(2c_1\eta_1)^{-1/4} & s_{12}(2c_1\eta_2)^{-1/4} \\ s_{21}(2c_2\eta_1)^{-1/4} & s_{22}(2c_2\eta_2)^{-1/4} \end{bmatrix} \tag{6.22}$$

where

$$s_{ij} = 1 \text{ or } -1 \quad \text{satisfying} \quad s_{11}s_{12}s_{21}s_{22} = -1, \quad \text{with totally 8 combinations,} \tag{6.23}$$

$$\eta_1 = \mu_1^4 + \sqrt{\mu_1^4/\mu_2^4}m \tag{6.24}$$

$$\eta_2 = \mu_2^4 + \sqrt{\mu_2^4/\mu_1^4}m \tag{6.25}$$

Solutions B1 - B8 are the eight and only eight solutions in case B. However, for these solutions $\mathbf{V} \neq \mathbf{PD}$. They are spurious solutions that cannot perform source separation but still satisfy the four equilibrium equations.

In a word, the equilibrium equation eq. (6.6) is determined to have exactly sixteen solutions, namely Solution A1 - A8 and Solution B1 - B8. Solutions A1 - A8 (Group A) are correct solutions that can perform source separation, while Solutions B1 - B8 (Group B) are spurious solutions that cannot perform source separation.

Remark 6 The set of equations eq. (6.7) to (6.10) are physically interpreted as follows. Eq. (6.8) and (6.9) represent the coupling of the two channels (the two rows $\mathbf{v}_1^T$ and $\mathbf{v}_2^T$). The coupling terms $h_1(y_1)y_2$ and $h_2(y_2)y_1$ in the algorithm are formed by the cubic nonlinearity on one channel and a linear counterpart of the other channel. They restrict the forth order cross-moment $E[y_1^3y_2]$ and $E[y_2^3y_1]$ to be zero in the equilibrium equation and determine the possible form of the solutions. Eq. (6.7) and (6.10) control the magnitude of each recovered signal separately by restricting the forth order (self) moment $E[y_i^4]$ to be c_i^{-1}. Equivalently, they control the magnitudes of the row vectors $\mathbf{v}_1^T$ and $\mathbf{v}_2^T$. This result demonstrates Corollary 2.

6.2.2 Stability of the equilibrium points

The Hessian matrix $\nabla_V^2 J(\mathbf{V})$ is checked for each equilibrium point to determine whether and under what condition it is a local minimum or saddle point. In Appendix A, it is proved that for source signals satisfying the following condition:

$$E[s_1^4]E[s_2^4] - [3(E[s_1^2])^2][3(E[s_2^2])^2] < 0 \tag{6.26}$$

Solutions A1 - A8 are minima and Solutions B1 - B8 are saddle points of $J(\mathbf{V})$. For source signals satisfying:

$$E[s_1^4]E[s_2^4] - [3(E[s_1^2])^2][3(E[s_2^2])^2] > 0 \tag{6.27}$$

it is proved that Solutions B1 - B8 are minima and Solutions A1 - A8 are saddle points of $J(\mathbf{V})$.

Remark 7 A signal s is called sub-Gaussian if the kurtosis $E[s^4] - 3(E[s^2])^2$ is negative, called super-Gaussian if the kurtosis is positive and called Gaussian if the kurtosis is zero. A super-Gaussian signal has sharply peaked pdf with long tail and a sub-Gaussian signal has flat pdf with short tail. The term $E[s_1^4]E[s_2^4]-[3(E[s_1^2])^2][3(E[s_2^2])^2]$ in eq. (6.26) and (6.27) is defined as the 'joint kurtosis' for the two signals. Two signals that satisfy eq. (6.26), i.e., with negative joint kurtosis, are called 'globally sub-Gaussian' [28]. Similarly, two signals that satisfy eq. (6.27), i.e. with positive joint kurtosis, are called 'globally super-Gaussian', and two signals with zero joint kurtosis are called 'globally Gaussian'.

6.2.3 An alternative proof for the stability of the equilibrium points

The following is a useful equation that compares the values of J for two equilibrium points of the information-theoretic ICA algorithm with cubic nonlinearity on any number of channels of signals:

LEMMA 10 For the information-theoretic ICA algorithms with cubic nonlinearity eq. (4.21), if $\mathbf{V}_A$ and $\mathbf{V}_B$ are two equilibrium points, the equation:

$$J(\mathbf{V}_A) - J(\mathbf{V}_B) = \ln \frac{|\det \mathbf{V}_B|}{|\det \mathbf{V}_A|} \tag{6.28}$$

always holds.

Proof

$$\begin{aligned} J(\mathbf{V}_A) - J(\mathbf{V}_B) &= E_{\mathbf{s}}\left[\ln \frac{p_{\mathbf{s}}(\mathbf{s})}{|\det \mathbf{V}_A| \prod_{i=1}^{n} g_i(\mathbf{v}_{Ai}^T \mathbf{s})}\right] \\ &\quad - E_{\mathbf{s}}\left[\ln \frac{p_{\mathbf{s}}(\mathbf{s})}{|\det \mathbf{V}_B| \prod_{j=1}^{n} g_j(\mathbf{v}_{Bj}^T \mathbf{s})}\right] \\ &= \ln \frac{|\det \mathbf{V}_B|}{|\det \mathbf{V}_A|} \\ &\quad - \sum_{i=1}^{n} E_{\mathbf{s}}[\ln g_i(\mathbf{v}_{Ai}^T \mathbf{s})] + \sum_{j=1}^{n} E_{\mathbf{s}}[\ln g_j(\mathbf{v}_{Bj}^T \mathbf{s})] \end{aligned} \tag{6.29}$$

For the cubic nonlinearity, $g_i(y_i) = C_i \exp(-c_i y_i^4/4)$. Then,

$$\begin{aligned} J(\mathbf{V}_A) - J(\mathbf{V}_B) &= \ln \frac{|\det \mathbf{V}_B|}{|\det \mathbf{V}_A|} \\ &\quad + \sum_{i=1}^{n} E_{\mathbf{s}}[c_i(\mathbf{v}_{Ai}^T \mathbf{s})^4] - \sum_{j=1}^{n} E_{\mathbf{s}}[c_j(\mathbf{v}_{Bj}^T \mathbf{s})^4] \end{aligned} \tag{6.30}$$

However, we have $E_s[c_i(\mathbf{v}_i^T \mathbf{s})^4] = 1$ for any equilibrium point by the self-coupling equilibrium equations. Hence the second term in the above equation equals to n and the third term equals to $-n$. They cancel each other and hence eq. (6.28) holds. □

We realize that equilibrium points of the same group have the same value of J, since

$$J(\mathbf{V}) = E_{\mathbf{s}}[\ln p_{\mathbf{s}}(\mathbf{s})] - \ln|\det \mathbf{V}| - \sum_{i=1}^{n} E_{\mathbf{s}}[\ln g_i(\mathbf{v}_i^T \mathbf{s})] \tag{6.31}$$

and each term has the same value for the equilibrium points of the same group. By symmetry (scaling parameters cannot affect separation capability) , equilibrium points of the same group must have the same status of being local minima or saddle points. Owing to the continuity in non-singular regions and J tending to positive infinity on the singular subspace, the equilibrium point group with lower J value must be local minima. Since there are two groups only, the other group that has higher J value must be saddle points. (Otherwise, there must be a third group of saddle points.) Hence, we determine the condition that $J(\mathbf{V}_A) < J(\mathbf{V}_B)$ by Lemma 10:

$$|\det \mathbf{V}_A| = (c_1 c_2 \mu_1^4 \mu_2^4)^{-1/4} \tag{6.32}$$

$$\begin{aligned} |\det \mathbf{V}_B| &= 2(4c_1 c_2 \eta_1 \eta_2)^{-1/4} \\ &= \sqrt{2}(c_1 c_2(\mu_1^4 \mu_2^4 + 2\sqrt{\mu_1^4 \mu_2^4} m + m^2))^{-1/4} \end{aligned} \tag{6.33}$$

The condition

$$J(\mathbf{V}_A) - J(\mathbf{V}_B) = \ln \frac{\sqrt{2}(\mu_1^4\mu_2^4)^{1/4}}{(\mu_1^4\mu_2^4 + 2\sqrt{\mu_1^4\mu_2^4}m + m^2)^{1/4}} < 0 \tag{6.34}$$

becomes:

$$\frac{\mu_1^4\mu_2^4}{\mu_1^4\mu_2^4 + 2\sqrt{\mu_1^4\mu_2^4}m + m^2} < \frac{1}{4} \tag{6.35}$$

$$3\mu_1^4\mu_2^4 - 2\sqrt{\mu_1^4\mu_2^4}m - m^2 < 0 \tag{6.36}$$

$$(3\sqrt{\mu_1^4\mu_2^4} + m)(\sqrt{\mu_1^4\mu_2^4} - m) < 0 \tag{6.37}$$

$$\sqrt{\mu_1^4\mu_2^4} - m < 0 \tag{6.38}$$

or

$$\mu_1^4\mu_2^4 - 9(\mu_1^2)^2(\mu_2^2)^2 < 0 \tag{6.39}$$

Hence the correct solution group A are local minima and spurious solution group B are saddle points if eq. (6.39) is satisfied. Conversely, the spurious solution group B are local minima and correct solution group A are saddle points if the term in eq. (6.39) is positive. Hence, the above argument constitute an alternative proof to the stability of the equilibrium points. The use of Lemma 10 with the simplicity of the 2-channel case avoids dealing with the complex Hessian matrix in the investigation of stability.

6.2.4 Convergence Analysis

The equilibrium points of the algorithm have been exhaustively found and the condition for stability of the equilibrium points has been investigated. By corollary 1, $\mathbf{V}$ will not diverge to infinity and hence must converge to one of the local minimum. Therefore, we have the following theorem on the global convergence behavior of the algorithm.

THEOREM 1 For the information-theoretic ICA algorithms eq. (4.18), including the natural gradient algorithm eq. (4.28) the gradient algorithm eq. (4.23), with the cubic nonlinearity eq. (6.1) acting on two mixtures of two source signals, $\mathbf{W}$ being initialized at some nonsingular point,

- If the two source signals satisfy eq. (6.26), then $\mathbf{V}$ will converge to one of the Solutions A1 - A8 eq. (6.15) and (6.16).
- If the two source signals satisfy eq. (6.27), then $\mathbf{V}$ will converge to one of the Solutions B1 - B8 eq. (6.22).

In a word, Theorem 1 means that the information-theoretic ICA algorithms with cubic nonlinearity can separate two globally sub-Gaussian sources but cannot separate two globally super-Gaussian sources.

Remark 8 The cubic nonlinearity is related to some forth order statistics of the signals. This is because the coupling terms $h_i(y_i)y_j$ pick up the forth order moments $E[y_i^3 y_j]$ in the algorithm. In light of this relation between the cubic nonlinearity and forth order statistics, there is no surprise that the condition on successful separation of the source signals depends on the joint kurtosis of the sources. Interesting enough, the condition on successful separation in this work is exactly the same as those in [59] and [17] with cubic nonlinearity or forth order moment, although all the three works use different network architectures, different learning principles, different cost function (if any) and different algorithms. Similar conditions are found in other algorithms using the cubic nonlinearity [14, 64].

6.3 Experiments on the 2-Channel Case

The experiments are aimed at demonstrating the theoretical results. The natural gradient descent algorithm eq. (4.28) with the cubic nonlinearity eq. (6.1) is used. It is chosen that $c_1 = c_2 = 1$. For all experiments, the learning rate is kept at 0.0001. The following mixing matrix is used:

$$\mathbf{A} = \begin{bmatrix} 1 & 0.6 \\ 0.7 & 1 \end{bmatrix} \tag{6.40}$$

The experiments are run for a number of scan through the data set long enough that $\mathbf{W}$ looks having converged to a stable point.

The performance of the separation is determined by how close to $\mathbf{PD}$ the matrix $\mathbf{V} = \mathbf{WA}$ is. The element v_{ij} determines the amplitude of source signal s_j goes into recovered signal y_j and v_{ij}^2 determines the power. The greatest v_{ij}^2 in a row in $\mathbf{V}$ is regarded as the power of the 'signal' and the sum of other v_{ij}^2 of the row is regarded as

the power of the 'interference'. Hence, we define the interference-to-signal power ratio of channel i in decibel (dB) unit as:

$$I/S_i = 10 \times \log_{10}\left(\frac{\sum_{j\neq k} v_{ij}^2}{v_{ik}^2}\right), \qquad k = \arg\max_l v_{il}^2 \tag{6.41}$$

We use the mean of the interference-to-signal ratio as the performance index of source separation.

6.3.1 Experiments on two sub-Gaussian sources

In this experiment, two channels of artificially generated independently and identically distributed (iid) source signals with uniform distribution in [-1,1] are used. Each channel consist of 100,000 data points. Statistics of the data set are:

$$\mu_1^2 = 0.3326, \quad \mu_1^4 = 0.1991, \quad \mu_2^2 = 0.3337, \quad \mu_2^4 = 0.2007 \tag{6.42}$$

Both channels are sub-Gaussian, the standardized joint kurtosis $\mu_1^4\mu_2^4/[(\mu_1^2)^2(\mu_2^2)^2]-9$ is -5.756 and they are obviously globally sub-Gaussian. We tried two initializations of $\mathbf{W}$. The first one is the identity matrix, which is a natural choice when no supplementary information is provided, and means $\mathbf{V}$ starts from the original mixture $\mathbf{A}$. The second initialization is at one of the spurious Solution B, $\mathbf{W}_{\text{init}} = \mathbf{V}_B\mathbf{A}^{-1}$, where

$$\mathbf{V}_B = \begin{bmatrix} -0.9851 & 0.9832 \\ 0.9851 & 0.9832 \end{bmatrix} \tag{6.43}$$

to test the stability of solution group B.

For the case $\mathbf{W}$ is initialized as an identity matrix, the system converges in 30,000 data points. The performance graph, interference-to-signal ration versus number of data points scanned, is plotted in Figure 6.2. For the case initialization is at the Solution B, the system converges in 200,000 data points. The 4-dimensional trajectories of the convergence are plotted in two 2-dimensional graphs, Figures 6.3 (a) and (b), each of which being the projection to two coordinates.

$\mathbf{V}$ in the two cases converges to two of the correct Solution A's:

$$\mathbf{V}_A = \begin{bmatrix} \pm 1.4970 & 0 \\ 0 & 1.4941 \end{bmatrix} \tag{6.44}$$

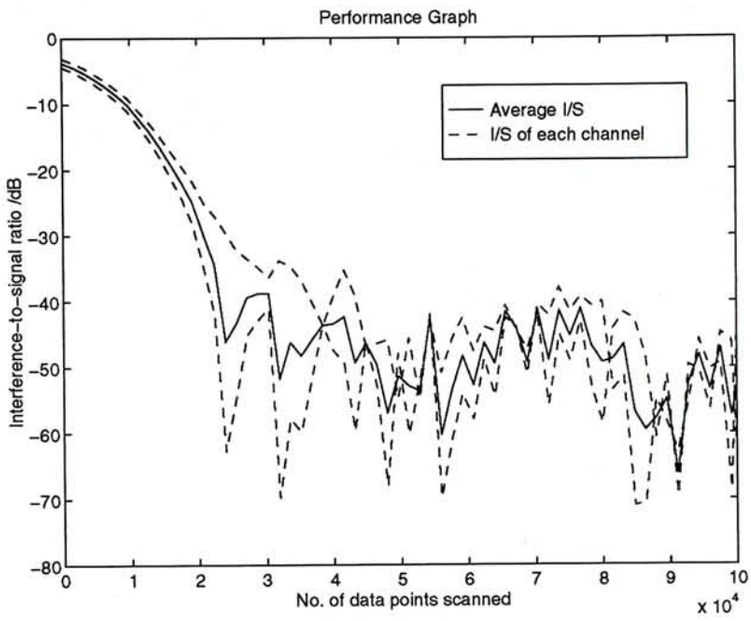


Figure 6.2: The performance graph of the algorithm with cubic nonlinearity acting on uniformly distributed sources.

The Interference-to-signal ratio reach -40dB in both case.

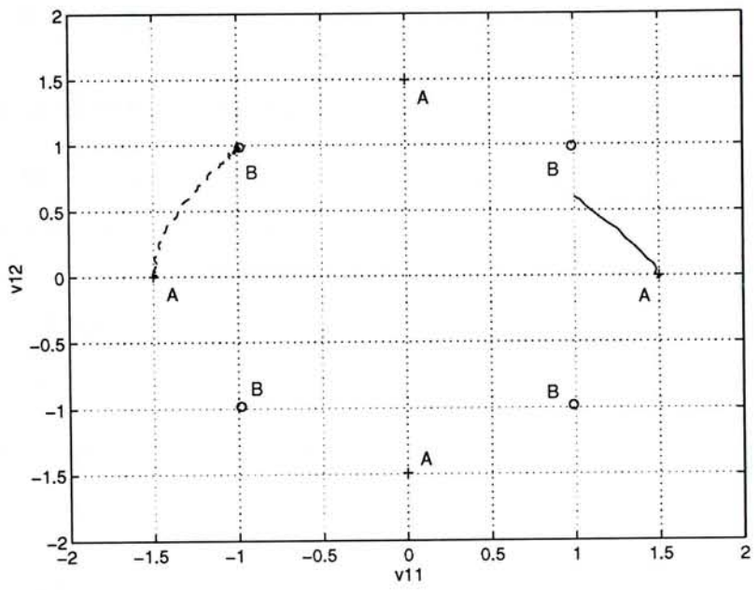
Hence it is experimentally verified that solution group A is stable and solution group B are saddle points in this case.

6.3.2 Experiments on two super-Gaussian sources

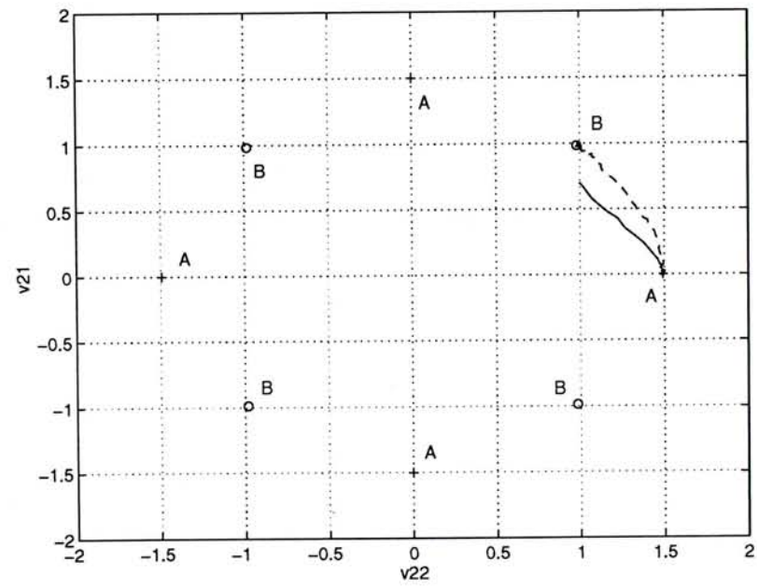
In this experiment, two channels of human speech signals are used. The first channel is recorded from a man telling a story and the second channel is recorded from a woman reading news. Both signals are recorded at 8kHz and consist of 100,000 data points (12.5 seconds). The signals are randomly permuted to get rid of non-stationarity. Statistics of the data set are:

$$\mu_1^2 = 0.0625, \quad \mu_1^4 = 0.0435, \quad \mu_2^2 = 0.2500, \quad \mu_2^4 = 0.3342 \tag{6.45}$$

Both signals are super-Gaussian. The standardized joint kurtosis $\mu_1^4\mu_2^4/[(\mu_1^2)^2(\mu_2^2)^2] - 9$ is 50.55 and obviously the source are globally super-Gaussian. We have also tried two initializations of $\mathbf{W}$. The first one is the identity matrix. The second initialization is



(a)



(b)

Figure 6.3: The trajectories of convergence of $\mathbf{V}$ of the information-theoretic ICA algorithm with cubic nonlinearity on uniformly distributed sources. Solid: $\mathbf{W}_{\text{init}} = \mathbf{I}$. Dashed: $\mathbf{W}_{\text{init}} = \mathbf{V}_B\mathbf{A}^{-1}$. Solution A's and B's are marked of by 'A' and 'B' respectively. The convergence points are solution A's.

at one of the correct Solution A, $\mathbf{W}_{\text{init}} = \mathbf{V}_A\mathbf{A}^{-1}$, where

$$\mathbf{V}_A = \begin{bmatrix} 2.1897 & 0 \\ 0 & 1.3152 \end{bmatrix} \tag{6.46}$$

to test the stability of solution group A.

For the case $\mathbf{W}$ is initialized as identity matrix, the system converges in 80,000 data points. For the case initialization is at the solution A, the system converges in 160,000 data points. The 4-dimensional trajectories of the convergence are plotted in two 2-dimensional graphs, Figures 6.4 (a) and (b), each of which being the projection to two coordinates.

$\mathbf{V}$ in both cases happens to converge to one of the spurious Solution B's:

$$\mathbf{V}_B = \begin{bmatrix} 1.6964 & 1.0187 \\ -1.6964 & 1.0187 \end{bmatrix} \tag{6.47}$$

Hence it is experimentally verified that solution group B is stable and solution group A are saddle points in this case.

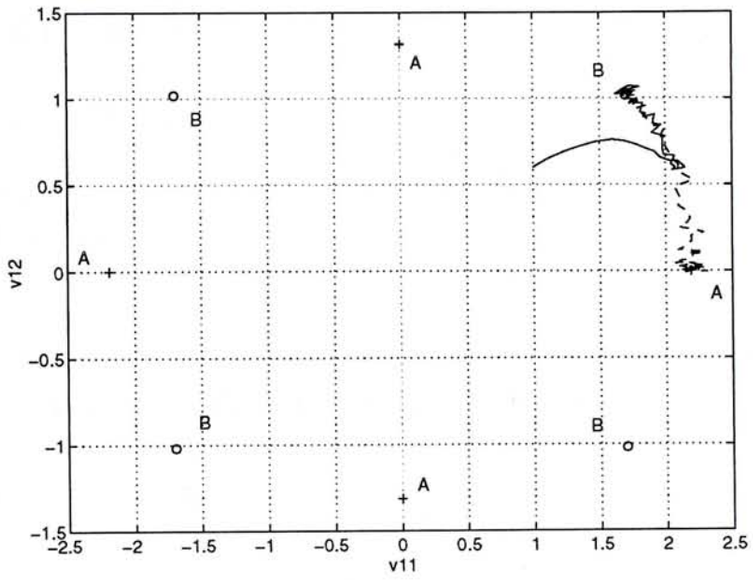
6.3.3 Experiments on one super-Gaussian source and one sub-Gaussian source which are globally sub-Gaussian

In this experiment, one signal recorded from a song at 8kHz, with the time index randomly permuted, is used as channel 1. One uniformly distributed signal used in the first experiment is used as channel 2. The statistics of the signals are:

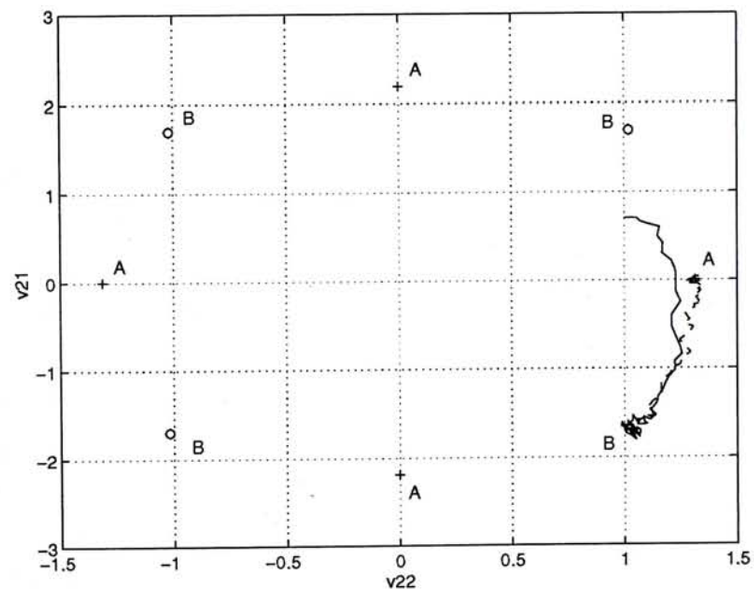
$$\mu_1^2 = 0.2532, \qquad \mu_1^4 = 0.2120, \qquad \mu_2^2 = 0.3326, \qquad \mu_2^4 = 0.1991. \tag{6.48}$$

The standardized joint kurtosis $\mu_1^4\mu_2^4/[(\mu_1^2)^2(\mu_2^2)^2] - 9$ is -3.048 and the two signals are globally sub-Gaussian. We also tried the two initialization that $\mathbf{W}_{\text{init}}=\mathbf{I}$ and $\mathbf{W}_{\text{init}}=\mathbf{V}_B\mathbf{A}^{-1}$. For the first initialization, $\mathbf{V}$ converged, after scanning 50,000 data points, to the solution A:

$$\mathbf{V}_A = \begin{bmatrix} 1.474 & 0 \\ 0 & 1.498 \end{bmatrix} \tag{6.49}$$



(a)



(b)

Figure 6.4: The trajectories of convergence of $\mathbf{V}$ of the information-theoretic ICA algorithm with cubic nonlinearity on permuted speech signals. Solid: $\mathbf{W}_{\text{init}} = \mathbf{I}$. Dashed: $\mathbf{W}_{\text{init}} = \mathbf{V}_A\mathbf{A}^{-1}$. The convergence points are solution B.

For the second initialization, $\mathbf{V}$ converged, after scanning 500,000 data points, to the solution A;

$$\mathbf{V}_A = \begin{bmatrix} 0 & 1.498 \\ 1.474 & 0 \end{bmatrix} \tag{6.50}$$

The Interference-to-signal ratio reached -35dB in both case.

Hence it is experimentally verified that solution group A is stable and solution group B are saddle points in this case.

6.3.4 Experiments on one super-Gaussian source and one sub-Gaussian source which are globally super-Gaussian

In this experiment, one uniformly distributed signal from the first experiment and one permuted speech signal from the second experiment is used. The statistics of the signals are:

$$\mu_1^2 = 0.2500, \qquad \mu_1^4 = 0.3342, \qquad \mu_2^2 = 0.3326, \qquad \mu_2^4 = 0.1991. \tag{6.51}$$

The standardized joint kurtosis $\mu_1^4\mu_2^4/[(\mu_1^2)^2(\mu_2^2)^2] - 9$ is 0.6239 and the two signals are globally super-Gaussian. We also tried the two initialization that $\mathbf{W}_{\text{init}}=\mathbf{I}$ and $\mathbf{W}_{\text{init}}=\mathbf{V}_A\mathbf{A}^{-1}$. For the first initialization, after scanning for about 2,000,000 data points, $\mathbf{V}$ became stabilized but still have some relatively large fluctuation. The slow and fluctuating convergence is plotted in Figure 6.5. A snapshot of $\mathbf{V}$ is

$$\mathbf{V} = \begin{bmatrix} 0.8703 & -1.1543 \\ 1.0088 & 0.9741 \end{bmatrix} \tag{6.52}$$

For the second initialization, after scanning for also for about 2,000,000 data points, $\mathbf{V}$ became stabilized but still have some relatively large fluctuation. A snapshot of $\mathbf{V}$ is

$$\mathbf{V} = \begin{bmatrix} 1.0094 & 0.9695 \\ -0.8684 & 1.1474 \end{bmatrix} \tag{6.53}$$

Both convergence points are a little deflected from the Solution B's:

$$\mathbf{V}_B = \begin{bmatrix} 0.9339 & \pm 1.0630 \\ \pm 0.9339 & 1.0630 \end{bmatrix} \tag{6.54}$$

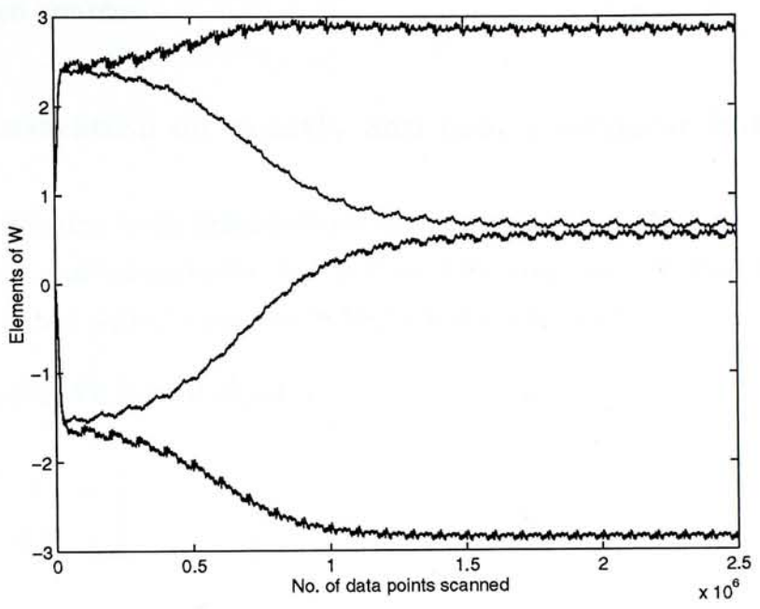


Figure 6.5: The convergence of $\mathbf{W}$ for two sources that are closed to globally Gaussian, showing the relatively large fluctuation.

The slow convergence and relatively large fluctuation occur might be because the $J(\mathbf{W})$ scalar field is relatively flat, that might be due to the magnitude of standardized joint kurtosis being small, i.e., the two sources are closed to globally Gaussian.

The above four experiments test different combination of sub-Gaussian or super-Gaussian sources and the experimental results are consistent to the theoretical analysis.

6.3.5 Experiments on asymmetric exponentially distributed signals

This experiment tests the information-theoretic ICA algorithm on sources with asymmetrical density. 2 channels of exponentially distributed ($\mu = 1$) sources are used. The samples are shifted to have zero means. Statistics of the data set are:

$$\mu_1^2 = 1.0278, \quad \mu_1^4 = 15.4697, \quad \mu_2^2 = 1.0711, \quad \mu_2^4 = 15.4276 \tag{6.55}$$

Obviously both signals are super-Gaussian. $\mathbf{W}$ was initialized to the identity matrix. After scanning for 80,000 data points, $\mathbf{V}$ converged to the spurious solution

$$\mathbf{V} = \begin{bmatrix} 0.404 & -0.403 \\ 0.404 & 0.403 \end{bmatrix} \tag{6.56}$$

Hence, this experiment verifies that the convergence analysis is valid for asymmetrically distributed sources.

6.3.6 Demonstration on exactly and nearly singular initial points

This part of experiment is to demonstrate the two cases that $\mathbf{W}_{\text{init}}$ is exactly singular, and $\mathbf{W}_{\text{init}}$ is only near singularity but not exactly singular. In this experiment, the uniformly distributed signal data set in Section 6.3.1 is used.

In the first case, $\mathbf{W}$ is initialized at

$$\mathbf{W}_{\text{init}} = \begin{bmatrix} 1 & 2 \\ 0.5 & 1 \end{bmatrix} \tag{6.57}$$

After stabilized, a snapshot of $\mathbf{W}$ is

$$\mathbf{W} = \begin{bmatrix} 0.3263 & 0.6526 \\ 0.3263 & 0.6526 \end{bmatrix} \tag{6.58}$$

which corresponds to:

$$\mathbf{V} = \begin{bmatrix} 0.7831 & 0.8484 \\ 0.7831 & 0.8484 \end{bmatrix} \tag{6.59}$$

which is neither in Solution Group A nor B, and satisfies $\det \mathbf{W} = 0$. The elements of $\mathbf{W}$ are plotted in Figure 6.6.

Then in the second case, $\mathbf{W}$ is initialized at a point near singularity but not exactly on the singular subspace:

$$\mathbf{W}_{\text{init}} = \begin{bmatrix} 1 & 2.0001 \\ 0.5 & 1 \end{bmatrix} \tag{6.60}$$

After convergence, a snapshot of the $\mathbf{V}$ is

$$\mathbf{V} = \begin{bmatrix} -0.0015 & 1.4846 \\ 1.4911 & -0.0090 \end{bmatrix} \tag{6.61}$$

which is one of the correct Solution A's.

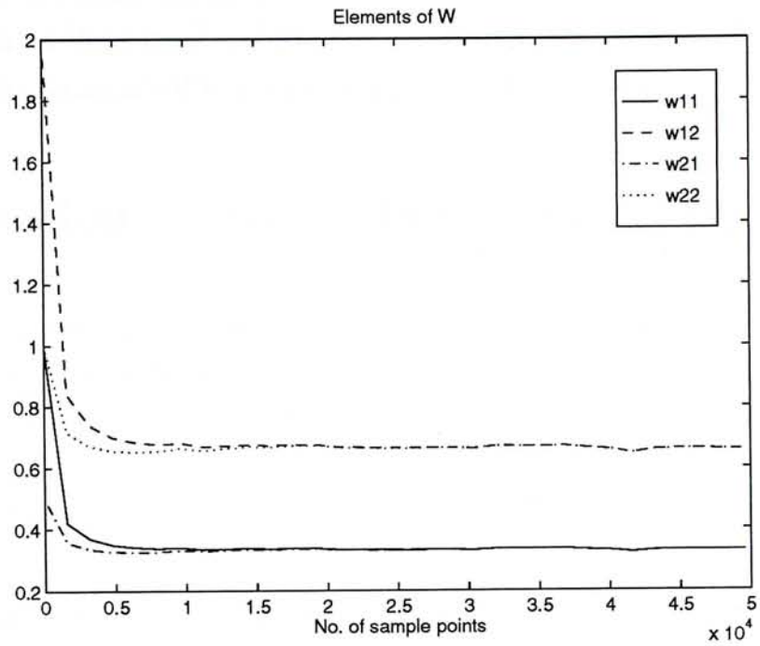


Figure 6.6: The graph of elements of $\mathbf{W}$ for the case $\mathbf{W}$ is initialized exactly on $\det \mathbf{W} = 0$.

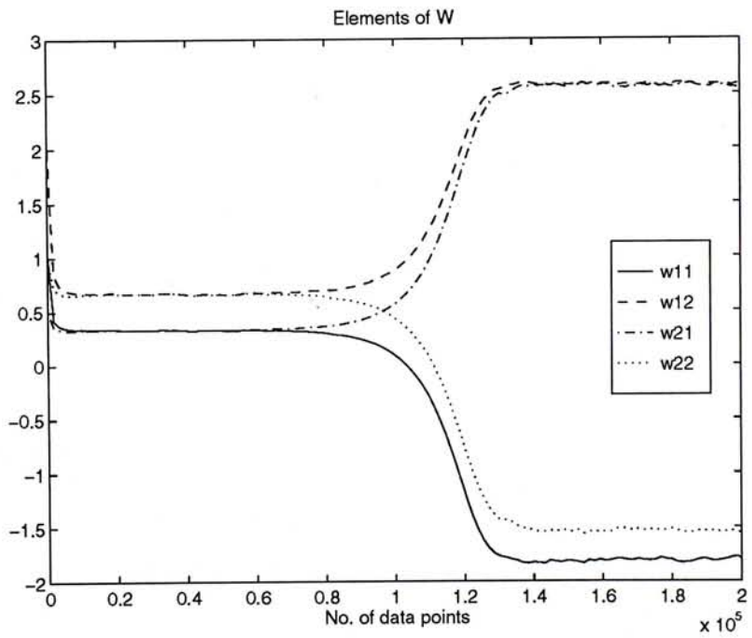


Figure 6.7: The graph of elements of $\mathbf{W}$ for the case $\mathbf{W}$ is initialized near $\det \mathbf{W} = 0$.

The elements of **W** verses the number of trained data points are plotted in Figure 6.7. We can see that this graph is similar to Figure 6.6 at the beginning but w_{ij} finally goes back to the correct solution. Hence it is demonstrated that **W** will finally leave the singular subspace if it is only near it but not exactly on it.

6.4 Theoretical Results on the 3-Channel Case

The 3-channel case is more complicated than the 2-channel case. The method we used in the previous section is difficult to scale up since the equilibrium matrix equation of high dimension is difficult to be solved exhaustively and the n^2-dimensional Hessian matrix is too complex to manipulate. However, Ruan and I have solved some equilibrium points and studied the stability of the group of correct solution. Although the analysis is not complete and global convergence behavior cannot be concluded, the partial results do provide some insight in high dimensional ($n > 2$) ICA problem.

We follow the method of analysis used in the 2-channel case, that is, try to solve the equilibrium points from the equilibrium equation in the **V**-parameter space. Then, we study the stability of the equilibrium points if possible.

6.4.1 Equilibrium points

In the 3-channel case, the equilibrium equation eq. (6.6) becomes 9 equations in 9 variables:

Self coupling equations

$$\begin{aligned}[(\nabla_{\mathbf{V}} J(\mathbf{V}))\mathbf{V}^T]_{11} &= 1 - c_1 E[y_1^4] \\ &= 1 - c_1\{v_{11}^4\mu_1^4 + v_{12}^4\mu_2^4 + v_{13}^4\mu_3^4 \\ &\quad + 2(v_{11}^2 v_{12}^2 m_{12} + v_{11}^2 v_{13}^2 m_{13} + v_{12}^2 v_{13}^2 m_{23})\} = 0\end{aligned} \tag{6.62}$$

$$\begin{aligned}[(\nabla_{\mathbf{V}} J(\mathbf{V}))\mathbf{V}^T]_{22} &= 1 - c_2 E[y_2^4] \\ &= 1 - c_2\{v_{21}^4\mu_1^4 + v_{22}^4\mu_2^4 + v_{23}^4\mu_3^4 \\ &\quad + 2(v_{21}^2 v_{22}^2 m_{12} + v_{21}^2 v_{23} r m_{13} + v_{22}^2 v_{23}^2 m_{23})\} = 0\end{aligned} \tag{6.63}$$

$$\begin{aligned}[(\nabla_{\mathbf{V}} J(\mathbf{V}))\mathbf{V}^T]_{33} &= 1 - c_3 E[y_3^4] \\ &= 1 - c_3\{v_{31}^4\mu_1^4 + v_{32}^4\mu_2^4 + v_{33}^4\mu_3^4 \\ &\quad + 2(v_{31}^2 v_{32}^2 m_{12} + v_{31}^2 v_{33} r m_{13} + v_{32}^2 v_{33}^2 m_{23})\} = 0\end{aligned} \tag{6.64}$$

Cross coupling equations:

$$\begin{aligned}[(\nabla_{\mathbf{V}} J(\mathbf{V}))\mathbf{V}^T]_{12} &= -c_1 E[y_1^3 y_2] \\ &= -c_1\{v_{11}v_{21}(v_{11}^2\mu_1^4 + v_{12}^2 m_{12} + v_{13}^2 m_{13}) \\ &\quad + v_{12}v_{22}(v_{11}^2 m_{21} + v_{12}^2\mu_2^4 + v_{13}^2 m_{23}) \\ &\quad + v_{13}v_{23}(v_{11}^2 m_{31} + v_{12}^2 m_{32} + v_{13}^2\mu_3^4)\} = 0\end{aligned} \tag{6.65}$$

$$\begin{aligned}[(\nabla_{\mathbf{V}} J(\mathbf{V}))\mathbf{V}^T]_{21} &= -c_2 E[y_2^3 y_1] \\ &= -c_2\{v_{21}v_{11}(v_{21}^2\mu_1^4 + v_{22}^2 m_{12} + v_{23}^2 m_{13}) \\ &\quad + v_{22}v_{12}(v_{21}^2 m_{21} + v_{22}^2\mu_2^4 + v_{23}^2 m_{23}) \\ &\quad + v_{23}v_{13}(v_{21}^2 m_{31} + v_{22}^2 m_{32} + v_{23}^2\mu_3^4)\} = 0\end{aligned} \tag{6.66}$$

$$\begin{aligned}[(\nabla_{\mathbf{V}} J(\mathbf{V}))\mathbf{V}^T]_{13} &= -c_1 E[y_1^3 y_3] \\ &= -c_1\{v_{11}v_{31}(v_{11}^2\mu_1^4 + v_{12}^2 m_{12} + v_{13}^2 m_{13}) \\ &\quad + v_{12}v_{32}(v_{11}^2 m_{21} + v_{12}^2\mu_2^4 + v_{13}^2 m_{23}) \\ &\quad + v_{13}v_{33}(v_{11}^2 m_{31} + v_{12}^2 m_{32} + v_{13}^2\mu_3^4)\} = 0\end{aligned} \tag{6.67}$$

$$\begin{aligned}[(\nabla_{\mathbf{V}} J(\mathbf{V}))\mathbf{V}^T]_{31} &= -c_3 E[y_3^3 y_1] \\ &= -c_3\{v_{31}v_{11}(v_{31}^2\mu_1^4 + v_{32}^2 m_{12} + v_{33}^2 m_{13}) \\ &\quad + v_{32}v_{12}(v_{31}^2 m_{21} + v_{32}^2\mu_2^4 + v_{33}^2 m_{23}) \\ &\quad + v_{33}v_{13}(v_{31}^2 m_{31} + v_{32}^2 m_{32} + v_{33}^2\mu_3^4)\} = 0\end{aligned} \tag{6.68}$$

$$\begin{aligned}[(\nabla_{\mathbf{V}} J(\mathbf{V}))\mathbf{V}^T]_{23} &= -c_2 E[y_2^3 y_3] \\ &= -c_2\{v_{21}v_{31}(v_{21}^2\mu_1^4 + v_{22}^2 m_{12} + v_{23}^2 m_{13}) \\ &\quad + v_{22}v_{32}(v_{21}^2 m_{21} + v_{22}^2\mu_2^4 + v_{23}^2 m_{23}) \\ &\quad + v_{23}v_{33}(v_{21}^2 m_{31} + v_{22}^2 m_{32} + v_{23}^2\mu_3^4)\} = 0\end{aligned} \tag{6.69}$$

$$\begin{aligned}[(\nabla_{\mathbf{V}} J(\mathbf{V}))\mathbf{V}^T]_{32} &= -c_3 E[y_3^3 y_2] \\ &= -c_3\{v_{31}v_{21}(v_{31}^2\mu_1^4 + v_{32}^2 m_{12} + v_{33}^2 m_{13}) \\ &\quad + v_{32}v_{22}(v_{31}^2 m_{21} + v_{32}^2\mu_2^4 + v_{33}^2 m_{23}) \\ &\quad + v_{33}v_{23}(v_{31}^2 m_{31} + v_{32}^2 m_{32} + v_{33}^2\mu_3^4)\} = 0\end{aligned} \tag{6.70}$$

where $m_{ij} = 3\mu_i^2\mu_j^2$.

Unfortunately, we cannot factorize the equilibrium equations into useful form such that the solutions could be solved exhaustively. (It seems that the method we use to solve the equilibrium points exhaustively in the 2-channel case is based on the simplicity of 2 channels and cannot be generalized to the n-channel case.) However, we can try to guess some forms of the solutions, substitute them into the equations and determine the solutions if possible. For example, to determine the correct solutions the recover all sources by $y_i = v_{ii}s_i$, we substitute all the off-diagonal elements to be zero and get:

Solution P1-P8

$$\mathbf{V} = \begin{bmatrix} \pm(c_1\mu_1^4)^{-1/4} & 0 & \\ 0 & \pm(c_2\mu_2^4)^{-1/4} & 0 \\ 0 & 0 & \pm(c_3\mu_3^4)^{-1/4} \end{bmatrix} \tag{6.71}$$

By the symmetry between channels ($\{c_i\}$ only affect the scale of the recovered signals and do not affect the symmetry), solution of the form Solution P1-P8 with different permutations of order of channels are still solutions to the equilibrium equation. The solutions of this type have one non-zero $(c_i\mu_j^4)^{-1/4}$ in row i, column j and other elements being zero. So we have totally $8 \times 3! = 48$ solution P's.

Then we investigate the solution that only one source is extracted but the other two are still mixed. For example, let y_1 recovers s_1, we put $v_{12} = v_{13} = v_{21} = v_{22} = 0$ into the equilibrium equations eq. (6.62) to (6.68). Using the same technique as those in the 2-channel case, we get,

Solution Q1-Q16

$$\mathbf{V} = \begin{bmatrix} \pm(c_1\mu_1^4)^{-1/4} & 0 & 0 \\ 0 & s_{22}(2c_2\eta_{23})^{-1/4} & s_{23}(2c_2\eta_{32})^{-1/4} \\ 0 & s_{32}(2c_3\eta_{23})^{-1/4} & s_{33}(2c_3\eta_{32})^{-1/4} \end{bmatrix} \tag{6.72}$$

where

$s_{ij} = 1$ or -1 satisfying $s_{22}s_{23}s_{32}s_{33} = -1$,

with totally 8 combinations,

$$\eta_{ij} = \mu_i^4 + \sqrt{\mu_i^4/\mu_j^4}m_{ij} \tag{6.73}$$

By symmetry again, solutions similar to Solution Q1-Q16 with different permutations of channels are also solutions to the equilibrium equations. Solution of this type have

a $\pm(c_i\mu_j^4)^{-1/4}$ element in row i,column j, other elements in row i and column j being zero, that let y_i recovers s_j. The remaining four elements have the form $s_{kl}(2c_k\eta_{l\bar{l}})^{-1/4}$ where k is the row index, l is the column index and $\bar{l} \neq i \neq l$ is the index of the remaining column. There are totally $16 \times 9 = 144$ solution Q's.

It is a regret that we cannot determine the solutions with all nine elements non-zero, which experiments have verified their existence. Hence the above solutions listed are by no means exhaustive.

6.4.2 Stability

In Appendix B, it is proved that the solution group P are stable if the three sources are pairwise globally sub-Gaussian:

$$E[s_i^4]E[s_j^4] - 9(E[s_i^2])^2(E[s_j^2])^2 < 0 \quad i,j = 1,2,3., \; i \neq j \tag{6.74}$$

and are saddle points if the above condition is not satisfied. The stability of solution group Q is difficult to determine since the Hessian matrix is too complicated.

As the equilibrium points have not been exhaustively found and the stability of solution group Q has not been determined, we can only conclude that the system *may* converge to a solution P if the three sources are pairwise globally sub-Gaussian (that depends on the initial point) and will not converge to solution P if any two sources are not globally sub-Gaussian.

6.5 Experiments on the 3-Channel Case

The experiments test the natural gradient algorithm with cubic nonlinearity on different combinations of 3 sources of different kurtosis. The learning rate is kept at 0.0001, $\mathbf{W}$ is initialized as an identity matrix and the following mixing matrix is used:

$$\mathbf{A} = \begin{bmatrix} 1 & 0.6 & 0.3 \\ 0.8 & 1 & 0.3 \\ 0.4 & 0.9 & 1 \end{bmatrix} \tag{6.75}$$

6.5.1 Experiments on three pairwise globally sub-Gaussian sources

In the first trial, the two channels of uniformly distributed sources used in Section 6.3.1 and a third channel of uniform $[-0.5, 0.5]$ distributed source are used in this experiment. The statistics of the third source are $\mu_3^2 = 0.0832$, $\mu_3^4 = 0.0125$. After scanning for 120,000 data points, $\mathbf{V}$ has successfully converged to one of the solution P's:

$$\mathbf{V} = \begin{bmatrix} 1.497 & 0 & 0 \\ 0 & 1.494 & 0 \\ 0 & 0 & 2.992 \end{bmatrix} \tag{6.76}$$

The interference-to-signal power ratio reaches -35dB.

In the second trial, the same two channels of uniformly distributed sources are used and the third channel is replaced by the permuted song signal used in Section 6.3.3. The sources are pairwise globally sub-Gaussian though the song signal is super-Gaussian. After scanning for 240,000 data points, $\mathbf{V}$ converged to one of the solution P's:

$$\mathbf{V} = \begin{bmatrix} 1.497 & 0 & 0 \\ 0 & 1.494 & 0 \\ 0 & 0 & 1.474 \end{bmatrix} \tag{6.77}$$

The interference-to-signal power ratio also reaches -35dB.

6.5.2 Experiments on three sources consisting of globally sub-Gaussian and globally super-Gaussian pairs

In the first trial, the first and third channels are song signals recorded at 8kHz and the time index of them are randomly permuted to remove non-stationarity. The second channel is iid uniformly distributed source. The statistics of the data are:

$$\begin{aligned} \mu_1^2 &= 0.2532 \quad \mu_2^2 = 0.3326 \quad \mu_3^2 = 0.2702 \\ \mu_1^4 &= 0.2120 \quad \mu_2^4 = 0.1991 \quad \mu_3^4 = 0.2541 \end{aligned} \tag{6.78}$$

The standardized joint kurtosis are: $\mu_1^4\mu_2^4/[(\mu_1^2)^2(\mu_2^2)^2]-9 = -3.0471$, $\mu_1^4\mu_3^4/[(\mu_1^2)^2(\mu_3^2)^2]-9 = 2.5113$, $\mu_2^4\mu_3^4/[(\mu_2^2)^2(\mu_3^2)^2] - 9 = -2.7370$. After scanning for 800,000 data points,

V converged to one of the solution Q:

$$\mathbf{V} = \begin{bmatrix} 1.0577 & 0 & 1.0109 \\ 0 & 1.4970 & 0 \\ -1.0577 & 0 & 1.0109 \end{bmatrix} \tag{6.79}$$

for example, a snapshot of it is:

$$\mathbf{V} = \begin{bmatrix} 1.0448 & 0.0131 & 1.0192 \\ 0.0105 & 1.4909 & -0.0084 \\ -1.0621 & 0.0180 & 1.0052 \end{bmatrix} \tag{6.80}$$

Only the uniformly distributed source is extracted out.

In the second trial, the first channel is artificially generated iid data drawn from the symmetrical, bimodal beta distribution beta(0.5, 0.5) shifted into [-0.5,0.5], (the pdf of which can be visualized from the histogram of y_1 in Figure 8.7 of Chapter 8). The second channel is one uniformly distributed signal used in Section 6.3.1 and the third channel of permuted speech signals from Section 6.3.2. The statistics of the sources are:

$$\begin{aligned} \mu_1^2 &= 0.1252 \quad \mu_2^2 = 0.3326 \quad \mu_3^2 = 0.0625 \\ \mu_1^4 &= 0.0234 \quad \mu_2^4 = 0.1991 \quad \mu_3^4 = 0.0435 \end{aligned} \tag{6.81}$$

The standardized joint kurtosis are: $\mu_1^4\mu_2^4/[(\mu_1^2)^2(\mu_2^2)^2]-9 = -6.307$, $\mu_1^4\mu_3^4/[(\mu_1^2)^2(\mu_3^2)^2]-9 = 7.6522$, $\mu_2^4\mu_3^4/[(\mu_2^2)^2(\mu_3^2)^2]-9 = 11.029$. Channel 1 and 2 are globally sub-Gaussian, channel 1 and 3 are globally super-Gaussian and channel 2 and 3 are globally super-Gaussian.

The simulation scan for 500,000 data points and the system become stable. A snapshot of **V** is:

$$\begin{bmatrix} 2.5528 & 0.0014 & 0.1142 \\ 0.1065 & 1.1112 & -1.6470 \\ -0.0936 & 1.1073 & 1.6015 \end{bmatrix} \tag{6.82}$$

which is close to a solution Q

$$\begin{bmatrix} 2.557 & 0 & 0 \\ 0 & 1.107 & -1.620 \\ 0 & 1.107 & 1.620 \end{bmatrix} \tag{6.83}$$

Only the beta distributed source can be extracted, the uniformly distributed source and the permuted speech signal are still mixed.

In the third trial, the two channels of uniformly distributed sources used in Section 6.3.1 and the first channel of permuted speech signal used in Section 6.3.2 are used in this experiment. The permuted speech signal with either one of the uniformly distributed sources are globally super-Gaussian. After scanning for 160,000 data points, $\mathbf{V}$ is stabilized and a snapshot of it is:

$$\mathbf{V} = \begin{bmatrix} 1.4642 & -0.0023 & 0.4935 \\ 0.1599 & 1.1086 & -1.5791 \\ -0.2081 & 1.0823 & 1.6200 \end{bmatrix} \tag{6.84}$$

No source can be extracted.

6.5.3 Experiments on three pairwise globally super-Gaussian sources

In the first trial, the two permuted speech signal used in Section 6.3.2 are used as channel 1 & 2 and the permuted song signal is used as channel 3. All of the three sources are super-Gaussian. After scanning for 500,000 data points, the system stabilizes and a snapshot of $\mathbf{V}$ is:

$$\mathbf{V} = \begin{bmatrix} 1.6641 & 1.0819 & 0.0883 \\ -1.2568 & 0.7323 & -1.1115 \\ -1.4289 & 0.6536 & 1.0801 \end{bmatrix} \tag{6.85}$$

No source can be extracted.

In the second trial, the two permuted speech signal used in Section 6.3.2 are used as channel 1 & 2 and the third channel is replaced by the uniformly distributed source used in Section 6.3.1. As previous calculation shown, the three channels are pairwise globally super-Gaussian. The system converges very slowly, using 2,400,000 data points, and a snapshot of $\mathbf{V}$ after it stabilized is:

$$\mathbf{V} = \begin{bmatrix} 1.6141 & 1.0266 & 0.4158 \\ -1.5908 & 1.0132 & -0.4218 \\ -1.0109 & -0.0192 & 1.3813 \end{bmatrix} \tag{6.86}$$

No source can be extracted.

In the above experiments, it is demonstrated that in the two cases that the three sources are pairwise globally sub-Gaussian, the system happens to converge to the correct solution P's; in the cases that one pair is globally sub-Gaussian and two pairs are globally super-Gaussian, the system may converge to one of the solution Q's or solution that all nine elements are non-zero; in the cases that three sources are pairwise globally super-Gaussian, the system converges to spurious solutions that all nine elements are non-zero.

Chapter 7

Nonlinearity and Separation Capability

The relationship between the nonlinearity of the algorithm and the distribution of the sources that it is capable to separate is discussed in this chapter. In Section 7.1, we compare several nonlinearities and their separation capability and argue that a 'loose matching' between the nonlinearity and source distribution is needed for successful separation. In Section refsepcapexpt, we present experiment finding and verification for the arguments in Section 7.1.

7.1 Theoretical Argument

The function $h_i(y_i)$ is the nonlinearity in algorithms eq. (4.18) that determines the properties and capability of the algorithm, as it determines the constraints on the higher order statistics of the recovered signals in the equilibrium equations of the algorithms. The nonlinearities $\{h_i(\cdot)\}$ follows from the choice of $\{g_i(\cdot)\}$ in eq. (4.21), which are the pdf's we manually assign to the unknown $\{p_{s_j}(s_j)\}$

In the information-theoretic ICA scheme, minimizing $J(\mathbf{W})$ w.r.t. $\mathbf{W}$ means tuning $\mathbf{W}$ in such a way that makes $p_{\mathbf{y}}(\mathbf{y})$ and $\prod_{i=1}^{n} g_i(y_i)$ as close to each other as possible. The network is said to be workable on a particular set of source signals if the minimization of $J(\mathbf{W})$ yields a correct solution $\mathbf{W}$, i.e. a minima $\mathbf{W}$ of $J(\mathbf{W})$ satisfy $\mathbf{WA} = \mathbf{PD}$ given by eq. (2.11). Whether such minimization can yield a correct solution $\mathbf{W}$ depends on the choice of the set of function $\{g_i(y_i)\}$.

7.1.1 Nonlinearities that strictly match the source distribution

If the source marginal densities *were* known, $g_i(s_i) = p_{s_i}(s_i)$, $i = 1, ..., n$, *would be* a perfect choice of $\{g_i(\cdot)\}$ since $g_i(y_i)$ is a manual pdf that we assign to $p_{s_i}(s_i)$. Considering the fundamental indeterminacy of scaling factor and channel order, we can inspire that $\{g_i(.)\}$ being in the set of scale family of $p_{s_i}(s_i)$ can also ensure separation. However, the above choices are of course impossible since $\{p_{s_i}(s_i), i = 1, ..., n\}$ is blind to us.

On the other hand, from eq. (4.11) $g_i(s_i) = p_{y_i}(y_i)$, $i = 1, ..., n$, is also a perfect choice since the cost function would reduce to the mutual information used in [2, 24]. The global minimization of mutual information can always yield a correct solution $\mathbf{W}$ because the mutual information attains its global minimum when

$$p_{\mathbf{y}}(\mathbf{y}) = \prod_{i=1}^{n} p_{y_i}(y_i) \tag{7.1}$$

which means that the components of $\mathbf{y}$ are independent. Hence, theoretically $g_i(y_i) = p_{y_i}(y_i)$ is a perfect choice but this choice needs some implementation technique as $p_{y_i}(y_i)$ is not known in advance and can only be estimated on-line. We call the assignment $g_i(y_i) = p_{y_i}(y_i)$, $i = 1, ..., n$, 'strict matching' between nonlinearities and source distributions. It should be noted that $\{p_{y_i}(y_i)\}$ is a member of the scale families of $\{p_{s_i}(s_i)\}$ if separation is successful ($\mathbf{y} = \mathbf{Vs} = \mathbf{PDs}$)

7.1.2 Nonlinearities that loosely match the source distribution

Though nonlinearities that strictly match the source distribution can always yield source separation, it is found that a much wider class of $\{g_i(y_i), i = 1, ..., n\}$ can perform separation on sources with a particular $\{p_{s_i}(s_i), i = 1, ..., n\}$ (Figure 4.2). One objective of this thesis is to argue that the choice of $\{g_i(y_i), i = 1, ..., n\}$ is much wider than $\{p_{y_i}(y_i), i = 1, ..., n\}$, or the scale family of $\{p_{s_j}(s_j)\}$. We assert that the use of a set of *prespecified* and *fixed* $g_i(y_i)$ can perform source separation on sources with *a particular class* of probability density function 'loosely matched' with $g_i(y_i)$, but not on sources with *any* distribution.

In the followings, we justify the argument of 'loose matching' in case-by-case studies of fixed nonlinearity, using theoretical analysis whenever possible and experiments otherwise:

(i) In [5], $f_i(y_i)$ are chosen to be logistic sigmoid $f_i(y_i) = \text{logsig}(y_i) = 1/(1 +$

$exp(-y_i)$) or hyperbolic tangent $f_i(y_i) = \tanh(y_i)$, etc. The corresponding $g_i(y_i)$ are sharply peaked and the corresponding $h_i(y_i)$ have reversed sigmoidal shape. It is suggested that they can perform source separation on sources with sharply peaked pdf, and experimentally verified on human speech signals [5]. In the experiment section, it is demonstrated that this nonlinearity cannot perform signal separation on some sub-Gaussian sources.

(ii) In the experiments in the next section, the *cubic root nonlinearity* $h_i(y_i) = -\sqrt[3]{y_i}$, $i = 1, ..., n$, can perform separation on mixtures of sources with sharply peaked densities but cannot separate uniformly distributed sources, which are sub-Gaussian. The behavior of this nonlinearity is similar to that of $\{h_i(y_i), i = 1, ..., n\}$ being reversed sigmoids.

(iii) In the previous chapter, the cubic nonlinearity $h_i(y_i) = -c_i y_i^3$, $i = 1, ..., n$ where $c_i > 0$. is both theoretically and experimentally investigated. In the 2-channel case we have proved that it can perform separation on the mixtures of two globally sub-Gaussian sources and cannot perform separation on the mixture of two globally super-Gaussian sources. We also assert that it can separate sub-Gaussian sources in higher-channel case.

(iv) Case (i) and (ii) are suggested to work on sources with sharply peaked pdf (possibly super-Gaussian) and Case (iii) works on two channels of sub-Gaussian sources. Moreover, we have the following theorem for another case that performs separation on one super-Gaussian source and one sub-Gaussian source.

THEOREM 2 Consider an information-theoretic ICA algorithm with $h_1(y_1) = -c_{11}y_1$ and $h_2(y_2) = -c_{23}y_2^3$ with $c_{11} > 0$ and $c_{23} > 0$ acting on two channels of signals. If one source is sub-Gaussian and the other source is super-Gaussian, and $\mathbf{W}$ is initialized at any non-singular point, $\mathbf{V}$ will converge to one of the following eight correct solutions for source separation:

$$\textbf{Solution } \mathbf{A}_{\mathrm{I}}\textbf{:} \quad \mathbf{V} = \begin{bmatrix} \pm(c_{11}E[s_1^2])^{-\frac{1}{2}} & 0 \\ 0 & \pm(c_{23}E[s_2^4])^{-\frac{1}{4}} \end{bmatrix} \tag{7.2}$$

$$\textbf{Solution } \mathbf{A}_{\mathrm{II}}\textbf{:} \quad \mathbf{V} = \begin{bmatrix} 0 & \pm(c_{11}E[s_2^2])^{-\frac{1}{2}} \\ \pm(c_{23}E[s_1^4])^{-\frac{1}{4}} & 0 \end{bmatrix} \tag{7.3}$$

such that the resulting y_2 recovers the sub-Gaussian source.

The proof of the Theorem 2 is provided in Appendix C.

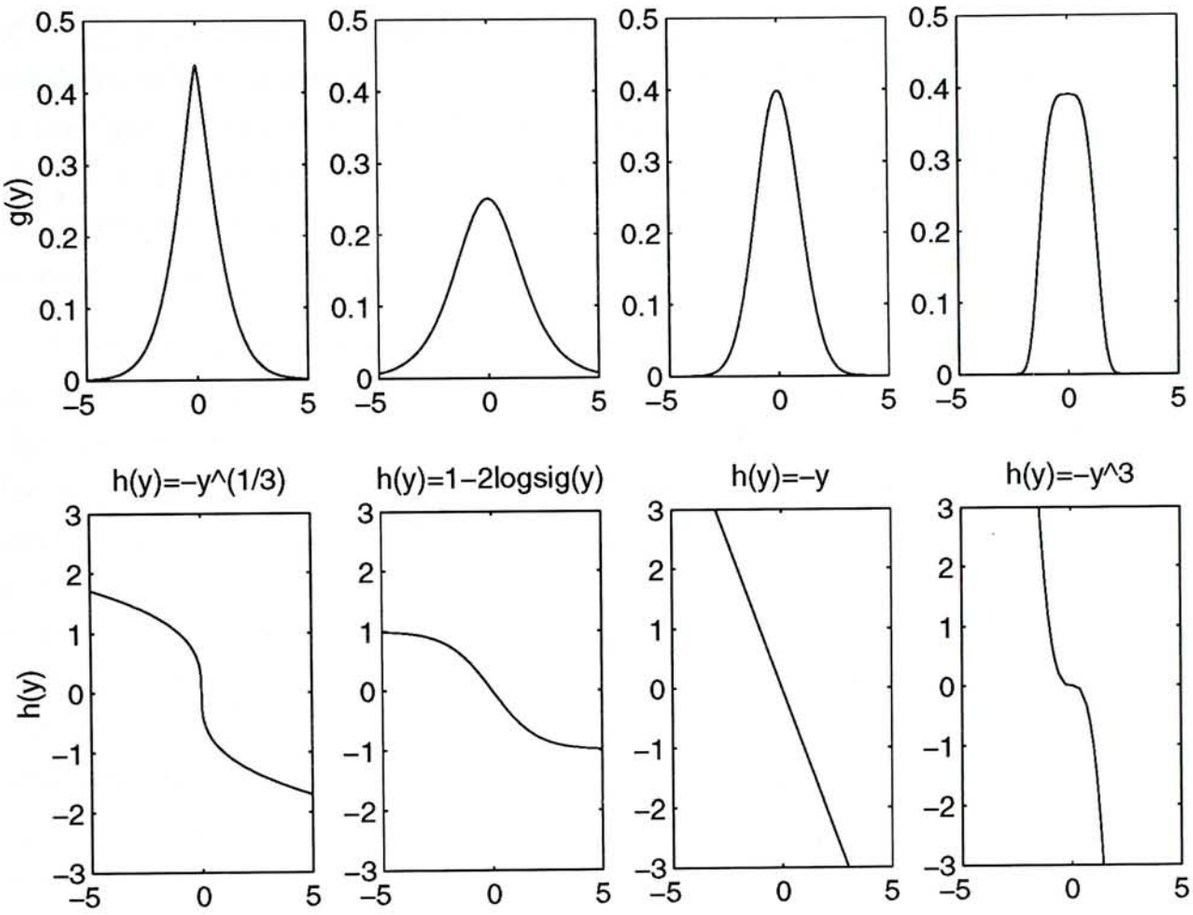


Figure 7.1: Several nonlinearities investigated.

$g_i(y_i)$	$\frac{1}{2(3/4)^{1/4}\gamma(3/4)} \times \exp(-\frac{3}{4}y_i^{4/3})$	$\frac{\exp(-y_i)}{(1+\exp(-y_i))^2}$	$\frac{1}{\sqrt{2\pi}}\exp(-y_i^2/2)$	$\frac{\sqrt{2}}{\gamma(1/4)}\exp(-y_i^4/4)$
$h_i(y_i)$	$-\sqrt[3]{y_i}$	$1 - 2\,\mathrm{logsig}(y_i)$	$-y_i$	$-y_i^3$
$\mu^4/(\mu^2)^2 - 3$	1.2216	1.2	0	-0.8118
Sub-Gaussian samples (uniformly distributed)	no	no	see Theorem 2	yes
Super-Gaussian samples (speech signals)	yes	yes	see Theorem 2	no

Table 7.1: Properties and separation capabilities of several nonlinearities

Figure 7.1 plots the $g_i(y_i)$ and $h_i(y_i)$ used in case (i) to (iv) above. All $h_i(y_i)$ in these four cases are monotonic decreasing, odd function which span only in the second and forth quadrants of the graph. The standardized kurtosis $\mu_i^4/(\mu_i^2)^2 - 3$, where $\mu_i^p = \int_{-\infty}^{\infty} y_i^p g_i(y_i) \mathrm{d}y_i$, that describe the sharp peakedness of $g_i(y_i)$, and the separation capability of several samples by the nonlinearities, are listed in Table 7.1. The columns in the figure and table are sorted in descending order of standardized kurtosis. From the figure, it can be inspected that the $g_i(y_i)$ in case (i) and (ii) more sharply peaked and have longer tails (have greater / more positive kurtosis) and the $g_i(y_i)$ in case (iii) is more flat and have shorter tail (have smaller / more negative kurtosis).

From the theorems developed and experiment results, it can be seen that a fixed nonlinearity can perform separation on the mixtures of sources with a class of densities that are similar to the shapes of $\{g(y)\}$. For example, the cubic nonlinearity has a flat $g_i(y_i)$ can separate sub-Gaussian sources and the nonlinearity in case (i) and (ii) have sharply peaked $g_i(y_i)$ can separate a class of sources with sharply peaked pdf. However, the fixed nonlinearity cannot separate sources whose density differ from $g_i(y_i)$ too much. For example, the cubic nonlinearity cannot separate super-Gaussian sources and the nonlinearity in case (i) and (ii) cannot separate uniformly distributed signals. Hence, we can conclude that a *'loose matching'* is required between $\{g_i(y_i)\}$ and the source densities. In case (iv), the cubic nonlinearity in channel 2 always selects the s_i with sub-Gaussian (flatter) density to recover. This is a vivid example that the nonlinearity extract source with density matched with itself.

In the Minimum Mutual Information approach [2], the adaptive algorithm developed has the form eq. (4.28) with nonlinearity

$$h_i(y_i) = -\frac{3}{4}y_i^{11} - \frac{25}{4}y_i^9 + \frac{14}{3}y_i^7 + \frac{47}{4}y_i^5 - \frac{29}{4}y_i^3 \tag{7.4}$$

and correspondingly,

$$g_i(y_i) = \exp(-\frac{1}{16}y_i^{12} - \frac{5}{8}y_i^{10} + \frac{7}{12}y_i^8 + \frac{47}{24}y_i^6 - \frac{29}{16}y_i^4) \tag{7.5}$$

The nonlinearities are plotted in Figure 7.2. The $h_i(y_i)$ function is odd but is not monotonic decreasing, and cross the first and third quadrant. The $g_i(y_i)$ function is not unimodal. In the experiments in the next section, the MMI algorithm can separate the uniformly distributed sources but cannot separate the permuted speech sources used in Section 6.3.2. This result is consistent to the assertion that a fixed nonlinearity can perform separation on sources with a class of distribution but not any distribution.

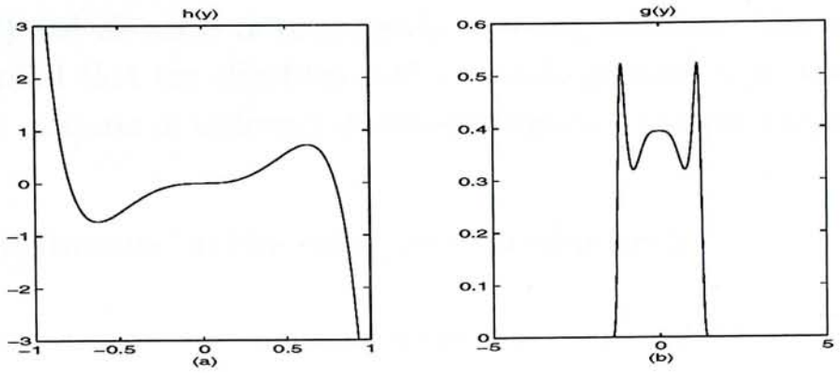


Figure 7.2: The nonlinearities used in the MMI algorithm.
(a) $h_i(y_i) = -\frac{3}{4}y_i^{11} - \frac{25}{4}y_i^9 + \frac{14}{3}y_i^7 + \frac{47}{4}y_i^5 - \frac{29}{4}y_i^3$,
(b) $g_i(y_i) = \exp(-\frac{1}{16}y_i^{12} - \frac{5}{8}y_i^{10} + \frac{7}{12}y_i^8 + \frac{47}{24}y_i^6 - \frac{29}{16}y_i^4)$

7.2 Experiment Verification

The experiments are aimed at providing empirical results or verifying theorems in this chapter. For all experiments, the learning rate is kept at 0.0001. The following mixing matrix is used:

$$\mathbf{A} = \begin{bmatrix} 1 & 0.6 \\ 0.7 & 1 \end{bmatrix} \tag{7.6}$$

The experiment is run for a number of scans through the data set long enough that $\mathbf{W}$ looks having converged to a stable point.

7.2.1 Experiments on reversed sigmoid

Sub-Gaussian sources

The reversed sigmoid nonlinearity $h_i(y_i) = 1 - 2\,\text{logsig}(y_i) = (\exp(-y) - 1)/(1 + \exp(-y))$, $i = 1, ..., n$, are used in both channels to act on the uniformly distributed sources used in Section 6.3.1. After scanning for 500,000 data points, $\mathbf{V}$ stabilized and a snapshot is:

$$\mathbf{V} = \begin{bmatrix} 2.0961 & -2.0869 \\ 2.0844 & 2.0762 \end{bmatrix} \tag{7.7}$$

Similar experiments have been done using $h_i(y_i) = -2\tanh(y_i)$ and $h_i(y_i) = -2y_i/(1+y_i^2)$ and all result in convergence to wrong solutions. Hence, it is experimentally verified that the algorithm with reversed-sigmoidal $h_i(y_i)$ cannot perform separation on mixtures of uniformly distributed signals, which are sub-Gaussian.

7.2.2 Experiments on the cubic root nonlinearity

This experiment tests the performance of the cubic root nonlinearity $h_i(y_i) = -\sqrt[3]{y_i}$, $i = 1, ..., n$, on super-Gaussian and sub-Gaussian sources.

Super-Gaussian sources

The data set is the permuted speech signals used in Section 6.3.2. The system converged in 40,000 data points and a snapshot of $\mathbf{V}$ is:

$$\mathbf{V} = \begin{bmatrix} 5.8386 & 0.0015 \\ 0.0230 & 2.6151 \end{bmatrix} \tag{7.8}$$

The average interference-to-signal power ratio reach -40 dB. Hence it is experimentally found that the cubic root nonlinearity can separate some super-Gaussian sources.

Sub-Gaussian sources

The data set is the uniformly distributed signals used in Section 6.3.1. After scanning for 240,000 data points, $\mathbf{V}$ converged and a snapshot of $\mathbf{V}$ is:

$$\mathbf{V} = \begin{bmatrix} 1.3843 & -1.3810 \\ 1.3841 & 1.3701 \end{bmatrix} \tag{7.9}$$

Hence it is experimentally found that the cubic root nonlinearity cannot separate some sub-Gaussian sources.

7.2.3 Experimental verification of Theorem 2

In this experiment, the nonlinearities used are $h_1(y_1) = -y_1$ and $h_2(y_2) = -y_2^3$ in Theorem 2. We use the permuted speech signal and uniformly distributed signal in

Section 6.3.4. The first channel is super-Gaussian and the second is sub-Gaussian. After scanning for 120,000 data points, $\mathbf{V}$ converged to one of the Solution A_I:

$$\mathbf{V}_{A_I} = \begin{bmatrix} 2 & 0 \\ 0 & 1.497 \end{bmatrix} \tag{7.10}$$

Then we swap the order of the channels of the sources. After scanning for 320,000 data points, $\mathbf{V}$ converged to one of the Solution A_{II}:

$$\mathbf{V}_{A_I} = \begin{bmatrix} 0 & 2 \\ -1.49 & 0 \end{bmatrix} \tag{7.11}$$

Hence, this experiment provides verification of Theorem 2.

7.2.4 Experiments on the MMI algorithm

This experiment tests the performance of the resulting nonlinearity derived in the MMI approach [2] eq. (7.4) on super-Gaussian sources and sub-Gaussian sources.

Super-Gaussian sources

We use the data set of the permuted speech signals used in Section 6.3.2. $\mathbf{W}$ fluctuates largely and shows no tendency of convergence in scanning 500,000 data points. The elements of $\mathbf{W}$ versus the number of data points scanned is plotted in Figure 7.3. Hence it is concluded that the algorithm cannot separate the permuted speech signals.

Sub-Gaussian sources

We use the uniformly distributed sources in Section 6.3.1. The system converge in 20,000 data points and a snapshot of $\mathbf{V}$ is:

$$\mathbf{V} = \begin{bmatrix} 1.3061 & -0.0052 \\ 0.0025 & 1.3038 \end{bmatrix} \tag{7.12}$$

The average interference-to-signal power ratio reach -40dB. Hence it is experimentally found that the MMI algorithm can separate some sub-Gaussian sources.

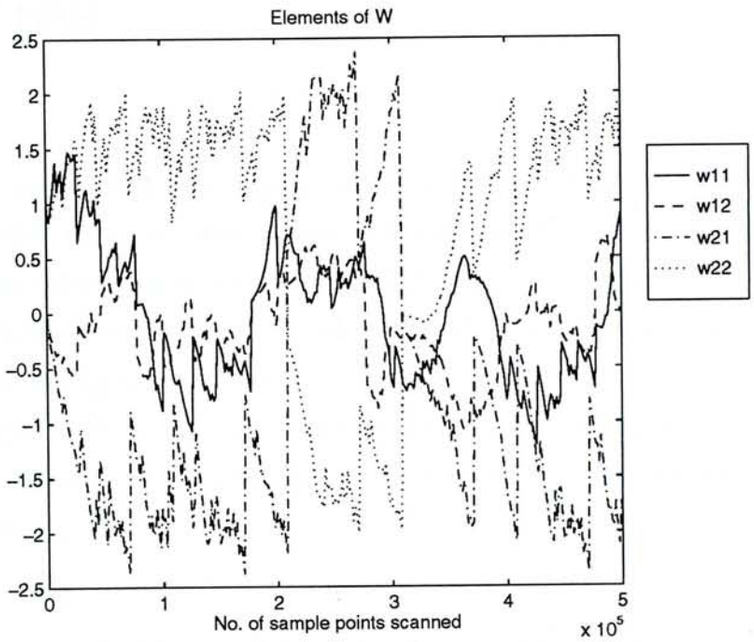


Figure 7.3: The fluctuate of **W** when the MMI algorithm is applied to permuted speech signals.

Chapter 8

Implementation with Mixture of Densities

In this chapter, we present the experimental verification of the information-theoretic ICA algorithm with mixture of densities and some analysis. In Section 8.1, the idea of implementing the information-theoretic ICA scheme with mixture of densities (Xu *et al* , 1997) will be reviewed. Then we will present the derivation of the algorithm. In Section 8.2, various experiment results will be provided that support the implementation with mixture of densities can adapt and separate sources of any distribution. In Section 8.3, we discuss the simplest form of mixture of densities that may still adapt any source density.

8.1 Implementation of the Information-theoretic ICA scheme with Mixture of Densities

The idea of implementing the information-theoretic ICA scheme with mixture of densities is proposed by Xu *et al* (1997). It is targeted at performing signal separation on sources with any distribution automatically. The method to achieve so is to allow $g_i(y_i)$ to be a 'flexible member' of a family of functions in the function space rather than a fixed function. Such designation of $g_i(y_i)$ as a flexible function is achieved by using a parameterized function for $\{g_i(y_i)\}$. The shape of $g_i(y_i)$ can be changed by the parameters. The parameters of the flexible function are adapted such that $g_i(y_i)$ approximate, or get close to, the marginal density $p_{y_i}(y_i)$ (though only loose matching is needed). As a family of functions takes up a greater subspace of the function space, the flexible $g_i(y_i)$ inside the family has a greater possible to achieve the loose matching

with $p_{y_i}(y_i)$. Thus it increases the possibility that minimizing J with this flexible $g_i(y_i)$ can achieve successful source separation.

8.1.1 The mixture of densities

The function $g_i(y_i)$ is chosen to be the mixture of densities because it can approximate any density function arbitrarily well if the number of components is sufficiently large. The mixture of densities is given by:

$$g_i(y_i) = \sum_{j=1}^{p_i} \alpha_{ij} \psi(u_{ij}), \qquad \sum_{j=1}^{p_i} \alpha_{ij} = 1 \tag{8.1}$$

where

$$u_{ij} = b_{ij}(y_i - a_{ij}) \tag{8.2}$$

and $\psi(\cdot)$ is some density function. p_i is the number of components in the mixture. α_{ij} is the weight of the component. b_{ij} control the variant of the j^{th} pdf component and a_{ij} is the bias, or location of the center, of the j^{th} pdf component.

Writing $\psi(u_{ij})$ in the form

$$\psi(u_{ij}) = b_{ij}\phi'(u_{ij}) \tag{8.3}$$

we have that the $g_i(y_i)$ with this $\psi(y_i)$ corresponds to the nonlinear transformation function in the form of mixture of CDF's:

$$f_i(y_i) = \sum_{j=1}^{p^i} \alpha_{ij} \phi(u_{ij}) \tag{8.4}$$

One of the choice for $\phi(u_{ij})$ is:

$$\phi(u_{ij}) = \text{logsig}(u_{ij}) = 1/(1 + \exp(-u_{ij})) \tag{8.5}$$

with

$$\phi'(u_{ij}) = \phi(u_{ij})(1 - \phi(u_{ij})) = \frac{\exp(-u_{ij})}{(1 + \exp(-u_{ij}))^2} \tag{8.6}$$

The constraint on α_{ij} in eq. (8.1) introduces computing complexity in the algorithm to be developed. Hence, we apply a transformation on α_{ij}:

$$\alpha_{ij} = \frac{\exp(\gamma_{ij})}{\sum_{k=1}^{p_i} \exp(\gamma_{ik})} \tag{8.7}$$

to get γ_{ij} which can take any real value. Now, $\{\gamma_{ij}, b_{ij}, a_{ij}, j = 1, ..., p_i\}$ is the set of parameters for the configuration of $g_i(y_i)$. The set of parameters of the system becomes $\{\mathbf{V}, \gamma, \mathbf{b}, \mathbf{a}\}$ and $J = J(\mathbf{V}, \gamma, \mathbf{b}, \mathbf{a})$, where $\gamma = \{\gamma_{ij}, i = 1, ..., n, j = 1, ..., p_i\}$ and the same notation applies to $\mathbf{b}$ and $\mathbf{a}$.

8.1.2 Derivation of the algorithms

Algorithm for the de-mixing matrix

The $h_i(y_i)$ nonlinearity is given by:

$$h_i(y_i) = \frac{1}{g_i(y_i)} \sum_{j=1}^{p_i} \alpha_{ij} b_{ij} \psi'(u_{ij}) \tag{8.8}$$

For $\phi(u_{ij}) = \text{logsig}(u_{ij})$, $\psi'(u_{ij}) = b_{ij}(1-2\phi(u_{ij}))\phi'(u_{ij}) = (\exp(-u_{ij})-1)\exp(-u_{ij})/(1+\exp(-u_{ij}))^3$. We adopt the natural gradient decent algorithm eq. (4.28) and plug eq. (8.8) into eq. (4.28) to obtain the update equation for tuning $\mathbf{W}$.

Algorithm for the tuning of parameters of the mixture of densities

The parameters $\{\gamma, \mathbf{a}, \mathbf{b}\}$ are tuned in the direction to minimize $J(\mathbf{V}, \gamma, \mathbf{a}, \mathbf{b})$. This minimization has the effect of approximating the flexible $g_i(y_i)$ to $p_{y_i}(y_i)$, $i = 1, ..., n$, because $J(\mathbf{W}, \gamma, \mathbf{a}, \mathbf{b})$ would decrease if the shape of function $g_i(y_i)$ is closer to the $p_{y_i}(y_i)$, and J would attain its lower bound 0 if $g_i(y_i) = p_{y_i}(y_i)$, $i = 1, ..., n$, and $\{y_i, i = 1, ..., n\}$ are mutually independent. Using Bell's explanation [5, 6], the entropy of the transformed variable z_i would be maximized if it is uniformly distributed, which occurs if $f_i(y_i)$ is the CDF of y_i, or equivalently, $g_i(y_i)$ is the pdf of y_i. Hence, minimizing $J(\mathbf{W}, \gamma, \mathbf{a}, \mathbf{b})$ (maximizing output entropy) w.r.t. $\{\gamma, \mathbf{a}, \mathbf{b}\}$ can make $g_i(y_i)$ more close to $p_{y_i}(y_i)$. With each $g_i(y_i)$ more closed to $p_{y_i}(y_i)$, (though the approximation may not be so good if p_i is small), it is more possible that loose matching is achieved and $\mathbf{W}$ with a minimum of $J(\mathbf{W})$ is a correct solution for signal separation.

Stochastic gradient algorithms are used to achieve the minimization. They are derived as follows.

The cost function can be written as:

$$\begin{aligned} J &= E_{\mathbf{y}}\left[\log \frac{p_{\mathbf{y}}(\mathbf{y})}{\prod_{i=1}^{n} g_i(y_i)}\right] \\ &= -E_{\mathbf{y}}\left[\sum_{i=1}^{n} \log g_i(y_i)\right] + E_{\mathbf{y}}[\log p_{\mathbf{y}}(\mathbf{y})] \end{aligned} \tag{8.9}$$

The parameter $\{\gamma_{ij}, a_{ij}, b_{ij}\}$ for channel i only engage in term $\log g_i(y_i)$ in the first term in the above equation. Differentiating J w.r.t. the parameters and replacing the expectation values by their instantaneous values, the algorithms for the adaptation of the nonlinearity are obtained as follows:

$$\begin{aligned} \Delta\gamma_{ij} \propto -\sum_{k=1}^{p_i} \frac{\partial J}{\partial \alpha_{ik}} \frac{\partial \alpha_{ik}}{\partial \gamma_{ij}} &= \sum_{k=1}^{p_i} \left\{\frac{\partial}{\partial \alpha_{ik}} \log(g_i(y_i))\right\} \left\{\frac{\partial}{\partial \gamma_{ij}} \frac{\exp(\gamma_{ik})}{\sum_{m=1}^{p_i} \exp(\gamma_{im})}\right\} \\ &= \frac{1}{g_i(y_i)} \sum_{k=1}^{p_i} b_{ik} \phi'(u_{ik}) \alpha_{ik} (\delta_{kj} - \alpha_{ij}) \end{aligned} \tag{8.10}$$

$$\begin{aligned} \Delta b_{ij} \propto -\frac{\partial J}{\partial b_{ij}} &= \frac{\partial}{\partial b_{ij}} \log(g_i(y_i)) \\ &= \frac{\alpha_{ij}}{g_i(y_i)} \{\phi'(u_{ij}) + \phi''(u_{ij}) b_{ij} (y_i - a_{ij})\} \\ &= \frac{\alpha_{ij}}{g_i(y_i)} \{\phi'(u_{ij}) + \phi''(u_{ij}) u_{ij}\} \end{aligned} \tag{8.11}$$

$$\begin{aligned} \Delta a_{ij} \propto -\frac{\partial J}{\partial a_{ij}} &= \frac{\partial}{\partial a_{ij}} \log(g_i(y_i)) \\ &= -\frac{1}{g_i(y_i)} \alpha_{ij} b_{ij}^2 \phi''(u_{ij}) \end{aligned} \tag{8.12}$$

δ_{ij} is the Kroniker delta function. The algorithm with mixture of densities is to update the parameters according to eq. (4.28), (8.10), (8.11) and (8.12) on the arrival of each data point.

It should be noted that the parameters $\{\gamma_{ij}, b_{ij}, a_{ij}\}$ for each channel cannot be initialized in the way that all γ_{ij}, $j = 1, ..., p_i$, are equal, all b_{ij}, $j = 1, ..., p_i$, are equal and all a_{ij}, $j = 1, ..., p_1$, are equal. If so, there would be 'symmetry' among the parameters. and components of different index (j) would be updated in the same way. Then, the shape of $g_i(y_i)$ would be greatly restricted and might not be able to adapt the source distribution.

Remark 9 In digital computer simulations, the division by $g_i(y_i)$ in eq. (8.8), (8.10), (8.11) and (8.12) may have precision error when the magnitude of the data point is

large, since $g_i(y_i)$ would be very small then. In some case, $g_i(y_i)$ is rounded to zero and the error division by zero may occur.

8.2 Experimental Verification on the Nonlinearity Adaptation

The experiments demonstrate that the flexible mixtures of densities are able to adapt themselves to approximate the marginal densities of the recovered signals. Moreover, it also demonstrates that the adaptive mixtures of densities can separate sub-Gaussian sources that single logsig$(\cdot)$ cannot separate. $\mathbf{W}$ is initialized as an identity matrix. The learning rates for $\mathbf{W}$, γ, $\mathbf{b}$, $\mathbf{a}$ are kept at 0.0001, 0.001, 0.01, 0.001 respectively, which results from empirical testing. The mixing matrix used is:

$$A = \begin{bmatrix} 1 & 0.6 \\ 0.7 & 1 \end{bmatrix} \tag{8.13}$$

In experiment 1, the algorithm is applied to two channels of uniformly distributed sources and the experimental convergence behavior is discussed. Experiment 2 tests the algorithm on two channels of human speech signals and experiment 3 tests the algorithm on 3 channels of different signals.

8.2.1 Experiment 1: Two channels of sub-Gaussian sources

Two channels of uniform distributed sources used in Section 6.3.1 are used in this experiment. We try using mixtures of densities with 7 components.

All γ_{ij} are initialized as $1/7$. $b_{i1}, \cdots, b_{i7}$ for each channel are initialized in the range $[10^{-0.3}, 10^{1.8}]$ such that $\log_{10} b_{ij}$ are in a regular interval in $[-0.3, 1.8]$, so that $g_i(y_i)$ is rich of components of different variants. All a_{ij} are initialized as 0. The simulation runs for 4 scans through the data set (400,000 points). The elements of $\mathbf{W}$ versus the number of data points trained are plotted in Figure 8.1. The performance graph, interference-to-signal power ratios versus the number of data points trained, are plotted in Figure 8.2 and the graph showing the histograms of $\{y_i\}$, $\{z_i\}$, and the shapes of the adapted $g_i(y_i)$ and $f_i(y_i)$ after 50,000 data points have been trained and 400,000 data points have been trained are plotted in Figure 8.3 and 8.4 respectively.

This convergence behavior can be described in two phases. In the first phase, approximately from the start to having been trained for 50,000 data points, $\mathbf{W}$ moves

from its initial value to a correct solution for separation. It can be seen from the performance graph that the system can already perform signal separation successfully right after trained for 50,000 data points. A snapshot of $\mathbf{W}$ after 50,000 data points trained is:

$$\mathbf{W} = \begin{bmatrix} 1.5358 & -0.9091 \\ -1.2807 & 1.8149 \end{bmatrix} \tag{8.14}$$

which corresponds to:

$$\mathbf{V} = \mathbf{W}A = \begin{bmatrix} 0.8994 & 0.0124 \\ -0.0102 & 1.0465 \end{bmatrix} \tag{8.15}$$

with average interference-to-signal power ratio -39dB. Figure 8.3 shows that $g_i(y_i)$ have already adapted itself to be close to the pdf of uniform distribution, and $f_i(y_i)$ have already adapted itself to be closed to the CDF of uniform distribution - a straight line from 0 to 1 across the range. Loose matching is already achieved.

However, $\mathbf{W}$ continues to increase in magnitude slowly afterward, keeping itself as a correct solution for separation. This can be called the second phase of the convergence and is reasoned as follows. As discussed in previous chapters, the magnitude of the recovered signals, or $\mathbf{W}$, is controlled by the scaling of the nonlinearity. However, the scaling of $\{g_i(y_i)\}$ is controlled by $\{\gamma, \mathbf{a}, \mathbf{b}\}$ and are free to change. Hence the system can continue to search for a better scaling after the more dominant process, source separation has already been achieved.

The increase in the scale of $\mathbf{W}$ (or $\mathbf{y}$) shown in the experimental result suggests that the $J(\mathbf{W}, \gamma, \mathbf{a}, \mathbf{b})$ has lower value if the scale is greater. This may be because the approximation of $p_{y_i}(y_i)$ by $g_i(y_i)$ is better if the scale is greater, as seen from Figure 8.4 that the shape of $g_i(y_i)$ is more close to the pdf of uniform distribution and $f_i(y_i)$ looks more like a slant straight line cross the range.

8.2.2 Experiment 2: Two channels of super-Gaussian sources

The two channels of human speech signals used in Section 6.3.2 are used in this experiment. Both channels are super-Gaussian. The initialization is the same as that in the previous experiment. The simulations using the algorithm with mixture of densities, and the algorithm with reversed sigmoids $h_i(y_i) = 1 - 2\,\mathrm{logsig}(y_i)$, $h_i(y_i) = -2\tanh(y_i)$ and $h_i(y_i) = -2y_i/(1+y_i^2)$ are run for 1 scan of the data set. All of them can perform

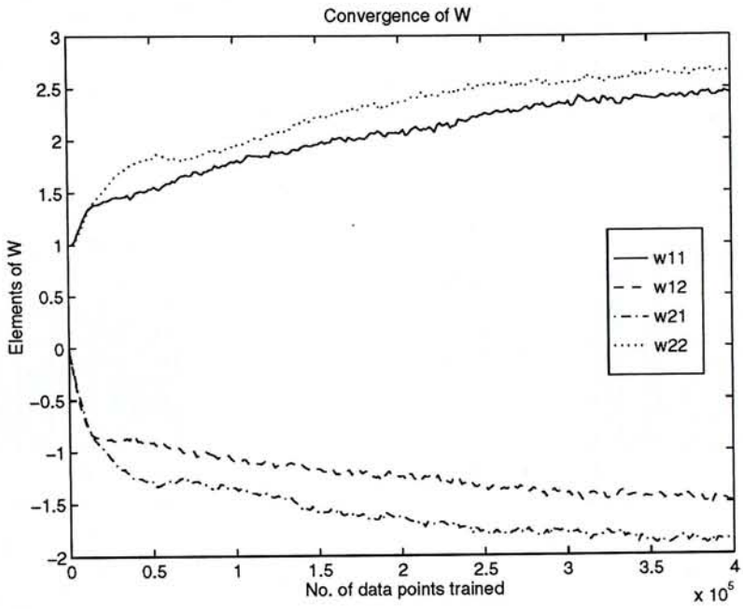


Figure 8.1: Convergence of **W** in trial 1 of experiment 1.

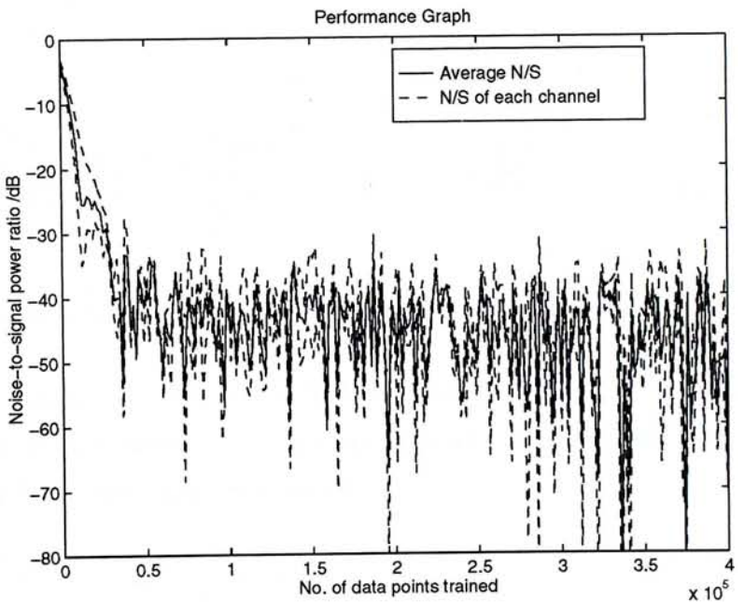


Figure 8.2: The performance graph of trial 1 of experiment 1.

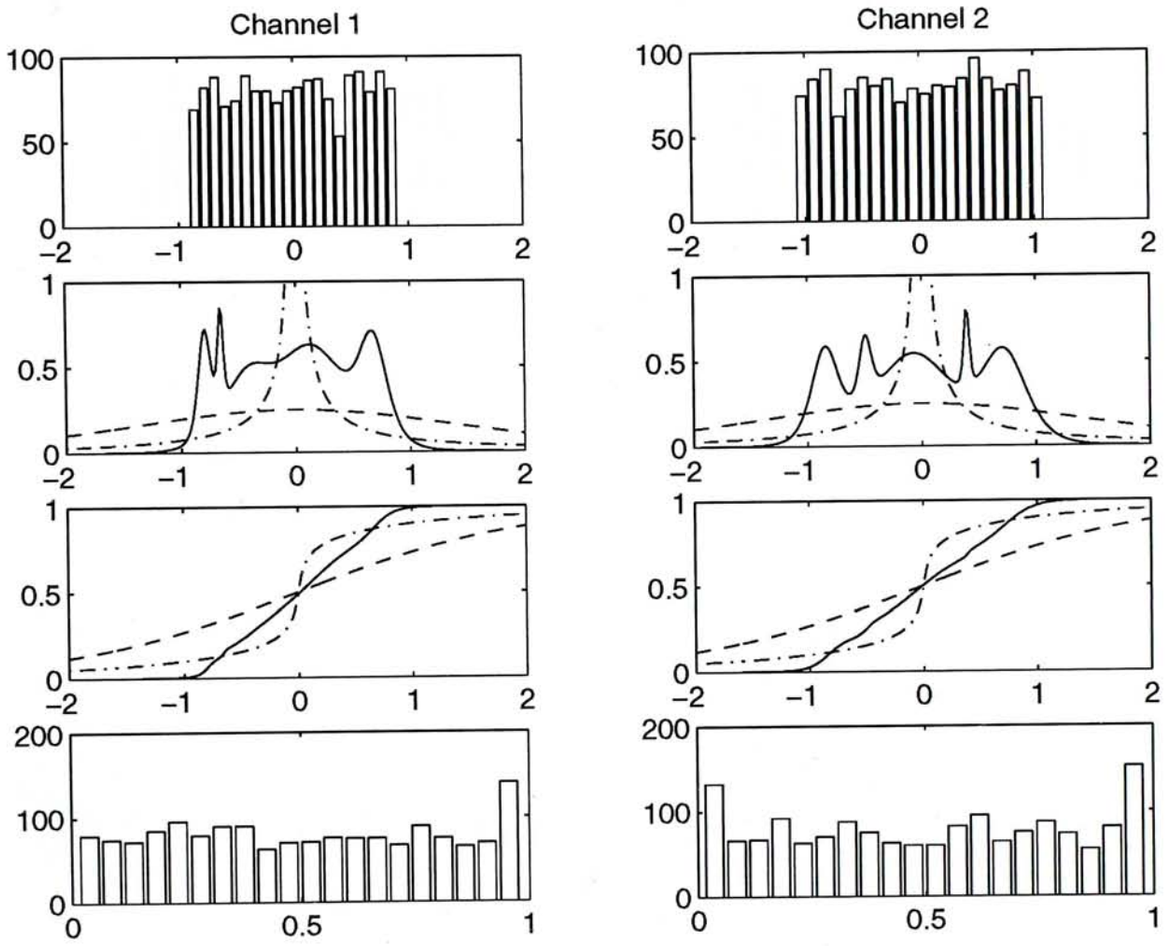


Figure 8.3: The result of trial 1 of experiment 1 after 50,000 data points are trained.

First row: histograms of y_i. Second and third row: $g_i(y_i)$ and $f_i(y_i)$ respectively. (—— adapted mixture of densities, $-\cdot-$ initial mixture of densities, $--$ $f_i(\cdot) = \text{logsig}(\cdot)$ for comparison.) Last row: histograms of z_i.

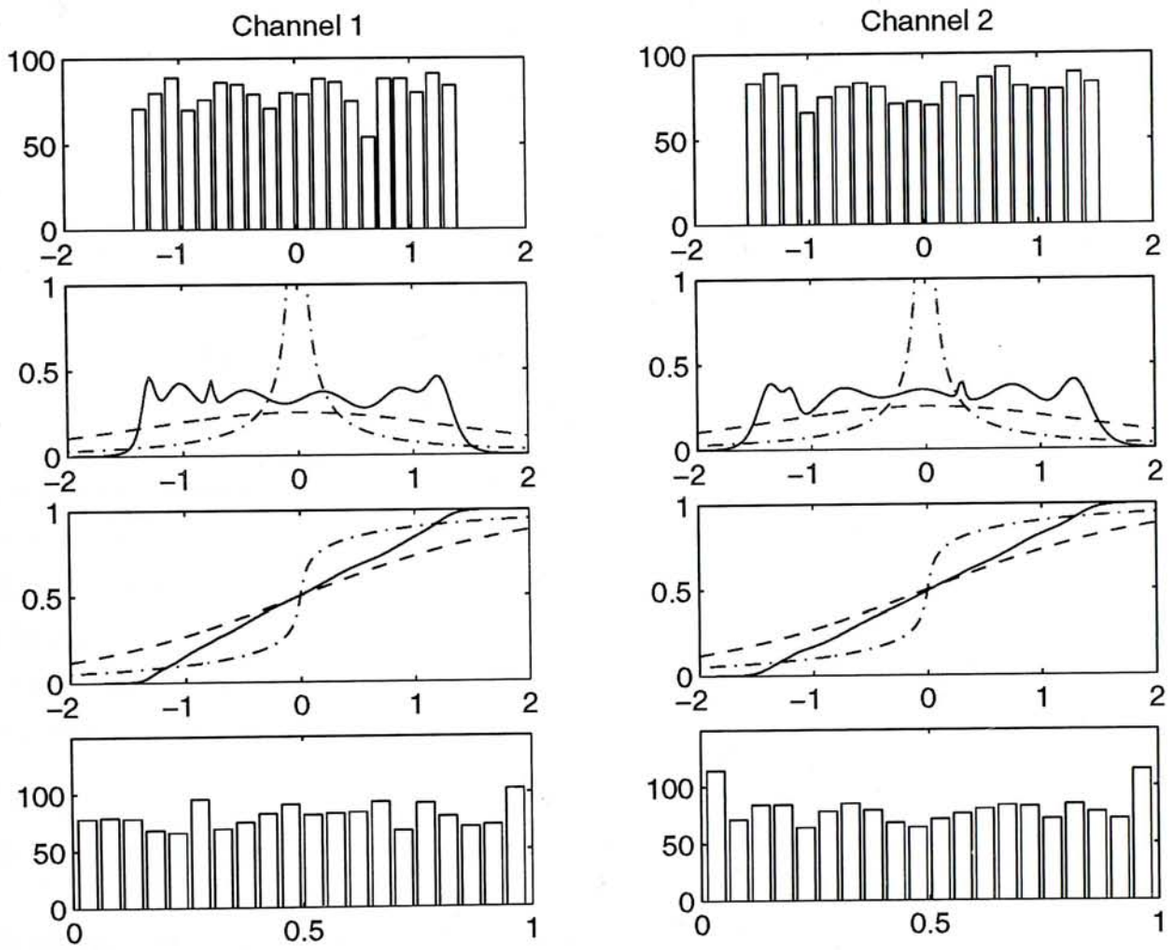


Figure 8.4: The result of trial 1 of experiment 1 after 400,000 data points are trained.

First row: histograms of y_i. Second and third row: $g_i(y_i)$ and $f_i(y_i)$ respectively. (—— adapted mixture of densities, $-\cdot-$ initial mixture of densities, $--$ $f_i(\cdot) = \text{logsig}(\cdot)$ for comparison.) Last row: histograms of z_i.

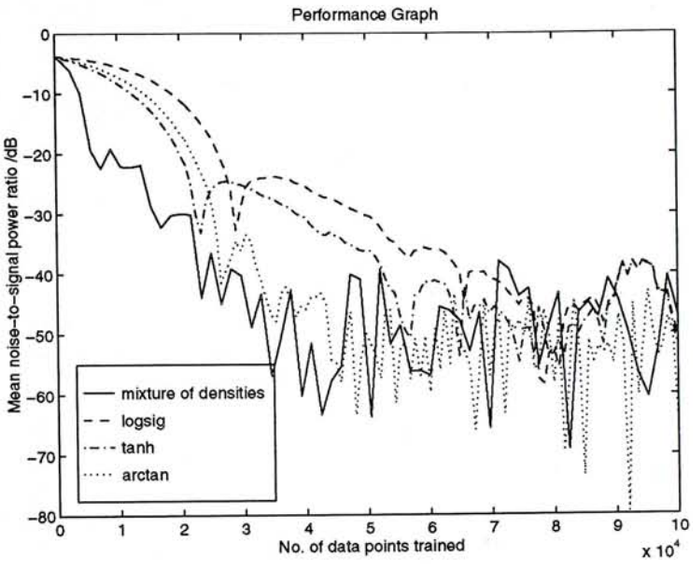


Figure 8.5: The performance graphs for Experiment 2.

separation successfully. The performance graph of them is plotted in Figure 8.5. It can be noticed that the mixture of densities has highest convergence speed among the four. The histograms of $\{y_i\}$ and $\{z_i\}$ and the shapes of $\{g_i(y_i)\}$ and $\{f_i(y_i)\}$ are plotted in Figure 8.6. It can be verified that every z_i is more or less uniformly distributed and hence $f_i(y_i)$ have been adapted to approximate the CDF of y_i, and equivalently, $g_i(y_i)$ have been adapted to approximate $p_{y_i}(y_i)$.

8.2.3 Experiment 3: Three channels of different signals

In this experiment, the data set used in Section 6.5.2, which consists of one beta$(0.5, 0.5)$ distributed source, one uniformly distributed source and one permuted speech signal, and the same mixing matrix, are used. We compare the result of the algorithm with mixture of densities and the algorithms with fixed nonlinearities.

Result of the algorithm with mixture of densities

The mixtures of densities with 5 components are used. All γ_{ij} are initialized as $1/5$ and all a_{ij} are initialized as 0. $b_{i1}, \cdots, b_{i5}$ of each channel are initialized in the interval $[10^{-0.3}, 10^{1.2}]$ such that $\log_{10} b_{ij}$ are at regular interval inside $[-0.3, 1.2]$. The simulation runs for one scan through the data set. All sources are successfully separated. The histograms of $\{y_i\}$ and $\{z_i\}$, and the shape of $\{g_i(y_i)\}$ and $\{f_i(y_i)\}$ are plotted in

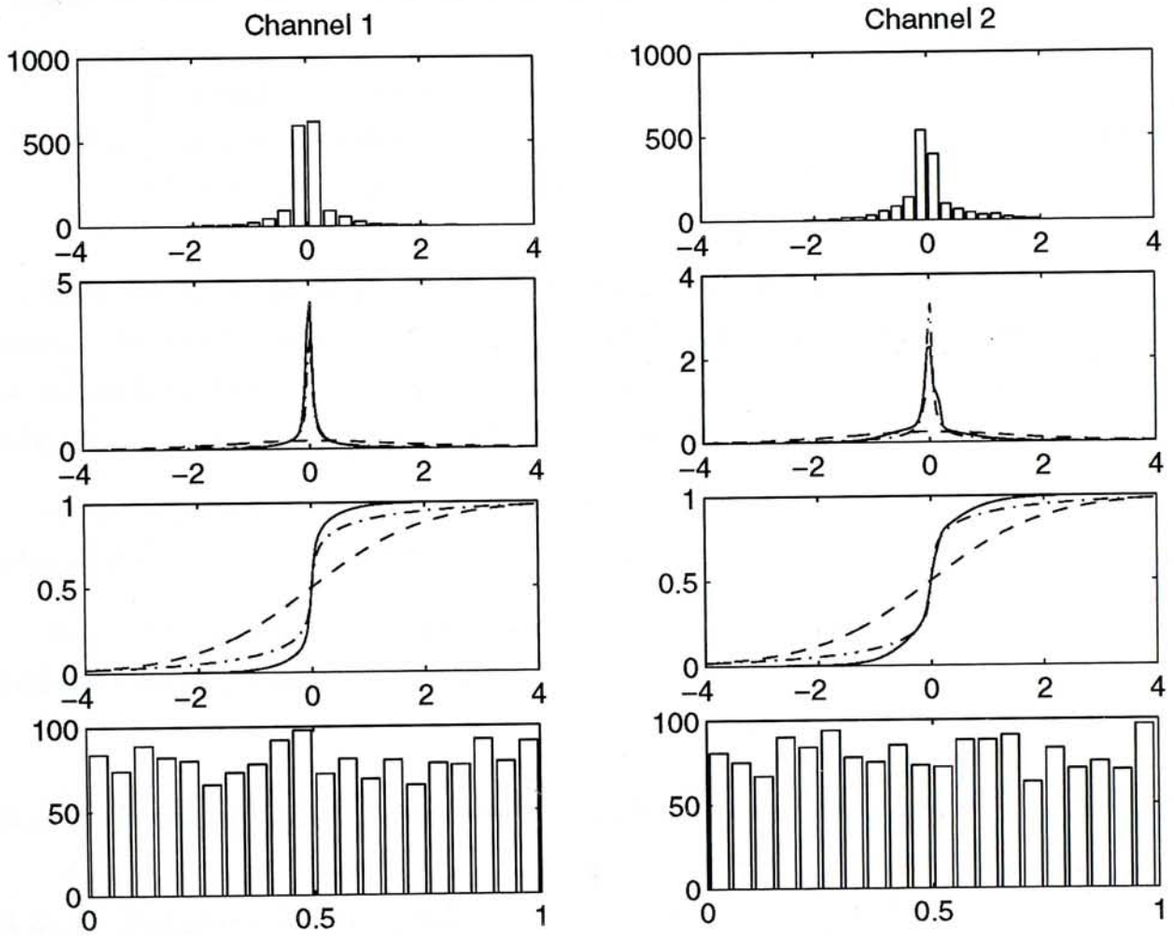


Figure 8.6: The result of Experiment 2.

First row: histograms of y_i. Second and third row: $g_i(y_i)$ and $f_i(y_i)$ respectively. (—— adapted mixture of densities, $- \cdot -$ initial mixture of densities, $- -$ $f_i(\cdot) = \text{logsig}(\cdot)$ for comparison.) Last row: histograms of z_i.

Figure 8.7.

Result of the algorithm with fixed nonlinearities

The INFORMAX algorithm [5] with $h_i(y_i) = 1 - 2\,\mathrm{logsig}(y_i)$ is tested on the data set. The simulation runs for 5 scan through the data set and the system have converged to a stable solution. The snapshot of $\mathbf{V}$ at the end of the simulation is:

$$\mathbf{V} = \begin{bmatrix} 3.4005 & -2.0905 & -0.0545 \\ 3.3719 & 2.0888 & -0.0260 \\ -0.0242 & -0.0157 & 9.3984 \end{bmatrix} \tag{8.16}$$

Only the third channel, the super-Gaussian human speech signal, can be separated. The sub-Gaussian uniformly distributed and beta-distributed sources cannot be separation. The same experiments have been done using $h_i(y_i) = -2\tanh(y_i)$ and $h_i(y_i) = -2y_i/(1 + y_i^2)$ and the results are similar.

In the experiment on the same data set in Section 6.5.2, only the beta distributed source can be extracted and the other two sources remain mixed.

Hence, the flexible mixture of densities is the only one that can separate these three different signals among the nonlinearities we tested.

8.3 Seeking the Simplest Workable Mixtures of Densities

8.3.1 Number of components

The above experiments show that the mixtures of five or seven densities can approximate the source densities 'quite well' (from inspection on the shapes of $\{g_i(y_i)\}$ and the histograms of $\{y_i\}$), and they can separate all sources we tried in the experiments. The number of components five or seven is arbitrarily chosen in the previous section for empirical trials and we would like to question on "what is the minimum number of components that can separate sources of any distribution?". This is an important question since the cost of implementation decreases with simplicity of the system. We try using two components in the mixtures of densities and initialize them in another way: all $\gamma_{ij} = 0.5$, all $b_{ij} = 1$ and $[a_{i1}, a_{i2}] = [-1, 1]$ for each channel. All the three data sets in the previous section are tested and the mixtures of two densities can perform

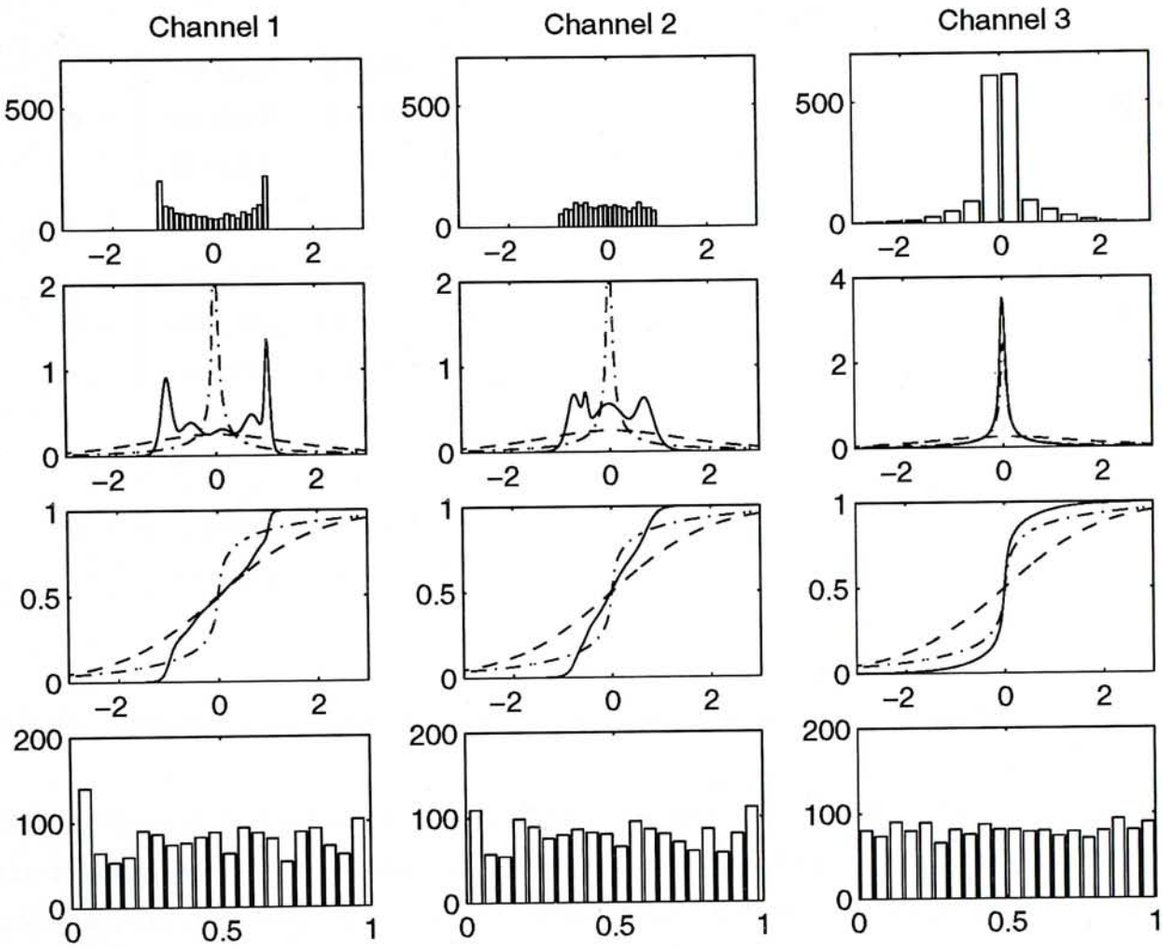


Figure 8.7: The result of experiment 3. Legends are same as those in Figure 8.3

First row: histograms of y_i. Second and third row: $g_i(y_i)$ and $f_i(y_i)$ respectively. (—— adapted mixture of densities, $-\cdot-$ initial mixture of densities, $--$ $f_i(\cdot) = \text{logsig}(\cdot)$ for comparison.) Last row: histograms of z_i.

separation in all cases. The result of the test on the three channels of different source in Experiment 3 is plotted in Figure 8.8 and a snapshot of the parameters is:

$$\mathbf{V} = \begin{bmatrix} 1.4211 & 0.0062 & -0.0469 \\ -0.0139 & 0.9969 & -0.0224 \\ -0.0088 & 0.0030 & 3.9505 \end{bmatrix} \tag{8.17}$$

$$\gamma = \begin{bmatrix} -0.0000 & 0.0000 \\ -0.0486 & 0.0486 \\ 0.1183 & -0.1183 \end{bmatrix} \tag{8.18}$$

$$\mathbf{a} = \begin{bmatrix} -0.4450 & 0.4373 \\ -0.5013 & 0.4892 \\ -0.0035 & 0.0613 \end{bmatrix} \tag{8.19}$$

$$\mathbf{b} = \begin{bmatrix} 7.6672 & 7.9065 \\ 5.5956 & 5.0061 \\ 13.3486 & 1.4946 \end{bmatrix} \tag{8.20}$$

8.3.2 Mixture of two densities with only biases changeable

The mixture of two densities is suspected to be still 'too flexible'. Firstly, the scale of the nonlinearity is not constrained as the discussion in Experiment 1 in Section 8.2.1 pointed out. Secondly, the empirical results discussed in nonlinearity and separation capability suggest that the 'loose matching' seems so loose that only the peakedness (kurtosis) of $g_i(y_i)$ is of concern. Hence we try using mixtures of two densities that only the kurtosis of whom are tunable and all other freedom is fixed.

One of the implementation of the above idea is to fix all $\gamma_{ij} = 0$ and fix all $b_{ij} = 1$. The biases (centers) of components $\{a_{ij}\}$ are allowed to change. The biases of components control the distance with the two bell-shape components and hence control the kurtosis of $g_i(y_i)$. The bias $\{a_{i1}, a_{i2}\}$ for each channel are initialized in a symmetrical way as $\{-1, 1\}$.

The mixture of two densities with only bias changeable is tested on the three data sets in Section 8.2. In all cases source separation is successful. The result of the

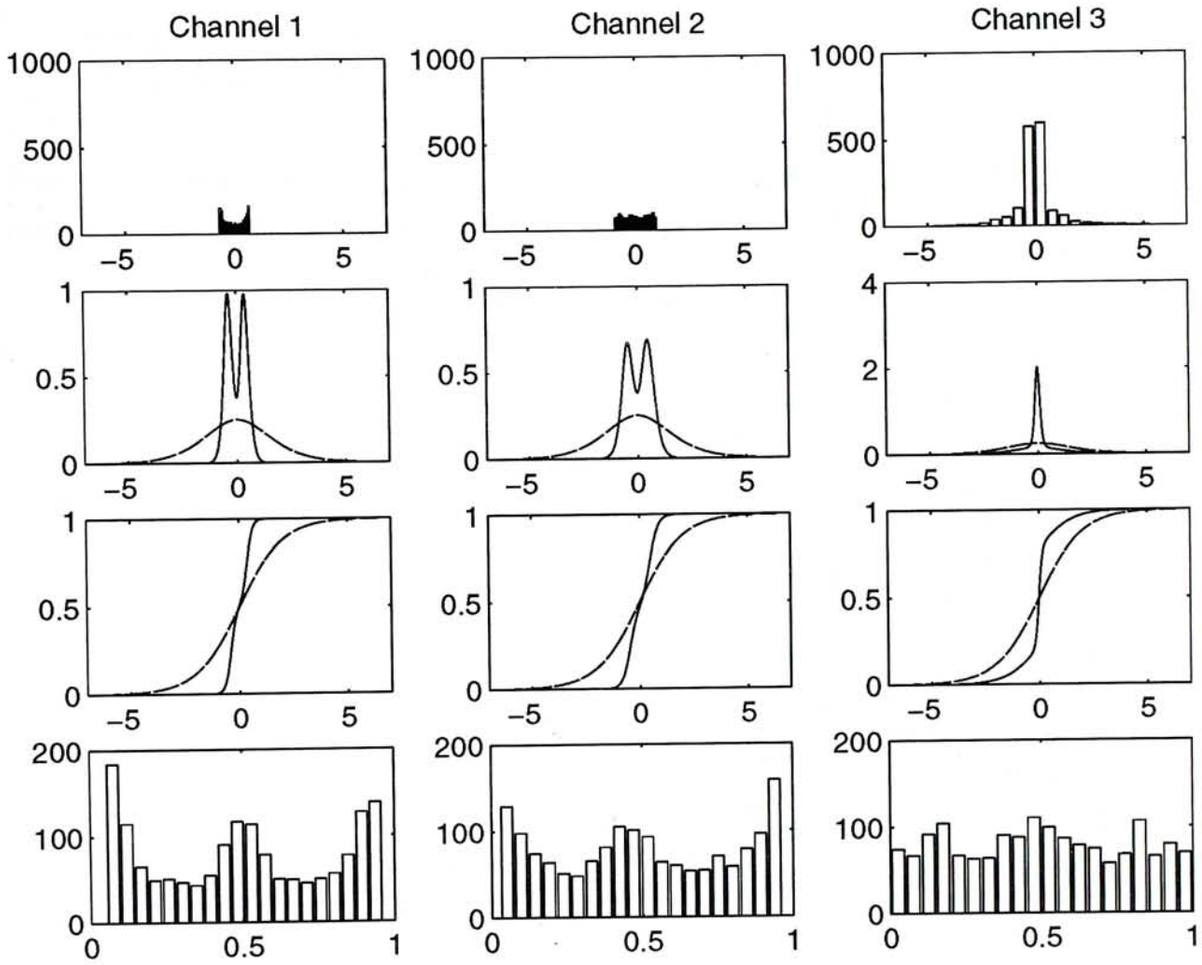


Figure 8.8: The experiment result of the algorithm with mixture of two densities.

experiment on the three channels of different signals are plotted in Figure 8.9. After the system has stabilized, a snapshot of $\mathbf{V}$ and $\mathbf{a}$ are:

$$\mathbf{V} = \begin{bmatrix} 10.2583 & 0.0205 & -0.1169 \\ -0.0085 & 5.0408 & -0.0813 \\ -0.0132 & -0.0095 & 9.3977 \end{bmatrix} \tag{8.21}$$

$$\mathbf{a} = \begin{bmatrix} -3.3378 & 3.3657 \\ -2.4460 & 2.4672 \\ 0.0241 & 0.0241 \end{bmatrix} \tag{8.22}$$

From the figures, the resulting $\{g_i(y_i)\}$ have the distances between the two centers corresponding to the kurtosis of the source as expected.

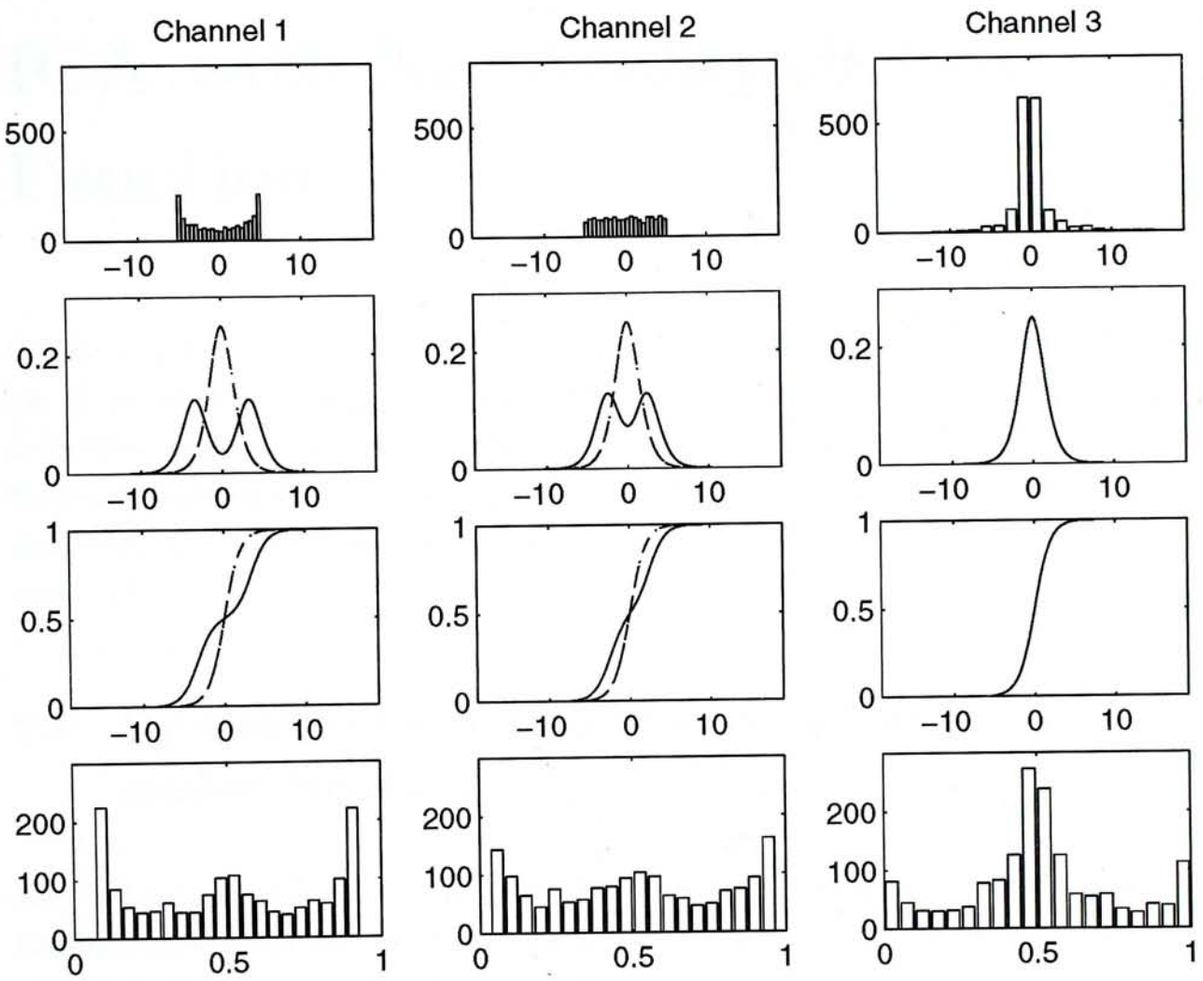


Figure 8.9: The experiment result of the algorithm with mixture of two densities that γ and $\mathbf{b}$ are fixed.

Chapter 9

ICA with Non-Kullback Cost Function

In this chapter, we leave the Information-theoretic ICA approach and consider the use of non-Kullback separation functionals in the ICA problem. In Section 9.1, we investigate the possibility of developing algorithms from ICA principles that use non-Kullback separation functionals proposed by Xu [73]. In Section 9.2, experiments that verify the performance and capability of one non-Kullback ICA algorithm will be presented.

9.1 Derivation of ICA Algorithms from Non-Kullback Separation Functionals

Xu [72] has suggested the use of three non-Kullback separation functionals besides the Kullback divergence in the YING-YANG learning theory. They are suggested to be applied to the ICA problem and are investigated in the coming sub-sections.

9.1.1 Positive Convex Divergence

The Convex divergence is given by:

$$F_s(M_1, M_2) = \zeta(1) - \int_{\mathbf{y}} p_{M_1}(\mathbf{y}) \zeta\left(\frac{p_{M_2}(\mathbf{y})}{p_{M_1}(\mathbf{y})}\right) \mathrm{d}\mathbf{y} \tag{9.1}$$

where $\zeta(\cdot)$ is a convex function in $(0, \infty)$. We use $\zeta(r) = r^\beta$, $0 < \beta < 1$ here and the Convex divergence is called the Positive Convex (PC) divergence. The case with $\beta =$

0.5 is of particular interest since it reduces to the Root-Iner-Product (RIP) divergence

$$F_s(p_{M_1}, p_{M_2}) = 1 - \int_{\mathbf{y}} \sqrt{p_{M_1}(\mathbf{y}) p_{M_2}(\mathbf{y})} \mathrm{d}\mathbf{y} \tag{9.2}$$

The PC divergence is suggested to be the cost function of a generalized version of the Minimum Mutual Information approach [73]. If we use $\prod_{i=1}^{n} g_i(y_i)$ in place of $p_{M_2}(\mathbf{y})$ as in the information-theoretic ICA approach, the cost function become:

$$\begin{aligned} J(\mathbf{W}) &= 1 - \int_{\mathbf{y}} p_{\mathbf{y}}(\mathbf{y}) \left(\frac{\prod_{i=1}^{n} g_i(y_i)}{p_{\mathbf{y}}(\mathbf{y})} \right)^{\beta} \mathrm{d}\mathbf{y} \\ &= 1 - \int_{\mathbf{x}} p_{\mathbf{x}}(\mathbf{x}) \left(\frac{|\det \mathbf{W}| \prod_{i=1}^{n} g_i(\mathbf{w}_i^T \mathbf{x})}{p_{\mathbf{x}}(\mathbf{x})} \right)^{\beta} \mathrm{d}\mathbf{x} \\ &= 1 - E_{\mathbf{x}} \left[\left(\frac{|\det \mathbf{W}| \prod_{i=1}^{n} g_i(\mathbf{w}_i^T \mathbf{x})}{p_{\mathbf{x}}(\mathbf{x})} \right)^{\beta} \right] \end{aligned} \tag{9.3}$$

The minimization of this cost function is suggested to be an ICA principle [73]. The derivative of this cost function w.r.t. w_{ij} is given by

$$\frac{\partial J(\mathbf{W})}{\partial w_{ij}} = -E_{\mathbf{x}} \left[\beta \left(\frac{|\det \mathbf{W}| \prod_{i=1}^{n} g_i(\mathbf{w}_i^T \mathbf{x})}{p_{\mathbf{x}}(\mathbf{x})} \right)^{\beta} \left(\frac{\mathrm{cof} w_{ij}}{\det \mathbf{W}} + h_i(\mathbf{w}_i^T \mathbf{x}) x_j \right) \right] \tag{9.4}$$

The gradient is given by:

$$\nabla_W J(\mathbf{W}) = -E_{\mathbf{x}} \left[\beta \left(\frac{|\det \mathbf{W}| \prod_{i=1}^{n} g_i(\mathbf{w}_i^T \mathbf{x})}{p_{\mathbf{x}}(\mathbf{x})} \right)^{\beta} ([\mathbf{W}^T]^{-1} + \mathbf{h}(\mathbf{W}\mathbf{x})\mathbf{x}^T) \right] \tag{9.5}$$

and the natural gradient is given by:

$$\begin{aligned} &\nabla_W J(\mathbf{W}) \mathbf{W}^T \mathbf{W} \\ &= -E_{\mathbf{x}} \left[\beta \left(\frac{|\det \mathbf{W}| \prod_{i=1}^{n} g_i(\mathbf{w}_i^T \mathbf{x})}{p_{\mathbf{x}}(\mathbf{x})} \right)^{\beta} (\mathbf{I} + \mathbf{h}(\mathbf{W}\mathbf{x})(\mathbf{W}\mathbf{x})^T) \right] \mathbf{W} \end{aligned} \tag{9.6}$$

We target at constructing an adaptive (natural) gradient decent algorithm to minimize the PC divergence. $p_{\mathbf{x}}(\mathbf{x})$ is a quantity that cannot be estimated from one sample, however, since $1/(p_{\mathbf{x}}(\mathbf{x}))^{\beta}$ occurs as a multiplicative scalar in the (natural) gradient, which does not affect the (natural) gradient direction, we can ignore it in the construction of the adaptive algorithm. Hence the update equation:

$$\Delta \mathbf{W} \propto \beta \left(|\det \mathbf{W}| \prod_{i=1}^{n} g_i(y_i) \right)^{\beta} (\mathbf{I} + \mathbf{h}(\mathbf{y})\mathbf{y}^T) \mathbf{W} \tag{9.7}$$

is suggested to be an algorithm to minimize the PC divergence.

PC divergence with mixture of densities

Apart from using pre-fixed $\{g_i(y_i)\}$, flexible mixtures of densities for $\{g_i(y_i)\}$ is also suggested to be used, so that they can adapt the source densities. The form of the mixture of densities is the same as that in Chapter 8. Plugging eq. (8.8) into eq. (9.7), eq. (9.7) becomes the update equation for learning $\mathbf{W}$ in the algorithm with mixture of densities.

The parameters $\{\boldsymbol{\gamma}, \mathbf{a}, \mathbf{b}\}$ are tuned in the direction to minimize $J(\mathbf{W}, \boldsymbol{\gamma}, \mathbf{a}, \mathbf{b})$ for the same reason explained in Chapter 8. The derivatives of the PC divergence in eq. (9.3) are calculated as follows:

$$\frac{\partial J}{\partial g_i(y_i)} = -E_{\mathbf{x}}\left[\beta \frac{(|\det \mathbf{W}| \prod_{k=1}^{n} g_k(\mathbf{w}_k^T\mathbf{x}))^{\beta}}{(p_{\mathbf{x}}(\mathbf{x}))^{\beta} g_i(\mathbf{w}_i^T\mathbf{x})}\right] \tag{9.8}$$

$$\frac{\partial J}{\partial \alpha_{ik}} = -E_{\mathbf{x}}\left[\beta \frac{(|\det \mathbf{W}| \prod_{l=1}^{n} g_l(\mathbf{w}_l^T\mathbf{x}))^{\beta}}{(p_{\mathbf{x}}(\mathbf{x}))^{\beta} g_i(\mathbf{w}_i^T\mathbf{x})} b_{ik}\phi'(u_{ik})\right] \tag{9.9}$$

$$\frac{\partial \alpha_{ik}}{\partial \gamma_{ij}} = \alpha_{ik}(\delta_{kj} - \alpha_{ij}) \tag{9.10}$$

$$\begin{aligned}\frac{\partial J}{\partial \gamma_{ij}} &= \sum_{k=1}^{p_i} \frac{\partial J}{\partial \alpha_{ik}} \frac{\partial \alpha_{ik}}{\partial \gamma_{ij}} \\ &= -\sum_{k=1}^{p_i} E_{\mathbf{x}}\left[\beta \frac{(|\det \mathbf{W}| \prod_{l=1}^{n} g_l(\mathbf{w}_l^T\mathbf{x}))^{\beta}}{(p_{\mathbf{x}}(\mathbf{x}))^{\beta} g_i(\mathbf{w}_i^T\mathbf{x})} \phi'(u_{ik}) b_{ik}\alpha_{ik}(\delta_{kj} - \alpha_{ij})\right]\end{aligned} \tag{9.11}$$

$$\frac{\partial J}{\partial b_{ij}} = -E_{\mathbf{x}}\left[\beta \frac{(|\det \mathbf{W}| \prod_{k=1}^{n} g_k(\mathbf{w}_k^T\mathbf{x}))^{\beta}}{(p_{\mathbf{x}}(\mathbf{x}))^{\beta} g_i(\mathbf{w}_i^T\mathbf{x})} \alpha_{ij}(\phi'(u_{ij}) + \phi''(u_{ij})u_{ij})\right] \tag{9.12}$$

$$\frac{\partial J}{\partial a_{ij}} = E_{\mathbf{x}}\left[\beta \frac{|\det \mathbf{W}| (\prod_{k=1}^{n} g_k(\mathbf{w}_k^T\mathbf{x}))^{\beta}}{(p_{\mathbf{x}}(\mathbf{x}))^{\beta} g_i(\mathbf{w}_i^T\mathbf{x})} \alpha_{ij} b_{ij}^2 \phi''(u_{ij})\right] \tag{9.13}$$

Since $1/(p_x(\mathbf{x}))^{\beta}$ occurs as a multiplicative scalar of the gradient, it is ignored and we obtain the stochastic update equations:

$$\Delta\gamma_{ij} \propto \beta |\det \mathbf{W}|^{\beta} \frac{(\prod_{l=1}^{n} g_l(\mathbf{w}_l^T\mathbf{x}))^{\beta}}{g_i(\mathbf{w}_i^T\mathbf{x})} \sum_{k=1}^{p_i} \phi'(u_{ik}) b_{ik}\alpha_{ik}(\delta_{kj} - \alpha_{ij}) \tag{9.14}$$

$$\Delta b_{ij} \propto \beta |\det \mathbf{W}|^{\beta} \frac{\left(\prod_{k=1}^{n} g_k(\mathbf{w}_k^T \mathbf{x})\right)^{\beta}}{g_i(\mathbf{w}_i^T \mathbf{x})} \alpha_{ij} [\phi'(u_{ij}) + \phi''(u_{ij}) u_{ij}] \tag{9.15}$$

$$\Delta a_{ij} \propto -\beta |\det \mathbf{W}|^{\beta} \frac{\left(\prod_{k=1}^{n} g_k(\mathbf{w}_k^T \mathbf{x})\right)^{\beta}}{g_i(\mathbf{w}_i^T \mathbf{x})} \alpha_{ij} b_{ij}^2 \phi''(u_{ij}) \tag{9.16}$$

Again we use $\phi(u_{ij}) = \text{logsig}(y_{ij}) = \frac{1}{1+\exp(-u_{ij})}$, corresponding to $\phi'(u_{ij}) = \frac{\exp(-u_{ij})}{(1+\exp(-u_{ij}))^2}$ and $\phi''(u_{ij}) = \frac{\exp(-u_{ij})(\exp(-u_{ij})-1)}{(1+\exp(-u_{ij}))^3}$.

9.1.2 L_p Divergence

The L_p divergence can also be used as separation functional of two densities in the YING-YANG learning scheme [72]:

$$F_s(M_1, M_2) = \int_{x,y} p(x) |\xi(p_{M_1}(y|x) p_{M_1}(x)) - \xi(p_{M_2}(x|y) p_{M_2}(y))|^p \mathrm{d}x\mathrm{d}y \tag{9.17}$$

where x is the 'real body' and y is the 'seed'. Using the same procedure as in Chapter 4 to apply the L_p divergence to the ICA problem, the cost function for the ICA problem becomes (Xu, not yet published):

$$\begin{aligned} J &= \int_{\mathbf{s},\mathbf{y}} p_{\mathbf{s}}(\mathbf{s}) \left| \xi(\delta(\mathbf{s}-\mathbf{y}) p_{\mathbf{y}}(\mathbf{y})) - \xi\left(\delta(\mathbf{y}-\mathbf{s}) \prod_{i=1}^{n} g_i(s_i)\right) \right|^p \mathrm{d}\mathbf{s}\mathrm{d}\mathbf{y} \\ &= \int_{\mathbf{y}} p_{\mathbf{y}}(\mathbf{y}) \left| \xi(p_{\mathbf{y}}(\mathbf{y})) - \xi\left(\prod_{i=1}^{n} g_i(y_i)\right) \right|^p \mathrm{d}\mathbf{y} \\ &= E_{\mathbf{y}}\left[\left| \xi(p_{\mathbf{y}}(\mathbf{y})) - \xi\left(\prod_{i=1}^{n} g_i(y_i)\right) \right|^p \right] \end{aligned} \tag{9.18}$$

Two special case of the L_p divergence are:

(Case 1) $p = 2$ and $\xi(r) = \sqrt{r}$ (Xu, not yet published):

The cost function becomes:

$$\begin{aligned} J &= E_{\mathbf{y}}\left[\left(\sqrt{p_{\mathbf{y}}(\mathbf{y})} - \sqrt{\prod_{i=1}^{n} g_i(y_i)} \right)^2 \right] \\ &= E_{\mathbf{x}}\left[\left(\sqrt{\frac{p_{\mathbf{x}}(\mathbf{x})}{|\det \mathbf{W}|}} - \sqrt{\prod_{i=1}^{n} g_i(\mathbf{w}_i^T \mathbf{x})} \right)^2 \right] \end{aligned} \tag{9.19}$$

We tried to investigate whether adaptive implementation of the minimization is possible. The derivative of J w.r.t. w_{ij} is:

$$\begin{aligned}\frac{\partial J(\mathbf{W})}{\partial w_{ij}} = &- E_{\mathbf{x}}\left[\left(\sqrt{\frac{p_{\mathbf{x}}(\mathbf{x})}{|\det \mathbf{W}|}} - \sqrt{\prod_{k=1}^{n} g_j(\mathbf{w}_k^T\mathbf{x})}\right)\right. \\ &\left.\left(\sqrt{\frac{p_{\mathbf{x}}(\mathbf{x})}{|\det \mathbf{W}|}}\frac{\operatorname{cof} w_{ij}}{\det \mathbf{W}} + \sqrt{\prod_{l=1}^{n} g_l(\mathbf{w}_l^T\mathbf{x})h_i(\mathbf{w}_i^T\mathbf{x})x_j}\right)\right]\end{aligned} \tag{9.20}$$

and the gradient is:

$$\begin{aligned}\nabla_W J(\mathbf{W}) = &- E_{\mathbf{x}}\left[\left(\sqrt{\frac{p_{\mathbf{x}}(\mathbf{x})}{|\det \mathbf{W}|}} - \sqrt{\prod_{k=1}^{n} g_k(\mathbf{w}_k^T\mathbf{x})}\right)\right. \\ &\left.\left(\sqrt{\frac{p_{\mathbf{x}}(\mathbf{x})}{|\det \mathbf{W}|}}\left[\mathbf{W}^T\right]^{-1} + \sqrt{\prod_{j=1}^{n} g_j(\mathbf{w}_j^T\mathbf{x})}\mathbf{h}(\mathbf{W}\mathbf{x})\mathbf{x}^T\right)\right]\end{aligned} \tag{9.21}$$

However, as $p_{\mathbf{x}}(\mathbf{x})$ occurs in some additive terms in the gradient and is not a multiplicative scalar, we cannot construct an adaptive algorithm to minimize the cost function.

(Case 2) $p = 1$ and $\xi(r) = r$ (Xu, not yet published):

The cost function becomes:

$$\begin{aligned}J &= \int_{\mathbf{y}} p_{\mathbf{y}}(\mathbf{y})\left|p_{\mathbf{y}}(\mathbf{y}) - \prod_{i=1}^{n} g_i(y_i)\right| d\mathbf{y} \\ &= \int_{\mathbf{x}} p_{\mathbf{x}}(\mathbf{x})\left|\frac{p_{\mathbf{x}}(\mathbf{x})}{|\det \mathbf{W}|} - \prod_{i=1}^{n} g_i(\mathbf{w}_i^T\mathbf{x})\right| d\mathbf{x} \\ &= E_{\mathbf{x}}\left[\left|\frac{p_{\mathbf{x}}(\mathbf{x})}{|\det \mathbf{W}|} - \prod_{i=1}^{n} g_i(\mathbf{w}_i^T\mathbf{x})\right|\right]\end{aligned} \tag{9.22}$$

As there is a $p_{\mathbf{x}}(\mathbf{x})$ in one additive term of J, the gradient must also consist of $p_{\mathbf{x}}(\mathbf{x})$ in some additive terms, which is not common multiplicative scalar of the gradient. Hence, we cannot construct an adaptive algorithm to minimize the cost function.

Since we cannot construct adaptive algorithms for the minimization of L_p divergence, the use of L_p divergence to the ICA problem is not investigated further.

9.1.3 De-correlation Index

The de-correlation index can also be used as the separation functional of the YING-YANG Machine [72]:

$$F_s(M_1, M_2) = 1 - \frac{\int_{x,y} p(x)\xi\left(p_{M_1}(y|x)p_{M_1}(x)\right)\xi\left(p_{M_2}(x|y)p_{M_2}(y)\right)\mathrm{d}x\mathrm{d}y}{\sqrt{\int_{x,y} p(x)\xi^2\left(p_{M_1}(y|x)p_{M_1}(x)\right)\mathrm{d}x\mathrm{d}y}\sqrt{\int_{x,y} p(x)\xi^2\left(p_{M_2}(x|y)p_{M_2}(y)\right)\mathrm{d}x\mathrm{d}y}} \tag{9.23}$$

For the case $\xi(r) = \sqrt{r}$, and using the procedures used in the previous section, the cost function for the ICA problem is derived as (Xu, not yet published):

$$\begin{aligned} J &= 1 - \frac{\int_{\mathbf{s},\mathbf{y}} p_{\mathbf{y}}(\mathbf{y})\delta(\mathbf{s}-\mathbf{y})\sqrt{p_{\mathbf{y}}(\mathbf{y})}\sqrt{\prod_{i=1}^n g_i(s_i)}\mathrm{d}\mathbf{s}\mathrm{d}\mathbf{y}}{\sqrt{\int_{\mathbf{s},\mathbf{y}} p_{\mathbf{y}}(\mathbf{y})\delta(\mathbf{s}-\mathbf{y})p_{\mathbf{y}}(\mathbf{y})\mathrm{d}\mathbf{s}\mathrm{d}\mathbf{y}}\sqrt{\int_{\mathbf{s},\mathbf{y}} p_{\mathbf{y}}(\mathbf{y})\delta(\mathbf{s}-\mathbf{y})\prod_{i=1}^n g_i(s_i)\mathrm{d}\mathbf{s}\mathrm{d}\mathbf{y}}} \\ &= 1 - \frac{\int_{\mathbf{y}} p_{\mathbf{y}}(\mathbf{y})\sqrt{p_{\mathbf{y}}(\mathbf{y})}\sqrt{\prod_{i=1}^n g_i(y_i)}\mathrm{d}\mathbf{y}}{\sqrt{\int_{\mathbf{y}} p_{\mathbf{y}}(\mathbf{y})p_{\mathbf{y}}(\mathbf{y})\mathrm{d}\mathbf{y}}\sqrt{\int_{\mathbf{y}} p_{\mathbf{y}}(\mathbf{y})\prod_{j=1}^n g_j(y_j)\mathrm{d}\mathbf{y}}} \\ &= 1 - \frac{\int_{\mathbf{x}} p_{\mathbf{x}}(\mathbf{x})\sqrt{p_{\mathbf{x}}(\mathbf{x})}\sqrt{\prod_{i=1}^n g_i(\mathbf{w}_i^T\mathbf{x})}\mathrm{d}\mathbf{x}}{C\sqrt{\int_{\mathbf{x}} p_{\mathbf{x}}(\mathbf{x})\prod_{j=1}^n g_j(\mathbf{w}_j^T\mathbf{x})\mathrm{d}\mathbf{x}}} \end{aligned} \tag{9.24}$$

where $C = \int_{\mathbf{x}}[p_x(\mathbf{x})]^2\mathrm{d}\mathbf{x}$.

The gradient of the cost function is:

$$\begin{aligned} &\nabla_W J(\mathbf{W}) = \\ &-\frac{1}{C\int_{\mathbf{x}} p_{\mathbf{x}}(\mathbf{x})\prod_{i=1}^n g_i(\mathbf{w}_i^T\mathbf{x})\mathrm{d}\mathbf{x}} \\ &\left\{\sqrt{\int_{\mathbf{x}} p_{\mathbf{x}}(\mathbf{x})\prod_{j=1}^n g_j(\mathbf{w}_j^T\mathbf{x})\mathrm{d}\mathbf{x}}\left[\frac{1}{2}\int_{\mathbf{x}} p_{\mathbf{x}}(\mathbf{x})\sqrt{p_{\mathbf{x}}(\mathbf{x})}\sqrt{\prod_{k=1}^n g_k(\mathbf{w}_k^T\mathbf{x})}\mathbf{h}(\mathbf{W}\mathbf{x})\mathbf{x}^T\mathrm{d}\mathbf{x}\right]\right. \\ &\left.-\left[\int_{\mathbf{x}} p_{\mathbf{x}}(\mathbf{x})\sqrt{p_{\mathbf{x}}(\mathbf{x})}\sqrt{\prod_{l=1}^n g_l(\mathbf{w}_l^T\mathbf{x})}\mathrm{d}\mathbf{x}\right]\left[\frac{\int_{\mathbf{x}} p_{\mathbf{x}}(\mathbf{x})\left(\prod_{p=1}^n g_p(\mathbf{w}_p^T\mathbf{x})\right)\mathbf{h}(\mathbf{W}\mathbf{x})\mathbf{x}^T\mathrm{d}\mathbf{x}}{2\sqrt{\int_{\mathbf{x}} p_{\mathbf{x}}(\mathbf{x})\prod_{q=1}^n g_q(\mathbf{w}_q^T\mathbf{x})\mathrm{d}\mathbf{x}}}\right]\right\} \end{aligned} \tag{9.25}$$

Since $p_{\mathbf{x}}(\mathbf{x})$ is not a multiplicative scalar to the gradient, we cannot construct an adaptive gradient descent algorithm. Hence, the use of de-correlation index on the ICA problem is not investigated further.

9.2 Experiments on the ICA Algorithm Based on Positive Convex Divergence

9.2.1 Experiments on the algorithm with fixed nonlinearities

The ICA algorithm based on PC divergence eq. (9.7) differs from the information-theoretic ICA algorithm eq. (4.28) that eq. (9.7) has an extra $\beta\left(|\det \mathbf{W}| \prod_{i=1}^{n} g_i(y_i)\right)^{\beta}$ factor. Xu [73] suggested that, as $g_i(y_i)$ has a bell shape, samples with large magnitude should have smaller effect in the algorithm based on PC divergence than in the information-theoretic algorithm, since the bell shape multiplier $g_i(y_i)$ will act as a small factor for large magnitude y_i. Therefore, we compare the results by the two algorithms on samples with some large magnitude 'outliers'. The samples with 'outliers' are formed by randomly selecting 5% of data points (with order of magnitude about 1) and superimposing ± 10 to them.

We test the hypothesis with six sets of source:

(1) 2 channels of beta(0.5,0.5) distributed sources, scaled and shifted to be within [-2,2] (sub-Gaussian)

(2) 2 channels of uniform(-1,1) distributed sources (sub-Gaussian)

(3) 2 channels of permuted human speech signals (super-Gaussian)

(4) The 2 channels of beta(0.5,0.5) distributed sources in (1) with outliers (super-Gaussian)

(5) The 2 channels of uniformly distributed sources in (2) with outliers (super-Gaussian)

(6) The 2 channels of permuted human speech signals in (3) with outliers (super-Gaussian)

The pdf of the sources are shown by the histograms of the data sets (Figure 9.1) and the statistics of the data set are shown in Table 9.1. It should be noted that the kurtosis of the sources become higher (more sharply peaked) after adding the outsiders. The sub-Gaussian beta distributed and uniformly distributed sources become super-Gaussian after adding the outliers.

In the experiments, the learning rate was fixed at 0.0001. $\beta = 0.5$ was used. 100,000

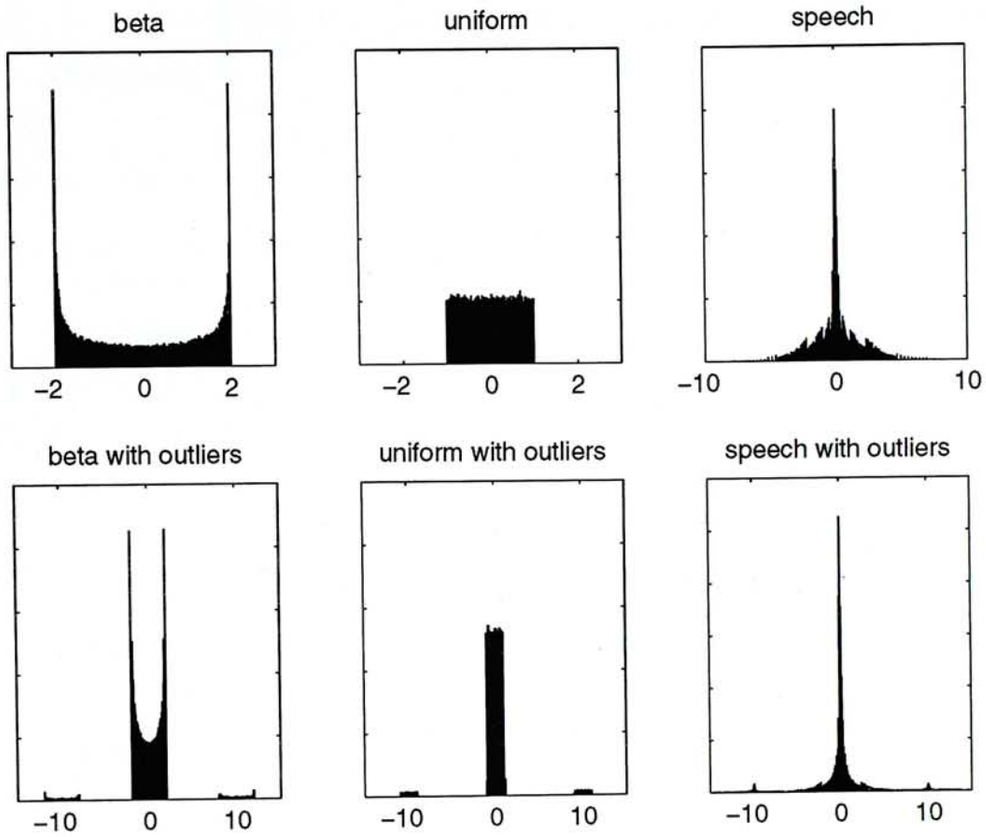


Figure 9.1: Histograms showing the pdf the six sets of sources. (They are of different scales and only the shapes are of reference.)

	without outliers			with outliers		
	beta	uniform	speech	beta	uniform	speech
$E[s_1^2]$	2.003	0.3326	1.000	6.877	5.202	5.872
$E[s_1^4]$	6.001	0.1991	11.129	551.9	496.8	530.0
$E[s_1^4]/E^2[s_1^2]-3$	−1.5042	−1.2002	8.129	8.6698	15.358	12.371
$E[s_2^2]$	1.996	0.3337	4.000	6.850	5.2032	8.810
$E[s_2^4]$	5.9846	0.2007	85.56	547.0	496.5	673.7
$E[s_2^4]/(E^2[s_2^2])-3$	−1.4978	−1.1977	2.3475	8.6575	15.3391	5.6799

Table 9.1: The statistics of the 6 sets of sources used in the experiment.

	without outliers			with outliers		
	beta	uniform	speech	beta	uniform	speech
Kullback divergence, $h_i(y_i) = -y_i^3$	ok	ok	fail	fail	fail	fail
PC divergence, $h_i(y_i) = -y_i^3$	ok	ok	fail	ok	ok	fail
Kullback divergence, $h_i(y_i) = 1 - 2\,\mathrm{logsig}(y_i)$	fail	fail	ok	ok	ok	ok
PC divergence, $h_i(y_i) = 1 - 2\,\mathrm{logsig}(y_i)$	fail	fail	ok	fail	fail	ok

Table 9.2: The results of the experiments.

data points are used in each channel of the data sets. The mixing matrix used is:

$$\mathbf{A} = \begin{bmatrix} 1 & 0.6 \\ 0.7 & 1 \end{bmatrix} \tag{9.26}$$

The cubic nonlinearity and the reversed sigmoid $h_i(y_i) = 1 - 2\,\mathrm{logsig}(y_i) = (\exp(-y_i) - 1)/(1 + \exp(-y_i))$ are used in the experiments. The results of the experiments are summarized in Table 9.2. For the information-theoretic ICA algorithm, cubic nonlinearity perform separation on sub-Gaussian sources and not on super-Gaussian sources, consistent to the theoretical proof in Chapter 6, and the reversed-sigmoid nonlinearity performed separation on super-Gaussian sources and not on sub-Gaussian sources as expected [5, 6].

The ICA algorithm based on PC divergence behaves like the information-theoretic algorithm for the sources without outliers. However, for the sources with outliers, the cubic nonlinearity can separate sources with the central patch of the pdf being flat although the overall kurtosis of the pdf is positive. The reversed sigmoid nonlinearity cannot separate sources with the central patch of the pdf being flat although the overall kurtosis of the pdf is positive. These two fact show that the algorithm based on PC divergence can, to some extent, ignore, or 'filter out', the outliers as expected. This behavior can be regarded as the 'robustness' of the algorithm based on the PC divergence to outliers in the sources. Experiments on $\beta = 0.2$ and $\beta = 0.8$ also yield similar results.

9.2.2 Experiments on the algorithm with mixture of densities

The ICA algorithm based on PC divergence with mixture of densities are tested on the three sets of data used in Section 8.2 - (1) two channels of uniformly distributed sources, (2) two channels of permuted speech signals, and (3) one beta(0.5,0.5) distributed source, one uniformly distributed source, and one permuted speech signals. The learning rates and initial settings are the same as those in Section 8.2. Experiments using $\beta = 0.5$ and $\beta = 0.2$ have been carried out.

All the signals in the three data sets can be separated in the experiments using $\beta = 0.2$. All signals in the first two data sets can be separated in the experiments using $\beta = 0.5$. In the experiment on data set (3), the precision problem remarked in Section 8.1.2 occurs and causes division by zero error in the simulation. Nevertheless, the experimental results are similar to those in Section 8.2 and empirically the ICA algorithm based on PC divergence with mixture of densities can perform separation on sources with any distribution. As the mixture of densities can, as experiments support, adapt source with any density by itself, the algorithm based on PC divergence seems not have casted any advantage to the algorithm based on Kullback divergence.

Chapter 10

Conclusions

The general information-theoretic ICA scheme proposed by Xu & Amari [74] suggests a new view that it is not necessary to approximate the marginal densities of the recovered signals (or sources) to the best but a wide class of nonlinearity can also perform source separation successfully. However, the condition on the choice of nonlinearity is not clear yet. Therefore, we carry out investigation on the information-theoretic ICA scheme [74], focusing on the relation between nonlinearity and separation capability of the algorithms.

Firstly, we give lemmas and corollaries on the properties of the cost function $J(\mathbf{V})$ in the information-theoretic ICA scheme. We have proved that $J(\mathbf{V})$ is continuous anywhere in the $n \times n$ dimensional $\mathbf{V}$-parameter space except on the singular subspace, where $J(\mathbf{V})$ tends to be infinitely large. We have proved that for odd, monotonic decreasing $h_i(y_i)$ functions, including the cubic, cubic root and reversed sigmoid nonlinearity we investigated, $J(\mathbf{V})$ is monotonic increasing along the radially outward direction beyond some finite point, and hence the general gradient descent algorithm will never diverge to infinity. We have also prove that the scale parameters of the nonlinearities can control the magnitude of the recovered signals and have no effect on the separation capability of the nonlinearity. In addition, there is no local maximum of $J(\mathbf{V})$ in the whole $\mathbf{V}$-parameter space. These results holds for the general n-channel case and are common to a large class of nonlinearities. Hence, they help much in further analysis of the information-theoretic ICA scheme.

We performed a detailed analysis of the information-theoretic ICA algorithm with cubic nonlinearity. In the 2-channel case, we have determined there are exactly 16 equilibrium points of the cost function, 8 being correct solutions for source separation and 8 being spurious solutions that cannot perform separation. Then, we investigate the stability of the equilibrium points by inspecting the Hessian matrix. We have found

that the correct solutions are stable if the two sources are globally sub-Gaussian, and are saddle points if the two sources are globally super-Gaussian. Meanwhile, the spurious solutions are found to be saddle points if the two sources are globally sub-Gaussian, and are stable if the two sources are globally super-Gaussian. We have also reached the same conditions on stability by an alternative method that directly compares the value of J and uses the simplicity of the 2-channel case. With the lemmas developed on continuity, singularity and asymptotic behavior of $J(\mathbf{V})$, we conclude the global convergence behavior of the information-theoretic ICA algorithms (general gradient descent algorithms on $J(\mathbf{W})$) in a theorem stating that the information-theoretic ICA algorithms with cubic nonlinearity can separate two globally sub-Gaussian sources but cannot separate two globally super-Gaussian sources. We have performed experiments on different sources and all of the experimental results are consistent to the theorem.

We have also determined two groups of equilibrium points of the information-theoretic ICA algorithm with cubic nonlinearity in the 3-channel case. One group is the correct solutions that all of the three sources are separated and the second group are solutions that only one source is extracted but the other two are still mixed. However, the determination of equilibrium points is not exhaustive. We have determined that the first group of solutions is stable if and only if the three sources are pairwise globally sub-Gaussian. The experimental results on three pairwise globally sub-Gaussian source is consistent to the analysis. Experimental results on other sources show that other solutions exist and the conditions on convergence to them are not so obvious.

Having investigated the cubic nonlinearity in details, we proceeds to investigate the more general relationship between the nonlinearities used in the scheme and the distribution of sources they are capable to separate. It is well known that if every $g_i(y_i)$ is equal to some (unknown) $p_{s_j}(s_j)$, or $p_{y_i}(y_i)$ (strict matching), the global minima of $J(\mathbf{V})$ must be correct solutions for separation. However, fixed nonlinearities can also perform separation on sources of a class of distribution. We compared the nonlinearity and separation capability in several cases. Firstly, from Bell & Sejnowski [5] we know $h_i(y_i)$ being several reversed sigmoids may separate super-Gaussian sources but not sub-Gaussian sources. Then, we experimentally found that the cubic root nonlinearity also may separate super-Gaussian sources but not sub-Gaussian sources. Thirdly, we have theoretically proved in the 2-channel case the cubic nonlinearity can separate two sub-Gaussian sources but cannot separate two super-Gaussian sources. In addition, we have proved in a theorem that if the algorithm using a linear $h_1(y_1)$ in one channel and cubic nonlinearity in the second channel is applied on one super-Gaussian source and one sub-Gaussian source, the second channel always recovers the sub-Gaussian source. Comparing the shape and kurtosis of $g_i(y_i)$ corresponding to the $h_i(y_i)$ in the

above cases and the type of source densities they can separate, we found that a fixed nonlinearity can separate sources with densities 'loosely matched' with $g_i(y_i)$ in terms of shape or kurtosis. This suggestion is intuitively sound and supported by theories and experiments we have so far.

We also carried out the implementation of the information-theoretic ICA scheme with mixture of densities, which was proposed by Xu *et al* [75]. In the experiments, we found that the flexible mixture of densities can adapt any source density and perform separation in all the experiments we tried. We found that the flexibility of the mixture of densities can be reduced and experimentally found that the mixture of two densities with only centers of components changeable can still perform separation on all the sources we tried.

Finally, we investigate the possibility of constructing adaptive ICA algorithms from the Bayesian YING YANG learning theory using non-Kullback separation functionals. We tried the Positive Convex divergence, L_p divergence and de-correlation index, and could only derive adaptive ICA algorithm based on the Positive Convex divergence. The resulting algorithm bear one more 'bell-shaped' factor on y_i, compared with the information-theoretic ICA algorithm and Xu [73] proposed that it would filter out (ignore to some extent) the large magnitude outliers in the source. The experiments we carried out verified this proposal.

There are still a number of open questions regarding the information-theoretic ICA scheme. For example, it is difficult to scale up the analysis method we used in analyzing cubic nonlinearity to the general n-channel case. The information-theoretic ICA algorithm with reversed sigmoid nonlinearity or cubic root nonlinearity is difficult to analyze since the nonlinear functions cannot be expanded in finite terms by Taylor series and the equilibrium points are difficult to determine. The convergence of the algorithm with mixture of densities to correct solution have not been theoretically proved yet. These questions worth further investigation.

Appendix A

Proof for Stability of the Equilibrium Points of the Algorithm with Cubic Nonlinearity on Two Channels of Signals

A.1 Stability of Solution Group A

This part is to prove that Solutions A1 - A8 eq. (6.15) and (6.16) are local minima of $J(\mathbf{V})$ of the information-theoretic ICA algorithms with cubic nonlinearity for two source signals that satisfy

$$\mu_1^4\mu_2^4 - [3(\mu_1^2)^2][3(\mu_2^2)^2] < 0 \tag{A.1}$$

and are saddle points for two source signals that satisfy

$$\mu_1^4\mu_2^4 - [3(\mu_1^2)^2][3(\mu_2^2)^2] > 0 \tag{A.2}$$

where $\mu_i^p = E[s_i^p]$.

Proof

For the cubic nonlinearity, $h_i'(y_i) = -3c_i y_i^2$. Substituting $h_i'(y_i)$ and $\mathbf{V}_{\text{A1-A4}}$ into eq. (5.17) of Lemma 5, we get the sufficient and necessary condition for Solution A1-A4 to be stable (local minima)

is:

$$\begin{aligned}\mu_1^2\mu_2^2 E_{s_1}[3c_1(c_1\mu_1^4)^{-1/2}s_1^2]E_{s_1}[3c_1(c_2\mu_2^4)^{-1/2}s_2^2] - (c_1\mu_1^4)^{1/2}(c_2\mu_2^4)^{1/2} \\ = (c_1c_2)^{1/2}(\mu_1^4)^{-1/2}(\mu_2^4)^{-1/2}\left[9(\mu_1^2)^2(\mu_2^2)^2 - \mu_1^4\mu_2^4\right] \\ > 0\end{aligned} \tag{A.3}$$

Hence, if eq. (A.1) is satisfied, then Solution A1-A4 are local minima. Otherwise, since there is no local maximum by Lemma 9, Solution A1-A4 would be saddle points. The case for Solution A5-A8 is similarly proved. □

A.2 Stability of Solution Group B

This part is to prove that Solution B1 - B8 eq. (6.22) are saddle points of $J(\mathbf{V})$ of the information-theoretic ICA algorithms with cubic nonlinearity for two source signals that satisfy

$$\mu_1^4\mu_2^4 - [3(\mu_1^2)^2][3(\mu_2^2)^2] < 0 \tag{A.4}$$

and are local minima for two source signals that satisfy

$$\mu_1^4\mu_2^4 - [3(\mu_1^2)^2][3(\mu_2^2)^2] > 0 \tag{A.5}$$

Proof The Hessian matrix $\nabla_V^2 J(\mathbf{V})$ is

$$\begin{aligned}\nabla_V^2 J(\mathbf{V}) = \mathbf{Q} = [q_{ij}] &= \begin{bmatrix} \frac{\partial^2 J(\mathbf{V})}{\partial v_{11}\partial v_{11}} & \frac{\partial^2 J(\mathbf{V})}{\partial v_{11}\partial v_{12}} & \frac{\partial^2 J(\mathbf{V})}{\partial v_{11}\partial v_{21}} & \frac{\partial^2 J(\mathbf{V})}{\partial v_{11}\partial v_{22}} \\ \frac{\partial^2 J(\mathbf{V})}{\partial v_{12}\partial v_{11}} & \frac{\partial^2 J(\mathbf{V})}{\partial v_{12}\partial v_{12}} & \frac{\partial^2 J(\mathbf{V})}{\partial v_{12}\partial v_{21}} & \frac{\partial^2 J(\mathbf{V})}{\partial v_{12}\partial v_{22}} \\ \frac{\partial^2 J(\mathbf{V})}{\partial v_{21}\partial v_{11}} & \frac{\partial^2 J(\mathbf{V})}{\partial v_{21}\partial v_{12}} & \frac{\partial^2 J(\mathbf{V})}{\partial v_{21}\partial v_{21}} & \frac{\partial^2 J(\mathbf{V})}{\partial v_{21}\partial v_{22}} \\ \frac{\partial^2 J(\mathbf{V})}{\partial v_{22}\partial v_{11}} & \frac{\partial^2 J(\mathbf{V})}{\partial v_{22}\partial v_{12}} & \frac{\partial^2 J(\mathbf{V})}{\partial v_{22}\partial v_{21}} & \frac{\partial^2 J(\mathbf{V})}{\partial v_{22}\partial v_{22}} \end{bmatrix} \\ &= \begin{bmatrix} \frac{v_{22}^2}{D^2} - E_{111} & -\frac{v_{21}v_{22}}{D^2} - E_{112} & -\frac{v_{12}v_{22}}{D^2} & \frac{v_{12}v_{21}}{D^2} \\ -\frac{v_{21}v_{22}}{D^2} - E_{112} & \frac{v_{21}^2}{D^2} - E_{122} & \frac{v_{11}v_{22}}{D^2} & -\frac{v_{11}v_{21}}{D^2} \\ -\frac{v_{12}v_{22}}{D^2} & \frac{v_{11}v_{22}}{D^2} & \frac{v_{12}^2}{D^2} - E_{211} & -\frac{v_{11}v_{12}}{D^2} - E_{212} \\ \frac{v_{12}v_{21}}{D^2} & -\frac{v_{11}v_{21}}{D^2} & -\frac{v_{11}v_{12}}{D^2} - E_{212} & \frac{v_{11}^2}{D^2} - E_{222} \end{bmatrix}\end{aligned} \tag{A.6}$$

where

$$D = \det \mathbf{V} \tag{A.7}$$

$$E_{ijk} = E[h_i'(\mathbf{v}_i^T\mathbf{s})s_j s_k] \qquad \text{and} \qquad h_i'(y_i) = -3c_i y_i^2 \tag{A.8}$$

For Solution B1 - B8,

$$(\det \mathbf{V})^2 = D^2 = [(2c_1\eta_1)^{-\frac{1}{4}}(2c_2\eta_2)^{-\frac{1}{4}} + (2c_1\eta_2)^{-\frac{1}{4}}(2c_2\eta_1)^{-\frac{1}{4}}]^2 = \frac{2}{c\eta} \tag{A.9}$$

where

$$\eta = \sqrt{\eta_1 \eta_2} = \sqrt{\mu_1^4 \mu_2^4} + m > 0$$
$$c = \sqrt{c_1 c_2} > 0 \tag{A.10}$$

and from eq. (A.6), the Hessian matrix is

$$\nabla^2_{\mathbf{V}} J(\mathbf{V}) = \mathbf{Q} = \frac{1}{2\sqrt{2}} \times$$
$$\begin{bmatrix} \sqrt{c_1}p_1 & -s_{21}s_{22}\sqrt{c_1}(\sqrt{\eta} + \frac{4m}{\sqrt{\eta}}) & -s_{12}s_{22}\sqrt{c}\sqrt{\eta_1} & s_{12}s_{21}\sqrt{c}\sqrt{\eta} \\ -s_{21}s_{22}\sqrt{c_1}(\sqrt{\eta} + \frac{4m}{\sqrt{\eta}}) & \sqrt{c_1}p_2 & s_{11}s_{22}\sqrt{c}\sqrt{\eta} & -s_{11}s_{21}\sqrt{c}\sqrt{\eta_2} \\ -s_{12}s_{22}\sqrt{c}\sqrt{\eta_1} & s_{11}s_{22}\sqrt{c}\sqrt{\eta} & \sqrt{c_2}p_1 & -s_{11}s_{12}\sqrt{c_2}(\sqrt{\eta} + \frac{4m}{\sqrt{\eta}}) \\ s_{12}s_{21}\sqrt{c}\sqrt{\eta} & -s_{11}s_{21}\sqrt{c}\sqrt{\eta_2} & -s_{11}s_{12}\sqrt{c_2}(\sqrt{\eta} + \frac{4m}{\sqrt{\eta}}) & \sqrt{c_2}p_2 \end{bmatrix} \tag{A.11}$$

where

$$p_1 = \sqrt{\eta_1} + \frac{6\mu_1^4}{\sqrt{\eta_1}} + \frac{2m}{\sqrt{\eta_2}}$$
$$p_2 = \sqrt{\eta_2} + \frac{6\mu_2^4}{\sqrt{\eta_2}} + \frac{2m}{\sqrt{\eta_1}} \tag{A.12}$$

The characteristic equation is:

$$
\begin{aligned}
0 = \det(\mathbf{Q} - \lambda\mathbf{I}) &= (q_{11}-\lambda)(q_{22}-\lambda)(q_{33}-\lambda)(q_{44}-\lambda) \\
&\quad - (q_{11}-\lambda)(q_{22}-\lambda)q_{43}q_{34} - (q_{33}-\lambda)(q_{44}-\lambda)q_{21}q_{12} \\
&\quad - (q_{11}-\lambda)(q_{44}-\lambda)q_{32}q_{23} - (q_{22}-\lambda)(q_{33}-\lambda)q_{41}q_{14} \\
&\quad - (q_{11}-\lambda)(q_{33}-\lambda)q_{42}q_{24} - (q_{22}-\lambda)(q_{44}-\lambda)q_{31}q_{13} \\
&\quad + (q_{11}-\lambda)(q_{32}q_{43}q_{24} + q_{42}q_{23}q_{34}) + (q_{22}-\lambda)(q_{31}q_{43}q_{14} + q_{41}q_{13}q_{34}) \\
&\quad + (q_{33}-\lambda)(q_{21}q_{42}q_{14} + q_{41}q_{12}q_{24}) + (q_{44}-\lambda)(q_{21}q_{32}q_{13} + q_{31}q_{12}q_{23}) \\
&\quad + q_{31}q_{42}q_{13}q_{24} + q_{41}q_{32}q_{23}q_{14} - q_{31}q_{42}q_{23}q_{14} - q_{41}q_{32}q_{13}q_{24} \\
&\quad - q_{21}q_{32}q_{43}q_{14} - q_{21}q_{42}q_{13}q_{34} - q_{31}q_{12}q_{43}q_{24} - q_{41}q_{12}q_{23}q_{34} \\
&\quad + q_{21}q_{12}q_{43}q_{34} \\
&= \lambda^4 - \lambda^3(q_{11}+q_{22}+q_{33}+q_{44}) \\
&\quad + \lambda^2\{(q_{11}q_{22}+q_{33}q_{44}+q_{11}q_{44}+q_{22}q_{33}+q_{11}q_{33}+q_{22}q_{44}) \\
&\qquad - (q_{43}q_{34}+q_{21}q_{12}+q_{32}q_{23}+q_{41}q_{14}+q_{42}q_{24}+q_{31}q_{13})\} \\
&\quad + \lambda\{-(q_{11}q_{22}q_{33}+q_{11}q_{22}q_{44}+q_{11}q_{33}q_{44}+q_{22}q_{33}q_{44}) \\
&\qquad + (q_{11}+q_{22})q_{43}q_{34} + (q_{33}+q_{44})q_{21}q_{12} + (q_{11}+q_{44})q_{32}q_{23} \\
&\qquad + (q_{22}+q_{33})q_{41}q_{14} + (q_{11}+q_{33})q_{42}q_{24} + (q_{22}+q_{44})q_{31}q_{13} \\
&\qquad - [q_{32}q_{43}q_{24} + q_{42}q_{23}q_{34} + q_{31}q_{43}q_{14} + q_{41}q_{13}q_{34} \\
&\qquad\quad + q_{21}q_{42}q_{14} + q_{41}q_{12}q_{24} + q_{21}q_{32}q_{13} + q_{31}q_{12}q_{23}]\} \\
&\quad + \{q_{11}q_{22}q_{33}q_{44} \\
&\qquad - (q_{11}q_{22}q_{43}q_{34} + q_{33}q_{44}q_{21}q_{12} + q_{11}q_{44}q_{32}q_{23} \\
&\qquad\quad + q_{22}q_{33}q_{41}q_{14} + q_{11}q_{33}q_{42}q_{24} + q_{22}q_{44}q_{31}q_{13}) \\
&\qquad + [q_{11}(q_{32}q_{43}q_{24} + q_{42}q_{23}q_{34}) + q_{22}(q_{31}q_{43}q_{14} + q_{41}q_{13}q_{34}) \\
&\qquad\quad + q_{33}(q_{21}q_{42}q_{14} + q_{41}q_{12}q_{24}) + q_{44}(q_{21}q_{32}q_{13} + q_{31}q_{12}q_{23})] \\
&\qquad + [q_{31}q_{42}q_{13}q_{24} + q_{41}q_{32}q_{23}q_{14} - q_{31}q_{42}q_{23}q_{14} - q_{41}q_{32}q_{13}q_{24} \\
&\qquad\quad - q_{21}q_{32}q_{43}q_{14} - q_{21}q_{42}q_{13}q_{34} - q_{31}q_{12}q_{43}q_{24} - q_{41}q_{12}q_{23}q_{34} \\
&\qquad\quad + q_{21}q_{12}q_{43}q_{34}]\}
\end{aligned}
\tag{A.13}
$$

The listed terms are simplified as follows:

$$
q_{11}+q_{22}+q_{33}+q_{44} = -\frac{1}{2\sqrt{2}}(\sqrt{c_1}+\sqrt{c_2})(p_1+p_2) \tag{A.14}
$$

$$
\begin{aligned}
&q_{11}q_{22}+q_{33}q_{44}+q_{11}q_{44}+q_{22}q_{33}+q_{11}q_{33}+q_{22}q_{44} \\
&= \frac{1}{8}\{c(p_1+p_2)^2 + (c_1+c_2)p_1p_2\}
\end{aligned}
\tag{A.15}
$$

$$
\begin{aligned}
&q_{43}q_{34}+q_{21}q_{12}+q_{32}q_{23}+q_{41}q_{14}+q_{42}q_{24}+q_{31}q_{13} \\
&= \frac{1}{8}\{c(\eta_1+\eta_2+2\eta) + (c_1+c_2)(\sqrt{\eta}+\frac{4m}{\sqrt{\eta}})^2\}
\end{aligned}
\tag{A.16}
$$

$$
\begin{aligned}
&q_{11}q_{22}q_{33}+q_{11}q_{22}q_{44}+q_{11}q_{33}q_{44}+q_{22}q_{33}q_{44} \\
&= \frac{1}{16\sqrt{2}}c(\sqrt{c_1}+\sqrt{c_2})p_1p_2(p_1+p_2)
\end{aligned}
\tag{A.17}
$$

$$(q_{11}+q_{22})q_{43}q_{34}+(q_{33}+q_{44})q_{21}q_{12}$$
$$=\frac{1}{16\sqrt{2}}c(\sqrt{c_1}+\sqrt{c_2})(\sqrt{\eta}+\frac{4m}{\sqrt{\eta}})^2(p_1+p_2) \tag{A.18}$$

$$(q_{11}+q_{44})q_{32}q_{23}+(q_{22}+q_{33})q_{41}q_{14}$$
$$=\frac{1}{16\sqrt{2}}c(\sqrt{c_1}+\sqrt{c_2})\eta(p_1+p_2) \tag{A.19}$$

$$(q_{11}+q_{33})q_{42}q_{24}+(q_{22}+q_{44})q_{31}q_{13}=\frac{1}{16\sqrt{2}}c(\sqrt{c_1}+\sqrt{c_2})(\eta_1p_2+\eta_2p_1) \tag{A.20}$$

$$q_{32}q_{43}q_{24}+q_{42}q_{23}q_{34}+q_{31}q_{43}q_{14}+q_{41}q_{13}q_{34}$$
$$+q_{21}q_{42}q_{14}+q_{41}q_{12}q_{24}+q_{21}q_{32}q_{13}+q_{31}q_{12}q_{23} \tag{A.21}$$
$$=-\frac{2}{16\sqrt{2}}c(\sqrt{c_1}+\sqrt{c_2})(\sqrt{\eta_1}+\sqrt{\eta_2})(\eta+4m)$$

$$q_{11}q_{22}q_{33}q_{44}=\frac{1}{64}c^2(p_1p_2)^2 \tag{A.22}$$

$$q_{11}q_{22}q_{43}q_{34}+q_{33}q_{44}q_{21}q_{12}=\frac{1}{64}c^2(\sqrt{\eta}+\frac{4m}{\sqrt{\eta}})^2(2p_1p_2) \tag{A.23}$$

$$q_{11}q_{44}q_{32}q_{23}+q_{22}q_{33}q_{41}q_{14}=\frac{1}{64}c^2\eta(2p_1p_2) \tag{A.24}$$

$$q_{11}q_{33}q_{42}q_{24}+q_{22}q_{44}q_{31}q_{13}=\frac{1}{64}c^2(\eta_1p_2^2+\eta_2p_1^2) \tag{A.25}$$

$$q_{11}(q_{32}q_{43}q_{24}+q_{42}q_{23}q_{34})+q_{22}(q_{31}q_{43}q_{14}+q_{41}q_{13}q_{34})$$
$$+q_{33}(q_{21}q_{42}q_{14}+q_{41}q_{12}q_{24})+q_{44}(q_{21}q_{32}q_{13}+q_{31}q_{12}q_{23}) \tag{A.26}$$
$$=-\frac{4}{64}c^2(\eta+4m)(\sqrt{\eta_1}p_2+\sqrt{\eta_2}p_1)$$

$$q_{21}q_{12}q_{43}q_{34}=\frac{1}{64}c^2(\sqrt{\eta}+\frac{4m}{\sqrt{\eta}})^4 \tag{A.27}$$

$$q_{31}q_{42}q_{13}q_{24}+q_{41}q_{32}q_{23}q_{14}-q_{31}q_{42}q_{23}q_{14}-q_{41}q_{32}q_{13}q_{24}=0 \tag{A.28}$$

$$-q_{21}q_{32}q_{43}q_{14}-q_{21}q_{42}q_{13}q_{34}-q_{31}q_{12}q_{43}q_{24}-q_{41}q_{12}q_{23}q_{34}=-\frac{4}{64}(\eta+4m)^2 \tag{A.29}$$

Using the following identities,

$$\sqrt{\frac{\eta_1}{\eta_2}}=\sqrt{\frac{\mu_1^4}{\mu_2^4}} \tag{A.30}$$

$$\frac{\mu_1^4}{\eta_1} + \frac{\mu_2^4}{\eta_2} = \frac{2\sqrt{\mu_1^4\mu_2^4}}{\eta} \tag{A.31}$$

$$\eta_1 + \eta_2 = \frac{\eta}{\eta - m}(\mu_1^4 + \mu_2^4) \tag{A.32}$$

$$\sqrt{\eta_1} + \sqrt{\eta_2} = \sqrt{\frac{\eta}{\eta - m}}\left(\sqrt{\mu_1^4} + \sqrt{\mu_2^4}\right) \tag{A.33}$$

$$\frac{1}{\sqrt{\eta_1}} + \frac{1}{\sqrt{\eta_2}} = \frac{\sqrt{\mu_1^4} + \sqrt{\mu_2^4}}{\sqrt{\eta(\eta - m)}} \tag{A.34}$$

$$\frac{\mu_1^4}{\sqrt{\eta_1}} + \frac{\mu_2^4}{\sqrt{\eta_2}} = \sqrt{\frac{\eta - m}{\eta}}\left(\sqrt{\mu_1^4} + \sqrt{\mu_2^4}\right) \tag{A.35}$$

we can simplify p_1p_2 and $(p_1 + p_2)$ as:

$$\begin{aligned} p_1p_2 &= \left(\sqrt{\eta_1} + \frac{6\mu_1^4}{\sqrt{\eta_1}} + \frac{2m}{\sqrt{\eta_2}}\right)\left(\sqrt{\eta_2} + \frac{6\mu_2^4}{\sqrt{\eta_2}} + \frac{2m}{\sqrt{\eta_1}}\right) \\ &= \eta + 2\left(6\sqrt{\mu_1^4\mu_2^4} + 2m\right) + 12m\left(\frac{\mu_1^4}{\eta_1} + \frac{\mu_2^4}{\eta_2}\right) + \frac{36\mu_1^4\mu_2^4}{\eta} + \frac{4m^2}{\eta} \\ &= 13\sqrt{\mu_1^4\mu_2^4} + 5m + 24m\frac{\sqrt{\mu_1^4\mu_2^4}}{\eta} + \frac{36\mu_1^4\mu_2^4}{\eta} + \frac{4m^2}{\eta} \\ &= \frac{1}{\eta}\left(49\mu_1^4\mu_2^4 + 42m\sqrt{\mu_1^4\mu_2^4} + 9m^2\right) \\ &= \frac{1}{\eta}\left(7\sqrt{\mu_1^4\mu_2^4} + 3m\right)^2 \\ &= \frac{1}{\eta}(7\eta - 4m)^2 \end{aligned} \tag{A.36}$$

$$\begin{aligned} p_1 + p_2 &= \sqrt{\eta_1} + \sqrt{\eta_2} + 6\left(\frac{\mu_1^4}{\sqrt{\eta_1}} + \frac{\mu_2^4}{\sqrt{\eta_2}}\right) + 2m\left(\frac{1}{\sqrt{\eta_1}} + \frac{1}{\sqrt{\eta_2}}\right) \\ &= \left(\sqrt{\frac{\eta}{\eta - m}} + 6\sqrt{\frac{\eta - m}{\eta}} + \frac{2m}{\sqrt{\eta(\eta - m)}}\right)\left(\sqrt{\mu_1^4} + \sqrt{\mu_2^4}\right) \\ &= \frac{7\eta - 4m}{\sqrt{\eta(\eta - m)}}\left(\sqrt{\mu_1^4} + \sqrt{\mu_2^4}\right) \end{aligned} \tag{A.37}$$

Also,

$$\sqrt{\eta_1}p_2 = \sqrt{\eta_2}p_1 = 7\eta - 4m \tag{A.38}$$

$$\eta_1p_2 + \eta_2p_1 = (7\eta - 4m)\left(\sqrt{\eta_1} + \sqrt{\eta_2}\right) = (7\eta - 4m)\sqrt{\frac{\eta}{\eta - m}}\left(\sqrt{\mu_1^4} + \sqrt{\mu_2^4}\right) \tag{A.39}$$

Now the coefficients of the characteristic polynomial

$$\det(\mathbf{Q} - \lambda\mathbf{I}) = \lambda^4 + A_3\lambda^3 + A_2\lambda^2 + A_1\lambda + A_0 \tag{A.40}$$

can be written as:

$$\begin{aligned}
A_3 &= -\frac{1}{2\sqrt{2}}(\sqrt{c_1} + \sqrt{c_2})(p_1 + p_2) \\
&= -\frac{1}{2\sqrt{2}}(\sqrt{c_1} + \sqrt{c_2})\frac{7\eta - 4m}{\sqrt{\eta(\eta - m)}}\left(\sqrt{\mu_1^4} + \sqrt{\mu_2^4}\right) \\
&= -\frac{1}{2\sqrt{2}}(\sqrt{c_1} + \sqrt{c_2})\frac{7\sqrt{\mu_1^4\mu_2^4} + 3m}{\sqrt{\eta(\eta - m)}}\left(\sqrt{\mu_1^4} + \sqrt{\mu_2^4}\right)
\end{aligned} \tag{A.41}$$

$$\begin{aligned}
A_2 &= \frac{1}{8}\left\{c\left[(p_1 + p_2)^2 - (\eta_1 + \eta_2) - 2\eta\right] + (c_1 + c_2)\left[p_1p_2 - \left(\sqrt{\eta} + \frac{4m}{\sqrt{\eta}}\right)^2\right]\right\} \\
&= \frac{1}{8}\left\{c\left[\frac{(7\eta - 4m)^2}{\eta(\eta - m)}\left(\sqrt{\mu_1^4} + \sqrt{\mu_2^4}\right)^2 - \frac{\eta}{\eta - m}(\mu_1^4 + \mu_2^4) - \frac{2\eta}{\eta - m}\sqrt{\mu_1^4\mu_2^4}\right]\right. \\
&\quad \left. +(c_1 + c_2)\left[\frac{1}{\eta}(7\eta - 4m)^2 - \frac{1}{\eta}(\eta + 4m)^2\right]\right\} \\
&= \frac{1}{8}\left\{c\frac{(7\eta - 4m)^2 - \eta^2}{\eta(\eta - m)}\left(\sqrt{\mu_1^4} + \sqrt{\mu_2^4}\right)^2 + (c_1 + c_2)\frac{(6\eta - 8m)(8\eta)}{\eta}\right\} \\
&= c\frac{(2\eta - m)(3\eta - 2m)}{\eta(\eta - m)}\left(\sqrt{\mu_1^4} + \sqrt{\mu_2^4}\right)^2 + 2(c_1 + c_2)(3\eta - 4m) \\
&= c\frac{(2\sqrt{\mu_1^4\mu_2^4} + m)(3\sqrt{\mu_1^4\mu_2^4} + m)}{\eta(\eta - m)}\left(\sqrt{\mu_1^4} + \sqrt{\mu_2^4}\right)^2 \\
&\quad + 2(c_1 + c_2)\left(3\sqrt{\mu_1^4\mu_2^4} - m\right)
\end{aligned} \tag{A.42}$$

$$
\begin{aligned}
A_1 &= -\frac{1}{16\sqrt{2}}c(\sqrt{c_1}+\sqrt{c_2})\left\{p_1p_2(p_1+p_2)-\left[\left(\sqrt{\eta}+\frac{4m}{\sqrt{\eta}}\right)^2+n\right](p_1+p_2)\right.\\
&\quad \left.-(\eta_1p_2+\eta_2p_1)-2(\sqrt{\eta_1}+\sqrt{\eta_2})(\eta+4m)\right\}\\
&= -\frac{1}{16\sqrt{2}}c(\sqrt{c_1}+\sqrt{c_2})\times\\
&\quad \left\{\left[\frac{1}{\eta}(7\eta-4m)^2-\frac{1}{\eta}(\eta+4m)^2-\eta\right]\frac{7\eta-4m}{\sqrt{\eta(\eta-m)}}\left(\sqrt{\mu_1^4}+\sqrt{\mu_2^4}\right)\right.\\
&\quad \left.-\frac{(7\eta-4m)\eta}{\sqrt{\eta(\eta-m)}}\left(\sqrt{\mu_1^4}+\sqrt{\mu_2^4}\right)-2\frac{(\eta+4m)\eta}{\sqrt{\eta(\eta-m)}}\left(\sqrt{\mu_1^4}+\sqrt{\mu_2^4}\right)\right\}\\
&= -\frac{1}{16\sqrt{2}}c(\sqrt{c_1}+\sqrt{c_2})\times\\
&\quad \frac{[8(6\eta-8m)-\eta](7\eta-4m)-(9\eta+4m)\eta}{\sqrt{\eta(\eta-m)}}\left(\sqrt{\mu_1^4}+\sqrt{\mu_2^4}\right)\\
&= -\frac{1}{16\sqrt{2}}c(\sqrt{c_1}+\sqrt{c_2})\frac{320\eta^2-640m\eta+256m^2}{\sqrt{\eta(\eta-m)}}\left(\sqrt{\mu_1^4}+\sqrt{\mu_2^4}\right)\\
&= -2\sqrt{2}c(\sqrt{c_1}+\sqrt{c_2})\frac{5\eta^2-10m\eta+4m^2}{\sqrt{\eta(\eta-m)}}\left(\sqrt{\mu_1^4}+\sqrt{\mu_2^4}\right)\\
&= -2\sqrt{2}c(\sqrt{c_1}+\sqrt{c_2})\frac{5\mu_1^4\mu_2^4-m^2}{\sqrt{\eta(\eta-m)}}\left(\sqrt{\mu_1^4}+\sqrt{\mu_2^4}\right)
\end{aligned}
\tag{A.43}
$$

$$
\begin{aligned}
A_0 &= \frac{1}{64}c^2\left\{(p_1p_2)^2-2\left[\left(\sqrt{\eta}+\frac{4m}{\sqrt{\eta}}\right)^2+n\right](p_1p_2)-(\eta_1p_2^2+\eta_2p_1^2)\right.\\
&\quad \left.-4(\eta+4m)\left(\sqrt{\eta_2}p_1+\sqrt{\eta_1}p_2\right)+\left(\sqrt{\eta}+\frac{4m}{\sqrt{\eta}}\right)^4-4(\eta+4m)^2\right\}\\
&= \frac{1}{64}c^2\left\{\frac{1}{\eta^2}(7\eta-4m)^4-\frac{2}{\eta^2}\left[(\eta+4m)^2+\eta^2\right](7\eta-4m)^2-2(7\eta-4m)^2\right.\\
&\quad \left.-4(\eta+4m)[2(7\eta-4m)+(\eta+4m)]+\frac{1}{\eta^2}(\eta+4m)^4\right\}\\
&= \frac{1}{64}c^2\left\{\frac{1}{\eta^2}\left[\left((49\eta^2-56m\eta+16m^2)-2(2\eta^2+8m\eta+16)\right)(7\eta-4m)^2\right.\right.\\
&\quad \left.\left.+(\eta^2+8m\eta+16m^2)^2\right]-2(7\eta-4m)^2-4(\eta+4m)(15\eta-4m)\right\}\\
&= \frac{1}{64}c^2\left[\frac{1}{\eta^2}\left(2206\eta^4-6032\eta^3m+4064\eta^2m^2\right)-158\eta^2-112m\eta+32m^2\right]\\
&= \frac{1}{64}c^2(2048\eta^2-6144m\eta+4096m^2)\\
&= 32c^2(\eta-2m)(\eta-m)\\
&= 32c^2\sqrt{\mu_1^4\mu_1^4}\left(\sqrt{\mu_1^4\mu_1^4}-m\right)
\end{aligned}
\tag{A.44}
$$

Or, putting together, the characteristic polynomial is:

$$\begin{aligned}\det(\mathbf{Q}-\lambda\mathbf{I}) = \lambda^4 \\ &+\lambda^3\left\{-\frac{1}{2\sqrt{2}}(\sqrt{c_1}+\sqrt{c_2})\frac{7\sqrt{\mu_1^4\mu_2^4}+3m}{\sqrt{\eta(\eta-m)}}\left(\sqrt{\mu_1^4}+\sqrt{\mu_2^4}\right)\right\} \\ &+\lambda^2\left\{c\frac{(2\sqrt{\mu_1^4\mu_2^4}+m)(3\sqrt{\mu_1^4\mu_2^4}+m)}{\eta(\eta-m)}\left(\sqrt{\mu_1^4}+\sqrt{\mu_2^4}\right)^2\right. \\ &\qquad \left.+2(c_1+c_2)\left(3\sqrt{\mu_1^4\mu_2^4}-m\right)\right\} \\ &+\lambda\left\{-2\sqrt{2}c(\sqrt{c_1}+\sqrt{c_2})\frac{5\mu_1^4\mu_2^4-m^2}{\sqrt{\eta(\eta-m)}}\left(\sqrt{\mu_1^4}+\sqrt{\mu_2^4}\right)\right\} \\ &+32c^2\sqrt{\mu_1^4\mu_1^4}\left(\sqrt{\mu_1^4\mu_1^4}-m\right)\end{aligned} \tag{A.45}$$

The eigenvalues are extremely difficult to solve out. However, the result can be proved without explicitly solving out the eigenvalues. This is done by inspecting the signs of the coefficients of the characteristic polynomial as follows.

As the Hessian matrix is symmetrical, all of the eigenvalues must be real. For super-Gaussian in-average signals, $\mu_1^4\mu_2^4 - m^2 > 0$, we have $A_3 < 0$, $A_2 > 0$, $A_1 < 0$ and $A_0 > 0$. Hence, every eigenvalue, being the root of the characteristic polynomial, must be positive. The Hessian matrix will be positive definite and therefore Solutions B1 - B8 are local minima.

For sub-Gaussian in-average signals, $\mu_1^4\mu_2^4 - m^2 < 0$, we have $A_0 < 0$. Noting that $A_0 = \lambda_1\lambda_2\lambda_3\lambda_4$, where λ_1, λ_2, λ_3, λ_4 are the eigenvalues, we get three eigenvalues must be of the same sign and the remaining eigenvalue must be of the opposite sign. Hence Solutions B1 - B8 are saddle points. □

Appendix B

Proof for Stability of the Equilibrium Points of the Algorithm with Cubic Nonlinearity on Three Channels of Signals

This part is to prove that the solution group P in Section 6.4 is a local minimum of $J(\mathbf{V})$ if

$$\mu_i^4\mu_j^4 - 9(\mu_i^2)^2(\mu_j^2)^2 < 0, \qquad i,j = 1,2,3\ i \neq j \tag{B.1}$$

where $\mu_i^p = E[s_i^p]$, and are saddle points otherwise.

Proof

The Hessian matrix of $J(\mathbf{V})$ is:

$$\nabla_{\mathbf{V}}^2 J(\mathbf{V}) = \mathbf{Q} =$$

$$\begin{bmatrix}
\frac{\bar{v}_{11}^2}{D^2} - E_{111} & \frac{\bar{v}_{11}\bar{v}_{12}}{D^2} - E_{121} & \frac{\bar{v}_{11}\bar{v}_{13}}{D^2} - E_{131} & \frac{\bar{v}_{11}\bar{v}_{21}}{D^2} & \frac{\bar{v}_{11}\bar{v}_{22}}{D^2} - \frac{\bar{v}_{23}}{D} & \frac{\bar{v}_{11}\bar{v}_{23}}{D^2} + \frac{\bar{v}_{22}}{D} & \frac{\bar{v}_{11}\bar{v}_{31}}{D^2} & \frac{\bar{v}_{11}\bar{v}_{32}}{D^2} + \frac{\bar{v}_{23}}{D} & \frac{\bar{v}_{11}\bar{v}_{33}}{D^2} - \frac{\bar{v}_{22}}{D} \\
\frac{\bar{v}_{12}\bar{v}_{11}}{D^2} - E_{121} & \frac{\bar{v}_{12}^2}{D^2} - E_{122} & \frac{\bar{v}_{12}\bar{v}_{13}}{D^2} - E_{132} & \frac{\bar{v}_{12}\bar{v}_{21}}{D^2} + \frac{\bar{v}_{23}}{D} & \frac{\bar{v}_{12}\bar{v}_{22}}{D^2} & \frac{\bar{v}_{12}\bar{v}_{23}}{D^2} - \frac{\bar{v}_{21}}{D} & \frac{\bar{v}_{12}\bar{v}_{31}}{D^2} - \frac{\bar{v}_{23}}{D} & \frac{\bar{v}_{12}\bar{v}_{32}}{D^2} & \frac{\bar{v}_{12}\bar{v}_{33}}{D^2} + \frac{\bar{v}_{21}}{D} \\
\frac{\bar{v}_{13}\bar{v}_{11}}{D^2} - E_{131} & \frac{\bar{v}_{13}\bar{v}_{12}}{D^2} - E_{132} & \frac{\bar{v}_{13}^2}{D^2} - E_{133} & \frac{\bar{v}_{13}\bar{v}_{21}}{D^2} - \frac{\bar{v}_{22}}{D} & \frac{\bar{v}_{13}\bar{v}_{22}}{D^2} + \frac{\bar{v}_{21}}{D} & \frac{\bar{v}_{13}\bar{v}_{23}}{D^2} & \frac{\bar{v}_{13}\bar{v}_{31}}{D^2} + \frac{\bar{v}_{22}}{D} & \frac{\bar{v}_{13}\bar{v}_{32}}{D^2} - \frac{\bar{v}_{21}}{D} & \frac{\bar{v}_{13}\bar{v}_{33}}{D^2} \\
\frac{\bar{v}_{21}\bar{v}_{11}}{D^2} & \frac{\bar{v}_{21}\bar{v}_{12}}{D^2} + \frac{\bar{v}_{23}}{D} & \frac{\bar{v}_{21}\bar{v}_{13}}{D^2} - \frac{\bar{v}_{22}}{D} & \frac{\bar{v}_{21}^2}{D^2} - E_{211} & \frac{\bar{v}_{21}\bar{v}_{22}}{D^2} - E_{221} & \frac{\bar{v}_{21}\bar{v}_{23}}{D^2} - E_{231} & \frac{\bar{v}_{21}\bar{v}_{31}}{D^2} & \frac{\bar{v}_{21}\bar{v}_{32}}{D^2} + \frac{\bar{v}_{13}}{D} & \frac{\bar{v}_{21}\bar{v}_{33}}{D^2} - \frac{\bar{v}_{12}}{D} \\
\frac{\bar{v}_{22}\bar{v}_{11}}{D^2} - \frac{\bar{v}_{23}}{D} & \frac{\bar{v}_{22}\bar{v}_{12}}{D^2} & \frac{\bar{v}_{22}\bar{v}_{13}}{D^2} + \frac{\bar{v}_{21}}{D} & \frac{\bar{v}_{22}\bar{v}_{21}}{D^2} - E_{221} & \frac{\bar{v}_{22}^2}{D^2} - E_{222} & \frac{\bar{v}_{22}\bar{v}_{23}}{D^2} - E_{232} & \frac{\bar{v}_{22}\bar{v}_{31}}{D^2} - \frac{\bar{v}_{13}}{D} & \frac{\bar{v}_{22}\bar{v}_{32}}{D^2} & \frac{\bar{v}_{22}\bar{v}_{33}}{D^2} + \frac{\bar{v}_{11}}{D} \\
\frac{\bar{v}_{23}\bar{v}_{11}}{D^2} + \frac{\bar{v}_{22}}{D} & \frac{\bar{v}_{23}\bar{v}_{12}}{D^2} - \frac{\bar{v}_{21}}{D} & \frac{\bar{v}_{23}\bar{v}_{13}}{D^2} & \frac{\bar{v}_{23}\bar{v}_{21}}{D^2} - E_{231} & \frac{\bar{v}_{23}\bar{v}_{22}}{D^2} + \frac{\bar{v}_{21}}{D} & \frac{\bar{v}_{23}^2}{D^2} - E_{233} & \frac{\bar{v}_{23}\bar{v}_{31}}{D^2} - \frac{\bar{v}_{12}}{D} & \frac{\bar{v}_{23}\bar{v}_{32}}{D^2} - E_{321} & \frac{\bar{v}_{23}\bar{v}_{33}}{D^2} - E_{331} \\
\frac{\bar{v}_{31}\bar{v}_{11}}{D^2} & \frac{\bar{v}_{31}\bar{v}_{12}}{D^2} - \frac{\bar{v}_{33}}{D} & \frac{\bar{v}_{31}\bar{v}_{13}}{D^2} + \frac{\bar{v}_{32}}{D} & \frac{\bar{v}_{31}\bar{v}_{21}}{D^2} - \frac{\bar{v}_{13}}{D} & \frac{\bar{v}_{31}\bar{v}_{22}}{D^2} + \frac{\bar{v}_{13}}{D} & \frac{\bar{v}_{31}\bar{v}_{23}}{D^2} - \frac{\bar{v}_{11}}{D} & \frac{\bar{v}_{31}^2}{D^2} - E_{311} & \frac{\bar{v}_{31}\bar{v}_{32}}{D^2} - E_{321} & \frac{\bar{v}_{31}\bar{v}_{33}}{D^2} - E_{331} \\
\frac{\bar{v}_{32}\bar{v}_{11}}{D^2} + \frac{\bar{v}_{31}}{D} & \frac{\bar{v}_{32}\bar{v}_{12}}{D^2} & \frac{\bar{v}_{32}\bar{v}_{13}}{D^2} - \frac{\bar{v}_{11}}{D} & \frac{\bar{v}_{32}\bar{v}_{21}}{D^2} + \frac{\bar{v}_{12}}{D} & \frac{\bar{v}_{32}\bar{v}_{22}}{D^2} & \frac{\bar{v}_{32}\bar{v}_{23}}{D^2} + \frac{\bar{v}_{11}}{D} & \frac{\bar{v}_{32}\bar{v}_{31}}{D^2} - E_{321} & \frac{\bar{v}_{32}^2}{D^2} - E_{322} & \frac{\bar{v}_{32}\bar{v}_{33}}{D^2} - E_{332} \\
\frac{\bar{v}_{33}\bar{v}_{11}}{D^2} - \frac{\bar{v}_{32}}{D} & \frac{\bar{v}_{33}\bar{v}_{12}}{D^2} + \frac{\bar{v}_{31}}{D} & \frac{\bar{v}_{33}\bar{v}_{13}}{D^2} & \frac{\bar{v}_{33}\bar{v}_{21}}{D^2} + \frac{\bar{v}_{12}}{D} & \frac{\bar{v}_{33}\bar{v}_{22}}{D^2} - \frac{\bar{v}_{11}}{D} & \frac{\bar{v}_{33}\bar{v}_{23}}{D^2} & \frac{\bar{v}_{33}\bar{v}_{31}}{D^2} - E_{331} & \frac{\bar{v}_{33}\bar{v}_{32}}{D^2} - E_{332} & \frac{\bar{v}_{33}^2}{D^2} - E_{333}
\end{bmatrix} \tag{B.2}$$

where $E_{ijk} = E[h_i'(\mathbf{v}_i^T\mathbf{s})s_js_k] = -3c_iE[(\mathbf{v}_i^T\mathbf{s})^2s_js_k]$, $\bar{v}_{ij} = \mathrm{cof}v_{ij}$ and $D = \det\mathbf{V}$.

For Solution P1-P8, the Hessian matrix become:

$$\mathbf{Q} = \begin{bmatrix}
\frac{1}{v_{11}^2} + c_1\mu_1^4 v_{11}^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & c_1 m_{12} v_{11}^2 & 0 & \frac{1}{v_{11}v_{22}} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & c_1 m_{13} v_{11}^2 & 0 & 0 & 0 & \frac{1}{v_{11}v_{33}} & 0 & 0 \\
0 & \frac{1}{v_{11}v_{22}} & 0 & c_2 m_{12} v_{22}^2 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \frac{1}{v_{22}^2} + c_2\mu_2^4 v_{22}^2 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & c_2 m_{23} v_{22}^2 & 0 & \frac{1}{v_{22}v_{33}} & 0 \\
0 & 0 & \frac{1}{v_{11}v_{33}} & 0 & 0 & 0 & c_3 m_{13} v_{33}^2 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & \frac{1}{v_{22}v_{33}} & 0 & c_3 m_{23} v_{33}^2 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{v_{33}^2} + c_3\mu_3^4 v_{33}^2
\end{bmatrix} \tag{B.3}$$

where $m_{ij} = 3\mu_i^2\mu_j^2$.

We prove that the Hessian matrix is positive definite under the stated condition by showing that all leading principal minors are positive (suggested by Jiong Ruan). The leading principal minors of order r is defined as:

$$p_r = \det \begin{bmatrix} q_{11} & \cdots & q_{1r} \\ \vdots & \ddots & \vdots \\ q_{r1} & \cdots & q_{rr} \end{bmatrix} \tag{B.4}$$

and $\mathbf{Q}$ is positive definite if and only if all of the $p_1, ..., p_{n^2}$ are positive (e.g., see [3]).

The leading principal minors are:

$$p_1 = q_{11} = \frac{1}{v_{11}^2} + c_1\mu_1^4 v_{11}^2 > 0 \tag{B.5}$$

$$p_2 = p_1(c_1 m_{12} v_{11}^2) > 0 \tag{B.6}$$

$$p_3 = p_3(c_1 m_{13} v_{11}^2) > 0 \tag{B.7}$$

$$p_4 = p_1(c_1^2 c_2) m_{13} v_{11}^4 v_{22}^2 [m_{12}^2 - \mu_1^4\mu_4] > 0 \quad \text{if} \quad \mu_1^4\mu_2^4 - 9(\mu_1^2)^2(\mu_2^2)^2 < 0 \tag{B.8}$$

$$p_5 = p_4 q_{55} = p_4(\frac{1}{v_{22}^2} + c_2\mu_2^4 v_{22}^2) > 0 \quad \text{if} \quad p_4 > 0 \tag{B.9}$$

$$p_6 = p_5(c_2 m_{23} v_{22}^2) > 0 \quad \text{if} \quad p_5 > 0 \tag{B.10}$$

$$\begin{aligned} p_7 &= p_6(c_3 m_{13} v_{33}^2) - \frac{1}{v_{11}^2 v_{33}^2}(c_2 m_{23} v_{22}^2)[c_1 c_2 m_{12}^2 v_{11}^2 v_{22}^2 - \frac{1}{v_{11}^2 v_{22}^2}] \\ &= (c_1^2 c_2^2 c_3) v_{vv}^4 v_{22}^4 v_{33}^2 m_{23} [m_{12}^2 - \mu_1^4\mu_2^4][m_{13}^2 - \mu_1^4\mu_3^4] \end{aligned} \tag{B.11}$$

is positive if $\mu_1^4\mu_3^4 - 9(\mu_1^2)^2(\mu_3^2)^2 < 0$ assuming $\mu_1^4\mu_2^4 - 9(\mu_1^2)^2(\mu_2^2)^2 < 0$. The last leading principal minor:

$$\begin{aligned} p_8 &= p_7(c_3 m_{23} v_{33}^2) \\ &\quad - q_{55}\frac{1}{v_{22}^2 v_{33}^2}[p_4(c_3 m_{13} v_{33}^2) - \frac{q_1 1}{v_{11}^2 v_{33}^2}(c_1 c_2 m_{12}^2 v_{11}^2 v_{22}^2 - \frac{1}{v_{11}^2 v_{22}^2})] \\ &= p_7(c_3 m_{23} v_{33}^2) - (c_1^2 c_2 c_3) v_{11}^4 v_{22}^2 v_{33}^2 (m_{12}^2 - \mu_1^4\mu_2^4)(m_{13}^2 - \mu_1^4\mu_3^4) \\ &= (c_1^2 c_2^2 c_3^2) v_{11}^4 v_{22}^4 v_{33}^4 (m_{12}^2 - \mu_1^4\mu_2^4)(m_{13}^2 - \mu_1^4\mu_3^4)(m_{23}^2 - \mu_2^4\mu_3^4) \end{aligned} \tag{B.12}$$

is positive if $\mu_2^4\mu_3^4 - 9(\mu_2^2)^2(\mu_3^2)^2 < 0$ assuming $\mu_1^4\mu_3^4 - 9(\mu_1^2)^2(\mu_3^2)^2 < 0$ and $\mu_1^4\mu_2^4 - 9(\mu_1^2)^2(\mu_2^2)^2 < 0$. Hence, the Hessian is positive definite and Solution P1-P8 are local minimum if the three channels are mutually Gaussian in-average. By Lemma 9, there is no local maximum and hence Solution P1-P8 are saddle points if the condition that the three channels are mutually Gaussian in-average is not satisfied.

By the symmetry between the channels, the other solutions of solution group P must have the same condition on stability. Therefore, the proposition is proved. □

Appendix C

Proof for Theorem 2

The equilibrium equation for the algorithm is $\nabla_{\mathbf{W}} J(\mathbf{W}) = [\nabla_{\mathbf{V}} J(\mathbf{V})]\mathbf{A}^{-1} = 0$, which implies, provided that $\det \mathbf{V} \neq 0$:

$$E_{\mathbf{s}}[\mathbf{I} + \mathbf{h}(\mathbf{Vs})(\mathbf{Vs})^T] = \mathbf{0} \tag{C.1}$$

or

Self-coupling equations:

$$1 - c_{11} E[y_1^2] = 1 - c_{11}(v_{11}^2 \mu_1^2 + v_{12}^2 \mu_2^2) = 0 \tag{C.2}$$

$$1 - c_{23} E[y_2^4] = 1 - c_{23}(v_{21}^4 \mu_1^4 + 6 v_{21}^2 v_{22}^2 \mu_1^2 \mu_2^2 + v_{22}^4 \mu_2^4) = 0 \tag{C.3}$$

Cross-coupling equations:

$$E[y_1 y_2] = v_{11} v_{21} \mu_1^2 + v_{12} v_{22} \mu_2^2 = 0 \tag{C.4}$$

$$E[y_2^3 y_1] = v_{11} v_{21}(v_{21}^2 \mu_1^4 + 3 v_{22}^2 \mu_1^2 \mu_2^2) + v_{12} v_{22}(v_{22}^2 \mu_2^4 + 3 v_{21}^2 \mu_1^2 \mu_2^2) = 0 \tag{C.5}$$

where $\mu_i^p = E[s_i^p]$. The cross-coupling equilibrium equations eq. (C.4) and (C.5) can be written as:

$$\begin{bmatrix} \mu_1^2 & \mu_2^2 \\ v_{21}^2 \mu_1^4 + 3 v_{22}^2 \mu_1^2 \mu_2^2 & v_{22}^2 \mu_2^4 + 3 v_{21}^2 \mu_1^2 \mu_2^2 \end{bmatrix} \begin{bmatrix} v_{11} v_{21} \\ v_{12} v_{22} \end{bmatrix} = \mathbf{0} \tag{C.6}$$

Denote the left matrix in eq. (C.6) as $\mathbf{M}'$, then $\det \mathbf{M}' = v_{22}^2\mu_1^2[\mu_2^4 - 3(\mu_2^2)^2] - v_{21}^2\mu_2^2[\mu_1^4 - 3(\mu_1^2)^2]$. Since one source is super-Gaussian and one source is sub-Gaussian, we have $\det \mathbf{M}' \neq 0$,[1] and hence

$$\begin{bmatrix} v_{11}v_{21} \\ v_{12}v_{22} \end{bmatrix} = \mathbf{0} \tag{C.7}$$

Coping it with the self-coupling equilibrium equations eq. (C.2) and (C.3), we get solution groups $\mathrm{A_I}$ and $\mathrm{A_{II}}$ eq. (7.2) and (7.3) exhaustively.

From Lemma 5, the sufficient and necessary condition for Solutions A1-A4 to be stable is:

$$\begin{aligned} & E[s_1^2]E[s_2^2][c_{11}]E[3c_{23}(v_{22}^* s_2)^2] - \frac{1}{{v_{11}^*}^2 {v_{23}^*}^2} \\ &= {v_{22}^*}^2 \left\{ 3c_{11}c_{23}E[s_1^2]\left(E[s_2^2]\right)^2 - \frac{1}{{v_{11}^*}^2 {v_{22}^*}^4} \right\} \\ &= {v_{22}^*}^2 c_{11}c_{23}E[s_1^2]\left\{ 3\left(E[s_2^2]\right)^2 - E[s_2^4] \right\} \\ &> 0 \end{aligned} \tag{C.8}$$

i.e., $kurt(s_2) = E[s_2^4] - 3\left(E[s_2^2]\right)^2 < 0$, or in other words, s_2 is sub-Gaussian.

Similarly, from Lemma 6, the sufficient and necessary condition for Solution A5-A8 to be stable is:

$$\begin{aligned} & E[s_1^2]E[s_2^2][c_{11}]E[3c_{23}(v_{21}^* s_1)^2] - \frac{1}{{v_{12}^*}^2 {v_{21}^*}^2} \\ &= {v_{21}^*}^2 \left\{ 3c_{11}c_{23}E[s_2^2]\left(E[s_1^2]\right)^2 - \frac{1}{{v_{12}^*}^2 {v_{21}^*}^4} \right\} \\ &= {v_{22}^*}^2 c_{11}c_{23}E[s_2^2]\left\{ 3\left(E[s_1^2]\right)^2 - E[s_1^4] \right\} \\ &> 0 \end{aligned} \tag{C.9}$$

i.e., $kurt(s_1) = E[s_1^4] - 3\left(E[s_1^2]\right)^2 < 0$, or s_1 is sub-Gaussian.

To summarize, we have:

s_1	s_2	Stable Solution	y_1	y_2
super-Gaussian	sub-Gaussian	A1-A4	s_1	s_2
sub-Gaussian	super-Gaussian	A5-A8	s_2	s_1

In all cases, the pdf of y_2 is flatter than that of y_1. Hence the theorem is proved. □

[1] Actually, if one source were exactly Gaussian and the other were non-Gaussian, we also have $\det \mathbf{M}' \neq 0$. However, processing only finite number of sample points, the empirical kurtosis would not be exactly zero but would be either a positive or negative number with small magnitude. Hence, the condition that one source is exactly Gaussian is not applicable in this case.

Bibliography

[1] S.-I. Amari, A. Cichocik and H. Yang, "Recurrent neural networks for blind separation of sources", in *Proceedings of the International Symposium on Nonlinear Theory and Its Applications (NOLTA 95)* (Las Vegas, USA, Dec 10-14, 1995), pp. 37-42, 1995.

[2] S.-I. Amari, A. Cichocki, H. Yang, "A new learning algorithm for blind separation of sources", in David S. Touretzky, Michael C. Mozer & d Michael E. Hasselmo, eds, *Advances in Neural Information Processing Systems 8* (MIT Press: Cambridge, MA. 1996), pp. 757-763, 1996.

[3] F. Ayres, JR, *Schaum's outline of theory and problem of Matrices, SI (Metric) Edition*, McGraw-Hill, Singapore, 1983.

[4] P. Bamford and N. Canagarajah, "Optimal adaptive decorrelator for signal separation", *Electronics Letters*, vol. 31, no. 14, pp. 1128-1129, July 1995.

[5] A.J. Bell and T.J. Sejnowski, "An information-maximatization approach to blind separation and blind deconvolution", *Neural Computation*, vol. 7, pp. 1129-1159, 1995.

[6] A.J. Bell and T.J. Sejnowski, "Fast blind separation based on information theory", in *Proceedings of the International Symposium on Nonlinear Theory and Its Applications (NOLTA 95)* (Las Vegas, USA, Dec 10-14, 1995), pp. 287-314, 1995.

[7] A.J. Bell and T.J. Sejnowski, "Edges are the 'independent components' of natural scenes", in *Advances in Neural Information Processing Systems 9*, MIT Press, 1997.

[8] A.J. Bell and T.J. Sejnowski, "Learning the higher-order structure of a natural sound", to appear in *Network: Computation in Neural Systems*.

[9] A. Belouchrani, and J.-F Cardoso, "Maximum likelihood source separation by the Expectation-Maximization technique: deterministic and stochastic implementation", in *Proceedings of the International Symposium on Nonlinear Theory and Its Applications (NOLTA 95)* (Las Vegas, USA, Dec 10-14, 1995), pp. 49-53, 1995.

[10] A. Belouchrani, K. Abed-Meraim, J.-F Cardoso and E. Moulines, "A blind source separation technique using second-order statistics", *IEEE Transactions on Signal Processing*, vol. 43, no. 2, pp. 434-443, 1997.

[11] Y. Cao, S. Sridharan, and M. Moody, "Co-talker separation using the 'cocktail party effect'", *Journal of the Audio Engineering Society*, vol. 44, no. 12, pp. 1084-1096, Dec 1996.

[12] J.-F. Cardoso, "Source separation using higher order moments" in *Proceedings of the 1989 International Conference on Acoustics, Speech and Signal Processing (ICASSP 89)*, pp. 2019-2112, 1989.

[13] J.-F. Cardoso, "Iterative techniques for blind source separation using only fourth order cumulants", in *Signal Processing VI Theories and Applications: Proceedings of the sixth European Signal Processing Conference (EUSIPCO 92)* (Brussels, Belgium, Aug 24-27, 1992), pp. 739-742, 1992.

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

[15] J.F. Cardoso and P. Comon, "Independent component analysis, a survey of some algebraic methods", in *Proc. ISCAS 96*, vol. 2, pp. 93-96, 1996.

[16] D.C.B. Chan, P.J.W. Rayner and S.J. Godsill, "Multi-channel blind signal separation by decorrelation", In *Proceedings of IEEE Signal Processing Society 1995 Workshop on Applications of Signal Processing to Audio and Acoustics* (New York, Oct 15-18, 1995).

[17] T. Chen, R. Chen, "A neural network approach to blind identification of stochastic and deterministic signals", in *Proceedings of the 28th Asilomar Conference on Signals, Systems & Computers*, pp. 892-896, 1994.

[18] C.C. Cheung and L. Xu, "Independent component analysis on two channels by the information-theoretic approach with cubic nonlinearity", submitted to *Neurocomputing*.

[19] C.C. Cheung and L. Xu, "Separation of two independent sources by the Information-theoretic approach with cubic nonlinearity", in *Proceedings of the*

1997 International Conference on Neural Networks (ICNN 97) (Houston, USA, Jun 9-12, 1997), pp. 2239-2244, 1997.

[20] A. Cichocki and L. Moszczynki, "New learning algorithm for blind separation of sources", *Electronics Letters*, vol. 28, no. 21, pp. 1986-1987, Oct 1992.

[21] A. Cichocki, W. Kasprzak and S.-I. Amari, "Multi-layer neural networks with a local adaptive learning rule for blind separation of source signals", in *Proceedings of the International Symposium on Nonlinear Theory and Its Applications (NOLTA 95)* (Las Vegas, USA, Dec 10-14, 1995) pp. 61-65, 1995.

[22] M.H. Cohen and A.G. Andreou, "Current-mode subthreshold MOS implementation of the Herault-Jutten autoadaptive network", *IEEE Journal of Solid-State Circuits*, vol. 27, no. 5, pp. 714-727, 1992.

[23] M.H. Cohen and A.G. Andreou, "Analog CMOS integration and experimentation with an autoadaptive independent component analyzer", *IEEE Transaction on Circuits Systems-II: Analog Digital Signal Process*, vol. 42, no. 2, pp. 65-77, 1995.

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

[25] P. Comon, "Contrasts for multichannel blind deconvolution", *IEEE Signal Processing Letters*, vol. 3, no. 7, pp. 209-211, July 1996.

[26] P. Comon, C. Jutten and J. Herault, "Blind separation of sources, part II: problems statement", *Signal Processing*, vol. 24, no. 1, pp. 11-20, 1991.

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

[28] Y. Deville and L. Andry, "Application of blind source separation techniques to multi-tag contactless identification systems", in *Proceedings of the International Symposium on Nonlinear Theory and its Applications (NOLTA 95)*, (Las Vegas, USA, Dec 10-14, 1995), pp. 73-78, 1995.

[29] M. Feder, A.V. Oppenheim, E. Weinstein, "Maximum likelihood noise cancelation using the EM algorithm", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 2, Feb 1989.

[30] J.C. Fort, "Stabilité de l'algorithme de séparation de sources de Jutten et Hérault", *Traitement du Signal*, vol. 8, no. 1, pp. 35-42, 1991.

[31] S.V. Gerven and D.V. Compernolle, "On the use of decorrelation in scalar signal separation", in *Proceedings of the 1994 International Conference on Acoustics, Speech and Signal Processing (ICASSP 94)*, pp. 57-60, 1994.

[32] S.V. Gerven and D.V. Compernolle, "Signal separation by symmetric adaptive decorrelation: stability, convergence, and uniqueness", *IEEE Transactions on Signal Processing*, vol. 43, no. 7, pp. 1602-1612, July 1995.

[33] J. Herault, C. Jutten and B. Ans, "Détection de grandeurs primitives dans un message composite par une architecture de calcul neuromimétique en apprentissage non supervisé", in *Actes du Xéme colloque GRETSI* (Nice, France, May 20-24 1985), pp. 1017-1022, 1985.

[34] A. Hyvarinen, "Simple one-unit neural algorithms for blind source separation and blind deconvolution", in *Progress in Neural Information Processing: Proceedings of the International Conference on Neural Information Processing (ICONIP 96)* (Hong Kong, Sept 24-27, 1996; Springer-Verlag: Singapore 1996), pp. 1201-1206, 1996.

[35] A. Hyvarinen, "A family of fixed-point algorithms for independent component analysis", in *Proceedings of the 1997 International Conference on Acoustics, Speech and Signal Processing (ICASSP 97)*, (Munich, Germany, April 1997).

[36] A. Hyvarinen and E. Oja, "Simple neuron models for independent component analysis", to appear in *International Journal of Neural Systems*.

[37] C. Jutten and J. Herault, "Blind separation of sources, part I: an adaptive algorithm based on neuromimetic architecture", *Signal Processing*, vol. 24, no. 1, pp. 1-10, 1991.

[38] C. Jutten and J. Herault, "Analog implementation of permanent unsupervised learning algorithm", in *Proceedings of the NATO Advanced Research Workshop on Neurocomputing*, (Les Arce, France, Feb 27 - Mar 3, 1989), pp. 145-155, 1989.

[39] C. Jutten and J.-F. Cardoso, "Separation of sources: really blind", in *Proceedings of the International Symposium on Nonlinear Theory and its Applications (NOLTA 95)*, (Las Vegas, USA, Dec 10-14, 1995), pp. 79-84, 1995.

[40] M. Kendall *Kendall's Advanced Theory of Statistics*, fifth ed., vol. 1, Charles Griffin & Company Limited, London, 1987.

[41] J. Karhunen, "Neural approaches to independent component analysis and source separation", invited paper, *Proceedings of the 4th European Symposium on Artificial Neural Networks (ESANN 96)* (Bruges, Belgium, April 24-26, 1996).

[42] J. Karhunen, L. Wang and R. Vigario, "Nonlinear PCA type approaches for source separation and independent component analysis", in *Proceedings of the 1995 IEEE International Conference on Neural Networks (ICNN 95)*, (Perth, Australia, Nov 27 - Dec 1, 1995), pp. 995-1000, 1995.

[43] J. Karhunen, A. Hyvarinen, R. Cigario, J. Hurri and E Oja, "Applications of neural blind separation to signal and image processing", to appear in *Proceedings of the IEEE 1997 International Conference on Acoustics, Speech, and Signal Processing (ICASSP 97)*, (Munich, Germany, Apr 21-24 1997)

[44] J.L. Lacoume (eds.), *Higher order statistics: Proceedings of the International Signal Processing Workshop on Higher Order Statistics* (Chamrousse, France, Jul 10-12, 1991).

[45] R. Linsker, "Local synaptic learning rules suffice to maximize mutual information in a linear network.", *Neural Computation*, vol. 4, pp. 691-702, 1992.

[46] T.W. Lee, A.J. Bell and R.H. Lambert, "Blind separation of delayed and convolved sources", accepted for publication in *Advances in Neural Information Processing Systems*, MIT Press, Cambridge MA, 1996

[47] T.W. Lee, A.J. Bell and R. Orglmeister, "Blind source separation of real world signals", to appear in *IEEE International Conference on Neural Networks (ICNN 97)* (Houston, USA, June 9-12, 1997).

[48] U. Lindgren, T. Wigren, and H. Broman, "On local convergence of a class of blind separation algorithms", *IEEE Transactions on Signal Processing*, vol. 43, no. 12, pp. 3054-3058, Dec 1995.

[49] R. Liu, F. Dong, and X. Ling, "A convergence analysis for neural networks with constant learning rates and non-stationary inputs", in *Proceedings of the 34th Conference on Decision & Control* (New Orleans, LA, Dec 1995), pp. 1278-1283, 1995.

[50] S. Makeig, T.P. Jung, A.J. Bell, and T.J. Sejnowski, "Independent component analysis of electroencephalographic data", in D. Touretzky, M. Mozer and M. Hasselmo (Eds.) *Advances in Neural Information Processing Systems 8*, MIT Press, Cambridge MA, pp. 145-151, 1996.

[51] H. Marsman, "A neural net approach to source separation problem", Master's Thesis, Department of Electrical Engineering, University of Twente, Netherlands, 1995.

[52] K. Matsuoka, M. Ohya and M. Kawamoto, "A neural net for blind separation of nonstationary signals", *Neural Networks*, vol. 8, no. 3, pp. 411-419, 1995.

[53] E. Moreau and O. Macchi, "New self-adaptive algorithms for source separation based on contrast functions", in *Proceedings of the IEEE Signal Processing Workshop on Higher Order Statistics*, pp. 215-219, 1993.

[54] E. Oja, J. Karhunen, L. Wang, and R. Vigarioi, "Principal and independent components in neural networks – recent developments", in *Proceedings of the 7th Italian Workshop on Neural Networks (WIRN 95)*, (Vietri sul Mare, Italy, May 1995).

[55] E. Oja and A. Hyvarinen, "Blind signal separation by neural networks", in *Progress in Neural Information Processing: Proceedings of the International Conference on Neural Information Processing (ICONIP 96)* (Hong Kong, Sept 24-27, 1996; Springer-Verlag: Singapore 1996), pp. 7-14, 1996.

[56] P. Pajunen, A. Hyvarinen, and J. Karhunen, "Nonlinear blind source separation by self-organizing maps", in *Progress in Neural Information Processing: Proceedings of the International Conference on Neural Information Processing (ICONIP 96)* (Hong Kong, Sept 24-27, 1996; Springer-Verlag: Singapore 1996), pp. 1207-1210.

[57] D.T. Pham, P. Garat and C. Jutten, "Separation of a mixture of independent sources through a maximum likelihood approach", in J. Vandewalle, R. Boite, M. Moonen, A. Oosterlinck (eds.), *Signal Processing VI: Theories and Applications* (Elsevier Science Publishers, 1992), pp. 771-774, 1992.

[58] J.C. Platt and F.Faggin, "Networks for the separation of sources that are superimposed and delayed", in Moody J.E *et al* (eds.) *Advances in Neural Information Processing System 4*, Morgan-Kaufmann, pp. 730-737, 1992.

[59] E. Sorouchyari, "Blind separation of sources, part III: stability analysis", *Signal Processing*, vol. 24 no. 1, pp. 21-29, 1991.

[60] H.L.N. Thi, C, Jutten, "Blind source separation for convolutive mixtures", *Signal Processing*, vol. 45, pp. 209-229, 1995.

[61] K. Torkkola, "Blind separation of convolved sources based on information maximization", in *IEEE workshop on Neural Networks for Signal Processing* (Kyoto, Japan, Sept 4-6, 1996).

[62] L. Tong, R.W. Liu, V.C. Soon, Y.F. Huang, "Indeterminacy and identifiability of blind identification", *IEEE Transactions on circuits and systems*, vol. 38, no. 5, pp. 499-509, May 1991.

[63] E.A. Vittoz and X. Arreguit, "CMOS integration of Herault-Jutten cells for separation of sources", in C. Mead and M. Ismail, eds., *Analog VLSI Implementation of Neural Systems*, Kluwer, Boston, pp. 57-84, 1989.

[64] L. Wang, J. Karhunen, and E. Oja, "A bigradient optimization approach for robust PCA, MCA, and source separation", in *Proceedings of the 1995 IEEE International Conference on Neural Networks (ICNN 95)* (Perth, Australia, Nov 1995), pp. 1684-1689, 1995.

[65] E. Weinstein, M. Feder and A.V. Oppenheim, "Multi-channel signal separation by decorrelation", *IEEE Transactions on speech and audio processing*, vol. 1, no. 4, pp. 405-413, Oct 1993.

[66] B. Widrow *et al* , "Adaptive noise cancelling: principle and application", *Proceedings of the IEEE*, vol. 63, no. 12, pp. 1692-1716, Dec 1975.

[67] L. Xu, "YING-YANG Machine: a Bayesian-Kullback scheme for unified learnings and new results on vector quantization", Keynote talk, in *Proceedings of the International Conference on Neural Information Processing (ICONIP 95)* (Beijing, China, Oct 30 - Nov 3, 1995), pp. 977-988, 1995.

[68] L. Xu, "YING-YANG Machine for temporal signals", Keynote talk, in *Proceedings of the International Conference on Neural Networks and Signal Processing 1995*, (Nanjing, China), vol. I, pp. 644-651, 1995.

[69] L. Xu, "A unified learning scheme: Bayesian-Kullback YING-YANG Machine", in David S. Touretzky, Michael C. Mozer & d Michael E. Hasselmo (eds.), *Advances in Neural Information Processing Systems 8* (MIT Press: Cambridge, MA. 1996), pp. 444-450, 1996.

[70] L. Xu, "Bayesian-Kullback YING-YANG Machine: reviews and new results", in *Progress in Neural Information Processing: Proceedings of the International Conference on Neural Information Processing (ICONIP 96)* (Hong Kong, Sept 24-27, 1996; Springer-Verlag: Singapore 1996), pp. 59-67, 1996.

[71] L. Xu, "Unsupervised, supervised and cosupervised Bayesian Ying-Yang learning: New developments", to appear in *Lecture Notes in Computer Science, Proceedings of the International Workshop on Theoretical Aspects of Neural Computation* (May 26-28, 1997, Hong Kong), Springer-verlag, 1997.

[72] L. Xu, "New advances on Bayesian YING-YANG learning system with Kullback and non-Kullback separation functionals", in *Proceedings of the 1997 International*

Conference on Neural Networks (ICNN 97) (Houston, USA, June 9-12, 1997), pp. 1942-1947, 1997.

[73] L. Xu, "Bayesian YING-YANG learning based ICA models", to appear in *Proceedings of the 1997 IEEE Workshop on Neural Network for Signal Processing, Special Session on Blind Signal Separation* (Amelia Island Plantation, Florida, Sept. 24-26, 1997).

[74] L. Xu and S.-I. Amari, "A general independent component analysis framework based on Bayesian-Kullback YING-YANG learning", in *Progress in Neural Information Processing: Proceedings of the International Conference on Neural Information Processing (ICONIP 96)* (Hong Kong, Sept 24-27, 1996; Springer-Verlag: Singapore 1996), pp. 1235-1239, 1996.

[75] L. Xu, C.C. Cheung, H.H. Yang and S.-I. Amari, "Independent component analysis by the information-theoretic approach with mixture of densities", in *Proceedings of the 1997 International Conference on Neural Networks (ICNN 97)* (Houston, USA, June 9-12, 1997), pp. 1821-1826, 1997.

[76] L. Xu, C.C. Cheung, J. Ruan and S.-I. Amari, "Nonlinearity and separation capability: Further justification for the ICA algorithm with mixture of densities", invited special session on Blind Signal Separation, in *Proceedings of the 5th European Symposium on Artificial Neural Networks (ESANN 97)*. (Bruges, Belgium, Apr 16-18, 1997), pp. 291-296, 1997.

[77] D. Yellin and E. Weinstein, "Criteria for multichannel signal separation", *IEEE Transactions on signal processing*, vol. 42, no. 8, Aug 1994.

[78] D. Yellin and E. Weinstein, "Multichannel signal separation: methods and analysis", *IEEE Transactions on signal processing*, vol. 44, no. 1, Jan 1996.