UNIVERSITY OF CALIFORNIA

Los Angeles

Non-Parametric Approaches to Unsupervised Structure Discovery

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Electrical Engineering

by

Riccardo Boscolo

2003

The dissertation of Riccardo Boscolo is approved.

Chiara Sabatti

Lieven Vandenberghe

Kung Yao

Vwani P. Roychowdhury, Committee Chair

University of California, Los Angeles
2003

Dedico questa tesi alla mia famiglia, fonte interminabile e insostituibile di sostegno.

Contents

List of Figures **vii**

List of Tables **xi**

1 Introduction **1**

1.1 Overview and Motivation 1

1.2 Outline 4

2 Information Theory and Unsupervised Learning **9**

2.1 Introduction 9

2.2 Differential Entropy 11

2.3 Relative Entropy and Mutual Information 13

2.4 Deviation from Gaussianity and Projection Pursuit 17

2.4.1 The curse of dimensionality 17

2.4.2 Centering and Sphering 19

2.4.3 Projection Pursuit Indices 20

3 Independent Component Analysis: Theory and Practice **27**

3.1 Theoretical Foundations 28

3.1.1 Model Identifiability 28

3.1.2 Uniqueness of the Separating Operator 30

3.2 Contrast Functions for ICA 34

3.2.1 InfoMax 34

3.2.2 Mutual Information Minimization 38

3.2.3 Maximum Likelihood Estimation 40

3.2.4 Negentropy Index 42

3.3 Cumulant Based Approximations 44
3.4 ICA Algorithms: State of the Art 46
3.4.1 Jade 47
3.4.2 InfoMax ICA and Extensions 47
3.4.3 Flexible Algorithms and Parametric ICA 48
3.4.4 Kernel Based Approaches 49
3.5 Challenges and Limitations 49

4 Non-Parametric ICA 51
4.1 Joint Estimation of the Unmixing Matrix and of the Distribution of the Source Signals 53
4.1.1 ICA Model and Separation Principle 53
4.1.2 Non-Parametric Kernel Density Estimation 54
4.1.3 Objective Function Derivation 55
4.2 Optimization and Global Convergence Issues 57
4.2.1 Optimization Algorithm 57
4.2.2 Analysis of the extrema of the cost function for N=2 sources 60
4.3 Simulation Experiments 67
4.3.1 Mixtures of Sources with Various Distributions 70
4.3.2 Skewed Sources 75
4.3.3 Convergence Properties 77
4.3.4 Bandwidth Parameter Sensitivity 77
4.3.5 Algorithmic Complexity 79
4.3.6 Large Scale Problems 80
4.4 Conclusions 85

5 An Extension of Comon's Identifiability Theorem 87
5.1 Introduction 88
5.2 Separation Principle and Objective Function Definition 89
5.3 Extending Comon's Identifiability Theorem 90
5.4 Conclusions 95

6 On the Uniqueness of the MI Minimum for Special Classes of Distributions 97
6.1 Introduction 98
6.2 Extrema for Mixtures of Two Nearly Gaussian Sources 100
6.3 An Extension of the Entropy Power Inequality 106

6.4 Conclusions 108

7 Learning Linear Non-Gaussian Networks 109

7.1 Introduction 110

7.2 Linear Non-Gaussian Networks 113

7.2.1 The Model 113

7.2.2 Network Properties 114

7.3 Learning The Network Structure 116

7.3.1 Model Identifiability 116

7.3.2 Mutual Information 118

7.3.3 The Optimization Framework 119

7.3.4 The Relaxation Graph 121

7.4 Simulation Experiments 123

7.5 Discussion 126

7.6 Conclusions 127

8 Learning Conditional Co-Expression Patterns in Gene Expression Data with Information Theoretic Exploratory Methods 131

8.1 Gene Expression Profiling Using DNA Microarray Data Technology 132

8.2 DNA Microarray Data Statistical Analysis 134

8.3 Conditionally Informative Clusters of Genes 137

8.3.1 Conditional Mutual Information as a Measure of Conditional Co-expression 138

8.3.2 Some Properties of the Cost Function 142

8.4 GeneScreen: Theory and Practice 145

8.4.1 Combinatorial Optimization Approach 146

8.4.2 A Moment Based Approximation of the Mutual Information 149

8.5 Simulation Results 153

8.5.1 Analysis of *Escherichia Coli* Expression Data 154

8.6 Discussion and Future Work 169

Bibliography 173

List of Figures

4.1 Graphical interpretation of the conditions on the extrema of the cost function (4.30). The curve $1/\tan(\theta_2 - \theta_1)$ is plotted for a fixed value of θ_2 (not to scale). 63

4.2 The nature of an extremum of the objective function is shown as a function of the second-order partial derivatives of the entropies of the two reconstructed sources (for $\theta_2 - \theta_1 = \text{const.}$). 64

4.3 Mixtures of a sub-gaussian signal and a super-gaussian signal. The figure shows a plot of the entropy of a generic reconstructed source as a function of the parameter θ. For these particular mixtures of unimodal sub-gaussian and super-gaussian sources, the entropy function does not present any spurious local minima. 65

4.4 Mixtures of a sub-gaussian signal and a super-gaussian signal. The overall cost function $L(W)$ is plotted as a function of (θ_1, θ_2). The plot clearly shows the set of four equivalent minima, corresponding to permutations or change of sign of the rows of the unmixing matrix. 66

4.5 Mixtures of two natural images. The entropy function is plotted as a function of $\theta_{1,2}$. In this case, the entropy presents several spurious local minima, which do not correspond to independent sources. Attempting the separation using a deflationary approach could result in a failure to reconstruct the sources. 68

4.6 Mixtures of two natural images. The overall cost function $L(W)$ is plotted as a function of (θ_1, θ_2), with a set of four equivalent global minima clearly appearing. The objective function is free from the spurious local minima encountered when observing the entropy function alone. At least in this case, the only values of $\theta_{1,2}$ that satisfy (4.30) are either (equivalent) global minima, or saddle points. 69

4.7 First simulation experiment. The results of attempting the separation of the six different sources listed in Table 4.2 are shown for various ICA algorithms (averaged over 1000 Monte Carlo simulations). The accuracy of the separation is measured in terms of median log signal-to-interference ratio (SIR). The relative performance for Source #1 and #2 is shown. 72

4.8 First simulation experiment. The results of attempting the separation of the six different sources listed in Table 4.2 are shown for various ICA algorithms (averaged over 1000 Monte Carlo simulations). The accuracy of the separation is measured in terms of median log signal-to-interference ratio (SIR). The relative performance for Source #3 and #4 is shown. 73

4.9 First simulation experiment. The results of attempting the separation of the six different sources listed in Table 4.2 are shown for various ICA algorithms (averaged over 1000 Monte Carlo simulations). The accuracy of the separation is measured in terms of median log signal-to-interference ratio (SIR). The relative performance for Source #5 and #6 is shown. 74

4.10 The figure shows the results of a set of simulation experiments aiming at evaluating the sensitivity of the proposed technique to the choice of the bandwidth parameter h. The error bars span between the 25 and the 75 percentiles of the SIR. This experiment seems to suggest that variations of such a parameter up to ±50% from the estimated optimal value do not considerably affect the separation performance. 78

4.11 The running time in terms of CPU seconds of various ICA algorithms is shown for a fixed number of sources (6) and variable number of samples. The methods capable of source adaptation are in general computationally more expensive, as the separation performance is paid in terms of running time. 82

4.12 The running time in terms of CPU seconds of various ICA algorithms is shown for a fixed number of samples (1000) and variable number of sources. The methods capable of source adaptation are in general computationally more expensive, as the separation performance is paid in terms of running time. 83

4.13 Large scale simulation. The median SIR (dB) achieved by Non-Parametric ICA is shown for the separation of a number of sources varying between 12 and 24 (averaged over the reconstructed signals), and a fixed number of samples ($M = 1000$). 84

7.1 An example of a belief network topology. 113

7.2 Relaxation graph obtained from the belief network of Figure 7.1. . 122

7.3 The synthetic data used in the simulation experiment was generated using a power exponential density ($k_4 = 1.52$) as a model for the local conditional probability density functions. A gaussian distribution is also shown for comparison. 124

7.4 Relaxation graph learned from data in the first simulation experiment. Because of the large sample size the weights of the graphical model are learned very accurately and a simple thresholding would results in a good estimate of the original model. 128

7.5 Relaxation graph learned from data in the second simulation experiment. The network topology is identified also in this case. However, the poor sampling results in larger errors in the estimation of the arc weights. 129

8.1 The plot shows the expression time-courses of three hypothetical genes, generated from synthetic data. 138

8.2 Scatter plots of the expression levels of the three hypothetical genes. 139

8.3 The co-expression pattern between gene-a and gene-b appears under conditioning by gene-c. When gene-c is up-regulated, gene-a and gene-b appear positively correlated. The opposite pattern appears when gene-c is under-expressed. No significant co-expression patterns appear when gene-c is expressed around the reference level. 141

8.4 Cluster of genes composed of L conditioning genes and M children nodes. This cluster represents the generic sub-network explored to identify conditional structure. 147

8.5 Co-expression pattern between the genes *yabM* and *ilvY*, when gene *tap* is the conditioning node. 157

8.6 Co-expression pattern between the genes *tap* and *ilvY*, when gene *yabM* is the conditioning node. 158

8.7 Co-expression pattern between the genes *hyfG* and *recN*, when gene *yciK* is the conditioning node. 159

8.8 Co-expression pattern between the genes *yciK* and *hyfG*, when gene *recN* is the conditioning node. 160

8.9 Co-expression pattern between the genes *yciK* and *recN*, when gene *hyfG* is the conditioning node. 161

8.10 Co-expression pattern between the genes *narL* and *ycdQ*, when gene *menB* is the conditioning node. 163

8.11 Co-expression pattern between the genes *narL* and *menB*, when gene *ycdQ* is the conditioning node. 164

8.12 Co-expression pattern between the genes *mbhA* and *recN*, when gene *yebF* is the conditioning node. 165

8.13 Co-expression pattern between the genes *ybgF* and *acpS*, when gene *nagB* is the conditioning node. 167

8.14 Co-expression pattern between the genes *yibP* and *acs*, when gene *otsB* is the conditioning node. 168

8.15 Network of high-scoring conditional association patterns, obtained by hierarchical grouping of clusters of genes. All the loops involving three edges in the graph can be associated to strong conditional co-expression patterns between the genes in the loop. 170

8.16 The histogram shown the 10% highest scores for the original *E.coli* dataset and a dataset obtained shuffling the data points for each gene expression time course. 171

List of Tables

4.1 Main steps of the Non-Parametric ICA algorithm 60

4.2 Distribution of the synthetic sources used in the first simulation experiment (see [30] for a description of the distributions generated with the Power Method). 71

4.3 The separation performance in terms of median SIR, as well as 25 and 75 percentiles, is shown for mixtures of four skewed sources (averaged over the sources), for various ICA algorithms. 76

4.4 Detailed analysis of the computational complexity of Non-Parametric ICA as a function of the number of sources (N) and the number of samples (M) 81

8.1 Experimental conditions and corresponding number of measurements for the *E.coli* dataset used in the simulations. 155

ACKNOWLEDGEMENTS

There are so many people whose presence has greatly influenced and changed my life in the last few years that these few words can hardly express my gratitude.

It was great to work together with my advisor, Professor Vwani Roychowdhury, whose vision always inspired me and whose guidance and support went certainly beyond his academic role. I would like to thank my former advisor Professor Helen Na for believing in me and giving me the chance to embark upon this wonderful journey. A special thanks must go to Professor Michael McNitt-Gray and Professor Matthew Brown for their support in a very difficult time, for a great work collaboration, and, I have to add, for the great time I had at the RSNA conference in Chicago in 2000. I would also like to thank Professor Hong Pan: I greatly benefited from the many discussion we had on our common research. I thank the members of my PhD committee, Professor Chiara Sabatti, Professor Lieven Vandenberghe and Professor Kung Yao, for the time they dedicated to me, discussing my research and providing valuable suggestions.

During the course of my PhD, I had the great opportunity of working at the HRL Laboratories in Malibu. I have to thank my colleagues at HRL for giving me the chance of being part of such an amazing work environment. Roy Matic, Yuri Owechko, Deepak Khosla, Mike Carrasco, Narayan Srinivasa and Swarup Medasani: you guys are the best. I hope I will always be so lucky to work with people like you guys.

I really do not think I will find the words to thank all the people that were close to me during these years. I would like to start by thanking my friends of a lifetime in Italy, Michele, Diego, Giancarlo e Luca. I have to thank you if our friendships remained strong, in spite of the distance .

I would like to thank all my wonderful friends and colleagues here at UCLA, you made working here a pleasure and an unforgettable experience. A special thanks to Merhdad Nouralishahi and Arash Behzad, to whom I owe the few words of Farsi I can speak, guys I hope I will not get in trouble when I try to use them. A special thanks also to Jesse Bridgewater for his valuable help with Linux and several other programming issues. I would like to thank Katherine Comanor, with whom I shared my cubicle in the last two and a half years: Katherine, I know it is not easy to cope with a person like me and I assure you you did a great job. I have to extend a very special thanks to Brien Alkire: Brien, without your help, putting together this thesis would have taken twice as long. Thanks for your friendship, I hope we will get again the chance to cast off for new and exciting sushi and sailing adventures. May the wind always blow at your back.

I met so many people here in the United States who not only became my new friends but soon also became like a family to me. Andrea Soligon, Marino De Terlizzi, Paola Acquaviva, Stefania Picconi, Pietro Musumeci, thanks for accompanying me during this great journey, you guys were amazing friends and I sincerely hope that our paths will always cross in the future. I also cannot forget my truly outstanding students from the GER101 class, you guys were the best EAPs we have seen in years, and that comes from a person who's been here quite some time. Of course I have not forgotten about you Jean, how could I? I will probably never find another person whose friendship is so true and wholehearted like the one Jean Hotta showed to me. Jean you were always there when I needed you and I thank you for that.

VITA

June 30, 1972	Born, Treviso, Italy.
1995	*Erasmus* fellowship, University of Patras, Greece.
1996	*EAP* fellowship, University of California, Los Angeles.
1997	Laurea Degree, Department of Electronics Engineering, University of Padova, Italy (*cum laude*).
1998-2003	Graduate Student Researcher, University of California, Los Angeles.
1999	M.S., Electrical Engineering, University of California, Los Angeles.
1999-2000	Software Engineer, HRL Laboratories, Malibu.

PUBLICATIONS

R. Boscolo (December,1999). *A Multiple Source Reconstruction Algorithm for 3D Computerized Ionospheric Tomography Based on Least Squares Estimation.* Master Thesis, University of California, Los Angeles.

R. Boscolo, H. Pan and V.P. Roychowdhury (December,2001), Non-Parametric ICA. In *Proceedings of the Third International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2001)*, pages 13–18, San Diego, California.

R. Boscolo, M.S. Brown, and M.F. McNitt-Gray (March,2002). Medical image segmentation using knowledge-guided robust active contours. *Radiographics*, 22:437–448.

R. Boscolo, H. Pan, and V.P. Roychowdhury (August,2002). Beyond Comon's Identifiability Theorem for Independent Component Analysis. In *Proceedings of* *the 2002 International Conference on Artificial Neural Networks (ICANN 2002)*, Lecture Notes in Computer Science, Springer-Verlag, 2415:1119–1124, Madrid, Spain.

R. Boscolo and V. P. Roychowdhury (April,2003). On the Uniqueness of the Minimum of the Information-Theoretic Cost Function for the Separation of Mixtures of Nearly Gaussian Signals. In *Proceedings of the 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, pages 137–141, Nara, Japan.

R. Boscolo, H. Pan, V. P. Roychowdhury (2003). Independent Component Analysis Based on Non-Parametric Density Estimation. *IEEE Transactions on Neural Networks*, in press.

Y. Owechko, N. Srinivasa, S. Medasani and R. Boscolo (June,2002). Vision-Based Fusion System for Smart Airbag Applications. In *Proceedings of the IEEE Intelligent Vehicle Symposium (IV'2002)*, Versailles, France.

ABSTRACT OF THE DISSERTATION

Non-Parametric Approaches to Unsupervised Structure Discovery

by

Riccardo Boscolo
Doctor of Philosophy in Electrical Engineering

University of California, Los Angeles, 2003

Professor Vwani P. Roychowdhury, Chair

For several years, blind signal or system identification problems have drawn the attention of researchers from various fields, including physicists, engineers, computer scientists, but also biologists, linguists, and economists. From a statistical perspective, this kind of problems can be seen as a special instance of unsupervised learning methods, which deal with the issue of finding a suitable representation of the data, when only a minimal knowledge on the generative process is available.

In particular, for the special case of blind signal separation of linearly mixed stationary signals, concepts from statistical information theory have allowed researchers to devise novel frameworks, in which solutions to the estimation problem that were not thought possible have been identified. The discovery of methods allowing the blind reconstruction of statistically independent components from linearly mixed data has led to the creation of an entire new field of research, known as Independent Component Analysis (ICA). In the last few years, ICA has found vast application as a data analysis tool (*e.g.* in biomedical and financial data analysis), as a feature extraction technique (*e.g.* in speech and image processing), as well as a pre-processing tool in the field of telecommunications (*e.g.*

wireless communications and multi-user detection in broadband systems). Moreover, the close connection between ICA and fields such as neural computation and unsupervised machine learning has opened new perspectives on statistical learning problems in general.

In this dissertation, we investigate and provide a solution to certain fundamental open problems in the field of blind signal separation. The first and most relevant is the design and implementation of a universal ICA method capable of reliably separating mixtures of signals with arbitrary statistical properties, lifting the conventional requirement of having to provide an approximate estimate of their distributions. This goal was achieved by designing a non-parametric estimation framework, capable of simultaneously learning the statistical distributions of the unknown signals, as well as the separating linear projection operator. The proposed method is shown to clearly outperform current state-of-the-art ICA techniques in terms of both accuracy of the separation and convergence properties. Such claim is substantiated by several sets of simulation experiments that were conducted comparing the proposed approach to six others popular ICA implementations.

An additional goal of this work is the investigation of certain fundamental theoretical properties of the estimation framework associated with finding the optimal solution to the ICA problem. In particular, we extended the conventional limitation that at the most one of the unknown signals can have a gaussian distribution. A novel result shows that by using the proposed framework the separation of all the non-gaussian signals is still possible, even in presence of multiple gaussian signals mixed to them. We further identified a class of distributions for which the proposed cost function has no local minima and we showed that such result can be used to extend well-known information theory inequalities, such as the entropy power inequality.

Finally, we applied the information-theoretic framework in the development of a class of unsupervised exploratory methods aiming at learning patterns of co-expression in gene expression measurement data. The proposed approach is capable of finding patterns of associations between genes by revealing clusters of genes which have a high degree of mutual information, only conditionally on the expression levels of other genes. We expect that the proposed method will serve as a useful aid for biologists in the discovery of unknown gene regulatory pathways.

Chapter 1

Introduction

1.1 Overview and Motivation

The aim of this dissertation is the investigation of novel approaches to statistical learning within an information-theoretic framework, as well as the application of such concepts to the problem of structure discovery in both synthetic and real world data.

The main focus will be on the class of *unsupervised learning* problems [38], and in particular, on the application to such problems of concepts deriving from statistical information theory. From a statistical perspective, unsupervised learning methods deal with the problem of finding a suitable representation for N random variables $\mathbf{x} = \{x_1, \ldots, x_N\}$, given a set of their observations.

In recent years, two closely related unsupervised learning problems have been the subject of an extensive investigation, leading to novel approaches that have combined concepts from the two fields of statistical learning and signal processing. The first, known in the literature as Projection Pursuit (PP) [31][40][32][49], deals with the problem of finding "interesting" lower-dimensional linear projections of high dimensional data, which can possibly reveal certain structure in the data. The second is the field of instantaneous linear blind signal separation [13][18][43][46], which deals with the problem of reconstructing a set of signals from their linear mixtures, when no other assumptions on the mixing process can be made besides its linearity. Both frameworks ultimately deal with the fundamental issue of finding a suitable representation of the data, by seeking a decomposition of the signals that satisfies certain properties.

As we will examine in detail in the following chapters, a large number of learning problems, including those we just mentioned, are centered around two concepts: *statistical independence*, and the related concept of *mutual information content* across random variables. Several learning problems are greatly simplified when statistical independence between all or at least groups of variables can be assumed. Independency has also an interpretation from a neural learning perspective. As it was argued by Barlow in [3], the equivalent concepts of independency and redundancy reduction can be seen as related to an optimum, in terms of efficiency, coding strategy in neurons.

A typical example where conditional independence has a central role is the problem of learning the topology and the associated parameters in Bayesian Networks [76][51][33][39][35]: the fundamental concept in learning with graphical models involves the idea of estimating the properties of a large system by decomposing the estimation problem into a set of smaller, hierarchically structured sub-problems. Consequently, learning the properties of a sub-network becomes independent from the remaining sub-problems, once the overall dependency structure of the network is known. Clearly, this is closely related to the issue of learning from high-dimensional data. The ultimate goal of graphical models, in fact, is to identify a factorial representation of the joint probability density function of a large set of random variables, when the direct estimation of such function is unfeasible because of the dimensionality. In Chapter 7, we will introduce a novel framework capable of simultaneously learning the structure of a bayesian network and the local probability density functions in polynomial time, when a specific model holds.

Several fundamental issues in the field of statistical unsupervised learning are still unsolved. For example, in linear blind signal separation, although the identifiability of the unknown sources under certain hypotheses has been demonstrated, two issues are still under investigation. The first is related to the choice of a suitable statistical model for the unknown signals whose reconstruction is the goal of the method. The second has to do with the properties of the resulting estimation problem, and in particular with the characteristics of the cost function associated to it. Chapter 4, 5 and 6 are dedicated to the introduction of a novel non-parametric estimation framework for blind signal separation and to the investigation of some of its properties. Analogous problems are encountered when attempting to solve the structure learning problem in hierarchical models: incorrect a-priori assumptions on the statistical properties of the unknown generating processes often lead to an erroneous interpretation of the data.

A legitimate question that should be posed is related to the practical relevance of conducting research on novel unsupervised learning techniques. There is evidence that for specific learning problems, such as blind signal separation, unsupervised approaches provide superior performance when compared to supervised approaches. In addition, it is often the case in many fields that the amount of measurement data available greatly exceeds our understanding of the generative process. The phenomenal increase in the amount of biological data that has become available in the last few years is a typical example. The capability of monitoring the expression levels of genes in living cells, using high-throughput standardized experimental procedures, has suddenly made available an enormous quantity of information, whose interpretation is just at its early stages. It will be one of the goals of this dissertation to demonstrate that information theoretic approaches and unsupervised learning can play a central role in shedding some light on complex biological systems (Chapter 8).

1.2 Outline

The dissertation is organized as follows. Chapter 2 provides a brief introduction on the field of statistical learning, clarifying the difference between supervised learning and unsupervised learning, which is the focus of this work. The role of information theory in statistical learning is clarified and several concepts that are fundamental for the rest of the discussion are introduced, such as the entropy of a random variable, the Kullback-Leibler distance between distributions, and the mutual information between two or more random variables. Some properties of these quantities are also described. The chapter also introduces the problem of learning from high-dimensional data and gives a brief review on Projection Pursuit, which constitutes the archetype of all dimensionality reduction techniques.

A detailed introduction to the field of blind signal separation is the focus of Chapter 3. In this chapter, we will review not only the foundations of Independent Component Analysis (ICA), but we will also show in a new and constructive derivation how, under certain assumptions, the apparently dissimilar theoretical frameworks on which most ICA algorithms are based, are indeed equivalent. The current state-of-the-art ICA implementations will be examined and the limitations of linear ICA as well as several open problems will be discussed.

In Chapter 4 we will introduce the main contribution of this dissertation, which consists of a novel blind signal separation framework, based on the non-parametric estimation of the probability density functions of the signals whose reconstruction is attempted. The proposed method is shown to clearly outperform current state-of-the-art ICA techniques in terms of both accuracy of the separation and convergence properties. Such claim is substantiated by several sets of simulation experiments that were conducted comparing the proposed approach to six others popular ICA implementations.

An extension to a fundamental theoretical result in Independent Components Analysis is introduced in Chapter 5. In this chapter, it is shown that the conventional restriction imposed by Comon's identifiability theorem [18], requiring at the most one of the unknown sources to be normally distributed can indeed be lifted. We prove in a novel and constructive proof that when several gaussian and non-gaussian signals are mixed together, the reconstruction of all the non-gaussian signals is always achieved by solving the same estimation problem as in the conventional framework.

One of the fundamental unresolved issues in Independent Component Analysis is to demonstrate in which instances a cost function based on the statistical mutual information can be affected by the presence of spurious local minima. In Chapter 2, we demonstrate the essential uniqueness of the separating solution: therefore, if the cost function is indeed affected by spurious local minima, these will unavoidably result in a failure to obtain the desired source separation. Chapter 6 is dedicated to the investigation of such issue. We were able to demonstrate that for a special class of source distributions, the information-theoretic cost function has no local minima, and the separation of the linearly mixed sources is guaranteed regardless of the initial guess.

Chapter 7 presents an application of the proposed non-parametric unsupervised learning approach. We show that for the special class of linear Bayesian networks characterized by local conditional probability density functions that are non-gaussian, the issue of learning the structure of the network can be cast as a blind signal separation problem. We also introduce certain novel concepts in the process of deriving the proposed method, such as the concept of relaxation graph and quasi-acyclicity of a graph.

The application of information-theoretic unsupervised learning to the discovery of structured patterns in biological signals is the topic of Chapter 8. In particular, we focus on the problem of learning patterns of interactions between genes in the bacterium *Escherichia Coli*. After a brief introduction on the statistical analysis of DNA microarray based measurements of gene expression levels, we discuss some of the issues related to the problem of learning patterns of co-expression between genes. We argue that, while traditional unsupervised learning techniques have found only limited application to this kind of data, current approaches based on simple pairwise correlations can be effectively extended to learning more complex interactions within clusters of genes. Some of the issues related to the computational complexity inherent in this kind of combinatorial exploratory methods are also examined.

In this chapter, we will also introduce *GeneScreen*, an integrated framework for the analysis of gene transcription data and the systematic exploration of patterns of conditional dependence between genes. The results obtained from the application of GeneScreen to gene expression data from real DNA microarray experiments are analyzed and thoroughly discussed at the end of the chapter. The clear patterns of conditional co-expression detected using the proposed approach demonstrate that the proposed approach can provide biologists with a valuable tool for investigating the unknown role of certain genes whose function is not completely understood.

In summary, the specific contributions of this dissertation are as follows:

1. The introduction of a novel framework for Independent Component Analy-

sis, based on the non-parametric kernel based estimation of the probability density functions of the signals whose reconstruction is attempted. The proposed method is truly blind to the particular distribution of the original sources, and does not require the selection of optimal working parameters. An extensive set of simulation experiments established the performance improvement resulting from the proposed method, when compared to other state-of-the-art ICA algorithms.

2. The extension of the classical identifiability theorem for ICA to mixtures of gaussian and non-gaussian signals. A novel and constructive proof is derived to show that, even in presence of multiple gaussian signals in the mixtures, all the non-gaussian signals can be accurately reconstructed by optimizing the conventional cost function based on the minimization of the mutual information between the reconstructed signals.

3. The investigation of the uniqueness of the minima of the information-theoretic cost function for ICA for a class of signals whose distribution follows a specific model. Namely, we demonstrate that for mixtures involving independently and identically distributed signals, whose distribution can be approximated by a Gram-Schmidt expansion involving only fourth-order Hermite polynomials, the resulting cost function is free from spurious local minima.

4. An extension of the entropy power inequality for the class of random variables obtained as linear combinations of independent random variables. We show that when the uniqueness of the minimum of the mutual information as a function of the linear mixing parameters can be established, a *converse entropy power inequality* holds. Specifically, when we consider statistically dependent random variables obtained as linear combinations of other inde-

pendent random variables, the entropy power inequality is always violated.

5. A novel application of blind signal separation to the problem of learning the structure of linear bayesian networks. We prove that for the special class of linear non-gaussian networks, a framework derived from ICA can be devised, which is capable of simultaneously learning the connectivity structure of the network as well as its local probability density models.

6. An application of information theoretic unsupervised learning to biological data is introduced. A combinatorial exploratory method based on the estimation of the *co-information* within clusters of genes is proposed as novel approach for learning conditional dependency structures in gene expression data from DNA microarray experiments.

7. The development of *GeneScreen*, a new tool for the statistical analysis of gene transcription data based on exploring patterns of conditional dependence between gene expression profiles. The proposed method was capable of consistently identifying patterns of conditional co-expression between genes in two organisms, *Escherichia Coli* and *Saccharomyces Cerevisiae.*

Chapter 2

Information Theory and Unsupervised Learning

2.1 Introduction

A fundamental problem in statistical learning, as well as in signal processing in general, is to find a suitable representation of the data, in order to unveil its inherent structure or simply find significative association patterns. We will refer to this problem as *learning from data* [38]. Any learning framework can generally be assigned to one of two broad classes, *supervised* or *unsupervised*.

In supervised learning problems, we make a distinction between input variables (also known as regressor variables) and output variables (also known as regressed variables). The goal is to find a model capable of predicting the output values that are associated to specific input values, based on some sample training data. Representative examples of supervised learning methods are classical statistical regression, least-squares and nearest-neighbor methods, linear discriminant analysis, kernel methods (*e.g.* support vector machines), as well as the corresponding neural-network based implementation. In this work we will not deal with the supervised learning problem and we invite the reader to refer to the rich literature on the subject, for example [38], for a thorough description of the field and the related applications.

We will focus, instead, on the class of *unsupervised learning* problems, and in particular, on the application to such problems of concepts derived from statistical information theory. From a statistical perspective, unsupervised learning methods deal with the problem of finding a suitable representation for N random variables $\mathbf{x} = \{x_1, \dots, x_N\}$, given a set of their observations. For small dimensional problems, $N \leq 3$, methods for estimating directly the joint probability density function (pdf) of $\mathbf{x}$ have been developed [86]. For higher dimensional problems, the direct estimation of the pdf is usually not feasible (with the exception of very special cases, *e.g.* when the random variables are statistically independent), and one must resort to alternative approaches. Examples of unsupervised learning techniques are cluster analysis (*e.g.* K-means algorithm), self-organizing maps, principal component analysis, projection pursuit, and independent component analysis. The reader can refer to [38] for a complete description of these methods.

An interesting perspective on unsupervised learning is provided by a class of methods that seek lower-dimensional projections of high-dimensional data. The idea behind these approaches is that, in some cases, it is possible to find lower-dimensional representations of the data that preserve the original information content, either because the data is intrinsically lower-dimensional, once a suitable representation is identified, or simply because some of the dimensions have a negligible information content. For example, in principal component analysis (PCA) [47] the goal is to find the best successive linear approximations of the data, based on the least-squares principle. Efficient algorithms for PCA have been derived based on well known matrix decomposition techniques, such as singular value decomposition (SVD) [37].

The meaning of information content and the general properties of linear projection methods will become clear once we introduce a few fundamental concepts in the following sections.

2.2 Differential Entropy

Given a scalar random variable x, its *entropy* gives a measure of its uncertainty, and is defined as [19]:

$$H(x) \triangleq -E[\log p_x(u)], \tag{2.1}$$

where $p_x(a)$ can be a continuous or discrete probability density function (pdf). For discrete random variables, we have:

$$H(x) = -\sum_{u_k \in \mathcal{X}} p_x(u_k) \log p_x(u_k), \tag{2.2}$$

where $\mathcal{X}$ is the alphabet of x. In the case of continuous random variables, the quantity in (2.1) is usually referred to as the *differential entropy*, and it is defined as:

$$H(x) = -\int_{-\infty}^{\infty} p_x(a) \log p_x(a) da. \tag{2.3}$$

It can be easily demonstrated that the differential entropy is translation invariant. Consider:

$$y = x + \alpha, \tag{2.4}$$

where α is a constant. We have:

$$F_y(u) = \Pr(y \leq u) = \Pr(x \leq u - \alpha) = F_x(u - \alpha), \tag{2.5}$$

where $F_x(u)$ and $F_y(u)$ are the cumulative distribution functions (cdf) of x and y, respectively. Hence:

$$f_y(u) = f_x(u - \alpha), \tag{2.6}$$

which clearly implies that:

$$H(y) = H(x + \alpha) = H(x). \tag{2.7}$$

Of particular interest is the differential entropy of the gaussian normal distribution:

$$n \sim \mathcal{N}(\mu, \sigma^2) \quad \Longleftrightarrow \quad p_n(u) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\dfrac{(u-\mu)^2}{2\sigma^2}}. \tag{2.8}$$

The expression for its differential entropy can be easily computed using the definition and it is given by:

$$H(n) = \frac{1}{2} \log 2\pi e \sigma^2, \tag{2.9}$$

which shows that for normally distributed random variables the entropy increases as the logarithm of the variance, thus supporting the intuition that the entropy can be viewed a measure of uncertainty of a random variable. In the next section we will re-derive a fundamental result showing that among all random variables with a fixed given variance, the normal *maximizes* the entropy. This is analogous to say that the state of a random process with known variance, is the least predictable when such process has a gaussian distribution, or equivalently that the normal is the least informative among all valid distributions.

2.3 Relative Entropy and Mutual Information

In this section, we will introduce the concept of statistical mutual information between random variables, which is pivotal to rest of the discussion. First let's introduce the closely related concept of *relative entropy*. The relative entropy is a measure of distance between two distributions [19] and, given two probability density functions $p_x(u)$ and $q_x(u)$, it is defined as:

$$D(p||q) \triangleq E_p \left[\log \frac{p_x(u)}{q_x(u)} \right] , \tag{2.10}$$

where E_p represents the expectation taken with respect to the distribution p_x. The quantity in (2.10) is also known as the *Kullback-Leibler* (KL) distance between $p_x(u)$ and $q_x(u)$. The relative entropy is always non-negative and is equal to zero if and only if $p = q$ almost everywhere.

In particular the Kullback-Leibler distance between two continuous random variables can be expressed as:

$$D(p||q) = \int_{-\infty}^{\infty} p_x(u) \log \frac{p_x(u)}{q_x(u)} du . \tag{2.11}$$

The *mutual information* of two random variables x and y is defined as the relative entropy between their joint distribution and the product of their marginal distributions, namely:

$$I(x,y) \triangleq D(p_{xy}||p_x p_y) = E_{xy} \log \frac{p_{xy}(u,v)}{p_x(u)p_y(v)} \ . \tag{2.12}$$

For continuous random variables, it can be written as:

$$I(x,y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p_{xy}(u,v) \log \frac{p_{xy}(u,v)}{p_x(u)p_y(v)} \ du\, dv. \tag{2.13}$$

The mutual information between two random variables can also be expressed in terms of their entropies. From (2.12), using Bayes's rule [19]:

$$\begin{aligned} I(x,y) &= E_{xy} \log \frac{p_{xy}(u,v)}{p_x(u)p_y(v)} \\ &= E_{xy} \log \frac{p_{x|y}(u|v)}{p_x(u)} \\ &= -E_{xy} \log p_x(u) + E_{xy} \log p_{x|y}(u|v) \\ &= -E_x \log p_x(u) - (-E_{xy} \log p_{x|y}(u|v)) \\ &= H(x) - H(x|y). \end{aligned} \tag{2.14}$$

Therefore, the mutual information can be viewed as the decrease in uncertainty on a given random variable, when the value of another random variable is known. The following fundamental theorem can be derived from Jensen's inequality [19]:

Theorem 1 (Non-negativity of the relative entropy) *: For any two probability density functions $p_x(a)$ and $q_x(a)$, it holds that:*

$$D(p||q) \geq 0 \tag{2.15}$$

with equality if and only if $p_x = q_x$ almost everywhere.

A fundamental property of the mutual information follows from this theorem. For any two random variables x and y, it holds that:

$$I(x, y) \geq 0, \tag{2.16}$$

with equality if and only if x and y are *statistically independent.* Combining (2.14) and (2.16), it follows that:

$$H(x|y) \leq H(x), \tag{2.17}$$

again with equality if and only if x and y are independent. Conditioning, therefore, decreases on average the uncertainty on a random process. The definition of mutual information can be extended to the case of N random variables in several ways [4]. For the purpose of this discussion we will use the following straightforward extension of definition (2.12):

$$\begin{aligned} I(x_1, \ldots, x_N) &\triangleq D\left(p_{\mathbf{x}} \,\middle\|\, \prod_{i=1}^{N} p_{x_i}\right) \qquad (2.18) \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p_{\mathbf{x}}(\mathbf{u}) \log \frac{p_{\mathbf{x}}(\mathbf{u})}{\prod_{i=1}^{N} p_{x_i}(u_i)} \, du_1 \ldots du_N. \end{aligned}$$

We will also write this quantity as $I(\mathbf{x})$. Once again $I(\mathbf{x}) \equiv 0$ if and only if $x_1, \ldots, x_N$ are independent random variables. By using the definition of relative entropy (2.10), we can now show a fundamental property of the entropy of the normal distribution [19].

Theorem 2 (Maximum entropy for a given variance) *: Among all distribution with a fixed variance σ^2, the gaussian has the maximum entropy. Consider an arbitrary zero-mean random variable x, such that:*

$$Ex^2 = \sigma^2, \tag{2.19}$$

and let:

$$n \sim \mathcal{N}(0, \sigma^2). \tag{2.20}$$

From the definition of relative entropy, we have that:

$$\begin{aligned} 0 &\leq D(x||n) \\ &= \int p_x \log(p_x/p_n) & (2.21) \\ &= -H(x) - \int p_x \log p_n & (2.22) \\ &= -H(x) - \int p_n \log p_n & (2.23) \\ &= -H(x) + H(n), & (2.24) \end{aligned}$$

where the equality $\int p_x \log p_n = \int p_n \log p_n$ *holds because* x *and* n *have the same moments up the second order.*

The theorem shows that among all distributions with a given variance, the normal has the largest degree of uncertainty. Therefore, the gaussian is the *least* informative among all distributions. One might recognize a connection between the law of large numbers and such property. We will show in the next chapter that, indeed, linearly mixing independent random variables results in an increase in the overall entropy or, equivalently, in the degree of gaussianity of the resulting variables. This result represent the foundation of exploratory methods based on information theoretic cost functions, and, in particular, of blind signal separation approaches based on minimizing the mutual information.

2.4 Deviation from Gaussianity and Projection Pursuit

We demonstrated in the previous section that the entropy can be used as a measure of uncertainty of a process. In particular, we showed that gaussianity can be associated to minimum information content. In this section, we will further develop the relationship between such principle and a class of linear unsupervised learning methods.

2.4.1 The curse of dimensionality

Consider a set of N random variables $\{x_1, \ldots, x_N\}$, and a matrix X (size: $N \times M$) of M independently drawn observations of such variables. We already mentioned in the introductory sections that learning the joint pdf of $\mathbf{x}$ becomes a very difficult problem when $N > 3$ and no prior assumptions on the distribution can be made [38]. In particular, when the number of dimensions N is large, learning the multivariate distribution of a random vector becomes an extremely challenging tasks. The following example due to Huber [40] elucidates the problem. Consider a sample of a 10-dimensional random vector uniformly distributed in the unit ball. The radius of the ball containing 5% of the points is $(0.05)^{0.1} = 0.74$, *i.e.* the ball is predominantly empty. Therefore, in order to learn the distribution from the sample data, an unreasonably large sample size is required. This problem is also known in the literature as the *"curse of dimensionality"*. Therefore one has to resort to alternative approaches in order to learn at least some properties of the unknown distribution.

A class of exploratory methods known as *Projection Pursuit* (PP) [58][50][40] [32][71] aims at finding linear, possibly lower-dimensional, projections that can reveal certain “interesting” features of the data. The choice of a linear operator is justified for several reasons: the results of a linear projection are easy to interpret and their properties can be studied analytically. Moreover, linear operators can reveal higher-dimensional structures present in the data by showing their lower-dimensional “shadows”, but they cannot artificially generate such structures [32].

The general approach in PP is to seek a linear projection operator $\mathbf{a}$ such that the resulting projected data:

$$Y = \mathbf{a}^T X, \tag{2.25}$$

maximizes some index of “interestingness”, also known as projection index:

$$\mathbf{a} = \arg\max_{\mathbf{a}} I(\mathbf{a}), \tag{2.26}$$

subject to certain constraints on $\mathbf{a}$. We will focus for now on one-dimensional projections, thus $\mathbf{a}$ is a $N \times 1$ column vector. The index should possess certain desirable properties: it should be a continuous function of $\mathbf{a}$, and possibly it should be differentiable with continuous derivatives, so that an automatic optimization routine can be designed. In general, for a given dataset, we are interested not only in the single projection that globally optimizes (2.26), but rather we seek a set of projections that result in local strong maxima of the index. An alternative approach, proposed by Friedman in [32] consists in finding one projection at the time by maximizing (2.26) and then removing the structure in the data associated to such optimum. The structure removal procedure should perform the task of making “un-interesting” the projection just found, according to the selected index, while preserving the structure along all other possible projection directions.

A well-known example of projection seeking technique is *Principal Component Analysis* (PCA) [47], where the index of interestingness is given by the variance of the resulting projected data. The first principal component is found by solving the problem:

$$\max_{\mathbf{a}} \ \ \mathrm{var}(\mathbf{a}^T X \mathbf{a}) \qquad (2.27)$$
$$\text{s.t.} \ \ \mathbf{a}^T\mathbf{a} = 1,$$

and the ith principal component is obtained by solving (2.27) with the additional constraints that it must must be orthogonal to the previous $i-1$ components. It can be shown [47] that an analytical closed form solution to the problem of finding the principal components of a matrix exist. This is simply obtained by computing the eigen-decomposition of its sample covariance matrix: the resulting eigen-vectors are the principal components and the corresponding eigen-values are equal to the principal variances.

2.4.2 Centering and Sphering

Translation invariance is one of the desirable features of the projection index. When the index indeed satisfies such property, it is generally more convenient to work with mean subtracted data, so that also any arbitrary linear projection will be zero mean. Given a matrix of observations X, we can compute the centered data matrix as follows:

$$X_c = (I_M - \mathbf{1}_M \mathbf{1}_M^T)X, \qquad (2.28)$$

where I_M is the $M \times M$ identity matrix and $\mathbf{1}_M$ is a column vector of size $M \times 1$, with entries all equal to one. In the previous section, we described PCA as an efficient method to extract maximally variant projection directions. Since we can use PCA to derive scale effects in a given dataset, PP indices often focus on data structures that are scale-invariant. Therefore, besides centering the data, it is a common practice to pre-process the data matrix in such a way that its sample covariance matrix is the identity matrix. This process is conventionally called "*sphering*" as the resulting principal axes of the data will be equal to the axis of an N-dimensional unitary sphere. Given the sample covariance matrix of the data:

$$S = \frac{X_c X_c^T}{M-1}, \tag{2.29}$$

the sphered data matrix is given by:

$$X_s = S^{-1/2} X_c, \tag{2.30}$$

where $S^{-1/2}$ is an inverse square root factor of S, which can be computed for example by taking the eigen-decomposition of S and replacing the eigenvalues with their inverse positive square roots. From (2.30), it is clear that sphering is equivalent to carrying out the principal component analysis of the centered data. When two or more observation vectors in the data matrix X are linearly dependent, the covariance matrix S will not be full-rank in general. In this particular case, one can show that it is still possible to obtain a sphered data matrix simply by reducing the dimensionality of the data to the number of linearly independent components: this is easily accomplished by taking the eigen-decomposition of S and considering only the projection directions associated to its non-zero eigenvalues.

2.4.3 Projection Pursuit Indices

Unless explicitly mentioned, we will assume that the data matrix is centered and sphered and we will simply refer to it as X, rather than X_s. What follows is a review of a selection of PP indices that have been proposed in the literature.

Friedman and Tukey suggested in [31] the following measure:

$$I_{FT}(\mathbf{a}) = s(\mathbf{a})d(\mathbf{a}). \tag{2.31}$$

The two terms composing the index are as follows:

$$s(\mathbf{a}) = \left[\sum_{i=pM}^{(1-p)M} \left(\mathbf{a}^T\mathbf{x}_i - \bar{X}_a\right)^2 / (1-2p)M \right]^{1/2} \tag{2.32}$$

$$\bar{X}_a = \sum_{i=pM}^{(1-p)M} \mathbf{a}^T\mathbf{x}_i / (1-2p)M, \tag{2.33}$$

where $\mathbf{x}_i$ is a column of the matrix X, representing a N-dimensional sample point, and p is a parameter allowing to omit points that lie at each of the projection extremes (the projections $\mathbf{a}^T\mathbf{x}_i$ are assumed to have been sorted for each value of $\mathbf{a}$).

$$d(\mathbf{a}) = \sum_{i=1}^{M}\sum_{j=1}^{M} f(r_{ij}(\mathbf{a}))\mathcal{U}\left(R - r_{ij}(\mathbf{a})\right), \tag{2.34}$$

where:

$$r_{ij}(\mathbf{a}) = |\mathbf{a}^T(\mathbf{x}_i - \mathbf{x}_j)|, \tag{2.35}$$

and:

$$\mathcal{U}(a) = \begin{cases} 1, & a > 0 \\ 0, & a \leq 0 \end{cases} \tag{2.36}$$

The function $f(r)$ should monotonically decreasing in the range $0 \leq r \leq R$, reducing to zero at $r = R$. The term $s(\mathbf{a})$ measures the "spread" of the data, while $d(\mathbf{a})$ measures the local density. The index (2.31) was then maximized via numerical optimization with respect to the parameters defining the projections. The spread term aimed at compensating the scale effects and, as argued separately by Huber [40] and Jones [50], its task is efficiently replaced by a pre-processing step where the data is sphered by using PCA. In [50] Jones re-derives the term $d(\mathbf{a})$ in (2.31) showing that it is minimized by a parabolic density function. Consequently, maximizing Friedman and Tukey's index is therefore equivalent to finding projections of the data that have the largest degree of departure from such distribution.

In [40] Huber classified projection indices into different categories according to their equivariance properties. In particular, he emphasized the importance of affine invariant indices, which are both translation and scale invariance. Starting with the work of Diaconis [25], followed by that of Huber [40], Friedman [32], and Jones [50], the concept that interestingness is related to that of non-normality became predominant. Most heuristic arguments on the validity of such criterion are fundamentally based on the unique properties of the gaussian distribution in relation to the law of large numbers and to the differential entropy. The following is a summary of a few arguments in favor of using non-gaussianity as a projection index (from [40, 32]):

- All projections of a multivariate normal distribution are normal. Therefore, evidence of non-normality in any projection provides evidence against multivariate joint normality.

- The multivariate normal density is elliptically symmetric and is totally specified by its linear structure (mean and covariance).

- Even when a few linear combinations of variables are highly structured (non-normal), most linear combinations will look normally distributed.
- For a fixed variance, the normal distribution has the least information (largest degree of uncertainty)
- For most high-dimensional point clouds, most low-dimensional projections are approximately normal.

Based on these considerations, different methods to measure departure from normality can be devised. In theory, any test statistic for testing normality would serve the purpose. On the other hand, depending on such choice, alternative types of distributions might be favored. Moreover, as we already mentioned, properties such as continuity and differentiability of the index are highly desirable when an automated projection pursuit method is sought.

In [32] Friedman proposed a PP index that measures deviation from normality by using the following transformation of the projected data Y:

$$Z = 2\Phi(Y) - 1, \tag{2.37}$$

where $\Phi(\cdot)$ is the gaussian cumulative distribution function given by:

$$\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-1/2t^2} dt. \tag{2.38}$$

It is easy to show that $-1 \leq Z \leq 1$ and Z will be uniformly distributed in such interval if Y follows a normal distribution. Therefore, the following projection index is suggested:

$$I_F(\mathbf{a}) = \int_{-1}^{1} \left[p_Z(u) - \frac{1}{2} \right] du. \tag{2.39}$$

The author also suggests a procedure for removing the structure associated to an index maximizing projection, while preserving the multivariate structure that is not captured by it.

A PP index based on estimating the differential entropy of the projected data was proposed by Jones *et al.* in [50]. We proved in a previous section that the entropy of a random variable with given variance is maximized when its distribution is gaussian. A natural choice for the projection index is thus given by:

$$I_{JS}(\mathbf{a}) = -H(Y) = \int_{-\infty}^{\infty} p_Y(u) \log p_Y(u) du, \tag{2.40}$$

subject to $\mathbf{a}^T\mathbf{a} = 1$. The estimation of (2.40) requires some form of estimate of the probability density function of the projected data Y. Jones proposes two alternative approaches in order to solve the problem. The first involves an approximation of such pdf using a Gram-Charlier expansion [55], which allows to re-write (2.40) as a polynomial function of the third and fourth order cumulants of the projected data. In particular for one-dimensional projections the following approximation is derived:

$$I_{JS}(\mathbf{a}) \approx (\kappa_3^2 + \kappa_4^2/4)/12, \tag{2.41}$$

where κ_3 and κ_4 are the third and fourth order cumulants [55] of the projected data, respectively, which, for a zero mean and unit-variance random variable Y are defined as:

$$\kappa_3 = E[Y^3], \tag{2.42}$$

$$\kappa_4 = E[Y^4] - 3. \tag{2.43}$$

Jones derives in [50] also an approximation of (2.40) to two-dimensional projections.

The second approach for the calculation of (2.40) suggested in [50] relies upon a kernel density estimate of the unknown probability density function of the projected data. Although the choice of such kernel is arbitrary, Jones' choice of using a gaussian kernel is justified for several reasons. First of all the normal kernel guarantees the smoothness of the projection index as a function of the projection direction, and, in addition, it allows a direct computation of its derivatives. More importantly, as shown by Silverman in [86], the gaussian kernel allows the use of an efficient algorithm for the computation of the projection index, based on the fast fourier transform (FFT) [80], which is particularly useful when dealing with large datasets.

In Chapter 4, we will examine all the details involved in using a kernel density estimator within an optimization procedure.

Chapter 3

Independent Component Analysis: Theory and Practice

For several years, blind signal or system identification problems have drawn the attention of researchers from various fields, including physicists, engineers, computer scientists, but also biologists, linguists, and economists. In the last decade, the specific problem involving the blind separation of linear mixtures of signals has seen a quite unexpected flourishing in the amount of research and literature dedicated to it. In this chapter, rather than having the ambition of pursuing the gargantuan task of reviewing the entire literature of blind signal separation, we will try to provide a concise but hopefully enlightening review of the foundations and most important theoretical results for the case of instantaneous linear mixtures of statistically independent signals. Since no theory can really be appreciated without showing how it can be applied, we will also provide a thorough review of the most significant algorithmic implementations that have been developed to solve such problem. Finally, we will discuss some of the questions that are left open in this field as well as some of its inherent limitations.

3.1 Theoretical Foundations

We will restrict our attention to the class of problems involving a vector of N stationary and *statistically independent* signals $\mathbf{s} = [s_1, \ldots, s_N]^T$, which are mixed by an unknown, full-rank mixing matrix A (size $N \times N$):

$$\mathbf{x} = A\mathbf{s}. \tag{3.1}$$

In the linear blind signal separation (BSS) literature, the signals s_i are conventionally referred to as the "sources", A as the "mixing matrix", and x_i as the "mixtures". The reconstruction of the original sources is attempted through a linear projection of the type:

$$\mathbf{y} = W\mathbf{x}, \tag{3.2}$$

where W is usually referred to as the separating or "un-mixing" matrix.

3.1.1 Model Identifiability

The fundamental question in blind signal separation is whether the model (3.1) is identifiable and under what conditions, considering that both the random vector $\mathbf{s}$ and the mixing matrix A are unknown. Clearly, the existence of a linear solution to the problem, given by $W = A^{-1}$, seems to suggest that seeking non-linear solutions to the identification problem posed by (3.1) may not be required.

The following fundamental result due to Darmois [23] and Skitovich [87], is behind the main identifiability result in BSS:

Theorem 3 (Darmois-Skitovich Theorem) *: Given a N-dimensional random vector $\mathbf{s} = [s_1, \ldots, s_N]^T$, with mutually statistically independent components, consider any two arbitrary linear combinations of such components:*

$$y_1 = a_1 s_1 + \ldots + a_N s_N \tag{3.3}$$

$$y_2 = b_1 s_1 + \ldots + b_N s_N. \tag{3.4}$$

If y_1 and y_2 are statistically independent, then it must hold that for every index i such that $a_i \neq 0$ then $b_i \neq 0$, s_i has a gaussian distribution.

By applying Theorem 3, Comon [18] proved the following theorem, which is at the basis of Independent Component Analysis (ICA):

Theorem 4 (Comon Identifiability Theorem) *: Let $\mathbf{s}$ be a N-dimensional vector with independent components, of which at most one is gaussian. Let C be an orthogonal $N \times N$ matrix and $\mathbf{y}$ the vector $\mathbf{y} = C\mathbf{s}$. Then, the following three properties are equivalent:*

i. The components y_i are pairwise independent.

ii. The components y_i are mutually independent.

iii. $C = \Lambda P$, where Λ is a diagonal matrix (with non-zero entries in the diagonal) and P is a permutation matrix.

The main result following Theorem 4 is that in order to estimate the unknown signals $\{s_1, \ldots, s_N\}$, it suffices to seek a linear projection operator W that minimizes the pairwise dependency between the reconstructed components. Such result states the **equivalence between blind signal separation and independent component analysis** when the linear model assumption holds. We will show at the end of this chapter that such equivalence no longer holds when such assumption is violated.

A fundamental indeterminacy affects the identifiability properties of the problem: the unknown sources can be reconstructed only up to scaling constants and a permutation of the signals. This fact can be immediately recognized as a consequence of Theorem 3: clearly, since the estimation criterion simply pursues a set of independent signals, any arbitrary permutation of such signals or multiplication by constant factors will equally satisfy the criterion. We will see in the following chapters that such indeterminacy is not overly restrictive. When suitable constraints are included in the selected contrast function, in fact, the relationship between mixture signals and reconstructed sources becomes uniquely determined.

Comon's identifiability principle (Theorem 4) reveals one of the limitations of ICA for blind signal separation, *i.e.* the requirements that at the most one of the original sources is gaussian. Cruces et al. [20] recently extended this result showing that when linear mixtures including an arbitrary number of gaussian and non-gaussian sources are considered, all the non-gaussian signals can be reconstructed, up to scaling and a permutation, using one of the conventional ICA contrast functions. In Chapter 5 of this dissertation, we will present certain novel results on the properties of the mutual information of the set of reconstructed signals, when multiple gaussian sources are considered.

3.1.2 Uniqueness of the Separating Operator

Theorem 4 shows under what conditions the estimation problem defined by (3.1) is identifiable. We will focus here on showing that once the indeterminacy due to scaling and permutations is taken into account, the linear projection operator resulting in the source separation is *unique*.

First let's consider a pre-processing step that is quite common in ICA and that will greatly simplify our discussion from now on. The idea is to subtract the mean of the mixtures $\mathbf{x}$ in such a way that:

$$\hat{x}_i = x_i - E[x_i] \implies E[\hat{x}_i] = 0, \qquad i = 1, \ldots, N \ , \tag{3.5}$$

and to sphere the data as in (2.30):

$$\tilde{\mathbf{x}} = S^{-1/2}\hat{\mathbf{x}}, \tag{3.6}$$

where $S^{-1/2}$ is an inverse square root factor of $S = E[\mathbf{x}\mathbf{x}^T]$, which can be computed from the covariance matrix of the mixture variables as detailed in Chapter 2. It is possible to show that if the original sources are at least uncorrelated and the mixing matrix is non-singular, then the covariance matrix of the mixture data is also non-singular, and this pre-processing step is always applicable. Therefore, from now on we will assume that the mixture data satisfies the following two properties:

$$\text{i.} \quad E[\mathbf{x}] = 0 \tag{3.7}$$

$$\text{ii.} \quad E[\mathbf{x}\mathbf{x}^T] = I_N. \tag{3.8}$$

We can now prove that when the mixture data satisfies such properties the linear operator resulting in the source separation must belong to the manifold of orthogonal matrices. If the reconstructed signals are mutually independent, it must hold in particular that:

$$E[\mathbf{y}\mathbf{y}^T] = \mathrm{diag}\{\lambda_1, \ldots, \lambda_N\}. \tag{3.9}$$

Since the scaling of the reconstructed signals is anyway arbitrary, we can make the non-restrictive assumption that $\lambda_i = 1$ $(i = 1, \ldots, N)$. Therefore, we have:

$$E[\mathbf{y}\mathbf{y}^T] = E[(W\mathbf{x})(W\mathbf{x})^T] = WE[\mathbf{x}\mathbf{x}^T]W^T = I. \tag{3.10}$$

Since $E[\mathbf{x}\mathbf{x}^T] = I$, it must hold that:

$$WW^T = I, \tag{3.11}$$

thus W must be *orthogonal*. We can now continue by proving that the separating matrix W is unique up to a permutation of its rows (notice that the scaling issue has already been included in the normalization of the reconstructed signals). Consider the combination of the mixing, sphering and unmixing systems:

$$\mathbf{y} = WS^{-1/2}A\mathbf{s}. \tag{3.12}$$

The sphering step, in particular, ensures that:

$$E[\mathbf{x}\mathbf{x}^T] = E[S^{-1/2}A\mathbf{s}\mathbf{s}^TA^T(S^{-1/2})^T] = S^{-1/2}AD_{ss}A^T(S^{-1/2})^T = I, \tag{3.13}$$

where $D_{ss} = E[\mathbf{s}\mathbf{s}^T]$ must be a diagonal matrix with non-zero entries in the diagonal. Since both matrices A and S are assumed non-singular, we can write:

$$D_{ss} = A^{-1}S^{1/2}(S^{1/2})^T(A^{-1})^T. \tag{3.14}$$

Since the random vector $\mathbf{s}$ has mutually independent components, from Theorem 4 and (3.12), it implies that $\mathbf{y}$ will have mutually independent components if and only if:

$$WS^{-1/2}A = \Lambda P \implies W = \Lambda P A^{-1}S^{1/2}, \tag{3.15}$$

where Λ is a diagonal matrix and P a permutation matrix. Now notice that, since $E[WW^T] = I$, we have that:

$$E[\Lambda P A^{-1} S^{1/2} (S^{1/2})^T (A^{-1})^T P^T \Lambda] = \Lambda P D_{ss} P^T \Lambda = \Lambda D_{ss} \Lambda = I. \tag{3.16}$$

Therefore it must hold that $[\Lambda]_{ii} = \pm [D_{ss}]_{ii}^{-1/2}$, hence (3.15) reduces to:

$$W = D_{ss}^{1/2} \hat{P}, \tag{3.17}$$

where $\hat{P}$ is a permutation matrix that can have positive as well as negative entries. In summary, we proved the following theorem.

Theorem 5 (Uniqueness Theorem) *: Given a N-dimensional random vector* $\mathbf{s}$ *with statistically independent components, let A be a full-rank $N \times N$ matrix and* $\mathbf{x}$ *the vector obtained as:*

$$\mathbf{x} = A\mathbf{s}. \tag{3.18}$$

The linear operator W reconstructing the original signals (up to a permutation and a sign inversion) is essentially unique and is given by the following expression:

$$W = D_{ss}^{1/2} \hat{P}, \tag{3.19}$$

where $D_{ss}^{1/2}$ is a square-root factor of the diagonal covariance matrix of $\mathbf{s}$*, and $\hat{P}$ is a generalized permutation matrix, where the non-zero entries can be both positive and negative.*

The importance of this result will become clear in the following sections, when several types of objective functions used to enforce the independence of the reconstructed signals will be examined. A fundamental issue in ICA is related to the presence of sub-optimal extrema of the selected cost function. Since the optimal separating operator is essentially unique, local minima of the resulting optimization procedure will, in general, result in a failure to produce the desired source separation.

3.2 Contrast Functions for ICA

In the previous section, we have determined that the blind separation of linearly mixed independent stationary signals can be achieved by seeking a linear projection operator that maximizes the statistical independence of the reconstructed signals. The next problem we are faced with is the selection of a measure of statistical independence and of its associated objective function (also referred to as the *"contrast function"*). In this section, we will review the most important criteria that have been investigated for ICA, and we will show that, when certain hypotheses are satisfied, all these criteria are analytically equivalent.

3.2.1 InfoMax

The first framework we are going to consider is the entropy maximization approach suggested by Bell and Sejnowski in [5], which is popularly known in the ICA community as *InfoMax.* InfoMax finds its origins in the "information preservation" principle described by Linsker in [63], which results in a network that maximizes the mutual information between its output and the signal portion of the input. Linsker's algorithm aims at extracting salient input features by maximizing the information transfer in low noise conditions. We have shown in Chapter 2 that the following equality holds:

$$I(\mathbf{y}, \mathbf{x}) = H(\mathbf{y}) - H(\mathbf{y}|\mathbf{x}), \tag{3.20}$$

where $\mathbf{x}$ is the network vectorial input and $\mathbf{y}$ is its vectorial output, and $I(\mathbf{y}, \mathbf{x})$ is their mutual information. When $\mathbf{x}$ and $\mathbf{y}$ are continuous random variables, the entropies $H(\mathbf{y})$ and $H(\mathbf{y}|\mathbf{x})$ are in general unbounded, thus the maximum of $I(\mathbf{y}, \mathbf{x})$ is not well-defined. When additive noise is considered in the model we have that:

$$\mathbf{y} = W\mathbf{x} + \mathbf{n}. \tag{3.21}$$

Thus, given the noise covariance matrix C_{nn}, it holds [70]:

$$H(\mathbf{y}|\mathbf{x}) = H(\mathbf{n}) \leq \frac{1}{2}\log(2\pi e C_{nn}) \tag{3.22}$$

satisfied with equality if the noise $\mathbf{n}$ is a gaussian process. In this case it is evident that:

$$\frac{\partial I(\mathbf{y}, \mathbf{x})}{\partial w_{ij}} = \frac{\partial H(\mathbf{y})}{\partial w_{ij}}. \tag{3.23}$$

since $H(\mathbf{n})$ does not depend on the network parameters. Nadal and Parga [70] showed that even in the *low-noise* limit case the problem of maximizing the mutual information, or equivalently the entropy H($\mathbf{y}$), has no solution for the class of unbounded transfer functions. In [70], a set of possible choices for the constraints that could be added to the optimization framework is examined. Moreover, it is shown that, in the case of non-linear transfer functions, for a suitable choice of the non-linearity, the maximization of the mutual information between inputs and outputs implies a factorial form for the output distribution. We will give a justification of this result below without giving a formal proof.

This result is used by Bell and Sejnowski to justify the InfoMax algorithm for ICA [5]. In particular, they propose the introduction of a fixed non-linearity, which has the function of clipping the signal outputs:

$$\mathbf{z} = \mathbf{f}(\mathbf{y}) = \mathbf{f}(W\mathbf{x}). \tag{3.24}$$

Typical choices for $f_i(u)$ are $\tanh(u)$ or the logistic function[1]. In general, such non-linear functions map the real line to the interval $(0, 1)$ and are monotonously increasing. Thus, if the f_i are differentiable they can be considered as the cumulative distribution functions of some probability density functions [12]. The importance of such property will be clarified below. In order to justify certain common choices for the non-linearity, one has to derive a different expression for $H(\mathbf{z})$. Using basic information theory equalities, we can write:

$$H(z_1, \ldots, z_N) = H(z_1) + \cdots + H(z_N) - I(z_1, \ldots, z_N). \tag{3.25}$$

The output entropy is, thus, maximized when, simultaneously, the marginal output entropies are maximized and the mutual information between the output variables is minimized. Because of the squashing non-linearity, the values of the output variables are lower and upper bounded. It can be shown that the distribution that maximizes the entropy for amplitude-bounded random variables is the uniform [19]. Therefore, the $H(z_i)$ are singularly maximized when[2]:

$$p_{z_i}(u_i) = \frac{p_{y_i}(u_i)}{\left|\frac{df_i(u_i)}{du_i}\right|} = 1, \tag{3.26}$$

which implies:

[1] f(u)= 1/(1+e^{-u})

[2] Assuming that each squashing function maps the real line to the interval $(0, 1)$.

$$p_{y_i}(u_i) = \left| \frac{df_i(u_i)}{du_i} \right| . \tag{3.27}$$

Thus, the maximization of the output entropy requires the derivative of each nonlinearity to be equal to the pdf of the corresponding output signal. When such hypothesis is indeed satisfied, the maximization of $H(\mathbf{z})$ implies the minimization of $I(\mathbf{z})$, which results, implicitly, in the minimization of $I(\mathbf{y})$, since the fixed nonlinearities cannot introduce dependence between the output variables (although the expression of $I(\mathbf{z})$ is, in general, different from the expression of $I(\mathbf{y})$ when not in the proximity of the global minimum).

We will now derive another expression for $H(\mathbf{z})$ that will be used to clarify the relationship between InfoMax and other contrast functions for ICA. Simply notice that for one-to-one mappings between random variables, it holds:

$$p_{\mathbf{z}}(\mathbf{u}) = \frac{p_{\mathbf{x}}(\mathbf{u})}{|\det J|} \tag{3.28}$$

where $J = \partial z_i / \partial x_j$ is the Jacobian of the transformation and in the case of (3.24) it is equal to:

$$J = \begin{bmatrix} f_1'(u_1) & & 0 \\ & \ddots & \\ 0 & & f_N'(u_N) \end{bmatrix} W \tag{3.29}$$

Therefore, since the first term on the right hand side of (3.29) is a diagonal matrix:

$$\det J = (\det W) \cdot \prod_{i=1}^{N} f_i'(u_i) \tag{3.30}$$

We can thus re-write H($\mathbf{z}$) as:

$$
\begin{aligned}
H(\mathbf{z}) &= -\int p_{\mathbf{z}}(\mathbf{u}) \log p_{\mathbf{z}}(\mathbf{u}) d\mathbf{u} && (3.31)\\
&= -\int p_{\mathbf{x}}(\mathbf{u}) \log \left(\frac{p_{\mathbf{x}}(\mathbf{u})}{|\det J|} \right) d\mathbf{u} && (3.32)\\
&= -\int p_{\mathbf{x}}(\mathbf{u}) \log \left(\frac{p_{\mathbf{x}}(\mathbf{u})}{|\det W \cdot \prod_{i=1}^{N} f_i'(a)|} \right) d\mathbf{u} && (3.33)\\
&= H(\mathbf{x}) + \int p_{\mathbf{x}}(\mathbf{u}) \log |\det W \cdot \prod_{i=1}^{N} f_i'(u_i)| \; d\mathbf{u}. && (3.34)
\end{aligned}
$$

The integral in (3.34) is well defined if and only if:

$$
\int_{-\infty}^{\infty} f_i'(u_i) \; du_i < \infty \quad , \quad i = 1, \ldots, N \qquad (3.35)
$$

Both $\tanh(u)$ and the logistic function satisfy this property since their derivatives integrate to 1. In particular when all the $f_i(u_i)$ are monotonically increasing, equation (3.34) can be rewritten as:

$$
H(\mathbf{z}) = H(\mathbf{x}) + \int p_{\mathbf{x}}(\mathbf{u}) \log \left(|\det W| \cdot \prod_{i=1}^{N} f_i'(u_i) \right) \; d\mathbf{u}. \qquad (3.36)
$$

Therefore, since $H(\mathbf{x})$ does not depend on the network parameters, maximizing $H(\mathbf{z})$ is equivalent to maximizing the second term in (3.36).

3.2.2 Mutual Information Minimization

A second approach to ICA is derived from Barlow's redundancy reduction principle, which was suggested by the author as a plausible coding strategy in neurons [3]. The goal of Barlow's approach is the design of a *factorial code*, where every output unit is statistically independent from every other unit. This approach is equivalent to a straightforward application of the mutual information as contrast function on the output signals, as suggested in [18].

In Chapter 2, we have shown that a valid definition of mutual information between a set of N random variables is given by the relative entropy of the joint probability density function and the product of its marginal pdfs:

$$I(\mathbf{y}) = D\left(p_y(\mathbf{u}) \,\middle\|\, \prod_{i=1}^{N} p_{y_i}(u_i)\right) = \int p_y(\mathbf{u}) \log\left(\frac{p_y(\mathbf{u})}{\prod_{i=1}^{N} p_{y_i}(u_i))}\right) \, d\mathbf{u} \ . \tag{3.37}$$

This quantity is always positive and it is equal to zero if and only if the joint pdf is equal to the product of the marginals. Following Theorem 4, the source separation can be obtained by seeking a linear projection of the type:

$$\mathbf{y} = W\mathbf{x}. \tag{3.38}$$

The expression that relates the pdfs of $\mathbf{y}$ and $\mathbf{x}$ is simply given by:

$$p_y(\mathbf{u}) = \frac{p_x(\mathbf{u})}{|\det W|}. \tag{3.39}$$

Consequently equation (3.37) can be written as:

$$\begin{aligned} D\left(p_{\mathbf{y}} \,\middle\|\, \prod_{i=1}^{N} p_{y_i}\right) &= \int p_{\mathbf{x}}(\mathbf{u}) \log\left(\frac{p_{\mathbf{x}}(\mathbf{u})}{|\det W| \cdot \prod_{i=1}^{N} p_{y_i}(u_i)}\right) d\mathbf{u} \\ &= -H(\mathbf{x}) - \int p_{\mathbf{x}}(\mathbf{u}) \log\left(|\det W| \cdot \prod_{i=1}^{N} p_{y_i}(u_i)\right) d\mathbf{u} \qquad (3.40) \end{aligned}$$

Clearly, the two objective functions (3.36) and (3.40) are equivalent if:

$$\boxed{\prod_{i=1}^{N} f_i'(u_i) = \prod_{i=1}^{N} p_{y_i}(u_i)} \tag{3.41}$$

and, in particular, when:

$$f_i(u) = \int_{-\infty}^{u} p_{y_i}(v)dv \qquad i = 1, \ldots, N. \tag{3.42}$$

Obradovic and Deco [74] derived similar conclusions on the equivalence between these two BSS principles. The importance of accurately estimating the marginal probability density functions $p_{y_i}(u_i)$ was recognized in [70]. Bell and Sejnowski also observed in their paper that different types of non-linearity were required in order to separate sources with different probability density functions [5].

3.2.3 Maximum Likelihood Estimation

In the previous sections, we clarified the relationship between the contrast functions and the distributions of the sources whose separation is attempted. We established that the equivalence between InfoMax and the mutual information minimization principle holds when the derivatives of the squashing non-linearities match the pdfs of the independent sources. The relationship between the maximum likelihood (ML) principle and the contrast functions examined so far will be the subject of this section.

Let's first recall how the ML principle can be applied to the estimation framework associated with the blind source separation problem. We will follow here the derivation formulated by Cardoso in [12]. Let us denote a set of independent realizations of the mixtures $\mathbf{x}$ of finite size M, as $\mathbf{x}_1, \ldots, \mathbf{x}_M$. If we assume the following parametric model for the density of $\mathbf{x}$:

$$\mathcal{P} = \{p_\theta(\mathbf{x}) | \theta \in \Theta\}, \tag{3.43}$$

then the normalized log-likelihood of a model, given the observations is equal to:

$$L_M(\theta) \triangleq \frac{1}{M} \log \prod_{m=1}^{M} p_\theta(\mathbf{x}_m) = \frac{1}{M} \sum_{m=1}^{M} \log p_\theta(\mathbf{x}_m). \tag{3.44}$$

Since (3.44) can be seen as the sample average of $\log p_\theta(\mathbf{x})$, by the law of large numbers, it converges in probability to its statistical expectation, when the sample size becomes infinite:

$$L_M(\theta) \xrightarrow{\mathcal{P}} L(\theta) \triangleq E[L_M(\theta)] = \int_{-\infty}^{\infty} p(\mathbf{x}) \log p_\theta(\mathbf{x}) d\mathbf{x}. \tag{3.45}$$

The expression above can be expanded as follows [12]:

$$L(\theta) = -D(p(\mathbf{x}) || p_\theta(\mathbf{x})) - H(\mathbf{x}). \tag{3.46}$$

Considering that the relative entropy is invariant under an invertible transformation of the sample space [19], (3.46) can be written as:

$$L(\theta) = -D\left(p(\mathbf{y}) ||\, p(\tilde{\mathbf{s}})\right) + \text{const.} \ , \tag{3.47}$$

where $\mathbf{y} = WA\mathbf{s}$ are the reconstructed mixtures, assuming the correct model for the density functions of the sources, and $\tilde{\mathbf{s}} = W\mathbf{x}_\theta$ are the estimated source distributions, assumed mutually independent. The following decomposition of the KL distance holds for any vector $\tilde{\mathbf{s}}$ with independent components [19]:

$$D\left(p(\mathbf{y})||\,p(\tilde{\mathbf{s}})\right) = D\left(p(\mathbf{y})||\,p(\tilde{\mathbf{y}})\right) + D\left(p(\tilde{\mathbf{y}})||\,p(\tilde{\mathbf{s}})\right), \tag{3.48}$$

where $p(\tilde{\mathbf{y}}) = \prod_{i=1}^{N} p(y_i)$ is the density function obtained by taking the product of the marginal densities of $p(\mathbf{y})$. Therefore $L(\theta)$ can be expressed as:

$$L(\theta) = -I(\mathbf{y}) - \sum_{i=1}^{N} D\left(p(\tilde{y_i})||\,p(\tilde{s_i})\right) + \text{const.} \tag{3.49}$$

Hence, the maximum likelihood principle and the mutual information minimization principle are equivalent when the model assumed for the density function of the source vector is exact, so that the second term on the right hand side of (3.49) is identically zero.

3.2.4 Negentropy Index

A final perspective to the blind source separation problem is, once again, related to the concept of *gaussianity*. We will show here that, when the mixture data is whitened using a sphering procedure of the type described in Chapter 2, the mutual information minimization principle is equivalent to the entropy index for projection pursuit, which measures deviation from gaussianity. This apparently surprising result has a straightforward interpretation: the mixtures of independent input signals $\{s_1, \ldots, s_N\}$ tend to have a larger entropy than the original signals, or from a statistical perspective tend to be 'more gaussian' than the original signals, where gaussianity is measured in terms of differential entropy.

This concept can be formalized by considering the *entropy power inequality* [19]. The entropy power of a scalar random variable s is defined as:

$$N(s) \triangleq \frac{1}{2\pi e}\, e^{2H(s)} \tag{3.50}$$

Given two independent random variables s_1 and s_2, the entropy power inequality states that:

$$N(s_1 + s_2) \geq N(s_1) + N(s_2), \tag{3.51}$$

with equality holding if and only if s_1 and s_2 are both normal. The inequality (3.51) can be used to prove the convexity of the entropy under a covariance preserving transformation, i.e. given $0 \leq \lambda \leq 1$, it holds that [24]:

$$H(\lambda s_1 + \sqrt{1-\lambda^2} s_2) \geq \lambda^2 H(s_1) + (1-\lambda^2) H(s_2), \tag{3.52}$$

and analogously:

$$H(-\sqrt{1-\lambda^2} s_1 + \lambda s_2) \geq (1-\lambda^2) H(s_1) + \lambda^2 H(s_2), \tag{3.53}$$

(note that $H(as) = H(s) + \log|a|$, a being a scalar parameter). Simply by adding (3.52) and (3.53), we obtain:

$$H(y_1) + H(y_2) \geq H(s_1) + H(s_2). \tag{3.54}$$

which, indeed, formalizes the idea that the marginal entropies increase on average when linearly mixing independent random variables. A natural question is, then, whether a contrast function minimizing the sum of the marginal entropies of the reconstructed signals can be used to seek independent components. The answer is found by considering the following expansion of the mutual information:

$$\begin{aligned} I(y_1, \ldots, y_N) &= \sum_{i=1}^{N} H(y_i) - H(y_1, \ldots, y_N) \\ &= \sum_{i=1}^{N} H(y_i) - H(x_1, \ldots, x_N) - \log|\det(W)|, \end{aligned} \tag{3.55}$$

where $\mathbf{y} = W\mathbf{x}$, and recalling that, when W is non-singular, it holds that $H(\mathbf{y}) = H(\mathbf{x}) + \log|\det(W)|$. We already proved in (3.11) that, when the mixture vector $\mathbf{x}$ is sphered (*i.e.* $E[\mathbf{x}\mathbf{x}^T] = I$), then W must be orthogonal. In this case, $\log|\det(W)| \equiv 1$, and the following two problems are equivalent:

$$\min_W I(y_1, \ldots, y_N) \iff \min_W \sum_{i=1}^{N} H(y_i) \ , \tag{3.56}$$

since $H(\mathbf{x})$ does not depend on W. Therefore, the source separation can indeed be obtained by minimizing the sum of the marginal entropies of the reconstructed signals, or equivalently by maximizing the deviation of the marginal densities of $\mathbf{y}$ from normality.

3.3 Cumulant Based Approximations

All the contrast functions for ICA described in the previous sections share the problem that some estimate of the unknown probability density functions of the source signals is required in the estimation framework. A large number of studies [13][44][45] have been directed at assessing the robustness of such estimation framework to incorrect assumptions on the source statistics.

A simple but effective solution to this problem consists in approximating the selected contrast function using high-order moments of the reconstructed signals. In general, even without explicitly referring to one of the separation principles mentioned above, it is possible to seek a linear projection with independent components simply by observing that the cross-cumulants [14] of the reconstructed signals are expected to be zero when such signals are indeed independent.

In [18] Comon derives an approximation of the mutual information based on the Edgeworth [53] expansion of the unknown density functions. In this paper the issue of selecting a suitable subset of all possible cross-cumulants [55] of the reconstructed signals is raised for the first time. Notice that, for example, the straightforward minimization of just third and fourth order cross-cumulants would require the computation of the optimal solution of a fairly large high-dimensional problem. Comon shows that a functional based on r-order ($r \geq 3$) cumulants only is a valid ICA contrast function provided that the reconstructed variables have at the most one null marginal cumulant of order r.

An effective solution to the problem of selecting a suitable set of high-order cross-cumulants is the one proposed by Cardoso for example in [14], which represents the basis for *JADE*, a well-known ICA implementations. The author derives approximations of the most important ICA contrast functions (e.g. maximum likelihood and minimum mutual information) based uniquely on second and fourth order cumulants. Because of the linear modeling assumption, Cardoso shows that by selecting a suitable set of high-order cumulants, the optimization of the resulting contrast function can be posed as a joint matrix diagonalization problem (namely N^2 matrices, each of size $N \times N$ for the separation of N signals). A Jacobi optimization [37] algorithm is proposed in order to estimate the optimal orthogonal unmixing matrix when operating on sphered data. Similarly to the cost function proposed in [18], this method requires as necessary condition that at the most one source signal has a zero fourth order cumulant.

A moment based method approximating the negentropy maximization principle is proposed by Girolami *et al.* in [36]. The authors show that the negentropy projection pursuit index accurately approximates the score function proposed by Bell and Sejnowski in [5], for the class of super-gaussian (positive kurtosis) distributions. Moreover, a generalization of the method to both sub-gaussian and super-gaussian signals is provided that employs an adaptive non-linearity, which is formed by estimating on-line the kurtosis of the latent variables. The idea of adapting the sign of the non-linearity during the estimation procedure is also the basis of the Extended InfoMax algorithm [61], which will be examined later on.

A thorough review of all moment based approaches for ICA is beyond the scope of this dissertation and we invite the reader to refer to the vast literature on the subject (e.g. [43]) for further details. Regardless of the specific choice for the contrast function, all moment based algorithms present several advantages and disadvantages when compared to approaches based on the estimation of the true density functions of the unknown sources. A major advantage of cumulant based techniques is that their computational complexity is rather unaffected by the sample size, since they generally rely on pre-computed statistics of the sample data, thus holding an edge in terms of computational payload when compared to other methods. However, restrictions on the statistical properties of the unknown sources, which depend on the specific choice of the set of cumulants, may limit their applicability as generic ICA frameworks. Speculations on the increased sensitivity of moment based approaches to the presence of outliers have been raised in the literature [2][41][61], but they are largely unsubstantiated.

3.4 ICA Algorithms: State of the Art

In this section, we will give a brief review of the algorithms representing the state-of-the-art in ICA. A more in depth analysis of numerous ICA frameworks can be found for example in [43][59][84].

3.4.1 Jade

Among all the cumulant based algorithms that have been developed for ICA, Cardoso's JADE [14, 15] has gained a vast popularity probably because of its capability of combining a generally reliable source separation, with an above average speed of execution. JADE is based on a Jacobi technique (namely, a sequence of planar Givens rotations) whose goal is the joint diagonalization of a maximal set of fourth-order cumulant matrices. Formally, the algorithm attempts the minimization of the following objective function [14]:

$$\phi_{JADE}(\mathbf{y}) \triangleq \sum_{ijkl \neq ijkk} \mathcal{R}^2_{ijkl}(\mathbf{y}), \tag{3.57}$$

where $\mathcal{R}_{ijkl}(\mathbf{y})$ is the fourth-order cross-cumulant of $\{y_i, y_j, y_k, y_l\}$, and is given by [15]:

$$\begin{aligned} \mathcal{R}_{ijkl}(\mathbf{y}) &\triangleq \mathrm{Cum}(y_i, y_j, y_k, y_l) \qquad (3.58) \\ &= E[y_i y_j y_k y_l] - E[y_i y_j]E[y_k y_l] - E[y_i y_k]E[y_j y_l] - E[y_i y_l]E[y_j y_k] \end{aligned}$$

The criterion has the restriction, common to other approaches based uniquely on fourth-order moments, that at the most one of the unknown sources is allowed to be zero-kurtotic.

3.4.2 InfoMax ICA and Extensions

The algorithm described by Bell and Sejnowski in [5] was characterized by a fixed choice of the squashing non-linearity, suitable only for the separation of highly-kurtotic signals. Lee et al. developed an algorithm capable of separating mixtures of super-gaussian and sub-gaussian sources [61] using a fixed non-linearity ($\tanh(u)$) and switching its sign according to the sign of the kurtosis of the underlying distribution. This approximate solution resulted in an improved capability of separating different types of sources although the criterion used for switching the non-linearity sign is heuristic.

3.4.3 Flexible Algorithms and Parametric ICA

More recently Karvanen *et al.* developed a maximum likelihood estimation framework capable of separating a wide class of source distributions [29]. These are modeled adaptively by either a Pearson model or an extended generalized lambda distribution. The most important aspect of this approach is that several types of sources that can be extracted by this algorithm are in general not distinguishable using a more conventional ICA approach. For example the method developed in [29] can reliably separate non-gaussian sources with identically zero kurtosis but non-zero skewness. One of the problems associated with this complex model is that the stability properties of the optimization procedure are unclear. Moreover, as explained by the authors, some heuristics are introduced in the algorithm in order to force the use of a *tanh* type of non-linearity when the source is clearly super-gaussian or sub-gaussian.

Alternative parametric approaches to Independent Component Analysis that employ a more flexible model for the pdf of the source signals have been introduced [1][54][89]. These methods usually consist of a parametric density estimation technique that alternates with a cost function optimization step in an iterative approximation framework. Although these approaches tend to outperform standard algorithms in specific cases (e.g. skewed sources), neither their convergence properties, nor their capability of modeling arbitrarily distributed sources, have been fully assessed.

3.4.4 Kernel Based Approaches

The recent introduction of kernel-based methods, such as Bach and Jordan's Kernel-ICA [2], demonstrate that finding a compromise between computational complexity, performance and strong convergence properties of a blind signal separation framework is still an open and challenging problem. The method introduced in [2] uses a contrast function based on canonical correlations in a reproducing kernel Hilbert space. In Chapter 4 we will compare this ICA framework with the proposed approach, showing that not only the proposed algorithm outperforms Kernel-ICA in terms of separation performance, but it also retains a significant advantage in terms of computational complexity and added capability of obtaining an estimate of the probability density functions of the source signals as a by-product of the estimation procedure.

3.5 Challenges and Limitations

The fundamental identifiability theorem for ICA proves that the reconstruction of independent signals from a set of their linear mixtures is a feasible estimation problem. In this chapter, we identified two fundamental issues that inherently limit such result, the first being the capability of accurately modeling the distributions of the unknown sources, the second being related to the properties of the resulting cost function and, in particular, to the problems associated with the risk of incurring in non-global optima of such contrast functions. One of the goal of this dissertation is to investigate such fundamental issues and provide novel solutions to such problems. This will be the topic, among others, of Chapters 4, 5, and 6.

Certain limitations are inherent in the ICA model and should be pointed out.

The most fundamental of such limitations is that Independent Component Analysis cannot be regarded as a first-order linear approximation of a non-linear estimation problem. Clearly, when the linearity assumption is violated, the identifiability theorem no longer holds and seeking independent components in the data is no longer equivalent to performing blind signal separation. This fact is better understood by observing that when the original independent signals are mapped through certain non-linear transformations it is often the case that several sets of independent components that are completely unrelated to the original ones can be identified. This is a consequence of the fact that there is an infinitely large number of non-linear mappings that preserve statistical independence.

Therefore, one of the open problems in ICA is to show whether a class of non-linear problems exist such that a suitable modification of the conventional estimation frameworks is still capable of providing the desired source separation result.

Chapter 4

Non-Parametric ICA

In recent years, Independent Component Analysis (ICA) algorithms have proven successful in separating linear mixtures of independent source signals [5][13] [14][18][20][36][42][61][70][74][79]. While most of the existing implementations have been tested and compared to each other using synthetic data, significant results on separating real world mixtures of signals have been reported as well [6][52] [60][66][67][68]. Many existing methods rely on simple assumptions on the source statistics and are characterized by well assessed convergence and consistency properties [45]. When such hypotheses hold strictly or are only moderately violated, most conventional ICA algorithms are capable of quickly and efficiently achieve the desired source separation. However, such algorithms can perform sub-optimally or even fail to produce the desired source separation, when the assumed statistical model is inaccurate [13].

A relevant example and a well-known ICA implementation is Hyvärinen's FastIca [41] (see Chapter 3, which requires the user to select a contrast function according to the hypothetical (but unknown) probability density functions (pdf) of the sources to be reconstructed. Such issues do not arise in the case of moment based implementations of blind signal separation algorithm (e.g. Cardoso's Jade [15]). However, these approaches usually rely exclusively on third or fourth order cross-cumulants in order to measure independency, and represent just an approximation of the mutual information minimization principle [14]. Clearly, when the separation of signals from real world data is attempted, such constraints are highly undesirable.

Alternative methods that employ a more flexible model for the pdf of the source signals have been introduced [1][54][89]. These methods usually consist of a parametric density estimation technique that alternates with a cost function optimization step in an iterative approximation framework. Although these approaches tend to outperform standard algorithms in specific cases (e.g. skewed sources), neither their convergence properties, nor their capability of modeling arbitrarily distributed sources, have been fully assessed. The recent introduction of kernel-based methods, such as Bach and Jordan's [2], demonstrate that finding a compromise between computational complexity, performance and strong convergence properties of a blind signal separation framework is still an open and challenging problem.

In this chapter, we introduce a novel non-parametric ICA algorithm that is truly "blind" to the particular underlying distributions of the mixed signals, especially when real world applications are sought. The proposed approach simultaneously estimates the unknown probability density functions of the source signals and the linear operator that allows the separation of the mixed signals (the so-called "unmixing matrix"). The resulting algorithm is non-parametric, data-driven, and does not require the definition of a specific model for the score functions.

4.1 Joint Estimation of the Unmixing Matrix and of the Distribution of the Source Signals

4.1.1 ICA Model and Separation Principle

The conventional generative model introduced in Chapter 3 is assumed, where N independent and stationary source signals $s_1, \ldots, s_N$ are mixed by an unknown, full-rank mixing matrix A (size $N \times N$), resulting in a set of mixtures given by $\mathbf{x} = A\mathbf{s}$. The reconstruction of the original sources is attempted through a linear projection of the type $\mathbf{y} = W\mathbf{x}$, with the assumption that at the most one of the sources has a gaussian density [18]. The basic principle behind most ICA frameworks is the minimization of the mutual information between the reconstructed signals [3], that is:

$$W_{opt} = \arg\min_{W} I(y_1, \ldots, y_N) \tag{4.1}$$

This principle is characterized by having the minimum asymptotic variance, as shown by Donoho in [26], and it can also be proved to be equivalent to the maximum likelihood (ML) principle when the source distributions are known [12][13]. Using basic information theory equalities[19], (4.1) can be written as:

$$\min_{W} \sum_{i=1}^{N} H(y_i) - \log|\det W| - H(\mathbf{x}). \tag{4.2}$$

Since the term $H(\mathbf{x})$ is a constant with respect to W, the objective function is reduced to:

$$
\begin{aligned}
L(W) &= \sum_{i=1}^{N} H(y_i) - \log|\det W| && (4.3) \\
&= -\sum_{i=1}^{N} E\left[\log p_{y_i}(\mathbf{w}_i \mathbf{x})\right] - \log|\det W| \;, && (4.4)
\end{aligned}
$$

where $\mathbf{w}_i$ is the ith *row* of the matrix W.

4.1.2 Non-Parametric Kernel Density Estimation

In order to evaluate the marginal entropies $H(y_i)$ in (4.3), a model for the distribution of the unknown signals is necessary. In a quite effective way, Cardoso shows in [13], that incorrect assumptions on such distributions can result in poor estimation performance, sometimes in a complete failure to obtain the source separation.

To tackle this issue, we propose a non-parametric model, where the probability density functions p_{y_i} are directly estimated from the data using a kernel density estimation technique [49][86]. The proposed approach allows a direct evaluation of the cost function and its derivatives, thus lifting the requirement of separating the optimization step from the step involving the re-estimation of the score functions, as in [54] or [89]. Given a batch of sample data of size M, the marginal distribution of an arbitrary reconstructed signal is approximated as follows:

$$
p_{y_i}(y_i) = \frac{1}{Mh} \sum_{m=1}^{M} \phi\left(\frac{y_i - Y_{im}}{h}\right), \quad i = 1, \ldots, M, \tag{4.5}
$$

where h is the kernel bandwidth and ϕ is the gaussian kernel:

$$\phi(u) \triangleq \frac{1}{\sqrt{2\pi}}\, e^{-\frac{u^2}{2}}\ . \tag{4.6}$$

The kernel centroids Y_{mi} are equal to:

$$Y_{im} = \mathbf{w}_i \mathbf{x}^{(m)} = \sum_{n=1}^{N} w_{in} x_{nm}\ . \tag{4.7}$$

where $\mathbf{x}^{(m)}$ is the mth *column* of the mixture matrix. This estimator is asymptotically unbiased and efficient, and it is shown to converge to the true pdf under several measures. Moreover, it is a continuous and differentiable function of the elements of the unmixing matrix W, with its gradient being given by:

$$\nabla p(y_i) = \frac{1}{Mh^2} \sum_{m=1}^{M} \mathbf{x}^{(m)} (y_i - \mathbf{w}_i \mathbf{x}^{(m)}) \phi \left(\frac{y_i - \mathbf{w}_i \mathbf{x}^{(m)}}{h} \right). \tag{4.8}$$

Using the kernel expansion of the source distributions, we can derive a closed form expression for the pdf estimate of the one-dimensional reconstructed signals, evaluated at the data points as:

$$p_{y_i}(\mathbf{w}_i \mathbf{x}^{(k)}) = \frac{1}{Mh} \sum_{m=1}^{M} \phi \left(\frac{\mathbf{w}_i \left(\mathbf{x}^{(k)} - \mathbf{x}^{(m)} \right)}{h} \right)\ . \tag{4.9}$$

4.1.3 Objective Function Derivation

The expectation in (4.4) can be approximated by its ergodic average, as follows:

$$L(W) \approx -\frac{1}{M} \sum_{i=1}^{N} \sum_{k=1}^{M} \log p_{y_i}(\mathbf{w}_i \mathbf{x}^{(k)}) - \log|\det W| \ , \tag{4.10}$$

resulting in the following cost function definition:

$$L(W) = -L_0(W) - \log|\det W|, \tag{4.11}$$

where $L_0(W)$ is obtained by replacing the marginal pdfs p_{y_i} with their kernel density estimates:

$$\begin{aligned} L_0(W) &= \sum_{i=1}^{N} E \ \log \left[\frac{1}{Mh} \sum_{m=1}^{M} \phi \left(\frac{y_i - Y_{im}}{h} \right) \right] \qquad (4.12) \\ &\approx \frac{1}{M} \sum_{i=1}^{N} \sum_{k=1}^{M} \log \left[\frac{1}{Mh} \sum_{m=1}^{M} \phi \left(\frac{\mathbf{w}_i \left(\mathbf{x}^{(k)} - \mathbf{x}^{(m)} \right)}{h} \right) \right] . \end{aligned}$$

The overall optimization problem can thus be posed as:

$$\min_W \ -\frac{1}{M} \sum_{i=1}^{N} \sum_{k=1}^{M} \log \left[\frac{1}{Mh} \sum_{m=1}^{M} \phi \left(\frac{\mathbf{w}_i \left(\mathbf{x}^{(k)} - \mathbf{x}^{(m)} \right)}{h} \right) \right] - \log|\det W| \tag{4.13}$$

$$\text{s.t. } ||\mathbf{w}_i|| = 1 \ , \quad i = 1, \ldots, N \, . \tag{4.14}$$

Given the sample data $\mathbf{x}^{(k)}, k = 1, \ldots, M$, the objective (4.13) is a non-linear function of the elements of the matrix W. The additional constraints (4.14) are introduced in order to restrict the space of possible solutions of the problem to be a finite set. Clearly, if a matrix W_0 is optimal according to (4.1), so is any other matrix obtained from W_0 by re-scaling or permuting its rows. The constraints (4.14) remove the degree of freedom given by the magnitude of the sources, thus limiting the solution space to all possible permutations of the reconstructed signals (a finite set).

Although it is not strictly required in the proposed algorithm, we can assume that the mixture data has been centered and sphered prior to attempting the reconstruction [49], thus the problem is reduced to the estimation of an orthogonal matrix [74]. Such pre-processing of the mixture data allows a further simplification in the design of the kernel density estimator, since all the reconstructed signals can be assumed to be zero-mean and unit variance random variables, due to the constraint (4.14). Therefore, the optimal value of the parameter h, which controls the smoothness of the functional, is a function of the sample size only ($h = 1.06M^{-1/5}$, [86]). Simulation experiments reported in section 4.3 show a relative insensitivity of the algorithm's performance for variations up to ±50% from the optimal value of the bandwidth parameter.

4.2 Optimization and Global Convergence Issues

4.2.1 Optimization Algorithm

The objective (4.13) is a smooth non-linear function of the elements w_{ij} of the unmixing matrix W. Its gradient can be computed using (4.8), as follows:

$$\begin{aligned} \nabla L(W) &= -\nabla L_0(W) - \nabla \log |\det(W)| \\ &= -\nabla L_0(W) - \left(W^T\right)^{-1}. \end{aligned} \tag{4.15}$$

If we define the following quantity:

$$Z_i(k,m) \triangleq \mathbf{w}_i\left(\mathbf{x}^{(k)} - \mathbf{x}^{(m)}\right)/h = \frac{1}{h}\sum_{j=1}^{N} w_{ij}(X_{jk} - X_{jm}) \tag{4.16}$$

we can compute the components of $\nabla L_0(W)$ as:

$$\frac{\partial L_0(W)}{\partial w_{ij}} = \sum_{k=1}^{M} \frac{-\sum_{m=1}^{M} \frac{\partial Z_i(k,m)}{\partial w_{ij}}\phi\left(Z_i(k,m)\right)}{h \cdot \sum_{k=1}^{M}\phi\left(Z_i(k,m)\right)} \tag{4.17}$$

$$= \sum_{k=1}^{M} \frac{-\sum_{m=1}^{M}\left(X_{jk} - X_{jm}\right) Z_i(k,m)\phi\left(Z_i(k,m)\right)}{h \cdot \sum_{k=1}^{M}\phi\left(Z_i(k,m)\right)}.$$

The constraints (4.14) can be enforced simply by operating the substitution:

$$\mathbf{w}_i = \frac{\tilde{\mathbf{w}}_i}{||\tilde{\mathbf{w}}_i||}, \quad i = 1, \ldots, N\,. \tag{4.18}$$

Using the transformation (4.18), the matrix W can be written as $W = \tilde{D}^{-1}\tilde{W}$, with:

$$\tilde{D} = \begin{bmatrix} ||\tilde{\mathbf{w}}_1|| & & 0 \\ & \ddots & \\ 0 & & ||\tilde{\mathbf{w}}_N|| \end{bmatrix}, \tag{4.19}$$

thus $\tilde{W} = \tilde{D}W$. Then:

$$\log|\det W| = -\sum_{i=1}^{N}\log||\tilde{\mathbf{w}}_i|| + \log|\det \tilde{W}|\,. \tag{4.20}$$

The derivatives with respect to $\tilde{w}_{ij}$ are thus computed as:

$$\frac{\partial(\log|\det W|)}{\partial \tilde{w}_{ij}} = -\frac{\tilde{w}_{ij}}{||\tilde{\mathbf{w}}_i||^2} + \left[(\tilde{W}^T)^{-1}\right]_{ij} . \tag{4.21}$$

When W is orthogonal ($W^{-1} = W^T$), we have:

$$(\tilde{W}^T)^{-1} = \tilde{D}^{-1}(W^T)^{-1} = \tilde{D}^{-2}\tilde{W} , \tag{4.22}$$

and the coefficients of the gradient (4.21) are all equal to zero. Therefore, as expected, the second term of the cost function (4.3) will no longer enter the optimization procedure when the matrix W is orthogonal. Applying the substitution as in (4.18), the components of $\nabla L_0(\tilde{W})$ can be computed as:

$$\frac{\partial L_0(\tilde{W})}{\partial \tilde{w}_{ij}} = \frac{1}{M}\sum_{k=1}^{M} \frac{-\sum\limits_{m=1}^{M}\left(X_{jk} - X_{jm} - \tilde{Z}_i(k,m)\tilde{w}_{ij}\right)\tilde{Z}_i(k,m)\phi\left(\tilde{Z}_i(k,m)\right)}{h\cdot\sum\limits_{m=1}^{M}\phi\left(\tilde{Z}_i(k,m)\right)} \tag{4.23}$$

where, analogously to (4.16), $\tilde{Z}_i(k,m)$ is defined as:

$$\tilde{Z}_i(k,m) \triangleq \tilde{\mathbf{w}}_i\left(\mathbf{x}^{(k)} - \mathbf{x}^{(m)}\right)/h = \frac{1}{h}\sum_{j=1}^{N}\tilde{w}_{ij}(X_{jk} - X_{jm}) \tag{4.24}$$

and $||\tilde{\mathbf{w}}_i||$ is arbitrarily chosen equal to one.

A natural choice for the optimization algorithm is the Quasi-Newton method [8][27], which provides a good compromise between fast convergence, and computational payload. A *backtracking* technique is adopted for the selection of the step size. The main steps of the proposed non-parametric ICA algorithm are shown in Table 4.1. The backtracking routine ensures convergence to the closest local minimum [73], even when the objective function is not convex.

Table 4.1. Main steps of the Non-Parametric ICA algorithm

```
NON-PARAMETRIC ICA
Initialize W, $\alpha$, $\beta$
Initialize the Hessian estimate $H := I_{M \times M}$
repeat
    1. Compute the search direction: $V := -H^{-1}\nabla L(W)$
    2. Backtracking: compute the step size
       $\mu := 1$
       while $L(W + \mu V) > L(W) + \alpha\mu\nabla L(W)^T V$
           $\mu := \beta\mu$
    3. Update $H^{-1}$
    4. Update W: $W := W + \mu V$
until $\sqrt{-V^T \nabla L(W)} \leq \epsilon$ (stopping criterion)
```

4.2.2 Analysis of the extrema of the cost function for N=2 sources

A well-known result in blind signal separation is that, given the assumption of linear and instantaneous mixing, the unmixing matrix is unique up to scaling and permutations [18]. Conventionally, the unmixing operator is estimated by minimizing a cost function derived from the mutual information measure (4.1). Although the global minimum of (4.1) is known to yield the desired source separation, no proof is available to show that such a function has no local minima. On the other hand, because of the uniqueness of the separation matrix (up to permutations and scaling), proved by Comon in [18], convergence to any solution other than the global would result in a failure to separate the source signals. As it was recently pointed out in [81] and [82], this specific issue is often overlooked in other ICA frameworks, where, instead, the main concern is whether convergence to a local minimum is obtained at all for an arbitrary initial guess [46].

The problem can be studied in detail in the case of mixtures of $N = 2$ sources. In this case the unmixing matrix W can be parametrized as follows (including implicitly the unit norm constraints on the rows of W):

$$W = \begin{bmatrix} \cos\theta_1 & \sin\theta_1 \\ \cos\theta_2 & \sin\theta_2 \end{bmatrix} . \tag{4.25}$$

With a slight abuse of notation we can write the cost function as:

$$L(\theta_1, \theta_2) = h(\theta_1) + h(\theta_2) - \log|\det(W)| \, , \tag{4.26}$$

where $\log|\det(W)| = \log|\sin(\theta_2 - \theta_1)|$, and $h(\theta_i)$ is defined as:

$$h(\theta_i) \triangleq H(y_{\theta_i}) \, , \quad y_{\theta_i} = \cos\theta_i x_1 + \sin\theta_i x_2 \tag{4.27}$$

Without loss of generality, we can assume the mixing matrix to be the 2x2 identity matrix, so that x_1= s_1 and x_2=s_2. The extrema of cost function (4.26) must, then, satisfy the following conditions:

$$\frac{\partial h(\theta_1)}{\partial \theta_1} + 1/\tan(\theta_2 - \theta_1) = 0, \tag{4.28}$$

$$\frac{\partial h(\theta_2)}{\partial \theta_2} - 1/\tan(\theta_2 - \theta_1) = 0, \tag{4.29}$$

or, equivalently:

$$\frac{\partial h(\theta_2)}{\partial \theta_2} = -\frac{\partial h(\theta_1)}{\partial \theta_1} = 1/\tan(\theta_2 - \theta_1)\,. \tag{4.30}$$

These conditions are graphically illustrated in Figure 4.1. In order to characterize the nature of these extrema, we can compute the Hessian of (4.26), obtaining:

$$\left[\frac{\partial^2 L}{\partial \theta^2}\right]_{ij} = \begin{bmatrix} \frac{\partial^2 h(\theta_1)}{\partial \theta_1^2} & 0 \\ 0 & \frac{\partial^2 h(\theta_2)}{\partial \theta_2^2} \end{bmatrix} + \frac{1}{\sin^2(\theta_2 - \theta_1)} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \tag{4.31}$$

The minima of the cost function (4.26) are found in correspondence of values of (θ_1, θ_2) that satisfy the first-order conditions (4.30), and simultaneously ensure that the Hessian (4.31) is positive semi-definite, which requires that (see Figure 4.2):

$$\frac{\partial^2 h(\theta_1)}{\partial \theta_1^2} + \frac{\partial^2 h(\theta_2)}{\partial \theta_2^2} + \frac{2}{\sin^2(\theta_2 - \theta_1)} \geq 0 \tag{4.32}$$

$$\frac{\partial^2 h(\theta_1)}{\partial \theta_1^2}\frac{\partial^2 h(\theta_2)}{\partial \theta_2^2} + \frac{1}{\sin^2(\theta_2 - \theta_1)}\left(\frac{\partial^2 h(\theta_1)}{\partial \theta_1^2} + \frac{\partial^2 h(\theta_2)}{\partial \theta_2^2}\right) \geq 0. \tag{4.33}$$

It can be easily verified that the cost function 4.26 is even and periodic both in θ_1 and θ_2 with period 2π, and that the conditions (4.30) through (4.33) are satisfied, in particular, when $\theta_1 = n\pi/2$ ($n \in \mathcal{Z}$), $\theta_2 = \theta_1 \pm \pi/2$, resulting in the source separation.

As an example, consider the mixture of a super-gaussian ($\kappa_4 = 1.0$) and a sub-gaussian source ($\kappa_4 = -1.0$), both unimodal. The entropy of an arbitrary linear projection of the mixtures is shown in Figure 4.3 as a function of $\theta_{1,2}$ (the function is symmetric with respect to the vertical axis). Clearly, in this simple example the entropy function has only minima corresponding to the optimal solutions ($\theta_{1,2} = 0, \pm\pi/2$), which satisfy conditions (4.30) and (4.33). Because of the independence

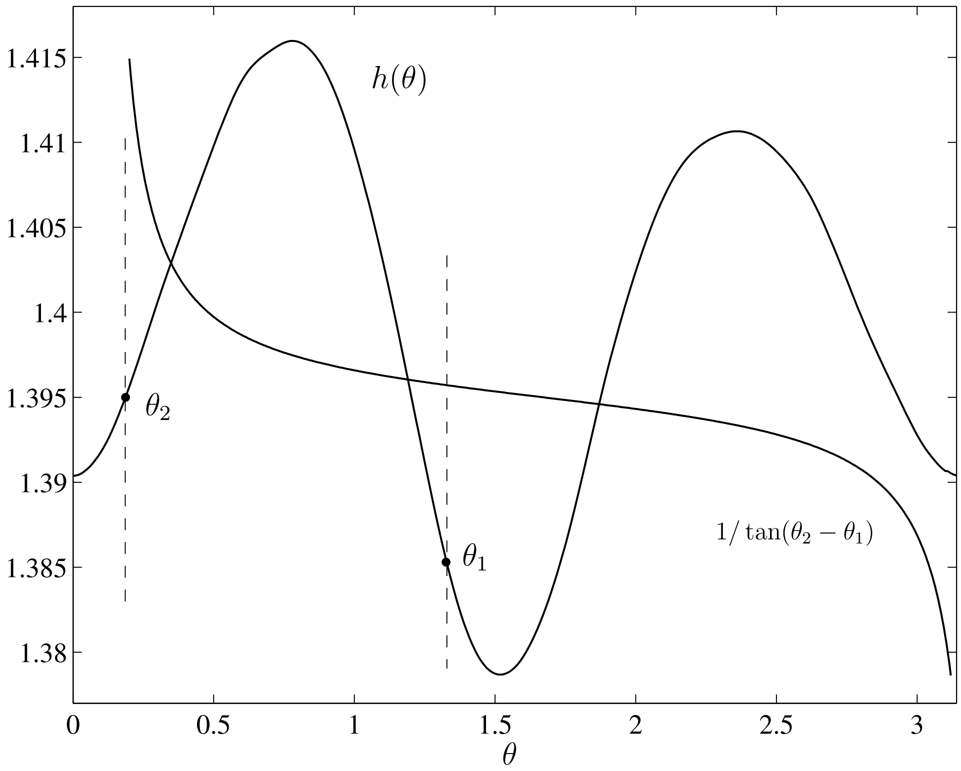


Figure 4.1. **Graphical interpretation of the conditions on the extrema of the cost function (4.30). The curve $1/\tan(\theta_2-\theta_1)$ is plotted for a fixed value of θ_2 (not to scale).**

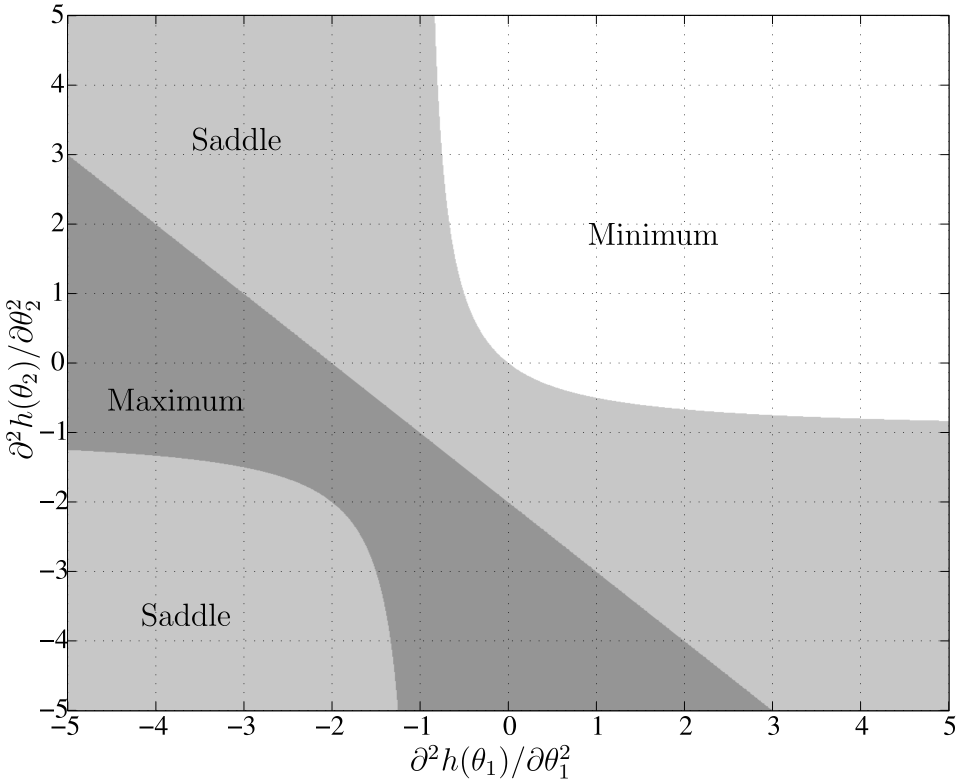


Figure 4.2. The nature of an extremum of the objective function is shown as a function of the second-order partial derivatives of the entropies of the two reconstructed sources (for $\theta_2 - \theta_1 = \text{const.}$).

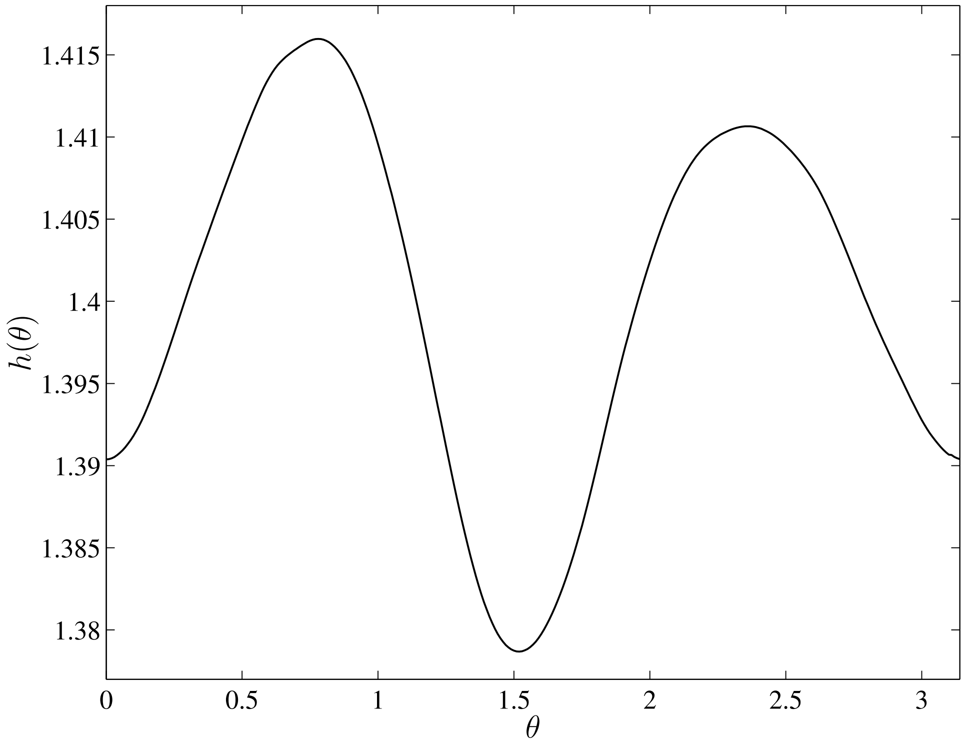


Figure 4.3. **Mixtures of a sub-gaussian signal and a super-gaussian signal. The figure shows a plot of the entropy of a generic reconstructed source as a function of the parameter θ. For these particular mixtures of unimodal sub-gaussian and super-gaussian sources, the entropy function does not present any spurious local minima.**

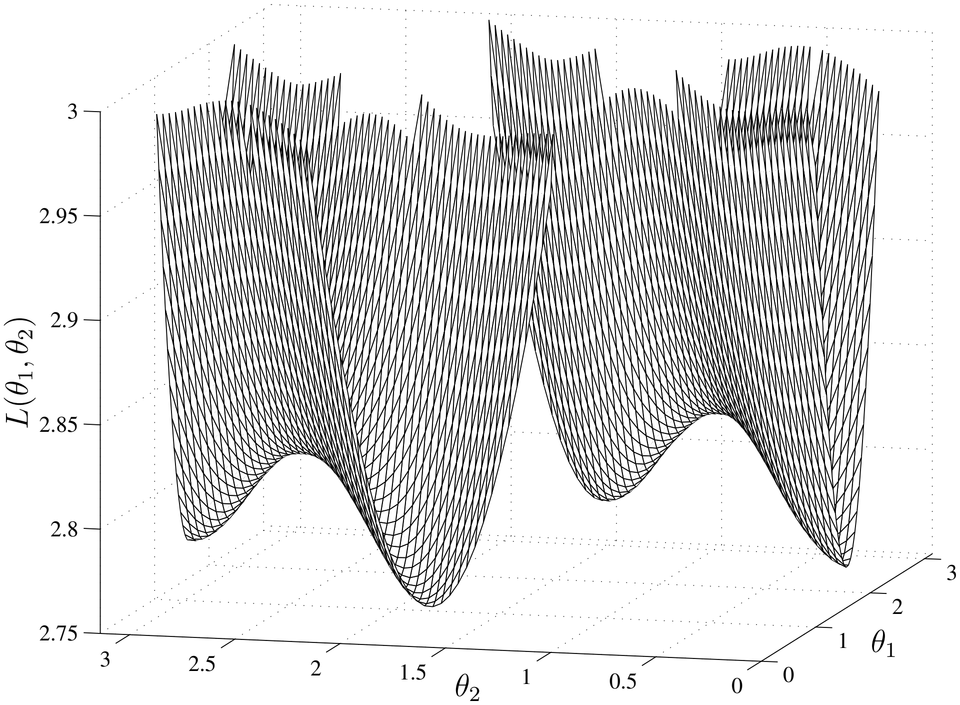


Figure 4.4. Mixtures of a sub-gaussian signal and a super-gaussian signal. The overall cost function $L(W)$ is plotted as a function of (θ_1, θ_2). The plot clearly shows the set of four equivalent minima, corresponding to permutations or change of sign of the rows of the unmixing matrix.

of the sources, the minima appear spaced by $\pi/2$, and correspond to the global optima of the overall cost function (see Figure 4.4).

The situation is quite different in the case of mixtures of sources characterized by a multimodal probability density function. An interesting example is given by mixtures of natural images, where each pixel is considered as a sample drawn from a distribution. This type of sources, in fact, tend to have a distribution that is "heavily" multimodal. In Figure 4.5, the entropy of a generic projection of a mixture of two images[1] is plotted as a function of θ. Although the entropy function shows minima at the optimal points ($\theta_{1,2} = 0, \pm\pi/2$), several spurious local minima appear in other locations. However, at least in this example, these minima do not satisfy the conditions in (4.30), and do not appear in the overall cost function, which, once again, has a unique set of equivalent global minima (cfr. Figure 4.6). The independence of the sources, in fact, imposes a special structure on the cost function, with the extrema of the entropy appearing in correspondence of orthogonal rows of the matrix W (a well known fact in the ICA theory). Other local spurious minima do not appear in the overall cost function because they do not satisfy the first-order constraints (4.30). Nevertheless, it is still an open problem to identify the class of distributions for which this property holds in general, as well as to show whether the same property applies for mixtures of $N > 2$ sources.

4.3 Simulation Experiments

A set of simulation experiments was conducted in order to investigate the performance of the proposed non-parametric method. The blind separation was attempted with each of the following algorithms: the Extended InfoMax ICA [61],

[1] The images can be downloaded at `http://www.ee.ucla.edu/~riccardo/ICA/images`.

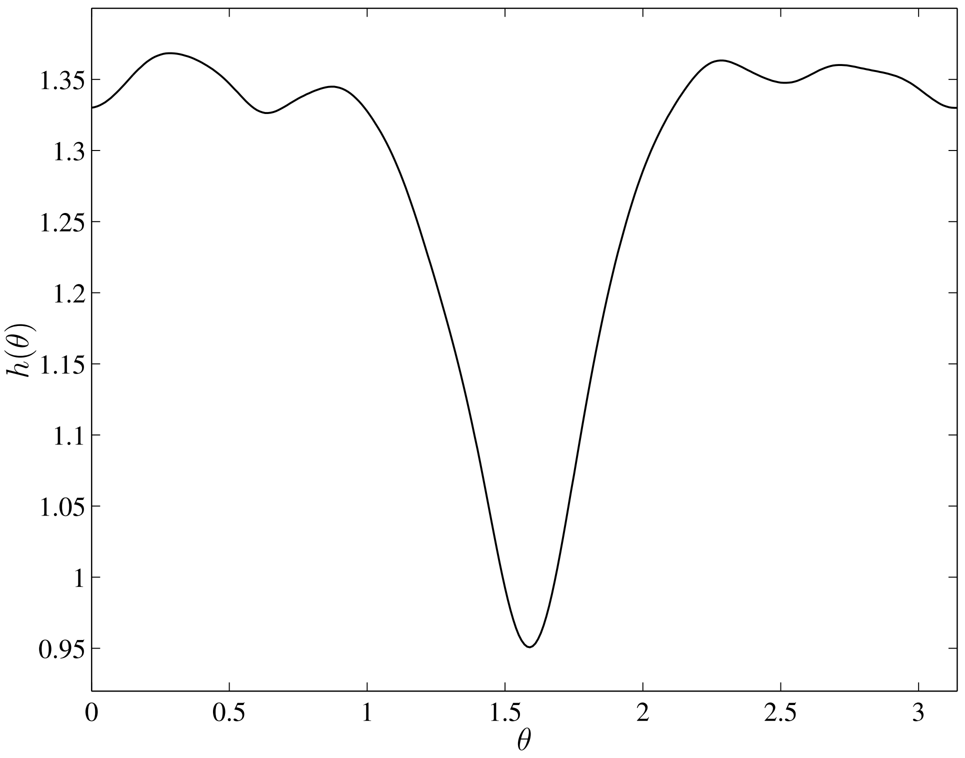


Figure 4.5. Mixtures of two natural images. The entropy function is plotted as a function of $\theta_{1,2}$. In this case, the entropy presents several spurious local minima, which do not correspond to independent sources. Attempting the separation using a deflationary approach could result in a failure to reconstruct the sources.

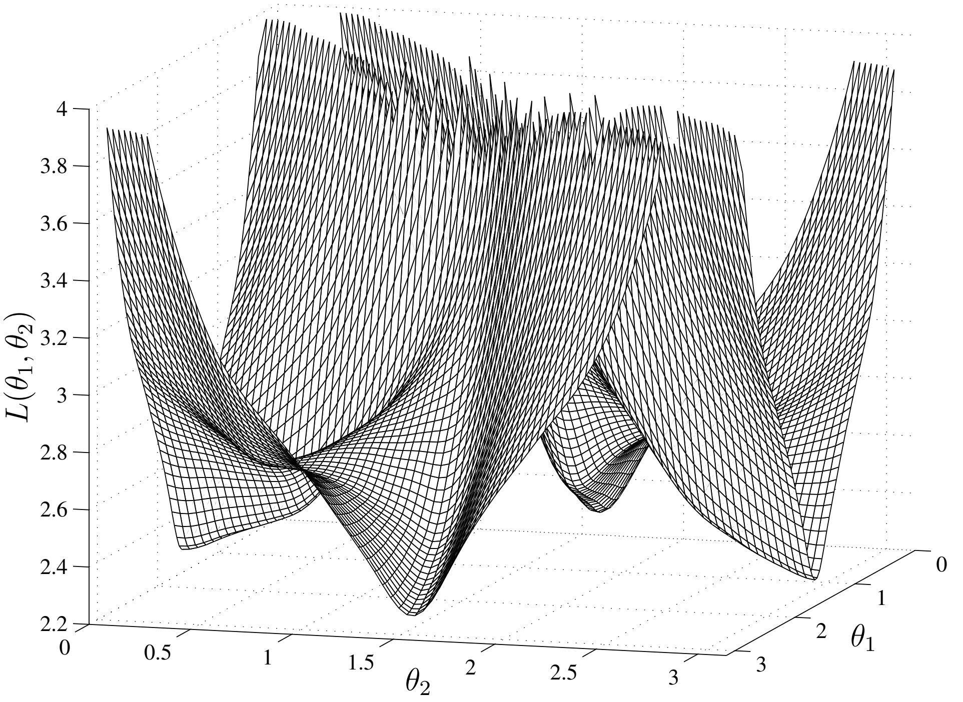


Figure 4.6. **Mixtures of two natural images. The overall cost function $L(W)$ is plotted as a function of (θ_1, θ_2), with a set of four equivalent global minima clearly appearing. The objective function is free from the spurious local minima encountered when observing the entropy function alone. At least in this case, the only values of $\theta_{1,2}$ that satisfy (4.30) are either (equivalent) global minima, or saddle points.**

FastIca [41], Jade [14], two so-called source adaptive methods, the Pearson model ICA [54] and the EGLD model ICA [29], Kernel-ICA [2] and the proposed approach[2]. The algorithms were all downloaded from the websites of respective authors, and in the case of FastIca, all the available contrast functions were tested, both in deflationary mode (sources extracted one at the time), and in simultaneous separation mode (all the sources separated simultaneously). Both versions of Kernel-ICA, KCCA and KGV, were tested in all the simulations.

4.3.1 Mixtures of Sources with Various Distributions

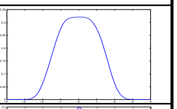
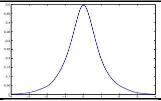
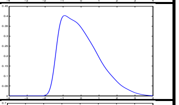
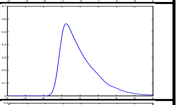
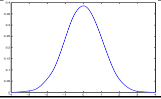
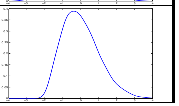
In a first experiment, 1000 realizations of six different sources, distributed as specified in Table 4.2, were independently generated, with sample sizes ranging between 500 and 5000, and mixed with randomly generated, full-rank (condition number ≤ 10) mixing matrices, noiselessly.

The separation performance was evaluated in terms of median SIR (Signal-to-Interference Ratio), defined as $10 \log_{10} \left(\sum_{m=1}^{M} s_m^2 / \sum_{m=1}^{M} (\hat{s}_m - s_m)^2 \right)$ (dB), where $\mathbf{s}$ is the original signal and $\hat{\mathbf{s}}$ is the reconstructed signal. The "interfering" components of the reconstructed signal are by definition those that are due to sources other than the one we are attempting to separate. The results of this first experiment are shown in Figures 4.7, 4.8, 4.9, and they clearly show the performance gain obtained with the non-parametric ICA algorithm. On the average, the *'gauss'* score function, when used in the simultaneous separation mode, resulted in the best overall performance for FastIca, and it is the only one reported for this first experiment. In general, SIR levels below the 8-10 db threshold are indicative of a failure in obtaining the desired source separation.

Although the gain is more consistent in the case of skewed sources (Source

[2] The Non-Parametric ICA algorithm can be downloaded at `http://www.ee.ucla.edu/~riccardo/ICA/npica.tar.gz`.

Table 4.2. Distribution of the synthetic sources used in the first simulation experiment (see [30] for a description of the distributions generated with the Power Method).

Source#	Source type	Skewness	Kurtosis	Pdf plot
1	Power Exponential (α= 2.0)	0.0	-0.8	
2	Power Exponential (α= 0.6)	0.0	2.2	
3	Power Method Distribution [a]	0.75	0.0	
4	Power Method Distribution [b]	1.50	3.0	
5	Normal Distribution	0.0	0.0	
6	Rayleigh Distribution (β= 1)	0.631	0.245	

[a] b=1.112, c=0.174, d=-0.050

[b] b=0.936, c=0.268, d=-0.004

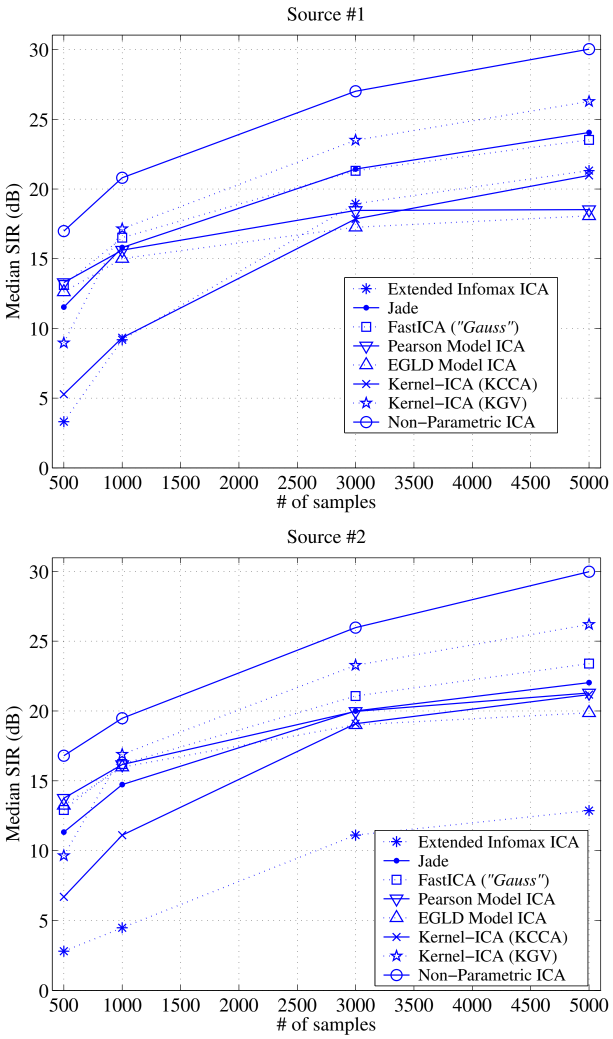


Figure 4.7. First simulation experiment. The results of attempting the separation of the six different sources listed in Table 4.2 are shown for various ICA algorithms (averaged over 1000 Monte Carlo simulations). The accuracy of the separation is measured in terms of median log signal-to-interference ratio (SIR). The relative performance for Source #1 and #2 is shown.

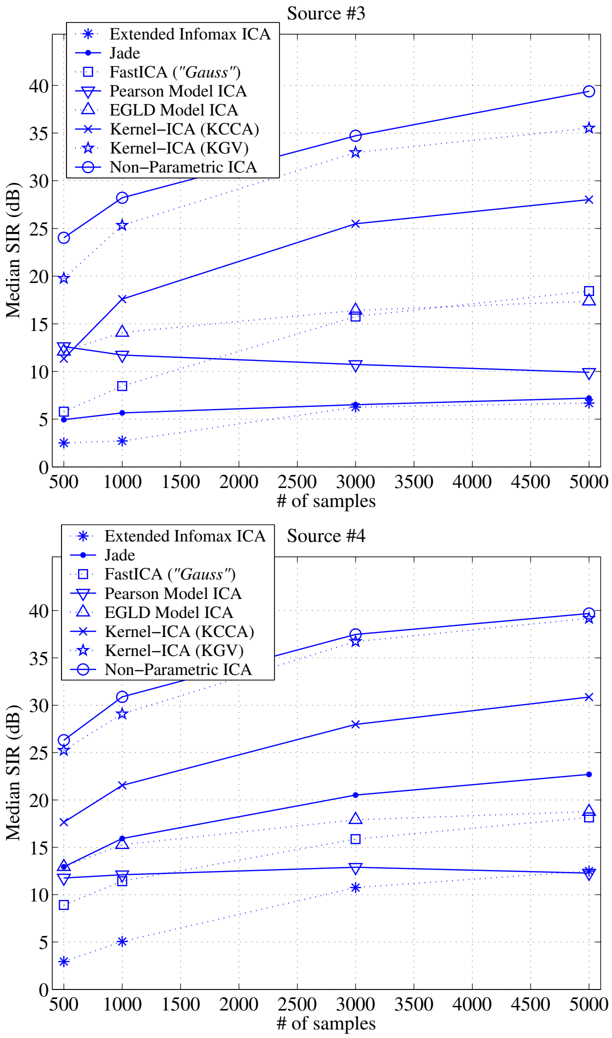


Figure 4.8. **First simulation experiment. The results of attempting the separation of the six different sources listed in Table 4.2 are shown for various ICA algorithms (averaged over 1000 Monte Carlo simulations). The accuracy of the separation is measured in terms of median log signal-to-interference ratio (SIR). The relative performance for Source #3 and #4 is shown.**

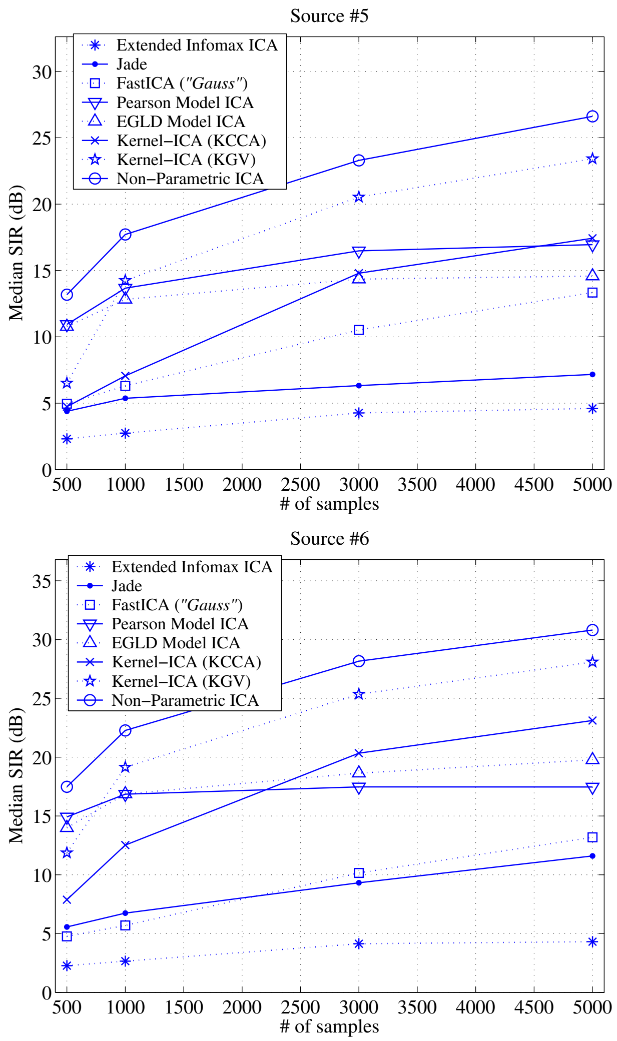


Figure 4.9. First simulation experiment. The results of attempting the separation of the six different sources listed in Table 4.2 are shown for various ICA algorithms (averaged over 1000 Monte Carlo simulations). The accuracy of the separation is measured in terms of median log signal-to-interference ratio (SIR). The relative performance for Source #5 and #6 is shown.

#3,#4, and #6), the separation improvement is substantial also for conventional sub-gaussian and super-gaussian sources (Source#1 and #2). Although KernelICA-KGV appears to somehow match the performance of the proposed method, Non-Parametric ICA still retains a performance gain of over 5 dB on average. It is interesting to notice that, although the "source-adaptive" algorithms tend to outperform more conventional ICA methods in the case of non-symmetric sources, they are often surpassed by traditional algorithms for symmetric sources.

The proposed technique delivers a consistent separation improvement for different sample sizes. In particular, the algorithm appears to be capable of learning the source statistics even when the sample size is very small (e.g. 500 samples), readily showing promising adaptive properties.

4.3.2 Skewed Sources

In a second simulation experiment, the specific sensitivity of each algorithm to the source skewness was investigated. Using the method described in [30], we generated samples drawn from four different sources, which are characterized by a very small kurtosis ($|\kappa_4| < 0.2$), and skewness ranging between 0.0 and 0.75. The experiment was conducted mixing all four sources with randomly generated mixing matrices, using 100 independent realizations of the signals, each consisting of 2000 samples. The results obtained with the various ICA algorithms are summarized in Table 4.3. The proposed method shows a noticeable performance improvement, confirming its capability of modeling arbitrarily distributed sources. Although FastICA resulted in the third highest median SIR, its performance is somehow biased by the choice of the score function *'skew'*, which assumes some a-priori knowledge about the nature of the mixed signals.

Table 4.3. The separation performance in terms of median SIR, as well as 25 and 75 percentiles, is shown for mixtures of four skewed sources (averaged over the sources), for various ICA algorithms.

Algorithm	$\hat{\text{SIR}}$	25%	75%
Extended InfoMax	**3.81**	2.89	5.90
Jade	**4.28**	3.03	6.38
FastIca (*'skew'*)	**18.94**	16.04	22.32
Pearson ICA	**14.97**	11.40	19.52
EGLD ICA	**16.73**	12.76	21.21
Kernel-ICA (KCCA)	**16.93**	13.89	20.54
Kernel-ICA (KGV)	**21.64**	17.86	25.10
Non-Parametric ICA	**23.40**	18.91	27.19

4.3.3 Convergence Properties

The convergence properties of the algorithms were empirically tested in a third simulation experiment. The goal was to measure the approximate number of data samples required by each method to achieve a median SIR of at least 20dB. For this purpose, we created mixtures of four independent sources with a super-gaussian ($\kappa_4 \approx 2.2$) symmetric pdf and we averaged the separation results over 100 simulations, for different sample sizes. The choice of standard super-gaussian sources guarantees that the experiment is unbiased, since all ICA algorithms under evaluation are capable of separating this type of signals accurately. Our results show that the proposed method is able to achieve the required quality of separation (20dB median SIR) with only 750 samples, performance matched by KernelICA-KGV. FastICA resulted in the second-best performance (1000 samples), when the score function was suitably chosen (in this case *'gauss'*).

4.3.4 Bandwidth Parameter Sensitivity

The sensitivity of the algorithm to the choice of the bandwidth parameter h in (4.5) was evaluated following the experimental setting used in the first simulation (sources generated according to Table 4.2). In a series of Monte Carlo simulations the bandwidth parameter was allowed to vary up to 50% from the optimal value, computed as a function of the sample size. The results displayed in Figure 4.10 show the obtained median SIR averaged across the 6 sources, for a sample size equal to 1000. The experiment seems to suggest that the separation performance is relatively insensitive to the particular choice of this parameter in a broad range of values.

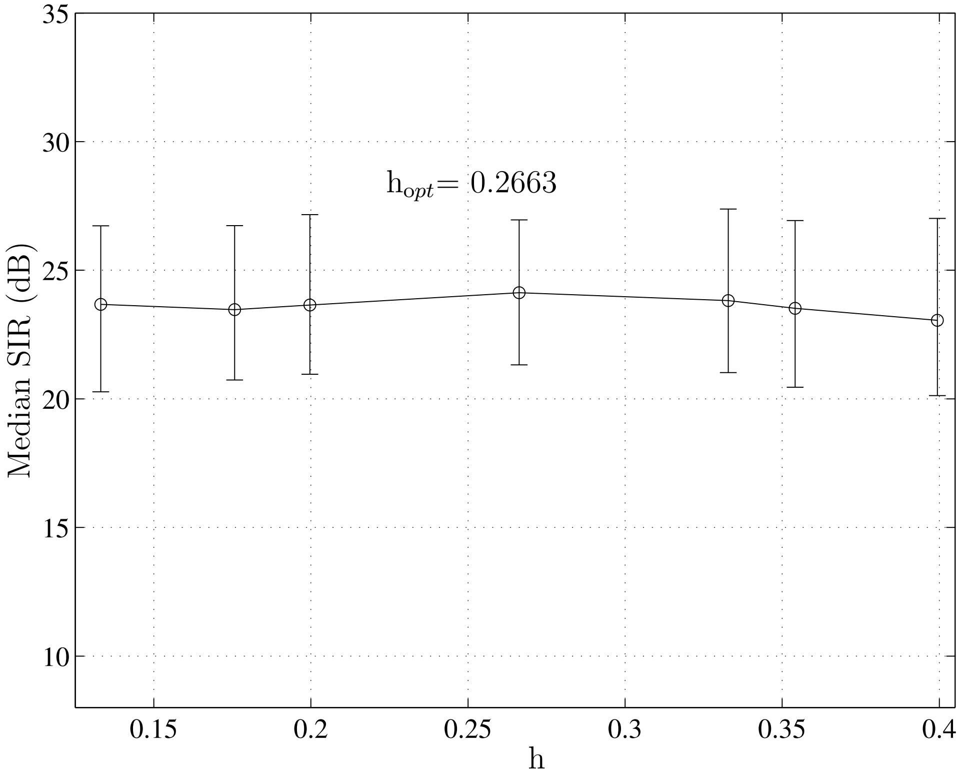


Figure 4.10. **The figure shows the results of a set of simulation experiments aiming at evaluating the sensitivity of the proposed technique to the choice of the bandwidth parameter h. The error bars span between the 25 and the 75 percentiles of the SIR. This experiment seems to suggest that variations of such a parameter up to ±50% from the estimated optimal value do not considerably affect the separation performance.**

4.3.5 Algorithmic Complexity

The introduction of a technique enabling the simultaneous estimation of the unmixing matrix and of the unknown pdfs of the sources is inevitably accompanied by an increase in its computational complexity. Regardless of the actual optimization algorithm, a brute force implementation of the proposed non-parametric method would require an amount of floating point operations proportional to $\mathcal{O}(M^2N)$ to evaluate the cost function and $\mathcal{O}(M^2N^2)$ to compute its derivatives, where N is the number of sources and M is the sample size. This compares unfavorably with fixed score function algorithms like FastIca whose computational complexity is on the order of $\mathcal{O}(MN)$ and $\mathcal{O}(MN^2)$, respectively, especially when the number of samples M is very large.

On the other hand, fast density estimation techniques based on the FFT algorithm can be developed, based on the observation that evaluating a density estimate is equivalent to computing the convolution of an unevenly sampled sequence with a gaussian kernel [86]. At the core of the proposed non-parametric method for ICA stands a fast density estimation algorithm of this type, which can perform the evaluation of the cost function and of its derivatives in a time proportional to $\mathcal{O}(M \log_2 MN)$ and $\mathcal{O}(M \log_2 MN^2)$, respectively, thus minimizing the additional payload required to achieve the increased separation performance and reliability. Table 4.4 shows a detailed derivation of the computational complexity of each step of the Non-Parametric ICA algorithm.

The median CPU time required to run the various ICA algorithms is shown in Figure 4.11 for a fixed number of sources (6) and a variable number of samples and in Figure 4.12 for a fixed number of samples (1000) and a variable number of sources[3]. Clearly, fixed contrast function or moment based ICA algorithms

[3] The simulations were all performed under Matlab©v.6.3, on a Dual Pentium IV 1.8Ghz PC with 512Mbytes of RAM, running Red Hat Linux v7.2.

are in general significantly faster than source adaptive methods. Although Non-Parametric ICA is among the algorithms characterized by a higher computational complexity, it is interesting to notice that it is on average one order of magnitude faster than Kernel-ICA.

4.3.6 Large Scale Problems

In a separate simulation, we investigated the properties of the proposed method for large scale problems. This was accomplished by creating mixtures of 12 up to 24 signals, randomly chosen among a set of sources, whose distributions included both unimodal and bimodal pdfs. The separation results obtained over 100 Monte Carlo simulations (Figure 4.13) demonstrate Non-Parametric ICA's capability of seamlessly handling large size problems. The decrease in median SIR which accompanies the increase in the problem size can be explained by considering that, while the sample size is kept constant ($M = 1000$ samples), the number of parameters that needs to be estimated ($N(N+1)/2$) increases approximately as the square of the number of sources. For example, the unmixing matrix has a total of 66 unique elements when $N = 12$, that number increasing to 276 for $N = 24$ sources.

In terms of convergence properties, we noticed only a marginal increase in the number of Newton steps required to achieve the desired separation accuracy, with the relative CPU time required to complete the routine closely matching the asymptotic computational complexity analysis described in Table 4.4.

Table 4.4. Detailed analysis of the computational complexity of Non-Parametric ICA as a function of the number of sources (N) and the number of samples (M)

Routine	Complexity
NpIca	
(a) Compute search direction	$\mathcal{O}(N^4)$
(b) Backtracking routine	
Cost function evaluation (*‘EstimateObjFFT’*)	
1. Data rebinning	$\mathcal{O}(NM)$
2. FFT of re-binned data	$\mathcal{O}(NM \log_2 M)$
3. FFTs multiplication	$\mathcal{O}(NM)$
4. Inverse FFT of pdf estimate	$\mathcal{O}(NM \log_2 M)$
5. Rebinning and entropies evaluation	$\mathcal{O}(NM)$
(c) Gradient computation *(‘EstimateGradFFT’)*	
1. Data rebinning	$\mathcal{O}(N^2M)$
2. FFT of re-binned data	$\mathcal{O}(N^2M \log_2 M)$
3. FFTs multiplication	$\mathcal{O}(N^2M)$
4. Inverse FFT of pdf derivative estimates	$\mathcal{O}(N^2M \log_2 M)$
5. Rebinning and gradient components evaluation	$\mathcal{O}(N^2M)$
(d) Inverse Hessian update	$\mathcal{O}(N^4)$
(e) Convergence criterion evaluation	$\mathcal{O}(NM)$
Overall computational complexity	$\mathcal{O}(N^4 + N^2M \log_2 M)$

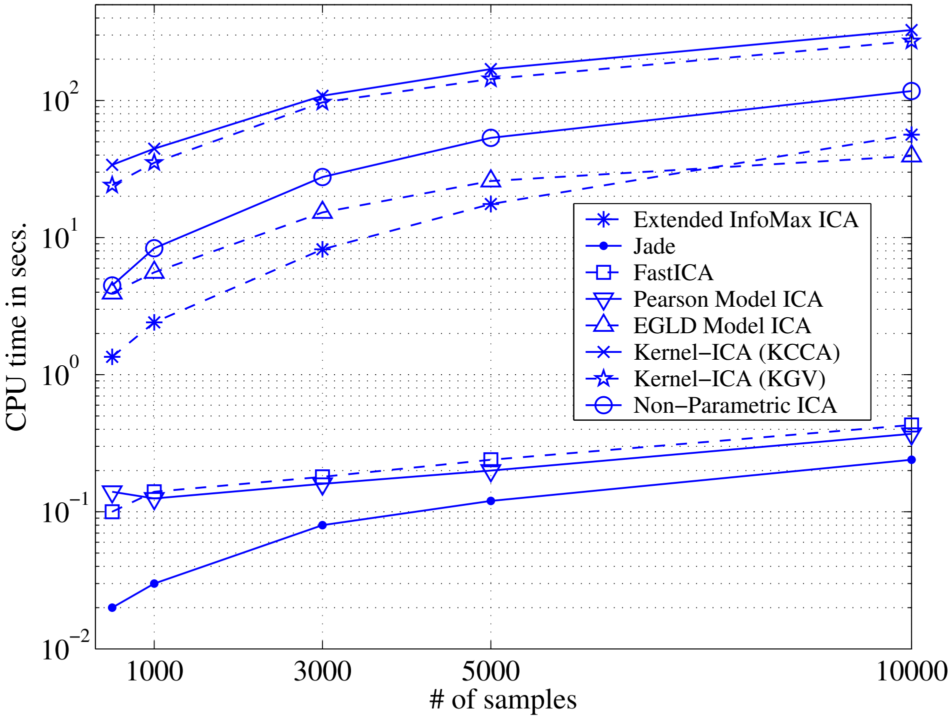


Figure 4.11. The running time in terms of CPU seconds of various ICA algorithms is shown for a fixed number of sources (6) and variable number of samples. The methods capable of source adaptation are in general computationally more expensive, as the separation performance is paid in terms of running time.

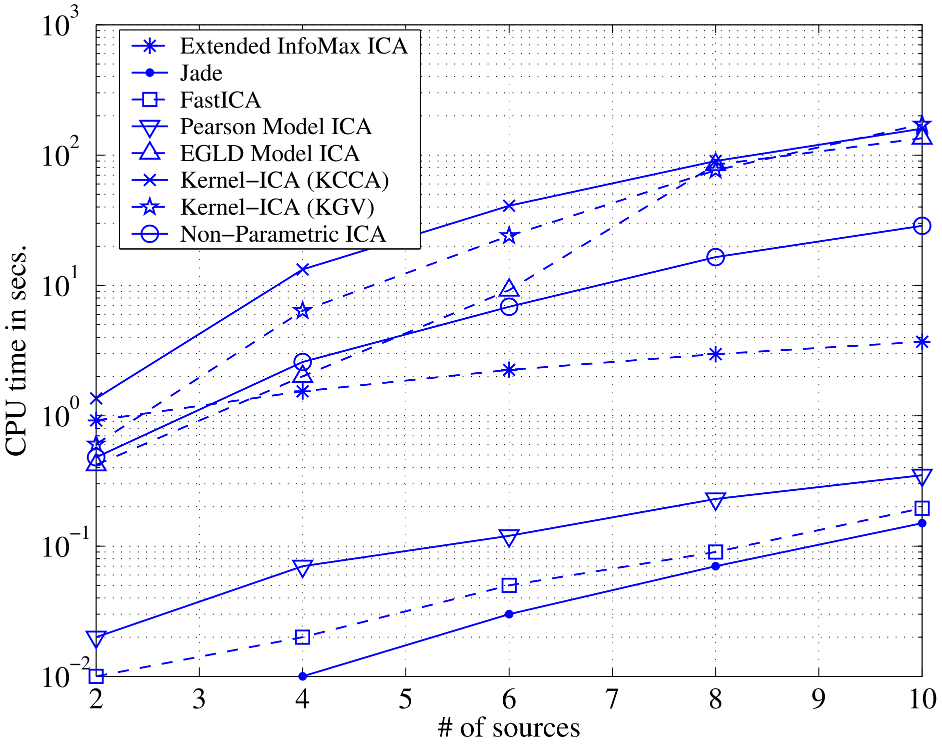


Figure 4.12. **The running time in terms of CPU seconds of various ICA algorithms is shown for a fixed number of samples (1000) and variable number of sources. The methods capable of source adaptation are in general computationally more expensive, as the separation performance is paid in terms of running time.**

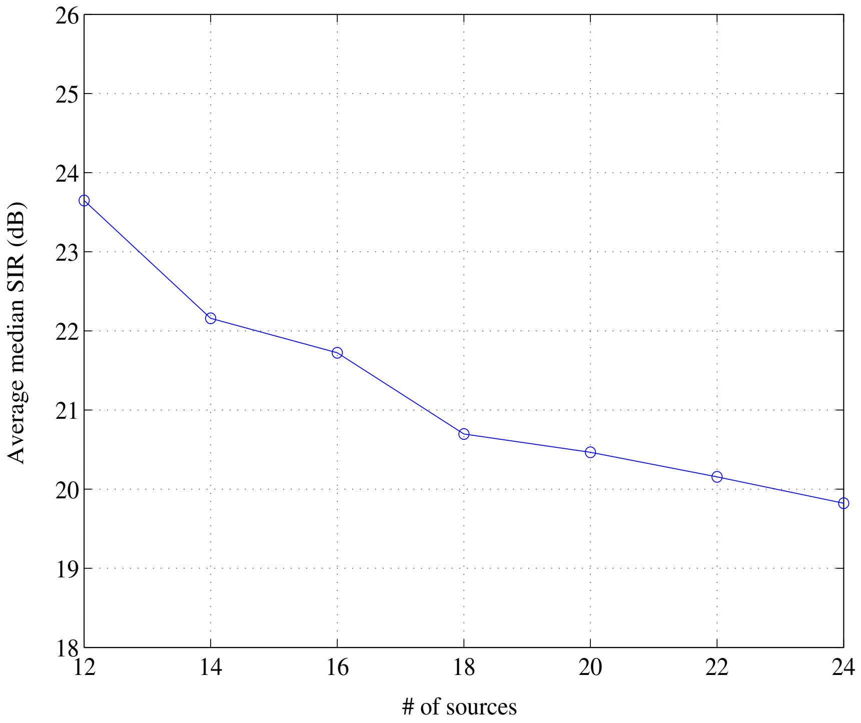


Figure 4.13. **Large scale simulation. The median SIR (dB) achieved by Non-Parametric ICA is shown for the separation of a number of sources varying between 12 and 24 (averaged over the reconstructed signals), and a fixed number of samples ($M = 1000$).**

4.4 Conclusions

A novel non-parametric independent component analysis algorithm was introduced. The proposed method is truly blind to the particular distribution of the original sources, and does not require the selection of optimal working parameters, or suitable non-linearities to act as contrast functions. The algorithm outperformed state-of-the-art ICA techniques in several simulation experiments, with different types of mixtures. The capability of modeling sources with arbitrary distribution, combined with the good convergence properties for small sample sizes, make the proposed approach a particularly attractive alternative to current ICA algorithms, especially for the analysis of real-world mixtures.

Chapter 5

An Extension of Comon's Identifiability Theorem

In this chapter, Comon's identifiability theorem for ICA (Theorem 4, Chapter 3) is extended to the case of mixtures where several gaussian sources are present. We show, in an original and constructive proof, that using the conventional mutual information minimization framework, the separation of all the *non-gaussian* sources is always achievable (up to scaling factors and permutations). In particular, we prove that a suitably designed optimization framework is capable of seamlessly handling both the case of one single gaussian source being present in the mixture (separation of all sources achievable), as well as the case of multiple gaussian signals being mixed together with non-gaussian signals (only the non-gaussian sources can be extracted).

5.1 Introduction

In his fundamental work [18], Comon showed that the separation of a set of stationary signals, instantaneously and linearly mixed, is always possible, as long as the mixing matrix has full rank, and at the most *one* of the original signals is gaussian distributed. This result is often cited in the literature as Comon's *identifiability theorem* for ICA, and it represents a well-known and widely mentioned result in the blind signal separation field. Most ICA algorithms are based on contrast functions that are a functional of the probability density functions of the unknown sources, such as the mutual information between the reconstructed signals, or its equivalent counterparts, i.e. the InfoMax principle, or the maximum likelihood (ML) principle.

In recent years, Cruces et al. [20][22][21] investigated several criteria for the extraction of a subset of sources from a linear mixture, both in the instantaneous case, and in the case of convolutive mixtures. In particular, it was shown in [22], that a suitably designed entropy minimizing framework can be used to extract the non-gaussian sources, from mixtures containing an arbitrary number of gaussian distributed signals. The authors also introduced a moment-based iterative algorithm that minimizes an approximation of the contrast function derived from this principle.

In this chapter, we derive a novel proof of Comon's identifiability theorem, and extend the theorem to the case of multiple gaussian sources being mixed with non-gaussian sources. All the results are derived by investigating the properties of the optimization problem associated with minimizing the mutual information between the reconstructed signals. In particular, we prove that, regardless of the number of gaussian sources in the mixture, the resulting objective function always has extrema that yield the separation of the non-gaussian sources (up to scaling and permutations), and the gaussian components are *irrelevant* in determining such extrema.

5.2 Separation Principle and Objective Function Definition

We make the conventional assumption (see Chapter 3) that N independent and stationary source signals $(s_1, \ldots, s_N)$ are mixed by an unknown, full-rank mixing matrix A, resulting in a set of mixtures given by $\mathbf{x} = A\mathbf{s}$. The reconstruction of the original sources is attempted from the mixture data through a linear projection of the type $\mathbf{y} = B\mathbf{x}$. Following the mutual information minimization principle, common to most ICA frameworks, we seek the matrix B, solution of the optimization problem:

$$B_{opt} = \min_B I(y_1, \ldots, y_N) \ , \tag{5.1}$$

where $I(\mathbf{y})$ is defined as in (2.19). Using basic information theory equalities, (5.1) becomes:

$$\min_B \sum_{i=1}^{N} H(y_i) - \log|\det B| - H(\mathbf{x}). \tag{5.2}$$

If we assume that the mixture data has been sphered (as in 2.30, Chapter 2), so that $E(\mathbf{x}\mathbf{x}^T) = I$, we can restrict the search space for the unmixing matrix B to the manifold of orthogonal matrices [13]. The problem can be simplified as:

$$\min_B \sum_{i=1}^{N} H(y_i) \tag{5.3}$$

$$\text{s.t. } BB^T = I\,, \tag{5.4}$$

since $\log|\det(B)| \equiv 1$, and $H(\mathbf{x})$ is a constant with respect to B. The equality constraints (5.4) define a sub-group of the Stiefel manifold for the case of square matrices. If we define $F(B) \triangleq \sum_{i=1}^{N} H(y_i)$, then the gradient of the cost function defined on such manifold is given by [27]:

$$\nabla_m F(B) \triangleq \nabla F(B) - B\nabla F(B)^T B\,. \tag{5.5}$$

where $\nabla F(B)$ is the conventional gradient of $F(B)$ in the Euclidean space:

$$\nabla F(B) \triangleq \left[\frac{\partial F(B)}{\partial b_{ij}}\right] = \begin{bmatrix} \nabla H(y_1) \\ \vdots \\ \nabla H(y_N) \end{bmatrix}. \tag{5.6}$$

The extrema of the optimization problem (5.3) are given by all the matrices that satisfy the condition:

$$\nabla_m F(B) = 0 \quad \Rightarrow \quad \nabla F(B) B^T = B\nabla F(B)^T\,, \tag{5.7}$$

since $BB^T = I$.

5.3 Extending Comon's Identifiability Theorem

In this section, an alternative proof of Comon's well-known theorem on ICA identifiability [18] is derived, and it is extended to the case where more than one gaussian source is present in the mixture. Under the modeling assumption of Section 5.2, we consider mixtures of N independent sources $s_1, \cdots, s_N$, with probability density function $f_{s_1}, \cdots, f_{s_N}$, M of which are gaussian distributed.

We make the further assumption that the mixing matrix A is the $N \times N$ identity matrix. This is not a restrictive assumption, since, if the mixture data is sphered, the solution spaces associated to any two full rank mixing matrices simply map to each other through an orthogonal transformation [74]. The generic reconstructed signal can be written as:

$$y_i = b_{i1}s_1 + b_{i2}s_2 + \ldots + b_{iN}s_N \qquad i = 1, \ldots, N\,, \tag{5.8}$$

and its differential entropy is given by:

$$H(y_i) = -\int_{-\infty}^{\infty} f_{y_i}(u) \log f_{y_i}(u) du\,, \tag{5.9}$$

where, because of the independence between the sources:

$$f_{y_i}(u) = \frac{1}{|b_{i1}|} f_{s_1}\left(\frac{u}{b_{i1}}\right) * \frac{1}{|b_{i2}|} f_{s_2}\left(\frac{u}{b_{i2}}\right) * \cdots * \frac{1}{|b_{iN}|} f_{s_N}\left(\frac{u}{b_{iN}}\right). \tag{5.10}$$

The components of the gradient of $H(y_i)$ with respect to $\mathbf{b}_i$ (ith row of B) can be computed as:

$$\frac{\partial H(y_i)}{\partial b_{ij}} = -\int_{-\infty}^{\infty} (1 + \log f_{y_i}(u)) \frac{\partial f_{y_i}(u)}{\partial b_{ij}} du \tag{5.11}$$

To make explicit the dependence of the entropy $H(y_i)$ on $\mathbf{b}_i$, define $h(\mathbf{b}_i) \triangleq H(y_i)$. In order to satisfy the first-order conditions given by (5.7), we must have that:

$$\begin{bmatrix} \nabla h(\mathbf{b}_1) \\ \vdots \\ \nabla h(\mathbf{b}_N) \end{bmatrix} \begin{bmatrix} \mathbf{b}_1^T & \cdots & \mathbf{b}_N^T \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_N \end{bmatrix} \begin{bmatrix} \nabla h(\mathbf{b}_1)^T & \cdots & \nabla h(\mathbf{b}_N)^T \end{bmatrix}. \tag{5.12}$$

The resulting set of equations is equivalent to the following set of $N(N-1)$ equalities:

$$\nabla h(\mathbf{b}_k)\mathbf{b}_l^T = \nabla h(\mathbf{b}_l)\mathbf{b}_k^T \qquad k,l = 1,\ldots,N \;\; (k \neq l)\,. \tag{5.13}$$

Using expression (5.11), we get:

$$\begin{aligned}\int_{-\infty}^{\infty} \log f_{y_k}(u)\left[b_{l1}\frac{\partial f_{y_k}(u)}{\partial b_{k1}} + \cdots + b_{lN}\frac{\partial f_{y_k}(u)}{\partial b_{kN}}\right] du = \\ = \int_{-\infty}^{\infty} \log f_{y_l}(u)\left[b_{k1}\frac{\partial f_{y_l}(u)}{\partial b_{l1}} + \cdots + b_{kN}\frac{\partial f_{y_l}(u)}{\partial b_{lN}}\right] du\,.\end{aligned} \tag{5.14}$$

The computation of $\partial f_{y_i}(u)/\partial b_{ij}$ can be efficiently carried out in the frequency domain. Using the conventional definition of *characteristic function* of a random variable [75]:

$$\Phi_X(\omega) \triangleq \mathcal{F}\{f_X(x)\} = \int_{-\infty}^{\infty} f_X(x)e^{-\jmath\omega x}dx\,, \tag{5.15}$$

we have from (5.10), using the convolution theorem:

$$\Phi_{y_i}(\omega) = \Phi_{s_1}(b_{i1}\omega)\Phi_{s_2}(b_{i2}\omega)\cdots\Phi_{s_N}(b_{iN}\omega) \qquad i = 1,\ldots,N. \tag{5.16}$$

If we assume that the pdfs f_{s_i} are continuous functions, with continuous derivatives almost everywhere, we can exchange the order of the integral and the derivative, and compute $\partial f_{y_i}(u)/\partial b_{ij}$ as follows:

$$\frac{\partial f_{y_i}(u)}{\partial b_{ij}} = \mathcal{F}^{-1}\left\{\omega\Phi_{s_1}(b_{i1}\omega)\cdots\Phi'_{s_i}(b_{ij}\omega)\cdots\Phi_{s_N}(b_{iN}\omega)\right\} \tag{5.17}$$

where $\mathcal{F}^{-1}$ denotes the inverse fourier transform operator. The conditions imposed by (5.14) are satisfied, in particular, when:

$$b_{l1}\frac{\partial f_{y_k}(u)}{\partial b_{k1}} + \cdots + b_{lN}\frac{\partial f_{y_k}(u)}{\partial b_{kN}} = 0 \qquad k,l = 1,\ldots,N \ \ (k \neq l)\,. \tag{5.18}$$

If we substitute (5.17) into (5.18), and, under the assumption that all the characteristic functions are non-zero for every ω, we divide by $\Phi_{s_1}(b_{k1}\omega)\cdots\Phi_{s_N}(b_{kN}\omega)$ the resulting expression, we obtain:

$$\frac{\omega b_{l1}\Phi'_{s_1}(b_{k1}\omega)}{\Phi_{s_1}(b_{k1}\omega)} + \ldots + \frac{\omega b_{lN}\Phi'_{s_N}(b_{kN}\omega)}{\Phi_{s_N}(b_{kN}\omega)} = 0 \qquad k,l = 1,\ldots,N \ \ (k \neq l). \tag{5.19}$$

Notice that if and only if f_{s_i} is a gaussian pdf it holds that:

$$\Phi'_{s_i}(\alpha\omega) = -\alpha\omega\Phi_{s_i}(\alpha\omega)\,. \tag{5.20}$$

Therefore in the special case where $M = N$, i.e. all the original sources have a gaussian distribution, (5.19) simplifies as:

$$-(b_{k1}b_{l1} + \ldots + b_{kN}b_{lN})\omega^2 = -\mathbf{b}_k\mathbf{b}_l^T\omega^2 = 0 \qquad k,l = 1,\ldots,N \ \ (k \neq l), \tag{5.21}$$

which are always satisfied because of the orthogonality constraints. Therefore, *if all sources are gaussian, the resulting objective is a constant with respect to the elements of an arbitrary orthogonal unmixing matrix*, and the separation is not possible.

When M is strictly less than N, in order to simplify the notation, we can assume that the first M sources, $(s_1,\ldots,s_M)$, are gaussian distributed. The equations in (5.19) can be simplified as:

$$-\omega^2(b_{l1}b_{k1} + \cdots + b_{lM}b_{kM}) + \frac{\omega b_{lM+1}\Phi'_{s_{M+1}}(b_{kM+1}\omega)}{\Phi_{s_{M+1}}(b_{kM+1}\omega)}\ldots + \frac{\omega b_{lN}\Phi'_{s_N}(b_{kN}\omega)}{\Phi_{s_N}(b_{kN}\omega)} = 0$$

$$k, l = 1, \dots, N \ \ (k \neq l) \tag{5.22}$$

The subset of orthogonal matrices that satisfy this set of equalities is given by:

$$B = \left[\begin{array}{c|c} Q & 0 \\ \hline 0 & P \end{array} \right] , \tag{5.23}$$

where Q is an arbitrary $M \times M$ orthogonal matrix, and P is a generalized permutation matrix. Notice, in fact, that:

$$\Phi'_{s_i}(b_{ij}\omega)\big|_{b_{ij}=0} = -\jmath E[s_i] = 0 \qquad i = 1, \dots, N \, , \tag{5.24}$$

if the sources are zero-mean[1]. This result shows that minima of the optimization problem, that was derived from the separation principle (5.1), appear in correspondence of matrices B that result in separation of the non-gaussian sources. Therefore, we proved the following theorem:

Theorem 6 (Extended ICA Identifiability Theorem) *Given N independent and stationary signals $s_1, \dots, s_N$, $M < N$ of which are gaussian distributed, the $N - M$ non-gaussian distributed signals can be reconstructed, up to scaling and permutations, from any linear mixture of the type $x = As$, where A is a full-rank $N \times N$ matrix, solving the following optimization problem:*

$$\min_B \sum_{i=1}^{N} H(y_i) \tag{5.25}$$

$$s.t. \ \ BB^T = I \, .$$

[1] In general this is not a restriction because the mean can always be removed during preprocessing of the mixtures.

Notice that in the summation (5.25), the index is up to N since *the number of non-gaussian sources is not assumed to be known a-priori*, thus preserving the "blindness" of the approach to the underlying distribution of the mixed signals.

5.4 Conclusions

An extension to the conventional identifiability theorem for ICA is introduced and rigorously proved. We show that, even when an arbitrary number of gaussian sources is included in the set of independent signals, the conventional mutual information minimization framework is still capable of separating all the non-gaussian signals, without requiring an ad-hoc ICA implementation. In particular, the main result of this Chapter is shown by investigating the properties of the extrema of the optimization problem derived from the separation principle.

Chapter 6

On the Uniqueness of the MI Minimum for Special Classes of Distributions

A large number of Independent Component Analysis (ICA) algorithms are based on the minimization of the statistical mutual information between the reconstructed signals, in order to achieve the source separation. While it has been demonstrated that a global minimum of such cost function will result in the separation of the statistically independent sources, it is an open problem to show that such cost function has a unique minimum (up to scaling and permutations of the signals). Without such result, there is no guarantee that the related ICA algorithms will not get stuck in local minima, and hence, return signals that are statistically dependent. In this chapter, we derive a novel result showing that for the special case of mixtures of two independent and identically distributed (i.i.d.) signals with symmetric, nearly gaussian probability density functions, such objective function has no local minima. This result is shown to yield a useful extension of the well-known entropy power inequality.

6.1 Introduction

In the classic independent component analysis (ICA) framework (see Chapter 3), a generative model is assumed where N independent stationary signals $\mathbf{s} = \{s_1, \ldots, s_N\}$ are mixed through a linear transformation $\mathbf{x} = A\mathbf{s}$. It has been shown (see Theorem 4) that, in absence of noise, there always exist an inverse linear transformation of the type $\mathbf{y} = B\mathbf{x}$, through which the reconstruction of the original signals is possible, up to an arbitrary scaling and permutations of the signals themselves. In particular, if we consider the statistical mutual information (2.19) between the reconstructed signals as a function of the unmixing matrix B, such a function has a global minimum, yielding the source separation [12][22].

Therefore, as we showed in Chapter 3, a vast number of independent component analysis frameworks attempt to find a solution to the following optimization problem:

$$B_{opt} = \arg\min_B I(y_1, \ldots, y_N) \tag{6.1}$$

or an approximate version thereof. When the mixture data $\mathbf{x}$ is sphered prior to the reconstruction ($E[\mathbf{x}\mathbf{x}^T] = I$), we proved in Chapter 3 that the unmixing matrix B must belong to the manifold of orthogonal matrices [74]. Using some basic information theory inequalities, the problem posed in (6.1) can be expressed as:

$$\min_B \sum_{i=1}^{N} H(y_i) \tag{6.2}$$

$$\text{s.t. } BB^T = I\,, \tag{6.3}$$

where $H(y_i)$ is the differential entropy, defined as in (2.3). The equality constraints (6.3) define a sub-group of the Stiefel manifold for the case of square matrices. If we define $F(B) \triangleq \sum_{i=1}^N H(y_i)$, then the gradient of the cost function defined on such manifold is given by [27]:

$$\nabla_m F(B) \triangleq \nabla F(B) - B\nabla F(B)^T B\,. \tag{6.4}$$

where $\nabla F(B)$ is the conventional gradient of $F(B)$ in the Euclidean space. The extrema of the optimization problem (6.2) are found in correspondence to all the matrices satisfying the condition:

$$\nabla_m F(B) = 0 \quad \Rightarrow \quad \nabla F(B)B^T = B\nabla F(B)^T\,. \tag{6.5}$$

Several ICA algorithms optimizing different approximated versions of the cost function (6.1) have been shown to possess good local convergence properties [41] [46]. Although the global minimum of (6.1) is known to yield the desired source separation [18], no proof is available to show that such a function has no local minima. On the other hand, because of the uniqueness of the separation matrix (up to permutations and scaling), that was proved in Chapter 3, Theorem 5, convergence to any solution other than the global would result in a failure to separate the source signals. The problem of convergence to sub-optimal solutions was recently investigated for example in [81] and in [82].

In this chapter, we address the fundamental problem of the uniqueness of the minimum (up to scaling and permutation of the solution) of the information-theoretic cost function in the case of linear mixtures. We show that in the case of mixtures of two symmetric i.i.d. nearly gaussian signals, such cost function is indeed free from spurious local minima. In addition, we derive an interesting connection between the problem defined by (6.2) and the well-known *entropy power inequality*, showing that, under the aforementioned hypotheses, not only this inequality does not hold for dependent random variables, but it is, in fact, always violated (converse entropy power inequality).

6.2 Extrema for Mixtures of Two Nearly Gaussian Sources

We consider the traditional linear framework, where we assume that the mixing matrix A is the 2×2 identity matrix and the original signals are zero-mean, and unit variance. The reconstructed signals can be written as:

$$y_1 = b_{11}s_1 + b_{12}s_2 \tag{6.6}$$

$$y_2 = b_{21}s_1 + b_{22}s_2. \tag{6.7}$$

The general case where the mixing matrix is not the identity matrix can be mapped to this special case through an orthogonal transformation [74], as long as the mixture data is sphered, thus preserving the characteristics of the solution space of (6.2) (in particular, the number of extrema). We restrict our analysis to those cases where the probability density functions of s_1 and s_2 are symmetric and they can be approximated using a Gram-Charlier [90] expansion of the type:

$$f_{s_i}(u) = g(u)\left(1 + \frac{\kappa_{4,s_i}}{24} H_4(u)\right) \quad i = 1, 2. \tag{6.8}$$

where $H_4(u)$ is the 4th order Chebyshev-Hermite polynomial and $g(u)$ is the zero-mean, unit-variance, normal probability density function. The probability density functions of y_1 and y_2, can be approximated as[1]:

$$f_{y_i}(u) \approx g(u)\left(1 + \frac{\kappa_{4,y_i}}{24} H_4(u)\right) \quad i = 1, 2. \tag{6.9}$$

The cumulants κ_{4,y_i} can be computed as:

$$\kappa_{4,y_1} = E[y_1^4] - 3 = b_{11}^4 \mu_{4,s_1} + 6 b_{11}^2 b_{12}^2 + b_{12}^4 \mu_{4,s_2} - 3 \tag{6.10}$$

$$\kappa_{4,y_2} = E[y_2^4] - 3 = b_{21}^4 \mu_{4,s_1} + 6 b_{21}^2 b_{22}^2 + b_{22}^4 \mu_{4,s_2} - 3 \tag{6.11}$$

where μ_{4,s_i} is the 4th order central moment of s_i.

The extrema of the cost function (6.2) must satisfy (6.5). For mixtures of two sources these conditions can be written as:

$$\nabla H(\mathbf{b}_1)\mathbf{b}_2^T = \nabla H(\mathbf{b}_2)\mathbf{b}_1^T\,, \tag{6.12}$$

where $\mathbf{b}_i$ is the ith row of B, and in order to make explicit the dependence of the entropy $H(y_i)$ on $\mathbf{b}_i$, we can define $H(\mathbf{b}_i) \triangleq H(y_i)$, $i = 1, 2$. Given that:

$$\frac{\partial H(b_i)}{\partial b_{ij}} = -\int_{-\infty}^{\infty} (1 + \log f_{y_i}(u)) \frac{\partial f_{y_i}(u)}{\partial b_{ij}} du \tag{6.13}$$

[1] Only the 8^{th} order term of this Gram-Charlier expansion is non-zero and it is neglected.

the identity (6.12) can be written as:

$$\int_{-\infty}^{\infty} \log f_{y_1}(u) \left[b_{21} \frac{\partial f_{y_1}(u)}{\partial b_{11}} + b_{22} \frac{\partial f_{y_1}(u)}{\partial b_{12}} \right] du = \tag{6.14}$$
$$= \int_{-\infty}^{\infty} \log f_{y_2}(u) \left[b_{11} \frac{\partial f_{y_2}(u)}{\partial b_{21}} + b_{12} \frac{\partial f_{y_2}(u)}{\partial b_{22}} \right] du \,.$$

Using (6.9) we can compute explicitly $(i = 1, 2)$:

$$\frac{\partial f_{y_i}(u)}{\partial b_{i1}} = g(u) \left(\frac{1}{6} b_{i1}^3 \mu_{4,s_1} + \frac{1}{2} b_{i1} b_{i2}^2 \right) H_4(u) \tag{6.15}$$

$$\frac{\partial f_{y_i}(u)}{\partial b_{i2}} = g(u) \left(\frac{1}{6} b_{i2}^3 \mu_{4,s_2} + \frac{1}{2} b_{i1}^2 b_{i2} \right) H_4(u) \tag{6.16}$$

Now define:

$$D_1(u, B) \triangleq \frac{1}{g(u)} \left[b_{21} \frac{\partial f_{y_1}(u)}{\partial b_{11}} + b_{22} \frac{\partial f_{y_1}(u)}{\partial b_{12}} \right] \tag{6.17}$$
$$= c_{4,y_1} H_4(u),$$

where:

$$c_{4,y_1} = \frac{1}{6} \left(b_{11}^3 b_{21} \mu_{4,s_1} + b_{12}^3 b_{22} \mu_{4,s_2} \right) + \tag{6.18}$$
$$+ \frac{1}{2} \left(b_{11} b_{12}^2 b_{21} + b_{11}^2 b_{12} b_{22} \right)$$

and:

$$D_2(u, B) \triangleq \frac{1}{g(u)} \left[b_{11} \frac{\partial f_{y_2}(u)}{\partial b_{21}} + b_{12} \frac{\partial f_{y_2}(u)}{\partial b_{22}} \right] \tag{6.19}$$
$$= c_{4,y_2} H_4(u).$$

where:

$$c_{4,y_2} = \frac{1}{6}\left(b_{11}b_{21}^3\mu_{4,s_1} + b_{12}b_{22}^3\mu_{4,s_2}\right) + \\ +\frac{1}{2}\left(b_{11}b_{21}b_{22}^2 + b_{12}b_{21}^2b_{22}\right), \tag{6.20}$$

The following integrals need to be evaluated:

$$\int_{-\infty}^{\infty} g(u)\log f_{y_i}(u) D_i(u,B)du \quad i = 1,2\,. \tag{6.21}$$

where:

$$\log f_{y_i}(u) = -\frac{1}{2}\log(2\pi) - \frac{u^2}{2}\log(e) + \\ +\log\left(1 + \frac{\kappa_{4,y_i}}{24}H_4(u)\right) \quad i = 1,2\,. \tag{6.22}$$

Substituting this expression in (6.21), we obtain:

$$\int_{-\infty}^{\infty} g(u)\left[-\frac{1}{2}\log(2\pi) - \frac{u^2}{2}\log(e) + \\ +\log\left(1 + \frac{\kappa_{4,y_i}}{24}H_4(u)\right)\right] D_i(u,B)du \tag{6.23}$$

Now notice that:

$$\int_{-\infty}^{\infty} g(u)H_4(u)du = 0\,, \tag{6.24}$$

and:

$$\int_{-\infty}^{\infty} u^2 g(u)H_4(u)du = 0\,. \tag{6.25}$$

The integral (6.23) simplifies as:

$$\int_{-\infty}^{\infty} g(u) \log \left(1 + \frac{\kappa_{4,y_i}}{24} H_4(u)\right) D_i(u) du. \tag{6.26}$$

Using the following useful indefinite integral:

$$\int g(u) D_i(u) du = -c_{4,y_i} H_3(u), \tag{6.27}$$

we can integrate (6.26) per parts. If we define $X_i(u) \triangleq \kappa_{4,y_i} H_4(u)/24$, we obtain:

$$\begin{aligned} &\int_{-\infty}^{\infty} g(u) \log \left(1 + X_i(u)\right) D_i(u) du = \\ &= c_{4,y_i} \int_{-\infty}^{\infty} g(u) H_3(u) \frac{X_i'(u)}{1 + X_i(u)} du, \end{aligned} \tag{6.28}$$

where $X_i'(u) = \kappa_{4,y_i} H_3(u)/6$. Using (6.28), we find that (6.14) reduces to:

$$\begin{aligned} &c_{4,y_1} \kappa_{4,y_1} \int_{-\infty}^{\infty} \frac{H_3^2(u)}{1 + \kappa_{4,y_1}/24\, H_4(u)} g(u) du = \\ &= c_{4,y_2} \kappa_{4,y_2} \int_{-\infty}^{\infty} \frac{H_3^2(u)}{1 + \kappa_{4,y_2}/24\, H_4(u)} g(u) du. \end{aligned} \tag{6.29}$$

In particular, when the sources are i.i.d. ($\mu_{4,s_1} = \mu_{4,s_2} \triangleq \mu_4$), we have that $k_{4,y_1} = k_{4,y_2} \neq 0$, and the two integrals on the left-hand-side and on the right-hand-side of (6.29) are always equal. Moreover, because their integrands are non-negative, these integrals are also strictly positive. Thus, the conditions for the gradient to be zero become simply:

$$c_{4,y_1} = c_{4,y_2} \tag{6.30}$$

We can now study the solutions of (6.30) in the space of orthogonal matrices. This is achieved by operating the substitution:

$$\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}. \tag{6.31}$$

Substituting in the expressions for c_{4,y_1} and c_{4,y_2}, we obtain:

$$c_{4,y_1} = -\frac{1}{6}\sin\theta\cos\theta\left[(\mu_4 - 3)(\cos^2\theta - \sin^2\theta)\right] \tag{6.32}$$

$$c_{4,y_2} = \frac{1}{6}\sin\theta\cos\theta\left[(\mu_4 - 3)(\cos^2\theta - \sin^2\theta)\right] \tag{6.33}$$

Thus, (6.30) is satisfied if and only if:

$$(\mu_4 - 3)\sin\theta\cos\theta\cos 2\theta = 0. \tag{6.34}$$

Because of the symmetry of the problem, it suffices to study the zeros of (6.34) in the interval $[0, \pi/2)$. The solutions found in $[\pi/2, 2\pi)$, correspond, in fact, to a permutation or sign change of the rows of B. In this interval, (6.34) has only two zeros, one for $\theta = 0$ corresponding to a minimum of (6.2), and one for $\theta = \pi/4$, corresponding to a maximum of the objective function, thus proving that (6.2) has no local minima.

6.3 An Extension of the Entropy Power Inequality

In this section we will illustrate the connection between the result we just proved and the well-known entropy power inequality (3.51). Let's recall the definition of entropy power of a scalar random variable from Chapter 3:

$$N(s) = \frac{1}{2\pi e}\, e^{2H(s)} \tag{6.35}$$

Given two independent random variables s_1 and s_2, the entropy power inequality (3.51) states that:

$$N(s_1 + s_2) \geq N(s_1) + N(s_2), \tag{6.36}$$

with equality holding if and only if s_1 and s_2 are both normal. The inequality (6.36) can be used to prove the convexity of the entropy under a covariance preserving transformation, i.e. given $0 \leq \lambda \leq 1$, it holds that [24]:

$$H(\lambda s_1 + \sqrt{1-\lambda^2} s_2) \geq \lambda^2 H(s_1) + (1-\lambda^2) H(s_2). \tag{6.37}$$

Now simply define:

$$\lambda = \cos\theta \;\;\Rightarrow\;\; \sqrt{1-\lambda^2} = \sin\theta \quad 0 \leq \theta \leq \pi/2 \tag{6.38}$$

Thus one can write:

$$H(\cos\theta\, s_1 + \sin\theta\, s_2) \geq \cos^2\theta\, H(s_1) + \sin^2\theta\, H(s_2) \tag{6.39}$$

and analogously:

$$H(-\sin\theta s_1 + \cos\theta s_2) \geq \sin^2\theta H(s_1) + \cos^2\theta H(s_2) \tag{6.40}$$

(note that $H(as) = H(s) + \log|a|$, a being a scalar parameter). Simply by adding (6.39) and (6.40) we obtain:

$$H(y_1) + H(y_2) \geq H(s_1) + H(s_2). \tag{6.41}$$

In particular (6.41) proves that the extremum corresponding to $\theta = 0$ is a global minimum of (6.2), regardless of the actual distributions of s_1 and s_2. The uniqueness of this minimum, proved in the previous section, extends the inequality theorem showing that there are no local minima of $H(y_1) + H(y_2)$, for $0 \leq \lambda < 1$.

This result can be used to show that a *converse entropy power inequality* holds, if certain hypotheses are satisfied. Define two random variables z_1 and z_2 as follows:

$$z_1 = \lambda y_1 + \sqrt{1-\lambda^2} y_2 \tag{6.42}$$

$$z_1 = \sqrt{1-\lambda^2} y_1 + \lambda y_2, \tag{6.43}$$

for $0 \leq \lambda < 1$. Because of the uniqueness of the minimum of $H(y_1) + H(y_2)$ in this interval, it follows that the following inequality never holds:

$$H(z_1) + H(z_2) \ngeq H(y_1) + H(y_2), \tag{6.44}$$

unless y_1 and y_2 are obtained from s_1 and s_2, solely through scaling or permutation. In other words, the entropy power inequality is *always* violated by two dependent random variables obtained through an orthogonal projection of independent random variables.

6.4 Conclusions

We introduced a novel result proving the uniqueness of the minimum of the information-theoretic cost function, for the special case of linear mixtures of independent and identically distributed signals with symmetric probability density functions. Such a result, the first of its kind, can be used to show that a converse entropy power inequality holds for this particular class of distributions. In process of deriving a proof for our result, we introduced a useful framework that can potentially be extended in order to investigate the problem for more general classes of distributions. In particular, the method can be used to study whether a converse entropy power inequality, proved for this special case, holds in general. So far, in fact, examples of source distributions for which the uniqueness property is systematically violated have not been identified.

Chapter 7

Learning Linear Non-Gaussian Networks

In this chapter, we show that the ICA framework we introduced in Chapter 4 can be extended to the problem of learning the structure of a specific class of bayesian networks, called *linear non-gaussian networks*. These are a special case of linear belief networks where the stochastic components are assumed to be non-gaussian. A new algorithm is derived for learning the network topology, as well as the local conditional probability distributions, from data. We show that when the modeling assumption holds exactly, the conventionally NP-hard problem of identifying the network structure is reduced to a continuous optimization problem, which is solved by a polynomial complexity algorithm. The general case is addressed introducing the concepts of *relaxation graph* and *relaxation matrix*, in a framework that permits effective handling of model mismatches or poor sampling. Simulation experiments with synthetically generated data show the effectiveness of the proposed technique in correctly reconstructing the network structure.

7.1 Introduction

A central problem in the field of statistical learning is represented by finding a suitable representation of the data that is both accurate and, at the same time, simple and sufficiently general. In this context, graphical models, and in particular *bayesian networks* [9][10][39][51][65][69][72][76][77] have established themselves as a powerful and effective framework for characterizing relationships of dependency, as well as independency, between variables in a model. These networks provide a sparse representation of the joint probability density function of the variables of interest, which is intuitively encoded by their topology. Therefore, under certain assumptions, causality relationships and inference problems can be effectively investigated. Bayesian networks have found vast applications, among others, in the fields of expert systems, fault diagnosis, optimal decision making, and in general in all those problems involving both data modeling and inference capabilities.

Traditionally, relatively small bayesian networks are built from a-priori knowledge, with the aid of an expert whose task is that of specifying the dependence relationships between the variables in a model (i.e. the network topology), as well as the probabilistic framework (i.e. the local conditional probability functions). The manual construction of these networks becomes a laborious and ultimately subjective task when the number of variables in the model becomes very large. On the other hand, in those situations where a large number of data samples is available, one might attempt to reconstruct the network purely from the data. Learning graphical models from measurement data has received the attention of several researchers in recent years [39]. In general, we must distinguish two separate problems. The goal of the first is to learn the local probability distributions when the network topology is given. This is classically known in statistical learning as the regression problem, representing the most general case of supervised learning. The second is the more challenging task of learning both the network topology and the local probability functions. The latter can be viewed as the more general problem of finding an optimal sparse representation of the joint probability density function of a set of variables, given a sample of these variables.

In a standard structure learning algorithm, once a candidate network topology is selected, its local probability distributions are estimated as regression functions and the structure is scored according to a *scoring metric* (data likelihood, MAP, minimal description length are just a few examples). The goal of the search algorithm is, therefore, to identify the network structure (or the ensemble of network structures) yielding the largest value of the selected scoring metric. Alternatively, the problem can be posed as a *constraint satisfaction* problem, where a candidate network structure that best matches the patterns of conditional mutual independence observed in the data is sought [78]. Unfortunately, even in the case when the maximum number of parents of any given node is constrained, the problem of searching over the set of all possible structures is NP-hard [39], in both approaches. Nevertheless, several approximate methods have been proposed, including heuristic algorithms based on greedy search with random restarts [16], simulated annealing, best-first search methods [57], and minimal description length (MDL) techniques [83]. Specifically, heuristic methods aiming at the discovery of patterns of conditional independence between sets of variables, based on the estimation of their mutual information [19], have been suggested [33, 34]. In particular, the "sparse candidate algorithm" proposed by Friedman et al. in [34] estimates the pairwise mutual information between the variables in the model to heuristically identify candidate nodes belonging to the same markov blanket. In the same work, the idea of using the pairwise conditional mutual information as a measure of modeling mismatch is also introduced.

In this chapter, we introduce a novel approach to learning bayesian networks, for a class of continuous graphical models which we will refer to as *linear non-gaussian networks*. The proposed model differs from standard linear gaussian networks (see for example [35]) in the fact that a strictly non-gaussian conditional probability density model is assumed at each node. Conventionally, gaussian local probability models are employed because of their simple formulation and because closed-form expressions of several types of estimators are available, such as the maximum likelihood (ML) and the maximum a-posteriori (MAP), when the prior distributions belong to the exponential family [7]. However, especially when one tries to model real world data, the normality assumption can turn out to be too restrictive. In Section 7.2, the relationship between the non-gaussianity assumption for the local distributions and the model identifiability is investigated. In particular, we will prove that, for a linear model where the non-gaussianity assumption holds, an optimal solution to the maximum likelihood problem associated with learning the network structure can be found in polynomial time, solving a continuous optimization problem.

In Section 7.3, we will show that the proposed framework represents a formalization of the approach defined in [34], which is capable of systematically identifying the optimal (according to the mutual information criterion) pattern of conditional independencies in the graphical model. The method provides an exact rather than a heuristic solution to the structure learning problem when the model is constrained to the class of linear non-gaussian networks, under several different scoring metrics. The novel concepts of *relaxation matrix* and *relaxation graph* are introduced, in order to deal with model mismatches, as well as poor sampling issues. A set of simulation experiments, discussed in Section 7.4 show that the proposed framework can be used to effectively learn the network topology, by systematically extracting information about patterns of conditional independence in the data.

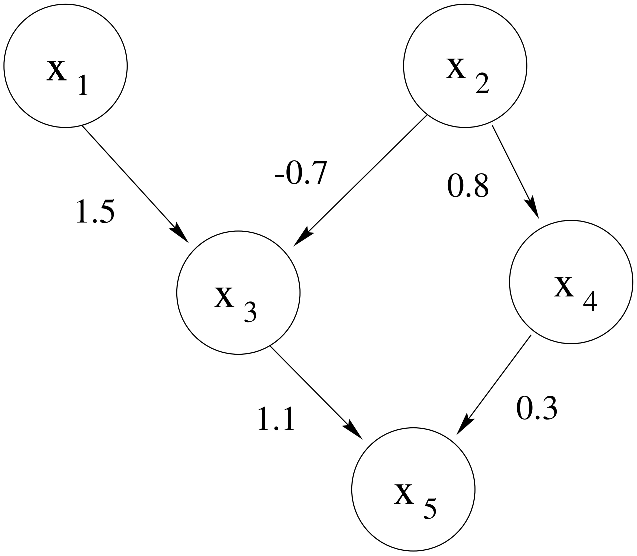


Figure 7.1. An example of a belief network topology.

7.2 Linear Non-Gaussian Networks

7.2.1 The Model

Given a set of N random variables $x_1, \ldots, x_N$, a *belief network* or *bayesian network* consists of a graph G encoding conditional statistical independence relationships among such variables, and of a set of conditional probability density functions (also known as local probability distributions) necessary to define the joint pdf over the set of variables [10][35][48][77]. Given a network structure G, we have that:

$$p(\mathbf{x}) = \prod_{j=1}^{N} p(x_j|\mathbf{Pa}_j), \tag{7.1}$$

where $\mathbf{Pa}_j$ is the set of parents of node j (an example is shown in Figure 7.1). In particular, we are interested in the case where the x_j's are continuous random variables and the following model for the conditional probability density functions $p(x_j|\mathbf{Pa}_j)$ holds:

$$p(x_j|\mathbf{Pa}_j) = f_{u_j}(t - \sum_{i \in \mathbf{Pa}_j} a_{ij}x_i), \quad j = 1, \ldots, N, \tag{7.2}$$

where $\{f_{u_1}, \ldots, f_{u_N}\}$ is a set of *independent* and *non-gaussian* pdfs. This is equivalent to assuming the following generative model:

$$x_j = u_j + \sum_{i \in \mathbf{Pa}_j} a_{ij}x_i \quad j = 1, \ldots, N, \tag{7.3}$$

where a_{ij} are arbitrary real coefficients. At each node j, realizations of the parent variables $\mathbf{Pa}_j$ are linearly combined to give the mean of the child node.

The model presents a close resemblance to the one discussed in [35], with the fundamental distinction that the local probability functions f_{u_j} are non-gaussian. It can also be viewed as a sparse linear regression model [10], where the noise is non-gaussian, and with the important difference that the topology of the relations regressor-regressed variables is not known a-priori.

7.2.2 Network Properties

To completely define the model one needs to provide the set of independent probability density functions f_{u_j} and the matrix A, which defines the topology of the network:

$$A \triangleq \begin{cases} a_{ij} \neq 0 & \exists \ \text{arc} \ \ i \rightarrow j \\ 0 & \text{otherwise} \end{cases} . \tag{7.4}$$

Certain properties of the network can be characterized noticing that this matrix is a *weighted adjacency matrix* (see for example [17] for a definition), which differs from a standard adjacency matrix in the fact that non-zero entries represent the weights associated with each directed arc. For example the matrix A associated with the graph of Figure 7.1 is given by:

$$A = \begin{bmatrix} 0 & 0 & 1.5 & 0 & 0 \\ 0 & 0 & -0.7 & 0.8 & 0 \\ 0 & 0 & 0 & 0 & 1.1 \\ 0 & 0 & 0 & 0 & 0.3 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{7.5}$$

The following result about the acyclicity of the graph holds. Define the matrix $\bar{A} \triangleq [|a_{ij}|]$, then the graph G is *acyclic* if and only if:

$$\text{trace}(\bar{A}^n) = 0 \qquad n = 1, \ldots, N. \tag{7.6}$$

The proof can be easily derived from a standard result of graph theory, which states that the ij element of the nth power of the adjacency matrix is equal to the number of paths between node i and node j in the directed graph [17]. Another property that we will not prove is that the graph defined by A is a minimal belief network with respect to equation (7.1), i.e. it is not possible to remove an arc from G without violating this equation. Therefore, a *minimal linear non-gaussian belief network* will be defined as the graph G where a directed arc between node i and node j exists if and only if $a_{ij} \neq 0$, and where the local probability distributions are given by (7.2).

7.3 Learning The Network Structure

In this section, the identifiability of a linear non-gaussian network is investigated. In particular, the fundamental problem of learning the network topology from data is addressed.

7.3.1 Model Identifiability

We can write the generative equations of the model (7.3) in a matrix-vector form as follows:

$$\mathbf{x} = A^T\mathbf{x} + \mathbf{u}, \tag{7.7}$$

where $\mathbf{x} = [x_1, \ldots, x_N]^T$ and $\mathbf{u} = [u_1, \ldots, u_N]^T$. From (7.7), we obtain:

$$(I - A^T)\mathbf{x} = \mathbf{u} \tag{7.8}$$

It is possible to show that if the directed graph G is acyclic, then the matrix $(I - A^T)$ has full rank. For any directed acyclic graph, in fact, it is always possible to define a partial ordering over the set of nodes such that if $i \in \mathbf{Pa}_j$, then $i < j$. Therefore, without loss of generality, we can assume that A is upper-triangular with zeros on the diagonal, and its eigenvalues are all zero. Consequently:

$$\mathrm{eig}(I - A^T) = \mathrm{eig}(I) - \mathrm{eig}(A^T) = [1, \ldots, 1]^T, \tag{7.9}$$

which proves that $(I - A^T)$ is full-rank. Therefore we can write:

$$\mathbf{x} = (I - A^T)^{-1}\mathbf{u} = B\mathbf{u}, \tag{7.10}$$

where $B \triangleq (I - A^T)^{-1}$. When the random variables u_j are statistically independent, the problem defined by (7.10) is known as independent component analysis [18] or independent factor analysis [1]. The goal is to obtain an estimate of B^{-1} without any prior knowledge on the structure of this matrix or on the distribution of the unknown factors u_j. A maximum likelihood (ML) estimator for problem (7.10) can be derived recalling that, given a matrix B and an estimate of the pdf $f_{\mathbf{u}} = \prod_{j=1}^{N} f_{u_j}$ we have that:

$$p(\mathbf{x}|B, f_{\mathbf{u}}) = |\det B|^{-1} f_{\mathbf{u}}(B^{-1}\mathbf{x}). \tag{7.11}$$

Assuming that M samples are drawn independently, the normalized log-likelihood $L(B, \mathbf{f})$ for the entire data can thus be written as:

$$L(B, \mathbf{f}) = \frac{1}{M} \log \prod_{k=1}^{M} |\det B|^{-1} f_{\mathbf{u}}(B^{-1}\mathbf{x}^{(k)}) \tag{7.12}$$

$$= \frac{1}{M} \sum_{k=1}^{M} \log f_{\mathbf{u}}(B^{-1}\mathbf{x}^{(k)}) - \log |\det B|,$$

where $\mathbf{x}^{(k)}$ is the kth data sample. The hypothesis under which the model described by (7.10) is identifiable are derived in [18]. When the contrast function used to reconstruct the unknown factors is a functional of the marginal pdfs f_{u_j} (the likelihood function and the mutual information are two examples), it can be proved that a necessary and sufficient condition for the model identifiability is that at the most one of the unknown factors follows a gaussian distribution. This well known result can be understood by observing in equation (7.12) that if the pdfs are gaussian, then the matrix that maximizes the likelihood function is defined only up to an orthogonal matrix, since uncorrelated gaussian variables are also independent.

7.3.2 Mutual Information

In this section, we will show that maximizing the log-likelihood function is equivalent to minimizing the errors introduced by making erroneous assumptions about the conditional mutual independence between variables in the network. A measure of the error introduced by assuming that a distribution of a random variable x is equal to $q(x)$, when the true distribution is $r(x)$, is given by the Kullback-Leibler distance:

$$D(r||q) = \int r(x) \log \frac{r(x)}{q(x)} dx. \tag{7.13}$$

We have in Chapter 2 that this measure is always non-negative and it is equal to zero if and only $r(x) = q(x)$. Therefore, the following scoring metric can be used to measure the discrepancy between the joint pdf of the variables and the estimated conditional independency model:

$$\begin{aligned}\mathcal{D} \triangleq D\left(p(\mathbf{x}) \,\middle\|\, \prod_{j=1}^{N} p(x_j|\mathbf{Pa}_j)\right) = \qquad (7.14)\\ = \int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{\prod_{j=1}^{N} p(x_j|\mathbf{Pa}_j)} d\mathbf{x}.\end{aligned}$$

In general, estimating this quantity is not possible. However, when the modeling assumption (7.3) holds we have that an estimate of $p(x_j|\mathbf{Pa}j)$ is given by $|\det B|^{-1}\tilde{f}_{\mathbf{u}}(B^{-1}\mathbf{x})$, where $\tilde{f}_{\mathbf{u}}$ is not known. Hence we can write:

$$\mathcal{D} = \int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{|\det B|^{-1}\tilde{f}_{\mathbf{u}}(B^{-1}\mathbf{x})} d\mathbf{x} \tag{7.15}$$

$$= -H(\mathbf{x}) - \int p(\mathbf{x}) \log \left(|\det B|^{-1}\tilde{f}_{\mathbf{u}}(B^{-1}\mathbf{x})\right) d\mathbf{x} \tag{7.16}$$

$$= -H(\mathbf{x}) - \log|\det B|^{-1} - \int p(\mathbf{x})\tilde{f}_{\mathbf{u}}(B^{-1}\mathbf{x}) d\mathbf{x}. \tag{7.17}$$

where $H(\mathbf{x}) = -\int p(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x}$ is the differential entropy of $\mathbf{x}$. If the distribution of the variables in the model is assumed to be stationary, then $H(\mathbf{x})$ is a constant. Moreover, if we approximate the integral in (7.17) with the summation over the data samples, we obtain:

$$\max_{B,\tilde{f}} \mathcal{D} = \max_{B,\tilde{f}} -\frac{1}{M} \sum_{k=1}^{M} \tilde{f}_{\mathbf{u}}(B^{-1}\mathbf{x}^{(k)}) + \log|\det B| \tag{7.18}$$

which is equivalent to maximizing the log-likelihood (7.12).

7.3.3 The Optimization Framework

Clearly, in order to maximize the likelihood function (7.12) an estimate of the (unknown) marginal pdfs f_{u_j} is required. We have solved an equivalent problem in Chapter 4, where we derived a non-parametric estimation technique, performing the joint estimation of the distributions of the unknown factors and of the inverse matrix. The quantity $f_{\mathbf{u}}(B^{-1}\mathbf{x}^{(k)})$, as well as its derivatives with respect to the elements of B^{-1} are computed from the sample data. This is achieved using a kernel density estimation technique to approximate the marginal densities f_{u_j}:

$$f_{u_j}(t) = \frac{1}{Mh} \sum_{m=1}^{M} \phi\left(\frac{t - \mu_{mj}}{h}\right), \quad j = 1, \ldots, N, \tag{7.19}$$

where h is the kernel bandwidth and ϕ is the gaussian kernel:

$$\phi(t) \triangleq \frac{1}{\sqrt{2\pi}}\, e^{-\frac{t^2}{2}} \ . \tag{7.20}$$

The kernel centroids μ_{mj} are computed as:

$$\mu_{mj} = \mathbf{z}_j \mathbf{x}^{(m)} \tag{7.21}$$

where $\mathbf{z}_j$ is the jth *row* of the matrix $Z \triangleq B^{-1}$. This estimator is asymptotically unbiased and efficient, and it is shown to converge to the true pdf under several measures. Moreover, it is a continuous and differentiable function of the elements of the matrix Z, its gradient being given by:

$$\nabla f_{u_j}(t) = \frac{1}{Mh^2} \sum_{m=1}^{M} \mathbf{x}^{(m)}(t - \mathbf{z}_j \mathbf{x}^{(m)}) \phi \left(\frac{t - \mathbf{z}_j \mathbf{x}^{(m)}}{h} \right). \tag{7.22}$$

Using the kernel expansion of the source distributions, we can derive a close form expression for the pdf of the one-dimensional reconstructed factors, evaluated at the data points as:

$$f_{u_j}(\mathbf{z}_j \mathbf{x}^{(k)}) = \frac{1}{Mh} \sum_{m=1}^{M} \phi \left(\frac{\mathbf{z}_j \left(\mathbf{x}^{(k)} - \mathbf{x}^{(m)} \right)}{h} \right). \tag{7.23}$$

Therefore, recalling that $f_{\mathbf{u}} = \prod f_{u_j}(t)$, we can re-write the likelihood function as follows:

$$\begin{aligned} L(B, \mathbf{f}) \quad = \quad & \log |\det Z| \; + \qquad\qquad (7.24) \\ & + \frac{1}{M} \sum_{j=1}^{N} \sum_{k=1}^{M} \log \left[\frac{1}{Mh} \sum_{m=1}^{M} \phi \left(\frac{\mathbf{z}_j \left(\mathbf{x}^{(k)} - \mathbf{x}^{(m)} \right)}{h} \right) \right]. \end{aligned}$$

We derived in Chapter 4 an efficient algorithm for the maximization of this non-linear cost function based on the Newton method. In general, a solution to problem (7.24) can be obtained only up to a permutation matrix. It is quite evident, in fact, looking at the expression of the likelihood function (7.12), that any permutation of the unknown factors results in the same value of the likelihood[1]. A generic expression for the estimate of B^{-1} can thus be written as:

[1] Notice that, unlike the independent component analysis problem, we do not have to deal with the scaling of the unknown factors. In fact, we are not directly interested in the actual distribution of the factors but only in the resulting expression for the local conditional pdfs.

$$Z = PB^{-1} = P(I - A^T), \tag{7.25}$$

where P is an arbitrary permutation matrix. In order to reconstruct an estimate of A one needs to identify the permutation matrix $\tilde{P}$, such that:

$$\tilde{Z} = \tilde{P}Z = (I - A^T), \tag{7.26}$$

or equivalently such that $\tilde{P}P = I$. Ideally, if the model holds exactly the matrix $\tilde{P}$ can be identified by observing that one just needs to rearrange the rows of Z to make it upper triangular. Thus, A can be estimated as:

$$\hat{A} = I - Z^T \tilde{P}^T. \tag{7.27}$$

7.3.4 The Relaxation Graph

In general, one might expect that some of the modeling assumptions will not hold strictly, for example because the sample size is not adequate, or just because a linear model does not accurately fit the data. As a consequence, a permutation matrix $\tilde{P}$ that makes the matrix $\tilde{Z}$ upper triangular is not guaranteed to exist in general.

In order to deal with model mismatches, we introduce the definitions of relaxation matrix and relaxation graph. Starting from a graph G and the corresponding matrix A, we can build a new graph G_ϵ obtained from G by simply adding directed arcs of strength ϵ until the resulting graph is complete (see example in Figure 7.2). We will refer to G_ϵ as the *relaxation graph* of G, and to the corresponding matrix A_ϵ, obtained from A substituting ϵ to the zero elements of A, as the *relaxation matrix*. Clearly, when $\epsilon \to 0$ and $G_\epsilon \to G$, since:

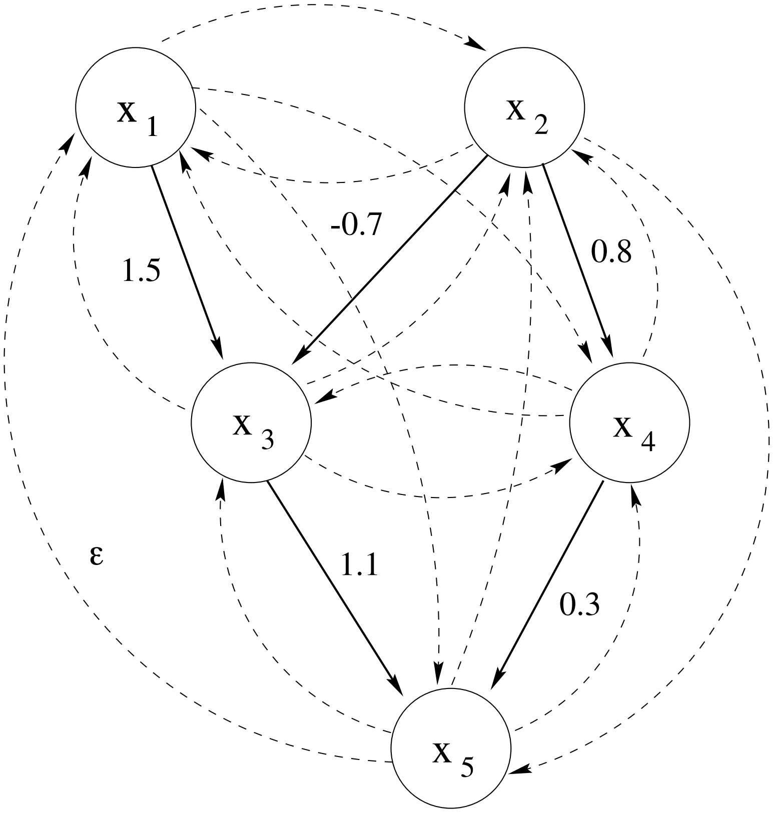


Figure 7.2. Relaxation graph obtained from the belief network of Figure 7.1.

$$\lim_{\epsilon \to 0} A_\epsilon = A \tag{7.28}$$

In general the acyclicity property is guaranteed if:

$$\mathrm{trace}({\bar{A}_\epsilon}^n) = \delta, \quad \delta \to 0, \quad n = 1, \ldots, N. \tag{7.29}$$

Even in the case when δ is small but non-zero, we can still talk about *quasi-acyclicity* to characterize the concept that, although loops might be present, they are statistically insignificant.

When the model does not strictly hold, we can still apply the linearization defined by (7.7) and try to estimate the optimal relaxation matrix. This is obtained by finding a matrix $\tilde{Z}_\epsilon = \tilde{P}_\epsilon Z$ that is "as lower triangular as possible" according to some measure, thus resulting in the essential acyclicity of the learned graph G_ϵ. We suggest the following optimality criterion for the estimation of such matrix:

$$\tilde{P}_\epsilon = \arg\min_{\tilde{P}_\epsilon} \sum_{i=1}^{N} \sum_{j=i+1}^{N} [\tilde{Z}_\epsilon]_{ij}^2, \tag{7.30}$$

which is a convex optimization problem defined on a discrete set. In general, solving the problem defined by (7.30) is NP-hard. On the other hand, when the linear model is sufficiently accurate, a simple greedy search algorithm can be used to estimate $\tilde{P}_\epsilon$. An example of heuristic method is a *top-down* approach, which consists of starting from identifying the best candidate for the first row of $\tilde{Z}_\epsilon$, simply by choosing the row of Z whose last $N-1$ elements have the smallest sum of the squares. Then one simply proceeds by sequentially identifying the next row down, choosing among the rows of Z that have not been yet selected. The algorithm has an asymptotic complexity that grows as $\mathcal{O}(N^2)$. This is the approach we followed in our simulations (section 7.4).

7.4 Simulation Experiments

The validity of the proposed learning technique was evaluated in two simulation experiments, using synthetic data generated according to the joint probability density function associated with the network structure of Figure 7.1. For the local conditional probability density functions, a super-kurtotic symmetric power exponential density ($\alpha = 0.6$), with a theoretical kurtosis $k_4 = 1.52$, was chosen (shown

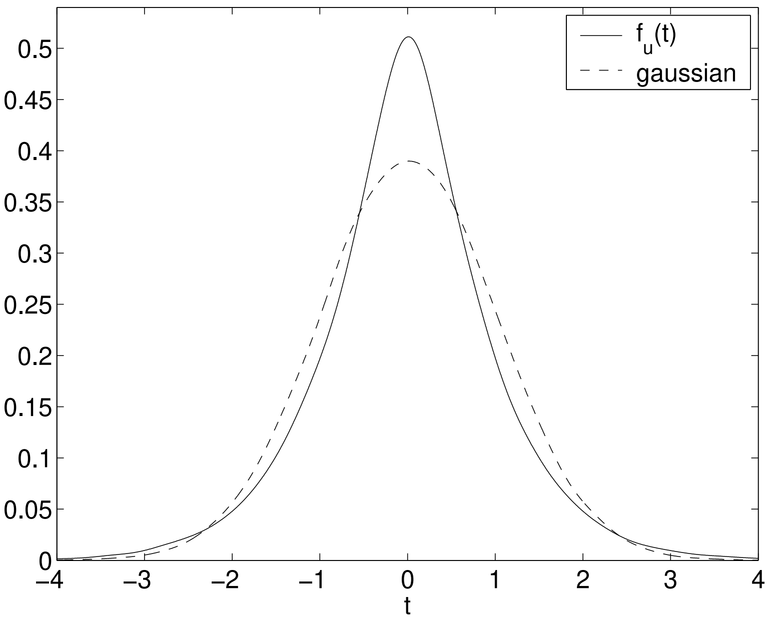


Figure 7.3. The synthetic data used in the simulation experiment was generated using a power exponential density ($k_4 = 1.52$) as a model for the local conditional probability density functions. A gaussian distribution is also shown for comparison.

in Figure 7.3, where a gaussian pdf is also shown for comparison). In the first experiment, 5000 independent samples were drawn for each independent factor u_j, and samples from the variables were generated according to the linear model. The independent factors, as well as an estimate of B^{-1}, were computed using the non-parametric method described in Section 7.3 (the algorithm is analogous to the one described in Table 4.1). In this case, because of the large sample size, the greedy search algorithm resulted in the optimal solution of problem (7.30). The following relaxation matrix was estimated (compare with (7.5)):

$$\hat{A}_\epsilon = \begin{bmatrix} 0.000 & 0.029 & \mathbf{1.496} & -0.038 & 0.027 \\ 0.015 & 0.000 & \textbf{-0.709} & \mathbf{0.801} & -0.028 \\ 0.047 & -0.036 & 0.000 & 0.030 & \mathbf{1.097} \\ 0.018 & 0.002 & -0.004 & 0.000 & \mathbf{0.306} \\ -0.043 & 0.016 & -0.006 & -0.015 & 0.000 \end{bmatrix} \tag{7.31}$$

Figure 7.4 shows the corresponding relaxation graph (not all the weights associated to the edges are shown). In the second simulation, only 300 samples randomly selected from the previous dataset were used in the experiment. This example is more challenging because, although the modeling assumption is not violated, the small sample size makes more difficult to identify the conditional independencies in the model. For this case, the algorithm estimated the following relaxation matrix, using the greedy search algorithm to identify the optimal permutation matrix:

$$\hat{A}_\epsilon = \begin{bmatrix} 0.000 & -0.057 & \mathbf{1.434} & 0.176 & 0.095 \\ 0.164 & 0.000 & \textbf{-0.831} & \mathbf{0.647} & -0.019 \\ 0.123 & -0.064 & 0.000 & -0.189 & \mathbf{1.066} \\ 0.083 & 0.001 & 0.195 & 0.000 & \mathbf{0.310} \\ -0.083 & 0.041 & 0.025 & 0.040 & 0.000 \end{bmatrix} \tag{7.32}$$

The corresponding relaxation graph is shown in Figure 7.5. Although the network topology is learned correctly, the poor sampling resulted in some moderate estimation errors, introducing some week loops in the network. Overall, these preliminary simulation experiments seem to suggest that the proposed method is fairly robust against minor model mismatches, for example due to poor sample sizes.

7.5 Discussion

In the previous section, we demonstrated that when the sample data is generated according to the non-gaussian linear model, then a simple greedy search algorithm generally results in the optimal solution. However, in presence of model mismatches and for networks with a large number of nodes, a more sophisticated relaxation technique needs to be devised in order to solve problem (7.30). On the other hand, the search space defined by this problem is already greatly reduced when compared to the original space, defined by all possible network topologies. Along this line, there are at least a few issues that deserve further investigation. To identify the relationships between the reduced search space defined by (7.30) and the original one is one of them. Even more crucial is to understand what kind of information about the underlying network structure can be retrieved, using the proposed approach, when the model does not hold at all (for example because non-linear relationships exist between the variables in the model). In particular, a key issue requiring further investigation is whether a similar method, allowing a systematic identification of the patterns of conditional independence between variables in the model, can be devised when a non-linear generative model is assumed.

Finally, although the method was derived using the data-likelihood as scoring function, alternative separable scoring metrics, such as maximum-a-posteriori, BIC [39], or minimal description length, could be adopted with minimal modifications to the framework.

7.6 Conclusions

A novel framework for learning the topology of linear non-gaussian networks was derived. We showed that for this specific class of linear belief network, the conventionally NP-hard problem of learning the structure from data, can be simplified to a continuous optimization problem that can be solved in polynomial time. Preliminary simulation results confirmed the validity of the proposed approach and seemed to suggest that the method is robust to poor sampling and modeling imperfections.

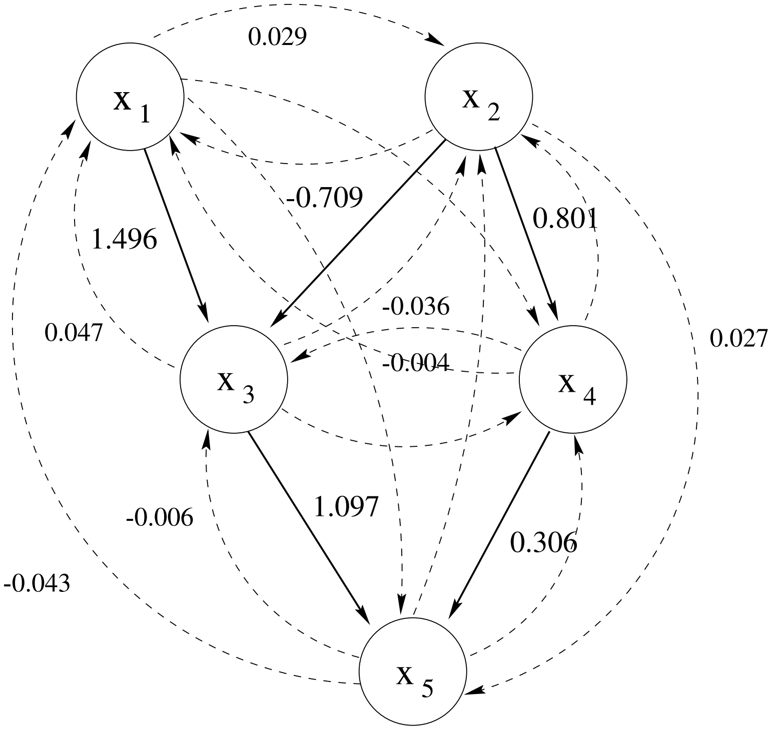


Figure 7.4. **Relaxation graph learned from data in the first simulation experiment. Because of the large sample size the weights of the graphical model are learned very accurately and a simple thresholding would results in a good estimate of the original model.**

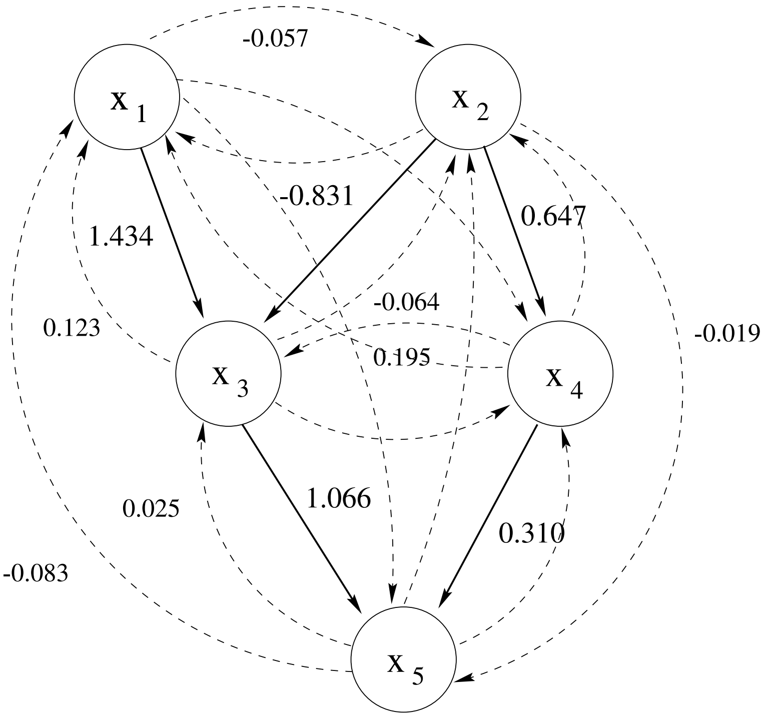


Figure 7.5. **Relaxation graph learned from data in the second simulation experiment. The network topology is identified also in this case. However, the poor sampling results in larger errors in the estimation of the arc weights.**

Chapter 8

Learning Conditional Co-Expression Patterns in Gene Expression Data with Information Theoretic Exploratory Methods

In this chapter, a novel application of information theoretic learning to biological problems will be introduced. Specifically, we will show how some of the methods that have been discussed in Chapter 2 and 3 can be applied to the study of expression levels of genes. We will first provide a concise introduction on in-vitro gene expression levels measurement using DNA microarray technology. It will be shown, then, how the statistical analysis of DNA microarray data presents several challenges from the point of view of statistical learning, and current approaches for analyzing such datasets will be briefly reviewed. In particular, we will focus on how information theoretic learning can contribute in pursuing the ultimate goal of such analysis, which is to shed some light on the complex interactions between gene expressions and on their regulatory mechanisms.

8.1 Gene Expression Profiling Using DNA Microarray Data Technology

In the last three decades, several advances in biotechnology have revolutionized the field of life science, providing biologists with new means of access to biological information. Among these are molecular cloning, automatic DNA sequencing, and polymerase chain reaction (PCR) [11]. More recently, DNA Microarray technology has further innovated the field with the introduction of an experimental technique allowing the simultaneous monitoring of the expression levels of all genes in a particular organism. Before the advent of microarray technology, molecular biologists were capable of measuring the expression levels of only a limited number of genes at the same time, generally by using experimental procedures that could not be automated or standardized. With the introduction of DNA microarrays, not only whole-genome expression profiling has become a common procedure, but the standardization of this technology has greatly reduced the costs involved in the experimental procedure.

Microarray chips are simply comprised of pre-arranged sets of DNA sequences, which are laid out on the chip in selected known locations. Such DNA sequences can be gene sequences, or parts thereof, or any kind of sequence which is part of a known genome. In general, a microarray chip can contain a few hundreds to as many as tens of thousands of these sequences. As more and more genomes are fully sequenced, the capability of building microarray chips capable of detecting the expression levels of the entire set of genes of a given organism has dramatically increased.

The basic principle behind microarray technology is that genes which code for proteins are transcribed into messenger RNA (mRNA) in the cell nucleus. The mRNA in turn is translated into a protein by ribosomes in the cytoplasm. Therefore, the average transcription level of a given gene can be assumed to be directly proportional to the concentration of the corresponding mRNA in the cell at a given time. In order to be used in microarray assays, the mRNA strains need to be extracted from the cells and purified. Since free mRNA molecules tend to become highly unstable and degrade very quickly, it is a well-established technique to reverse transcribe them back into more stable DNA strains. This procedure results in the synthesis of cDNA (complementary-DNA) strings, so called because their sequences are the complement of the original mRNA sequences.

The microarray chips are designed in such a way that specific cDNA sequences (also referred to as "cDNA probes") bind selectively to specific sites. In order to measure the amount of cDNA that binds in any given site, the cDNA samples are labeled usually with fluorescent dyes. The array holds hundreds or thousands of spots, each of which contains a different DNA sequence. If a probe contains a cDNA whose sequence is complementary to the DNA on a given spot, that cDNA will hybridize to the spot, where it will be detectable by its fluorescence. In this way, every spot on an array is an independent assay for the presence of a different cDNA probe.

It is a common practice, also in order to reduce measurement noise, to conduct comparative hybridization experiments, where the amounts of mRNA is measured in two different cell populations. In this case, the same microarray chip can be used, but the cDNA probes from the two populations are labeled using dyes of different colors. Once the cDNA probes have been hybridized to the array and any loose probe has been washed off, the array must be scanned to determine how much of each probe has bound to each spot. This is achieved through a device which detects the intensity of emitted light at each spot, after it is excited using a laser.

8.2 DNA Microarray Data Statistical Analysis

Before the advent of microarray technology, the analysis of the relatively small amount of transcription data produced could easily be managed by applying simple statistical analysis tools. The sudden surge of available biological data, deriving from the introduction of microarray technology, was accompanied by a parallel increase in the demand of analysis tools that could deal with the massive amount of data produced.

The first issue one has to deal with when analyzing gene expression data is the identification of all the sources of noise. In the case of DNA microarray, we can distinguish between two major sources of error [62]: the first is biological and can be attributed to the inherent variability of cell populations used during an experiment. We must keep in mind, in fact, that when the expression levels are measured as time series, the same cell culture cannot be re-used to obtain multiple measurements. Therefore, several cultures are grown simultaneously and for each one the mRNA is extracted at a given time to obtain a sample in the time-series. The second source of error derives from the measurement process itself. Several steps are required in order to obtain a data sample: mRNA extraction and purification, complementary DNA synthesis and labeling, hybridization of cDNA to the DNA arrays, and imaging of the hybrids. The efficiency of each step is unknown and the errors introduced at each step are multiplicative. In general, it can be shown that if the detected cDNA abundance in each spot is normalized against the total image intensity of all the spots detected, it is possible, in theory, to correct for the errors introduced by the unknown purification efficiency as well as for the unknown efficiency of the scanning device [62]. On the other hand, the efficiency factors associated with the labeling and the hybridization processes are gene dependent and cannot be compensated. A further improvement in terms of accuracy could be achieved by performing comparative studies, where the gene expression levels in condition 1 is compared to those in condition 2, by simultaneously measuring the abundance of mRNA from the two cultures on the same slide. However, the benefits of performing such comparison are limited by likely differences in total mRNA levels in the two cultures, which may be quite substantial considering that the growth rate of the organism might vary considerably.

Since the measured transcription levels are relative, due to the normalization step, the resulting distribution is highly skewed, with the down-regulated genes assuming values between 0 and 1 and the up-regulated genes between 1 and plus infinity. Therefore, it is a common practice to work with the logarithm of the relative expression levels rather than with the absolute values, thus resulting in a symmetric distribution of the measurement data.

Although DNA microarray technology resulted in the breakthrough capability of obtaining high-throughput gene expression measurements, their statistical analysis presents several challenges. Even when the noise present in the data is contained within reasonable levels, we are faced with the additional issues of having to deal with frequent missing values and poor sampling. Although the sampling characteristics of the problem may be improved through repeat experiments and averaging, it turns out that the repeatability of the experiments is itself an open problem, since several unpredictable factors intervene in the measurement process. For example a large variability in the experimental data can result from assaying cell cultures grown at different times, from using different types of microarray chips and optical readers, or simply from different laboratory personnel performing the experiments. Only a few of these sources of variability can be tightly controlled.

Recently, Tseng *et al.* [88] have developed an experimental framework capable of assessing statistical confidence intervals for each single gene. The method is based on repeat experiments with independent cultures measured on different slides, and on the wide use of calibration experiments. The resulting data is analyzed using a hierarchical Bayesian model where a Markov-Chain Monte Carlo method is used to compute confidence intervals.

As a result of both the limited sample size and the measurement noise, most standard statistical learning approaches have hardly found application to the analysis of gene expression data. In general, the complete genome of even a simple organism, such as a prokaryote, is made up of several thousands of genes. As a consequence, any attempt to build a model of such network of genes will have to deal somehow with the problem of learning the statistical properties of the network, when only few data samples are available. Therefore, the conventional learning framework in which the set of parameters that need to be estimated is considerably smaller than the available sample data is somehow reversed. Several attempts to adapt well-known statistical learning frameworks, such as Bayesian Networks [34], Support Vector Machines (SVM) [38], K-means clustering or Self-Organizing Maps (SOM) [56], have led to results whose biological interpretation remains unclear.

For this reason, members of the biology community have resorted to simpler analysis tools, which are widely accepted mainly because of their straightforward biological interpretation: among these are Pearson correlations and its extension to gene clustering by hierarchical agglomeration [28]. The idea consists of finding patterns of co-expressions between pairs of genes, or in simple terms of identifying which genes profiles appear to be concomitantly up or down regulated. A simple technique based on measuring correlations between gene expression profiles, paired with genome sequencing information, was demonstrated to be a reliable predictor of gene operons [85], being capable of accurately identifying the structural genes belonging to the same operon.

8.3 Conditionally Informative Clusters of Genes

A consequence of adopting pairwise correlations in order to identify genes which are co-expressed is that issues such as noise level or poor sampling become less restrictive, simply because the learning problem is extremely simplified. Even when a few samples are available in a DNA assay time-course experiment, it is still possible to get accurate estimates of the cross-correlation between gene profiles. In other terms, the learning problem is reduced to estimating the statistical properties of only two variables at the time, greatly reducing the complexity of the estimation problem from the case when learning the relationships between all genes simultaneously is the goal. The gene expression analysis tools we have developed represent an extension of the idea of investigating the patterns of co-expression within sub-networks of genes, and, at the same time, of retaining a biologically meaningful interpretation of the results.

The basic idea consists of identifying groups of genes whose expression levels are co-expressed *only conditionally on the expression levels of other genes.* We will show that, although the proposed approach is capable of performing a more traditional cross-correlation analysis, we will choose to explicitly ignore simple dependencies, in order to focus on the novel aspect of detecting patterns of co-expression that appear under statistical conditioning.

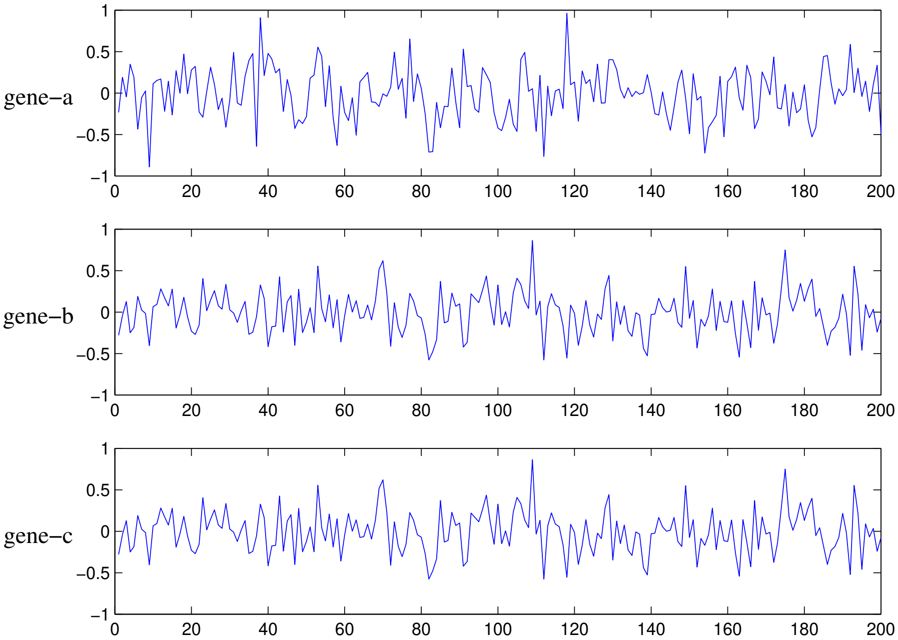


Figure 8.1. The plot shows the expression time-courses of three hypothetical genes, generated from synthetic data.

8.3.1 Conditional Mutual Information as a Measure of Conditional Co-expression

Figure 8.1 shows the time-courses of three genes whose expression levels where generated from synthetic data in order to simulate a case of conditional co-expression. The scatter plots of gene-a, gene-b, and gene-c in figure 8.2 show no clear pattern of dependency between these genes. On the other hand, when the points in the scatter plot of gene-a versus gene-b are color coded according to the value of gene-c[1] a clear pattern of linear correlations appears. This is shown in figure 8.3: when gene-c is up-regulated the other two genes appear to be posi-

[1] In this case the expression levels of gene-c are quantized according to three discrete levels.

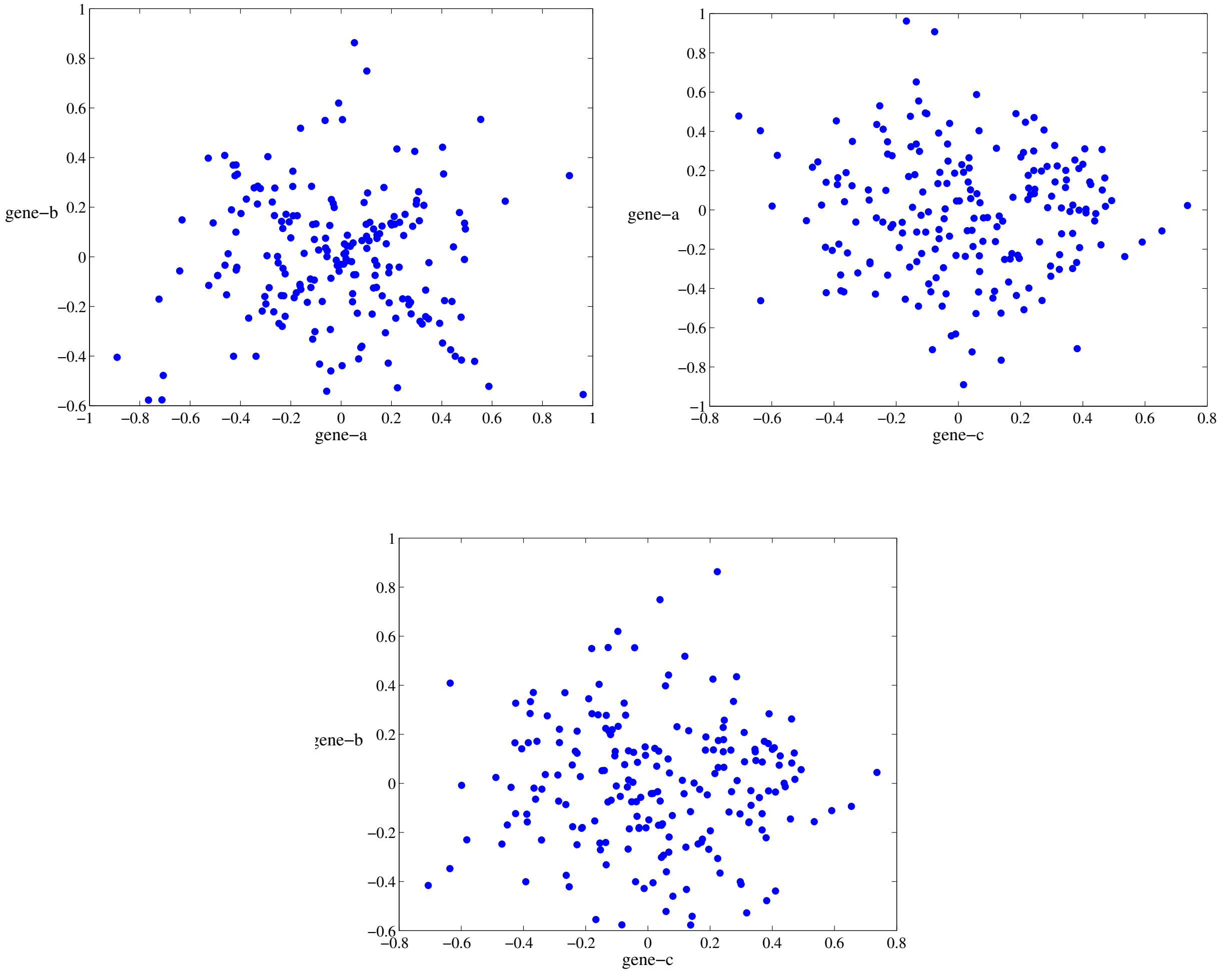


Figure 8.2. **Scatter plots of the expression levels of the three hypothetical genes.**

tively correlated, while when gene-c is down-regulated, they appear as negatively correlated. No correlation pattern appears when gene-c is baseline or when its values are not considered at all. This is a simple example which could be explained phenomenologically as follows: when the expression levels of gene-c is high a mechanism is triggered that closely relates the expression levels of gene-a and gene-b. On the other hand, when gene-c is under-expressed the opposite trend between gene-a and gene-b is sustained. However, when the level of gene-c is around the reference level, the remaining two genes appear to be acting independently from each other.

The main question is then how to select a measure that will capture such control mechanism. Ideally, we seek a cost function which is small or zero when no conditional structure is present in the data, while at the same time it tends to assume large values when the data shows strong dependencies under conditioning. A natural measure of independency, and therefore also of dependency, is given by the mutual information as defined in Chapter 2. The expression (2.12), has a straightforward extension to the class of conditional distribution functions. Let us consider first the definition of mutual information between two random variables x_1 and x_2 conditioned on the value of a third random variable y:

$$I(x_1, x_2|y) \triangleq D\left(p_{x_1,x_2|y} \middle\| \, p_{x_1|y}p_{x_2|y}\right). \tag{8.1}$$

In the case of continuous random variables (8.1) can be expressed as:

$$I(x_1, x_2|y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{x_1,x_2,y}(u, v, w) \log\left(\frac{p_{x_1,x_2|y}(u, v|w)}{p_{x_1|y}(u|w)p_{x_2|y}(v|w)}\right) dudvdw. \tag{8.2}$$

This definition can be extended to the mutual information of M random variables $\mathbf{x} = [x_1, \ldots, x_M]^T$, conditioned on a separate set of L variables $\mathbf{y} = [y_1, \ldots, y_L]^T$,

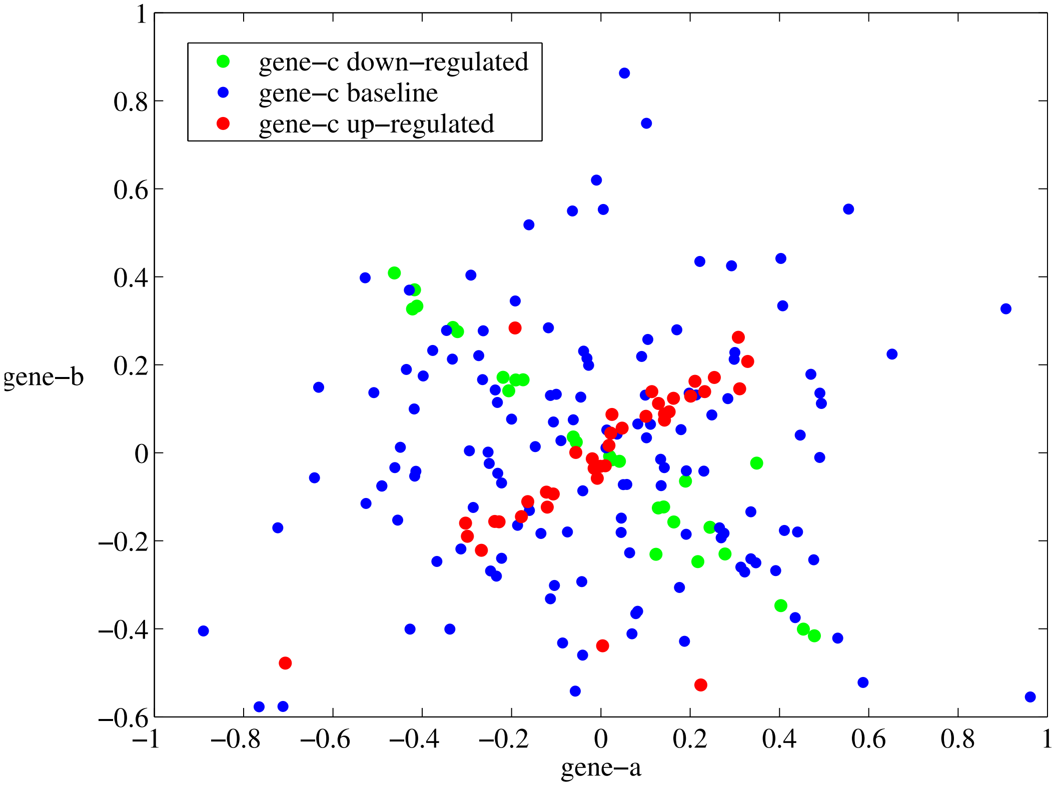


Figure 8.3. The co-expression pattern between gene-a and gene-b appears under conditioning by gene-c. When gene-c is up-regulated, gene-a and gene-b appear positively correlated. The opposite pattern appears when gene-c is under-expressed. No significant co-expression patterns appear when gene-c is expressed around the reference level.

in a manner analogous to the definition of (2.19):

$$I(\mathbf{x}|\mathbf{y}) = E_{p_{\mathbf{y}}}\left[I(\mathbf{x}|\mathbf{y}=\mathbf{w})\right] = \int_{-\infty}^{\infty} p_{\mathbf{x},\mathbf{y}}(\mathbf{u},\mathbf{v}) \log \frac{p_{\mathbf{x}|\mathbf{y}}(\mathbf{u}|\mathbf{v})}{\prod_{i=1}^{M} p_{x_i|\mathbf{y}}(u_i|\mathbf{v})} d\mathbf{u}, \tag{8.3}$$

This expression provides us with a measure of the expected mutual information of $\mathbf{x}$ conditionally on the value of $\mathbf{y}$. Evidently, when $\mathbf{x}$ and $\mathbf{y}$ are statistically independent, we have trivially that:

$$I(\mathbf{x}|\mathbf{y}) = I(\mathbf{x}) \int_{-\infty}^{\infty} p_{\mathbf{y}}(\mathbf{w}) d\mathbf{w} = I(\mathbf{x}). \tag{8.4}$$

Recalling that we are after certain structure in the data that appears only under conditioning, this result prompts us with the idea of adopting the following cost function for our framework:

$$\mathcal{L}(\mathbf{x}|\mathbf{y}) \triangleq I(\mathbf{x}|\mathbf{y}) - I(\mathbf{x}). \tag{8.5}$$

Clearly, we have that $\mathcal{L}(\mathbf{x}|\mathbf{y}) = 0$ when $\mathbf{x}$ and $\mathbf{y}$ are independent. In this case, even if a cluster of genes possesses a high information content, *i.e.* $I(\mathbf{x})$ is large, such structure appears regardless of the set of conditioning variables. On the other hand, $\mathcal{L}(\mathbf{x}|\mathbf{y})$ is a large positive number when the information content is significantly increased under conditioning. This is the case of interest in our framework. Notice that the quantity in (8.5) might assume negative values and it is not lower-bounded in general.

8.3.2 Some Properties of the Cost Function

In this section, we derive certain properties of the cost function (8.5), and we show some analogies between the proposed cost function and the concept of *co-information* [4]. Consider a simple network with a single parent node x_0 and two children nodes x_1 and x_2. The conditional information content of this network, according to (8.5), is simply given by:

$$\mathcal{L}(x_1, x_2|x_0) = I(x_1, x_2|x_0) - I(x_1, x_2). \tag{8.6}$$

Let us consider the first term on the right hand side of (8.6):

$$\begin{aligned}
I(x_1, x_2|x_0) &= \int_{-\infty}^{\infty} p_{x_0}(w) I(x_1, x_2|x_0 = w) dw && (8.7) \\
&= \int_{-\infty}^{\infty} p_{x_0}(w) \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{x_1,x_2|x_0}(u, v|w) \log \frac{p_{x_1,x_2|x_0}(u, v|w)}{p_{x_1|x_0}(u|w)p_{x_2|x_0}(v|w)}\, dudvdw \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{x_0,x_1,x_2}(w, u, v) \log \frac{p_{x_1,x_2|x_0}(u, v|w)p_{x_0}(w)^2}{p_{x_1|x_0}(u|w)p_{x_2|x_0}(v|w)p_{x_0}(w)^2}\, dudvdw \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{x_0,x_1,x_2}(w, u, v) \log \frac{p_{x_0,x_1,x_2}(w, u, v)p_{x_0}(w)}{p_{x_0,x_1}(w, u)p_{x_0,x_2}(w, v)}\, dudvdw \\
&= -H(x_0, x_1, x_2) - H(x_0) + H(x_0, x_1) + H(x_0, x_2) && (8.8)
\end{aligned}$$

Therefore, recalling that $I(x_1, x_2) = H(x_1)+H(x_2)-H(x_1, x_2)$, we have that (8.6) is equal to:

$$\mathcal{L}(x_1, x_2|x_0) = -H(x_0) - H(x_1) - H(x_2) + H(x_0, x_1) + H(x_0, x_2)+ \tag{8.9}$$

$$+H(x_1, x_2) - H(x_0, x_1, x_2).$$

From this expression we can notice that when considering a simple network with one conditioning node and two children nodes, the cost function (8.5) is indeed equal to minus the *co-information* between the three random variables. The general definition of co-information of N random variables is given by [4]:

$$\mathcal{C}(\mathbf{x}) = \sum_{E_j \subseteq E_N} q_j H(\mathbf{x}_{E_j}), \tag{8.10}$$

where E_j is the power set of j and q_j is the Möbius inversion function, defined as:

$$q_j = -(-1)^{|E_j|} = \begin{cases} 1 & \text{if } |E_j| \text{ is odd} \\ -1 & \text{if } |E_j| \text{ is even} \end{cases}, \tag{8.11}$$

where $|E_j|$ is the cardinality of E_j. The interesting aspect is represented by the fact that the co-information gives a measure of the total information content shared by all the random variables, unlike the conventional mutual information which includes all the information shared by the variables two at the time. Therefore, maximizing (8.5) is equivalent to seeking clusters whose representatives simultaneously share the least amount of information between each other. We can thus write:

$$\max_{x_0, x_1, x_2} \mathcal{L}(x_1, x_2 | x_0) = \min_{x_0, x_1, x_2} \mathcal{C}(x_0, x_1, x_2). \tag{8.12}$$

Notice that expression (8.9) is not altered if we exchange the variables x_0, x_1, or x_2. Hence, it holds that:

$$\mathcal{L}(x_1, x_2 | x_0) = \mathcal{L}(x_0, x_2 | x_1) = \mathcal{L}(x_0, x_1 | x_2) \tag{8.13}$$

Thus, the information content of the sub-network does not change if we exchange one of the children nodes with the parent node. The following theorem generalizes this property.

Theorem 7 (Symmetricity Property of the Cost Function $\mathcal{L}$) . *Let* $\mathbf{x}$*,* $\mathbf{y}$*, and* $\mathbf{z}$ *be three continuous random vectors, whose probability density functions are null only on sets of measure zero. Given the functional* $\mathcal{L}$ *defined as:*

$$\mathcal{L}(\mathbf{x}, \mathbf{y}|\mathbf{z}) = I(\mathbf{x}, \mathbf{y}|\mathbf{z}) - I(\mathbf{x}, \mathbf{y}), \tag{8.14}$$

the following holds:

$$\mathcal{L}(\mathbf{x}, \mathbf{y}|\mathbf{z}) = \mathcal{L}(\mathbf{x}, \mathbf{z}|\mathbf{y}) = \mathcal{L}(\mathbf{y}, \mathbf{z}|\mathbf{x}). \tag{8.15}$$

Therefore, the cost function (8.5), can be simply denoted as $\mathcal{L}(\mathbf{x}, \mathbf{y}, \mathbf{z})$*, where the order of the random vectors is irrelevant.*

8.4 GeneScreen: Theory and Practice

The framework outlined in the previous sections finds direct application to the analysis of gene transcription levels measured using DNA microarray assays. When designing a practical implementation of the algorithm seeking clusters that maximize the cost function (8.5), several issues must be carefully considered:

1. A direct evaluation of the cost function (8.5) requires an estimate of the multi-variate probability density function of all the N variables included in the cluster. We have already discussed in Chapter 2 the difficulties involved in obtaining such estimate when $N \geq 3$.

2. Estimates of the *conditional* density functions might simplify the issues associated with the dimensionality of the problem, but these are in turn quite difficult to obtain, in the case of continuous random variables.

3. The noise level in the data might significantly limit the number of parameters that can be learned with a certain accuracy.

4. The experimental procedure is affected by inherent limits in the number of samples per gene that can be measured in a given span of time. Hence, the sampling characteristics of any microarray assay are generally poor in the time-domain. This issue imposes a further limitation on the capability of estimating complex joint probability density functions, due to the limited sample size.

5. For a sub-network of a given size, the optimization of the cost function (8.5) requires a search through all possible combinations of nodes choosing among the set of genes included in the experiment. We will soon see that such number of combinations can be quite large if the parameters of the search algorithm are not chosen properly, quickly yielding to an intractable computational cost.

Let us examine in detail how we dealt with these issues.

8.4.1 Combinatorial Optimization Approach

The goal is to identify a list of sub-networks that yield large values of the cost function (8.5). Such goal can be achieved by selecting exhaustively sub-networks (see Figure 8.4) made up of all possible combinations of L genes as conditioning variables (which we will refer to as parent nodes), and all possible combinations of M genes among the remaining ones as conditioned variables (also known as children nodes), and to evaluate the corresponding value of the cost function (8.5). Clearly, the cost of this combinatorial approach increases quite rapidly with the total number of genes assayed in the experiment, and it is a non-linear function

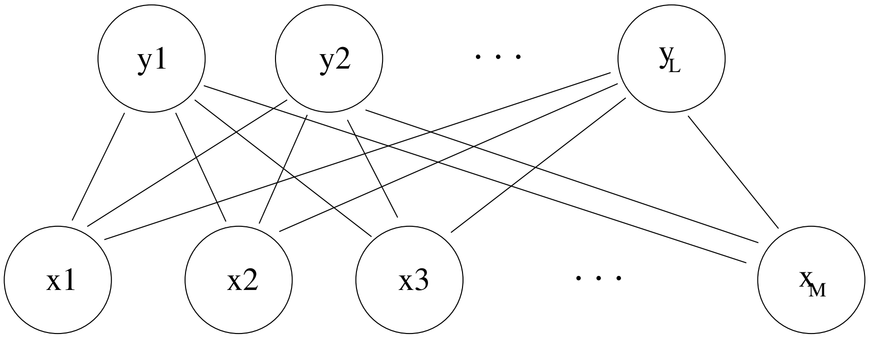


Figure 8.4. Cluster of genes composed of L conditioning genes and M children nodes. This cluster represents the generic sub-network explored to identify conditional structure.

of M and L. When a total number of N gene expression profiles are measured in the experiment, the total number of possible sub-networks with L parent nodes and M children nodes is given by the following expression:

$$\mathcal{K}(N, M, L) = \begin{pmatrix} N \\ L \end{pmatrix} \begin{pmatrix} N - L \\ M \end{pmatrix} = \frac{N!}{M!L!(N - L - M)!}. \tag{8.16}$$

For example, when dealing with $N = 2,000$ genes, a choice of $L = 3$ and $M = 5$ will result in $3.5 \cdot 10^{23}$ possible combinations! In general, for small values of M and L, we have that:

$$\mathcal{K}(N, M, L) \approx \mathcal{O}(N^{M+L}). \tag{8.17}$$

Hence, unless a technique is devised that allows for efficient pruning of non-informative clusters, the problem will result computationally tractable only for very small values of M and L. In addition, as it will be discussed more in details in the next section, for large values of M and L we will unavoidably incur in the problem of having to estimate high-dimensional multivariate statistics of the data, thus requiring a particularly large number of samples in order to get a robust estimate.

These constraints clearly suggest that a simple framework in which a sub-network involving only three genes (one parent node and two children nodes) should be the subject of an initial investigation and validation of the proposed approach. From the symmetric expression of the cost function given in (8.9), it is possible to show that the computational complexity associated with evaluating the co-information content of each possible sub-network, when $L = 1$ and $M = 2$ simplifies as:

$$\begin{aligned}\mathcal{K}(N,2,1) &= N(N-1)(N-2) + N(N-1) + N && (8.18)\\ &= N^3 - 2N^2 + N && (8.19)\\ &= \mathcal{O}(N^3), && (8.20)\end{aligned}$$

for a total number of N genes assayed. As an example, when $N = 2,000$, approximately $8 \cdot 10^9$ possible combinations need to be considered, and the corresponding cost function evaluated. This kind of task can be completed in a reasonable amount of time (a few days) by any modern off-the-shelf single-processor machine. It is also clear that the algorithm could be easily parallelized to run on clusters of processors, since the evaluation of the cost function for a given sub-network is an independent task.

8.4.2 A Moment Based Approximation of the Mutual Information

In the previous section, we have seen how computational complexity considerations have suggested to restrict our focus on very simple clusters involving one conditioning node and two conditioned nodes. Different considerations, involving the feasibility of computing a consistent estimate of the conditional mutual information, show that this is indeed a sensible choice.

The expression of the cost function given in (8.9) suggests that some kind of estimate of the multivariate joint probability density function of the three variables in the cluster is required in order to evaluate the corresponding entropies. However, considering that the typical experimental setting in DNA microarray assays produces between 5 and 20 samples per gene, the poor sampling properties generally discourage the use of fancy density estimators such as parametric models or kernel methods. Therefore, in the design of a practical implementation of the principle (8.12), we opted for the use of a moment based approximation of the information theoretical quantities involved in the calculation of the cost function.

Let us first examine an expression equivalent to (8.9) that is used as a starting point to define our working approximation:

$$\begin{aligned}\mathcal{L}(x_0, x_1, x_2) &= I(x_1, x_2|x_0) - I(x_1, x_2) \qquad (8.21)\\ &= H(x_1|x_0) + H(x_2|x_0) - H(x_1, x_2|x_0) - H(x_1) - H(x_2) + H(x_1, x_2).\end{aligned}$$

This expression suggests that a method to compute univariate and bivariate entropies must be devised. A moment based approximation of the univariate entropy is obtained by approximating the marginal probability density function of each variable using a Gram-Charlier expansion [90]. Recalling that, as we proved in Chapter 2, the entropy is shift invariant, we can assume that all the sample data has been mean subtracted. Moreover, since it holds that $H(ax) = H(x) + \log(|a|)$, where a is a deterministic constant parameter, we can re-scale each variable to be unit-variance and compute the entropy estimate as follows:

$$H(x_i) = H(\tilde{x}_i) + \log(\sigma_i), \qquad i = 1, 2 \tag{8.22}$$

where σ_i is the standard deviation of x_i, and $\tilde{x}_i \triangleq x_i/\sigma_i$ is unit-variance. A Gram-Charlier approximation of $p_{\tilde{x}_i}(u)$, including moments up to the fourth order is given by the following expression:

$$p_{\tilde{x}_i}(u) = g(u)\left(1 + \kappa_{3,i}H_3(u)/6 + \kappa_{4,i}H_4(u)/24\right), \qquad i = 1, 2 \tag{8.23}$$

where $H_3(u)$ and $H_4(u)$ are the 3rd and 4th order Chebyshev-Hermite polynomial [55], respectively, $g(u)$ is the zero-mean, unit-variance, normal probability density function, and $\kappa_{3,i}$ and $\kappa_{4,i}$ are the third and fourth order cumulants of $\tilde{x}_i$. For a zero mean, unit-variance random variable, these can be computed as follows:

$$\kappa_{3,i} = E[\tilde{x}_i^3] \tag{8.24}$$

$$\kappa_{4,i} = E[\tilde{x}_i^4] - 3. \tag{8.25}$$

By substituting the approximation of $p_{\tilde{x}_i}(u)$ considered in (8.23) in the definition of entropy (2.3), we can compute the following approximation:

$$H(\tilde{x}_i) \approx \frac{1}{2}\log(2\pi e) - (\kappa_{3,i}^2 + \kappa_{4,i}^2/4)/12, \qquad i = 1, 2. \tag{8.26}$$

which is consistent with the fact that the entropy of a random variable with a given variance is maximum if the variable is normally distributed. The maximum of (8.26) is indeed attained when $\kappa_3 = \kappa_4 = 0$ and is equal to the entropy of a unit variance gaussian random variable. A similar approximation can be obtained for the bivariate entropy. The derivation is simplified if the data is pre-whitened as described in (2.30), so that the resulting variables are uncorrelated. Following the notation of Chapter 2, we denote the sphering matrix as $S^{-1/2}$, where S is the sample covariance matrix of x_1 and x_2, and $S^{-1/2}$ is an inverse square root factor of S. Hence, if we define:

$$\begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \end{bmatrix} \triangleq S^{-1/2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \tag{8.27}$$

we have that:

$$H(x_1, x_2) = H(\hat{x}_1, \hat{x}_2) + \frac{1}{2} \log |\det(S)|, \tag{8.28}$$

where S is always full rank unless x_1 and x_2 are linearly dependent. A detailed derivation of an approximation of $H(\hat{x}_1, \hat{x}_2)$ can be found for example in [49], and is based on a bivariate Gram-Schmidt expansion of the corresponding probability density function. The resulting expression for the approximated entropy is given by:

$$\begin{aligned} H(\hat{x}_1, \hat{x}_2) \approx \log(2\pi e) - \frac{1}{12} \Bigg[\kappa_{30}^2 + 3\kappa_{21}^2 + 3\kappa_{12}^2 + \kappa_{03}^2 + \frac{1}{4}(\kappa_{40}^2 + \\ + 4\kappa_{31}^2 + 6\kappa_{22}^2 + 4\kappa_{13}^2 + \kappa_{04}^2) \Bigg], \end{aligned} \tag{8.29}$$

where the bivariate cross-cumulants can be computed as follows from the sample data:

$$\kappa_{30} = E[\hat{x}_1^3] \tag{8.30}$$

$$\kappa_{03} = E[\hat{x}_2^3] \tag{8.31}$$

$$\kappa_{21} = E[\hat{x}_1^2 \hat{x}_2] \tag{8.32}$$

$$\kappa_{12} = E[\hat{x}_1 \hat{x}_2^2] \tag{8.33}$$

$$\kappa_{40} = E[\hat{x}_1^4] - 3 \tag{8.34}$$

$$\kappa_{04} = E[\hat{x}_2^4] - 3 \tag{8.35}$$

$$\kappa_{31} = E[\hat{x}_1^3 \hat{x}_2] \tag{8.36}$$

$$\kappa_{13} = E[\hat{x}_1 \hat{x}_2^3] \tag{8.37}$$

$$\kappa_{22} = E[\hat{x}_1^2 \hat{x}_2^2] - 1 \tag{8.38}$$

By combining (8.22) and (8.28), the following approximation of $I(x_1, x_2)$ is obtained, which involves only cross cumulants up to the fourth order:

$$\begin{aligned}
I(x_1, x_2) &= H(x_1) + H(x_2) - H(x_1, x_2) && (8.39)\\
&= H(\tilde{x}_1) + \log(\sigma_1) + H(\tilde{x}_2) + \log(\sigma_2) + H(\hat{x}_1, \hat{x}_2) + \frac{1}{2}\log|\det(S)| &&\\
&= \frac{1}{12}\Bigg\{ - \Bigg[\kappa_{3,1}^2 + \kappa_{3,2}^2 + \frac{1}{4}(\kappa_{4,1}^2 + \kappa_{4,2}^2)\Bigg] + \Bigg[\kappa_{30}^2 + 3\kappa_{21}^2 + && (8.40)\\
&\quad +3\kappa_{12}^2 + \kappa_{03}^2 + \frac{1}{4}(\kappa_{40}^2 + 4\kappa_{31}^2 + 6\kappa_{22}^2 + 4\kappa_{13}^2 + \kappa_{04}^2)\Bigg]\Bigg\} + && (8.41)\\
&\quad + \log(\sigma_1) + \log(\sigma_2) - \frac{1}{2}\log|\det(S)|. && (8.42)
\end{aligned}$$

An analogous expression involving *conditional* cross-cumulants of the variables can be used to estimate the conditional mutual information $I(x_1, x_2|x_0)$.

8.5 Simulation Results

We developed *GeneScreen* as a collection of computational statistic routines, whose objective is to process gene expression data (typically from DNA microarray time-course experiments), extracting significant gene association patterns. The basic idea behind the technique used in GeneScreen is that gene expression time-courses, which appear to be unrelated when observed as a whole, often present a very defined structure when they are explained by a common cause. The unsupervised learning framework outlined in the previous sections aims precisely at this goal. For a given microarray assay experiment, all possible unique combinations of three genes are considered and the co-information is used to assign a score to each such combination. The highest scoring clusters are recorded in order to be further evaluated. GeneScreen includes a set of tools devoted at pre-processing of the transcription data, performing a series of tasks which include pruning the set of genes according to a user defined criterion (*e.g.* their sample variance), correcting for univariate and bivariate outliers, or accounting for missing values. GeneScreen is implemented in C++ and it was included in DARPA's BioSpice, since its very first release.

In order to evaluate the effectiveness of the proposed approach in unveiling hidden dependencies between gene transcription levels, we considered several datasets from experiments involving whole-genome assays of the gene expression levels of the bacterium *Escherichia Coli* (*E.coli*), a prokaryote. Further details on the exact nature of the experiments conducted are provided along with the results of our simulations in the next section.

8.5.1 Analysis of *Escherichia Coli* Expression Data

The dataset considered for our first set of simulations[2] consists of a collection of mutliple DNA microarray assays conducted on *E.coli*, including a total of 4291 genes and 72 sample points per gene. Table 8.1 provides a basic listing of the experimental conditions. A detailed description of the actual experimental setting can be found in [85].

The dataset comprises a variety of experimental conditions, some including the perturbation of specific regulatory mechanisms, others affecting genes controlled by different sets of transcriptional regulators. The resulting large oscillations in the expression levels of several genes ensure that the dataset provides enough variability to allow the consistent detection of specific patterns, in a statistically meaningful way.

DNA microarray data is conventionally expressed as the logarithm (usually in base 10) of the ratio between the estimated transcription level and and a reference value. Therefore, a log-ratio value of zero indicates that the gene is expressed at similar levels as the reference. On the other hand, a value of 0.3 or above is equivalent to a 2-fold increase in the transcription level. We will conventionally refer to gene levels that show at least a 2-fold increase as *up-regulated* or over-expressed. Equivalently, when the log-ratio level is -0.3 or less, the gene shows at least a 2-fold decrease in the transcription level and will be referred to as *down-regulated*, or under-expressed.

This distinction is particularly relevant when we consider that, in order to score different clusters of genes, a set of conditional entropies need to be evaluated. Due to the small sample size, this is most efficiently achieved by discretizing the expression levels of the parent node into three levels, according to whether the

[2] This dataset was kindly provided by professor James Liao's group at the Chemical Engineering Department at UCLA.

Experiment number	Condition	Number of measurements
2	Ihf+ versus ihf-	1
1	Minimal versus rich media	1
24–46	Tryptophan Regulation	23
5–15	NtrC regulation	11
16	Heat shock	1
61-66	Xylose fermentation	6
47–60	LexA regulation	14
67–69	SocRS regulation	3
70–72	MarRAB regulation	3
17–23	Transition from glucose to acetate	7
3–4	Balanced growth in acetate versus growth in minimal medium	2

Table 8.1. Experimental conditions and corresponding number of measurements for the *E.coli* dataset used in the simulations.

gene is down-regulated, close to the reference level (baseline), or up-regulated. The choice of the discretization levels is arbitrary and will, in general, affect the outcome of the exploratory analysis. In GeneScreen, such choice is ultimately left to the user although the default thresholds of -0.3 for down-regulation and 0.3 for up-regulation were used throughout our analysis. Such choice is dictated by considerations that are both biological and statistical. The goal is clearly to select a level at which the up- or down-regulation can be robustly established. Due to the large measurement error affecting the data, it is widely accepted that at least a 2-fold increase or decrease in the measured expression level is required in order to establish up or down regulation.

In figure 8.5 and 8.6 two different views of the highest scoring network for this dataset are shown. The three genes belonging to this highly informative cluster are *tap*, *yabM* and *ilvY*.

The *tap* gene, is one of four methyl-accepting chemotaxis proteins (MCPs) in *E. coli.* Its product, Tap functions as a conventional signal transducer, enabling the cell to respond chemotactically to dipeptides. *YabM* is a probable efflux transporter for sugars such as lactose and IPTG [64]. Cells over-expressing *yabM* show decreased accumulation of lactose and IPTG. *YabM* is a member of the major facilitator superfamily (MFS) of transporters, and its physiological significance and regulation are still unclear. *IlvY* is a very well studied transcription regulator, involved in the control of the parallel isoleucine-valine biosynthetic pathway. Figure 8.5 shows that when the expression level of *tap* is generally low, *yabM* and *ilvY* tend to be negatively correlated. On the other hand, when *tap* is up-regulated, both *yabM* and *ilvY* are strongly under-expressed. A similar mechanism can be deducted from figure 8.6: when *yabM* is down-regulated we assist to a simultaneous down-regulation of *ilvY* and up-regulation of *tap.* The same kind of conditional structure tends to appear when *tap* is replaced by *mbhA* in

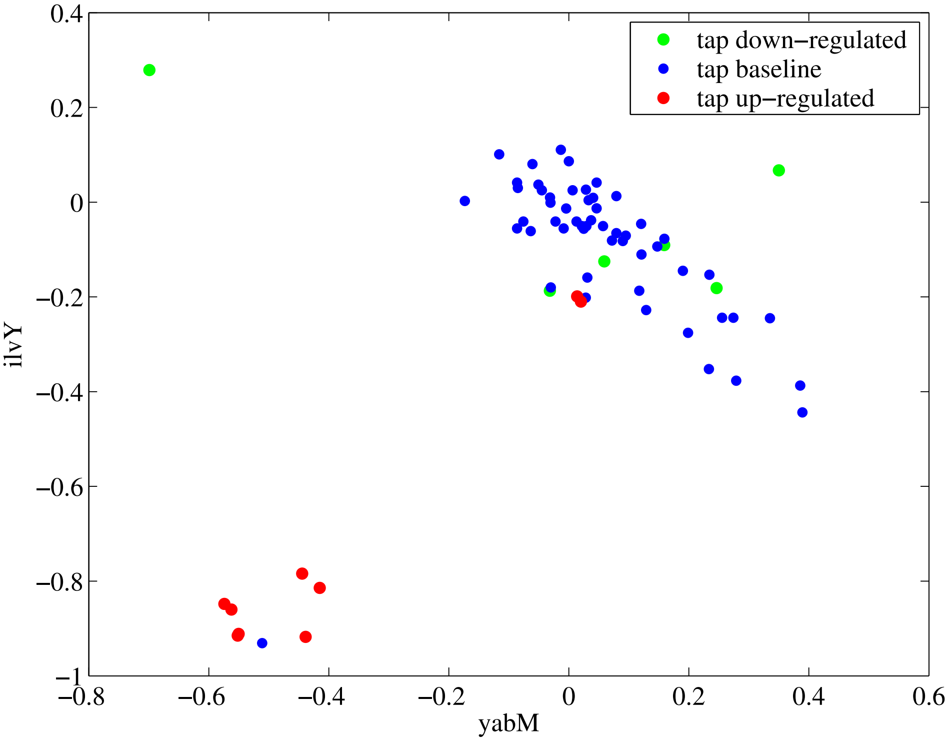


Figure 8.5. **Co-expression pattern between the genes *yabM* and *ilvY*, when gene *tap* is the conditioning node.**

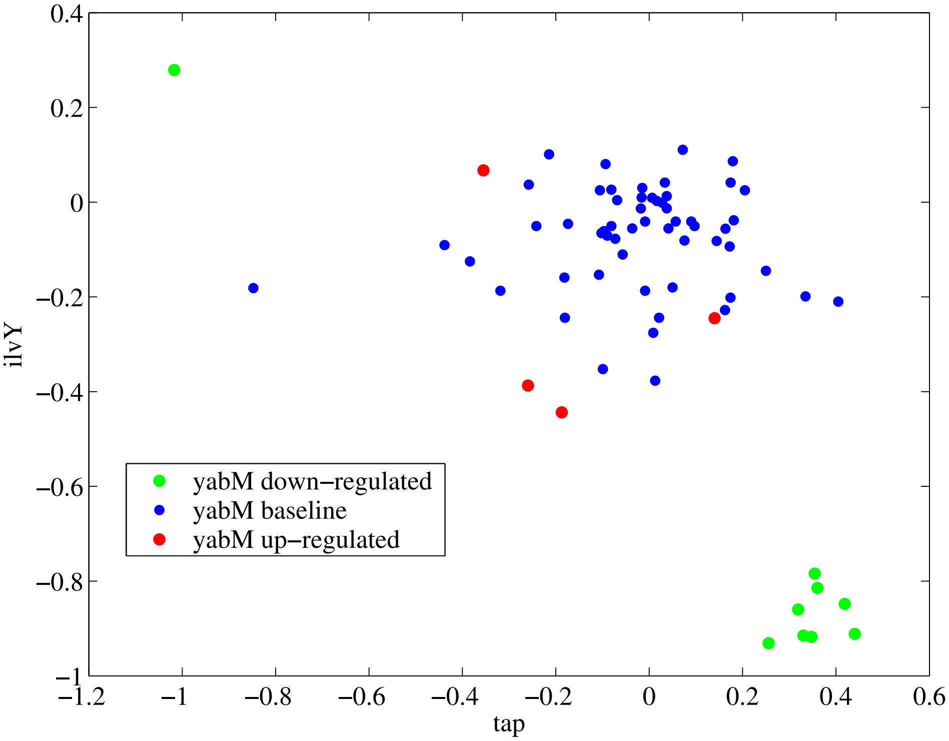


Figure 8.6. **Co-expression pattern between the genes *tap* and *ilvY*, when gene *yabM* is the conditioning node.**

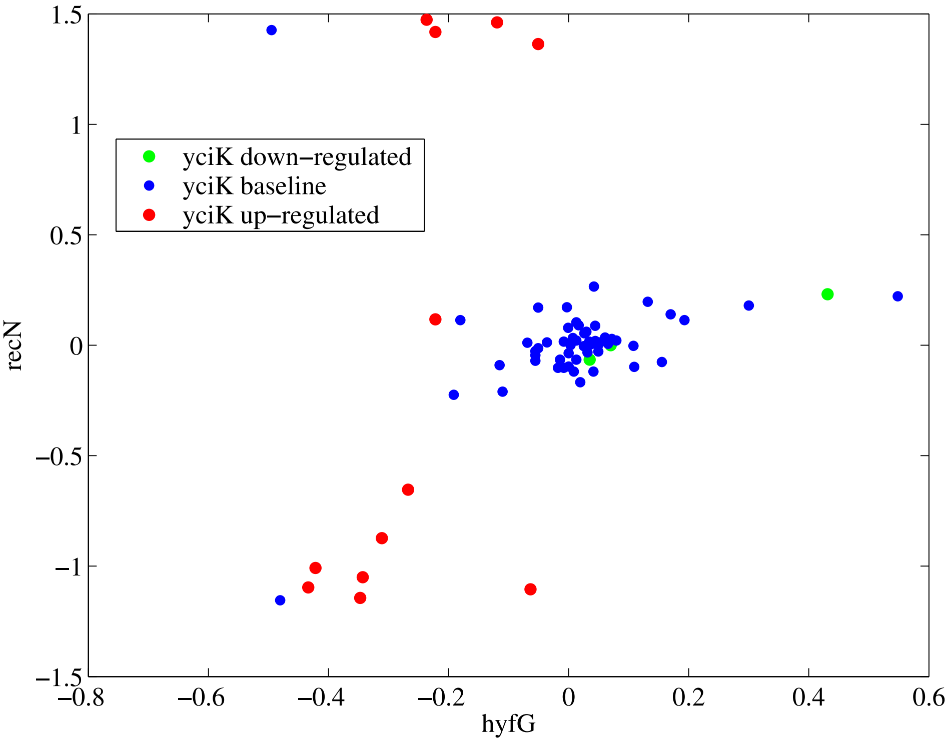


Figure 8.7. Co-expression pattern between the genes *hyfG* and *recN*, when gene *yciK* is the conditioning node.

the sub-network. This does not come out as a surprise since *mbhA* is a putative motility protein which is also thought to be involved in chemotaxis.

A cluster of genes presenting a very peculiar structure is the one involving *yciK*, *hyfG*, and *recN*, and it is shown in figures 8.7, 8.8 and 8.9.

Let us consider figure 8.7 first: this plot shows that when *yciK* is generally low, the expression levels of *recN* are always close to zero, signifying that the gene is always expressed to levels close to the reference. At the same time, *hyfG* varies from being also around the reference level to slightly over-expressed. On the other hand, when *yciK* is up-regulated the mutual behavior of the other two

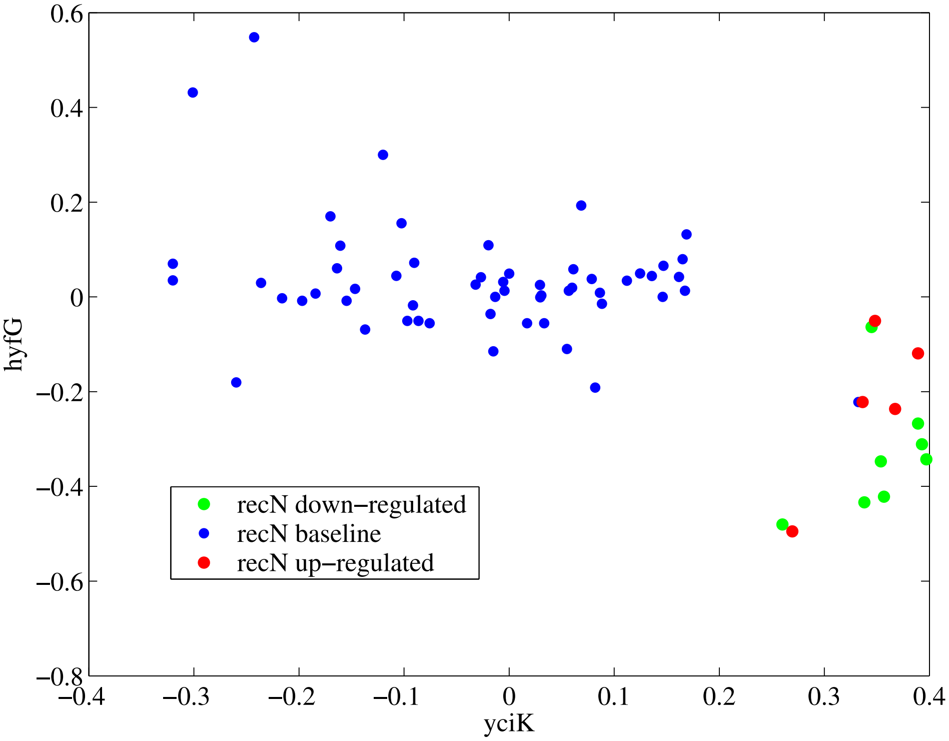


Figure 8.8. **Co-expression pattern between the genes *yciK* and *hyfG*, when gene *recN* is the conditioning node.**

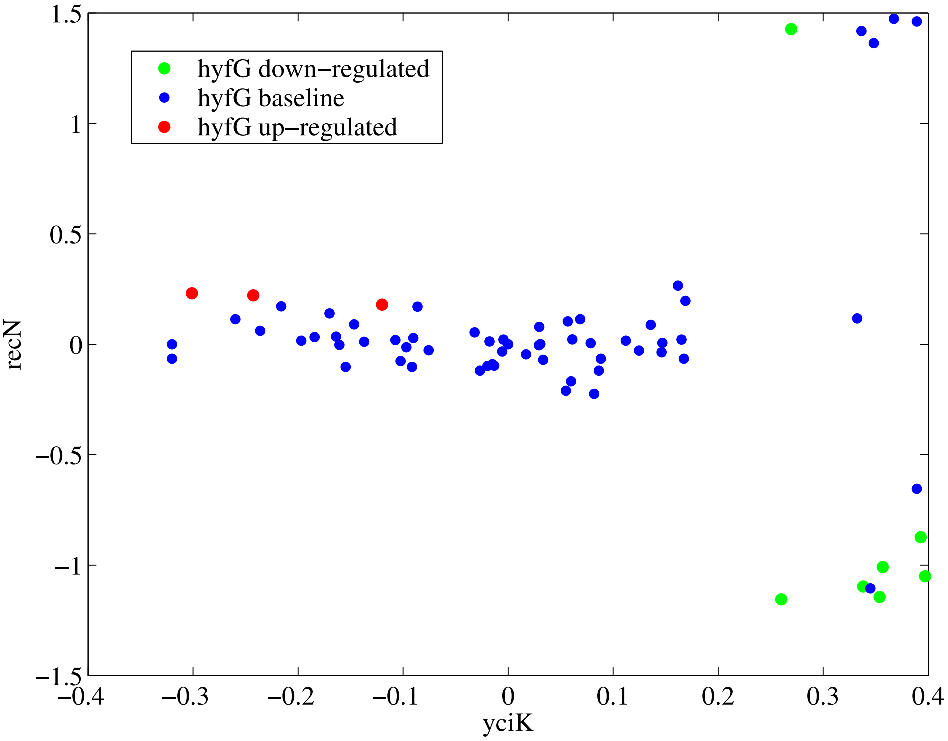


Figure 8.9. **Co-expression pattern between the genes *yciK* and *recN*, when gene *hyfG* is the conditioning node.**

genes changes radically. When *hyfG* is down-regulated, *recN* is strongly under-expressed, while when *hyfG* is generally close to the reference level, *recN* becomes generally over-expressed. The gene *yciK* is thought to be a putative oxidoreductase enzyme. *HyfG* is part of the hyf operon and its product is a catalytic sub-unit of hydrogenase-4, potential formate and transcriptional activator, involved in the oxidation of ferredoxin (anaerobic respiration). Finally, *recN* is a protein used in recombination and DNA repair. Figure 8.8 gives a different perspective on the interactions taking place in this cluster. It shows that when *recN* is around the reference levels, *hyfG* is generally slightly over-expressed or flat, while *yciK* is free to vary in a wide range of values. However, when *recN* is either over-expressed or under-expressed we assist to a simultaneous substantial down-regulation of *hyfG* and over-expression of *yciK*.

The interaction between the genes *narL*, *hmsR* (also known as *ycdQ*), and *menB* is shown in figures 8.10 and 8.11.

The gene *narL* is one of two response regulators in *E. coli* affecting anaerobic respiration. The second is the product of the *narP* gene. In the presence of nitrate the NarL protein can be phosphorylated. In this activated state phospho-NarL can act as both an activator of nitrate and nitrite reductase transcription and as a repressor of fumarate reductase transcription. Both actions switch anaerobic respiration to utilization of either nitrate or nitrite as electron acceptors. The product of *hmsR* is a putative uncharacterized transport protein, belonging to the family of vectorial glycosyl polymerization (VGP). The gene *menB* is involved in the anaerobic respiration pathway as well. Its product, the enzyme naphthoate synthase catalyzes a major step in menaquinone (also known as vitamin K2) biosynthesis, the formation of the bicyclic ring system. By observing the plots in figure 8.10 and 8.11 we can deduct that *hmsR* and *menB* show very small variations from the reference level in general, while *narL* alternates between

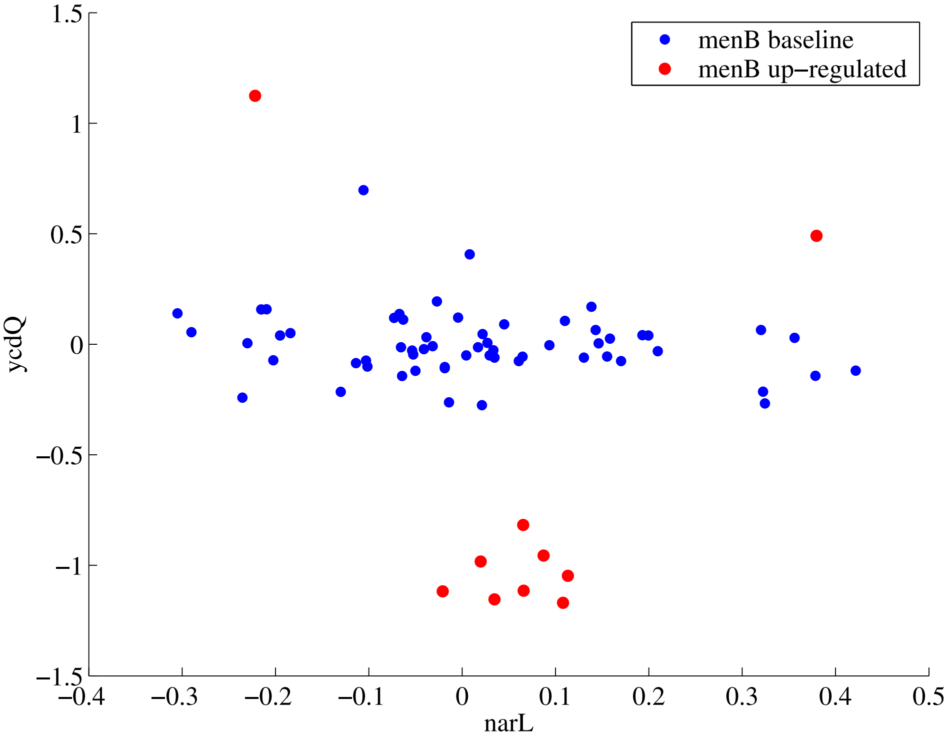


Figure 8.10. **Co-expression pattern between the genes *narL* and *ycdQ*, when gene *menB* is the conditioning node.**

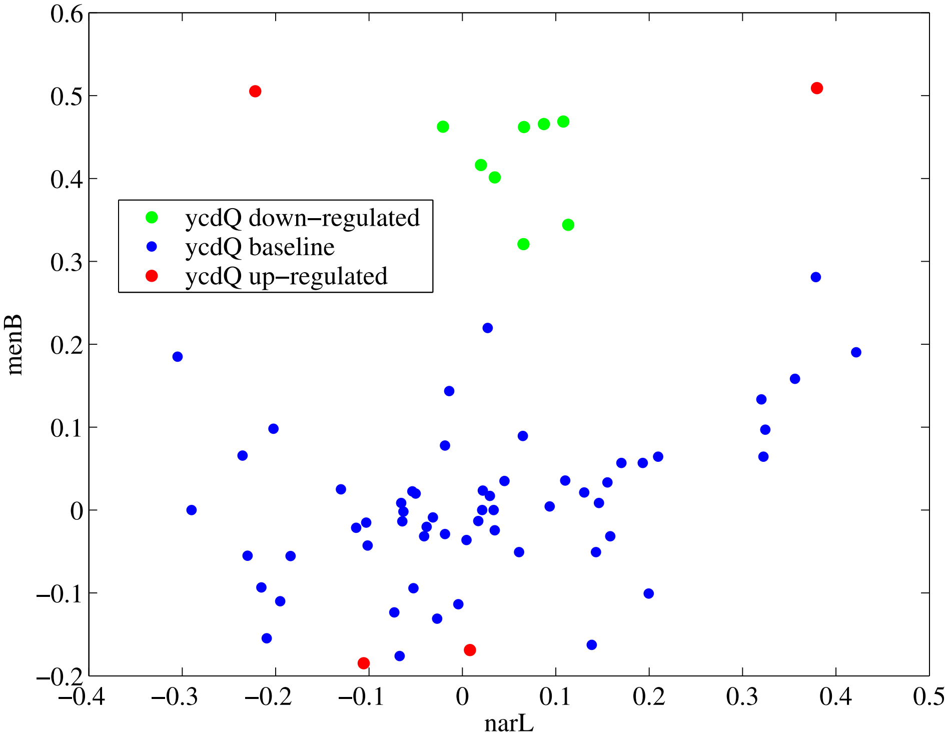


Figure 8.11. **Co-expression pattern between the genes *narL* and *menB*, when gene *ycdQ* is the conditioning node.**

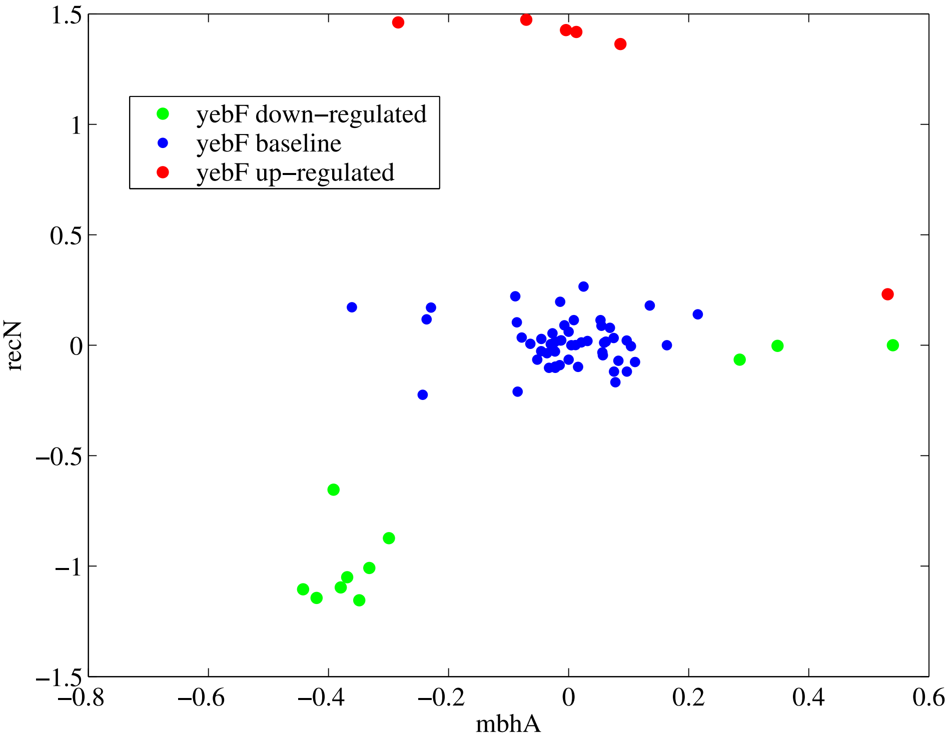


Figure 8.12. Co-expression pattern between the genes *mbhA* and *recN*, when gene *yebF* is the conditioning node.

over-expression and under-expression according to the specific condition. On the other hand, when *menB* becomes over-expressed the patterns of co-expression are radically different with *hmsR* being strongly under-expressed and *narL* sticking to values close to the reference level.

The conditional structure shown in figure 8.12, involves two genes that have been examined already, *i.e.* *mbhA* and *recN*, and a third gene *yebF*, whose product is a hypothetical protein with an unknown function.

This case is particularly compelling, since by using the proposed exploratory method we were capable of detecting interactions between genes whose function is well understood and genes whose function is totally unknown. Figure 8.12 shows that when *yebF* is within a small interval around the reference level, both *mbhA* and *recN* appear also stable around their reference values. However, the up-regulation of *yebF* is accompanied by a significative over-expression of *recN*, with no significant variations in *mbhA*, while its down-regulation is associated to a concomitant down-regulation of both *yebF* (marginal) and *recN* (more significant).

The last two clusters of genes whose significance we are going to briefly discuss are shown in figure 8.13 and 8.14. The first involves the genes *ybgF*, an hypothetical protein belonging to the same transcription group as the gene *tolB*, *acpS* whose product is an enzyme involved in the initial steps of the fatty acid biosynthesis, and *nagB* whose product is a subunit of glucosamine-6-phosphate deaminase, an enzyme involved in the glucosamine catabolism. The detected association pattern (figure 8.13) shows that *ybgF* and *acpS* are positively correlated in general, and tend to be slightly up-regulated when *nagB* is up-regulated or around the reference level. On the other hand, to a down-regulation of *nagB* corresponds a significative under-expression of both *ybgF* and *acpS*. The second cluster, shown in figure 8.14 involves the genes *acs*, whose product acetyl-CoA synthetase is involved in the acetate degradation pathway, *otsB* whose product is an enzyme involved in the trehalose biosynthesis which has not been yet fully characterized, and *yibP* a putative membrane protein. The association pattern between these genes shows that *acs* and *yibP* are characterized by a strong positive correlation when *otsB* is not up- or down-regulated. However, especially when *otsB* is over-expressed such pattern of correlation disappears, with *acs* often become strongly over-expressed (up to 10-fold) even when *yibP* is down-regulated.

An alternative way to interpret the results of the proposed exploratory technique is given by the hierarchical grouping of the discovered association patterns.

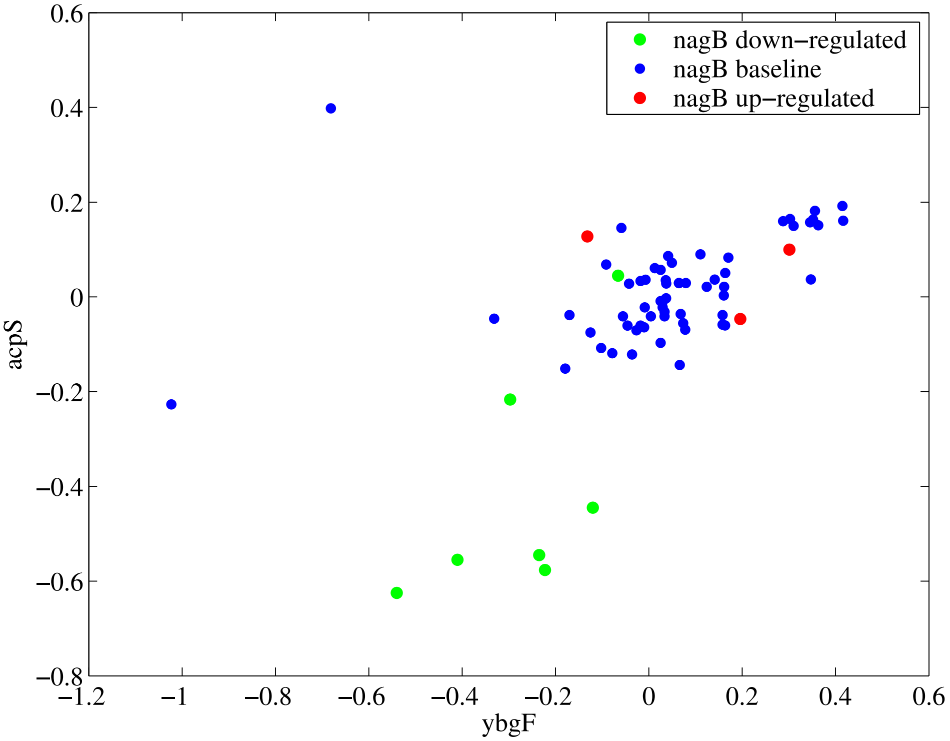


Figure 8.13. **Co-expression pattern between the genes *ybgF* and *acpS*, when gene *nagB* is the conditioning node.**

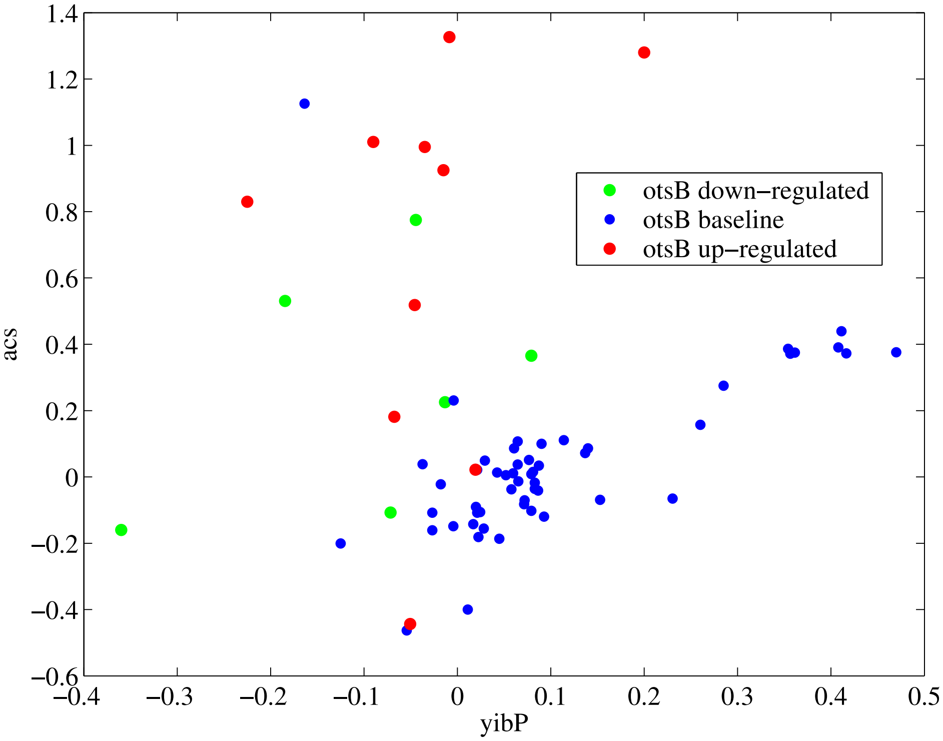


Figure 8.14. **Co-expression pattern between the genes *yibP* and *acs*, when gene *otsB* is the conditioning node.**

Figure 8.15 shows an example of such representation where genes belonging to the same functional group have been highlighted using the same shade of color.

The statistical significance of the proposed exploratory method was assessed by comparing the results obtained with the actual dataset to the results obtained by randomly shuffling the data points in each gene time course. By applying this technique the multivariate properties of data set are altered, while the marginal univariate distribution of each gene is preserved. Figure 8.16 shows a histogram comparing the distribution of the top 10% resulting scores for the original dataset and for the shuffled one. Such results demonstrate that the detected conditional co-expression patterns are statistically significant and can not be the product of an arbitrary arrangement of the data. A similar technique can be used to assess confidence levels for the statistical significance of the score assigned to a given cluster of genes.

8.6 Discussion and Future Work

Preliminary simulation experiments, have demonstrated that GeneScreen is capable of identifying significant association patterns between genes, which appear otherwise unrelated when their expression profiles are compared pair-wise. The identification of such high-level dependencies represents an important step towards the systematic discovery of regulatory mechanisms in the cell. On the other hand, several issues are still under investigations A method for accurately assessing the statistical relevance of a particular high-scoring cluster should be devised. Moreover, the relationship between a specific co-expression pattern and the underlying experimental setting should be investigated.

Patterns of co-expression involving multiple conditioning nodes and children nodes represent the next level of improvement of the proposed approach. Consid-

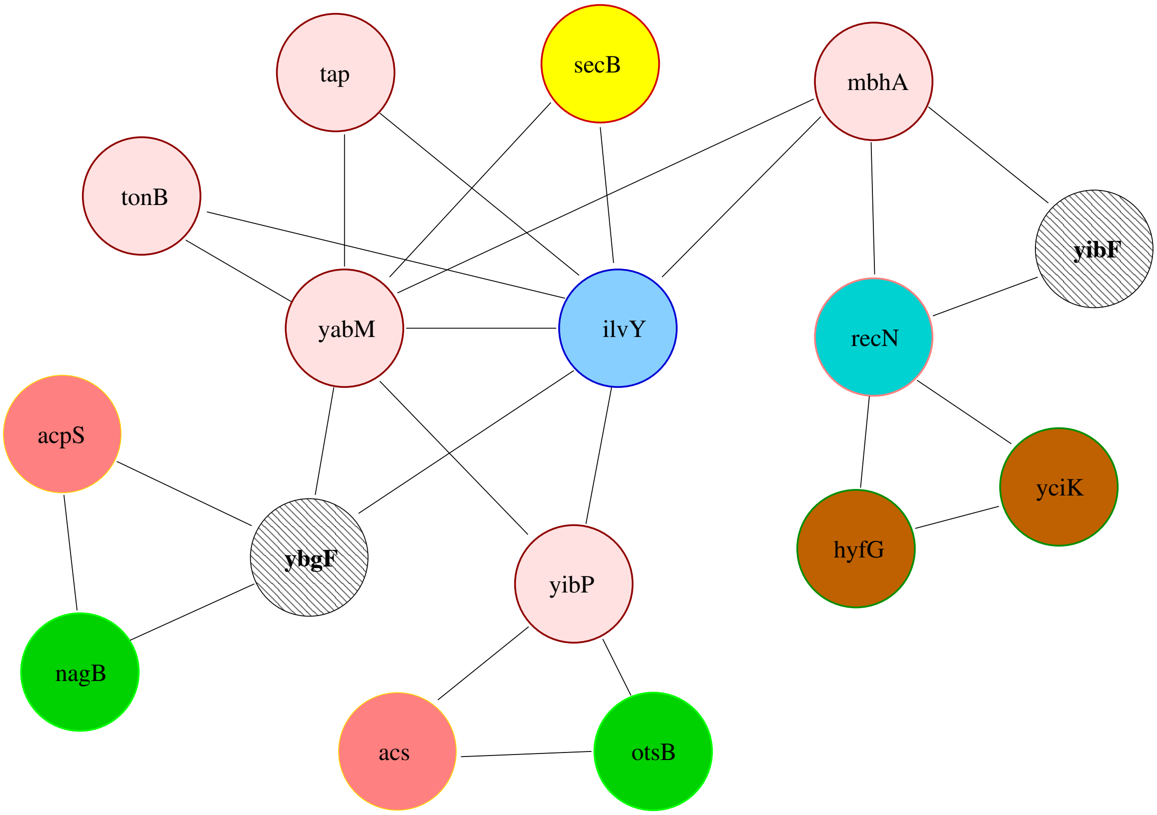


Figure 8.15. Network of high-scoring conditional association patterns, obtained by hierarchical grouping of clusters of genes. All the loops involving three edges in the graph can be associated to strong conditional co-expression patterns between the genes in the loop.

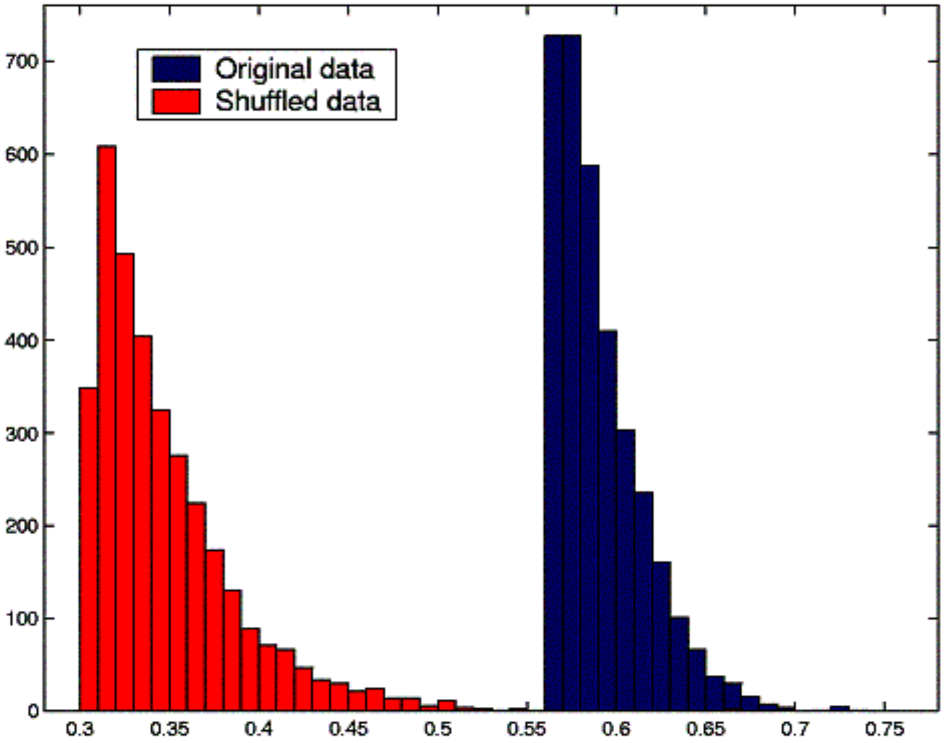


Figure 8.16. **The histogram shown the 10% highest scores for the original *E.coli* dataset and a dataset obtained shuffling the data points for each gene expression time course.**

ering the computational complexity issues arising from the resulting combinatorial exploratory technique, efficient methods for pruning the search space should be devised, possibly based on a-priori biological assumptions.

Bibliography

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

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

[3] H.B. Barlow. Unsupervised learning. *Neural Computation*, 1:295–311, 1989.

[4] A. J. Bell. The co-information lattice. In *Proc. 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, pages 921–926, Nara, Japan, April 2003.

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

[6] A.J. Bell and T.J. Sejnowski. The 'independent components' of natural scenes are edge filters. *Vision Research*, 37:3327–3338, 1997.

[7] J.M. Bernardo and A.F.M. Smith. *Bayesian Theory.* John Wiley & Sons, 1994.

[8] S. Boyd and L. Vandenberghe. *Convex Optimization*. Course reader for EE364 (Stanford) and EE236B (UCLA). `http://www.ee.ucla.edu/ee236b/reader.pdf`.

[9] W. Buntine. A guide to the literature on learning probabilistic networks from data. *IEEE Trans. On Knowledge and Data Engineering*, 8(2):195–210, 1996.

[10] Wray L. Buntine. Operations for learning with graphical models. *Journal of Artificial Intelligence Research*, 2:159–225, December 1994.

[11] A.M. Campbell and L.J. Heyer. *Discovering Genomics, Proteomics, and Bioinformatics*. Benjamin/Cummings, 2002.

[12] J.-F. Cardoso. Infomax and maximum likelihood for source separation. *IEEE Letters on Signal Processing*, 4(4):112–114, April 1997.

[13] J.-F. Cardoso. Blind signal separation: statistical principles. *Proceedings of the IEEE. Special issue on blind identification and estimation*, 9(10):2009–2025, October 1998.

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

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

[16] D.M. Chickering. Learning equivalence classes of Bayesian-network structures. *Journal of Machine Learning Research*, 2:445–498, February 2002.

[17] Nicos Christofides. *Graph Theory: An Algorithmic Approach*. Academic Press, 1975.

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

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

[20] S. Cruces, A. Cichocki, and L. Castedo. An iterative inversion approach to blind source separation. *IEEE Trans. Neural Networks*, 11(6):1423–1437, November 2000.

[21] S. Cruces, A. Cichocki, and S. i. Amari. Criteria for the simultaneous blind extraction of arbitrary groups of sources. In T-W. Lee, T-W. Jung, S. Makeig, and T.J. Sejnowski, editors, *Proceedings of the 3rd International Conference on Independent Component Analysis and Blind Signal Separation*, San Diego, California, USA, December 2001.

[22] S. Cruces, A. Cichocki, and S. i. Amari. The minimum entropy and cumulant based contrast functions for blind source extraction. In J. Mira and A. Prieto editors, editors, *Bio-Inspired Applications of Connectionism, Lecture Notes in Computer Science, Springer-Verlag. [6th International Work-Conference on Artificial and Natural Neural Networks (IWANN'2001)]*, volume II, pages 786–793, Granada, Spain, June 2001.

[23] G. Darmois. Analyse générale des liaisons stochastiques. *Rev. Inst. Internat. Statistique*, 21:2–8, 1953.

[24] A. Dembo and T.M. Cover. Information theoretic inequalities. *IEEE Trans. On Information Theory*, 37(6):1501–1518, 1991.

[25] P. Diaconis and D. Freedman. Asymptotics of graphical projection pursuit. *Ann. Statist.*, 12:793–815, 1984.

[26] D. Donoho. On minimum entropy deconvolution. pages 565–608, New York, 1981.

[27] A. Edelman, T.A. Arias, and S.T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.*, 20(2):303–353, 1999.

[28] M.B. Eisen, P.T. Spellman, P.O. Brown, , and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95:14863–14868, December 1998.

[29] J. Eriksson, J. Karvanen, and V. Koivunen. Source distribution adaptive maximum likelihood estimation of ICA model. In P. Pajunen and J. Karhunen Editors, editors, *Proceedings of Second International Workshop on Independent Component Analysis and Blind Signal Separation*, pages 227–232, Helsinki, 2000.

[30] A.I. Fleishman. A method for simulating non-normal distributions. *Psychometrika*, 43(4):521–532, December 1978.

[31] J. Friedman and J.W. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. on Computers*, C-23(9):881–889, 1974.

[32] J.H. Friedman. Exploratory projection pursuit. *Journal of the American Statistical Association*, 82(397):249–266, March 1987.

[33] N. Friedman and M. Goldszmidt. Discretizing continuous attributes while learning bayesian networks. In *Proc. 13th International Conference on Machine Learning (ICML)*, pages 157–165, 1996.

[34] N. Friedman, I. Nachman, and D. Peér. Learning bayesian network structure from massive datasets: The ”sparse candidate” algorithm. pages 206–215.

[35] Dan Geiger and David Heckerman. Learning gaussian networks. Technical Report MSR-TR-94-10, Microsoft, Redmond, WA, 1994.

[36] M. Girolami and C. Fyfe. An extended exploratory projection pursuit network with linear and non-linear anti-hebbian lateral connections applied to the cocktail party problem. *Neural Networks*, 10(9):1607–1618, 1997.

[37] G.H. Golub and C.F. Van Loan. *Matrix Computations (3*rd *ed.)*. The Johns Hopkins University Press, Baltimore and London, 1996.

[38] T. Hastie, R. Tibshirani, and J.H. Friedman. *The Elements of Statistical Learning*. Springer Verlag, New York, 2001.

[39] D. Heckerman. A tutorial on learning with bayesian networks, 1995.

[40] P.J. Huber. Projection pursuit. *Annals of Statistics*, 13, Issue 2:435–475, June 1985.

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

[42] A. Hyvärinen. Survey on independent component analysis. *Neural Computing Surveys*, 2:94–128, 1999.

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

[44] S. i. Amari and J.-F. Cardoso. Blind source separation; semiparametric statistical approach. volume 45, pages 2692–2700, December 1997.

[45] S. i. Amari, T.-P. Chen, and A. Cichocki. Stability analysis of learning algorithms for blind source separation. *Neural Networks*, 10(8):1345–1351, 1997.

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

[47] J.E. Jackson. *A User's Guide to Principal Components*. Wiley-Interscience, New York, 1991.

[48] F.V. Jensen. *An introduction to Bayesian Theory*. Springer, New York, 1996.

[49] M.C. Jones. *The Projection Pursuit Algorithm for Exploratory Data Analysis*. PhD thesis, University of Bath, School of Mathematics, 1983.

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

[51] M.I. Jordan, editor. *Learning in Graphical Models*. The MIT Press, 1999.

[52] T.-P. Jung, C. Humphries, T.-W. Lee, M.J. McKeown, V. Iragui, S. Makeig, and T.J. Sejnowski. Removing electroencephalographic artifacts by blind source separation. *Psychophysiology*, 37:163–178, 2000.

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

[54] J. Karvanen, J. Eriksson, and V. Koivunen. Pearson system based method for blind separation. In P. Pajunen and J. Karhunen Editors, editors, *Proceedings of Second International Workshop on Independent Component Analysis and Blind Signal Separation*, pages 585–590, Helsinki, 2000.

[55] M.G. Kendall and A. Stuart. *The Advanced Theory of Statistics. Volume I: Distribution Theory (4th ed.).* Griffin, London, 1977.

[56] T. Kohonen. Self-organizing formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59.

[57] R. Korf. Linear-space best first search. *Artificial Intelligence*, 62:41–78, 1993.

[58] J.B. Kruskal. Toward a practical method which helps uncover the structure of the set of multivariate observations by finding the linear transformation which optimizes a new 'index of condensation'. In R. C. Milton and J. A. Nelder, editors, *Statistical Computation*, pages 427–440. Academic Press, New York, 1969.

[59] T.-W. Lee. *Independent Component Analysis: Theory and Applications.* Kluwer Academic Publishers, Boston.

[60] T.-W. Lee, A.J. Bell, and R. Orglmeister. Blind source separation of real-world signals. In *IEEE Proceedings ICNN*, pages 2129–2135, Houston, TX, 1997.

[61] T.-W. Lee, Mark Girolami, and Terrence J. Sejnowski. Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. *Neural Computation*, 11(2):417–441, 1999.

[62] J.C. Liao and C. Sabatti. Microanalysis of dna microarrays. *ASM News*, 68(9):432–437, 2002.

[63] R. Linsker. Local synaptic learning rules suffice to maximize mutual information in a linear network. *Neural Computation*, 4(5):691–702, 1992.

[64] J.Y. Liu, P.F. Miller, M. Gosink, and E.R. Olson. The identification of a new family of sugar efflux pumps in escherichia coli. *Mol. Microbiology*, 31(6):1845–1851, 1999.

[65] D.J.C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992.

[66] S. Makeig, T.-P. Jung, A.J. Bell, D. Ghahremani, and T.J. Sejnowski. Blind separation of event-related brain responses into independent components. *Proceedings of the National Academy of Sciences USA*, 94:10979–10984, 1997.

[67] M.J. McKeown, T.-P. Jung, S. Makeig, G. Brown, S.S. Kindermann, T.-W. Lee, and T.J. Sejnowski. Spatially independent activity patterns in functional magnetic resonance imaging data during the stroop color-naming task. *Proceedings of the National Academy of Sciences USA*, 95:803–810, 1998.

[68] M.J. McKeown, S. Makeig, G. Brown, T-P. Jung, S.S. Kindermann, A.J. Bell, V. Iragui, and T.J. Sejnowski. Analysis of fMRI by blind separation into independent spatial components. *Human Brain Mapping*, 6(3):160–188, 1998.

[69] K. Murphy. Learning bayes net structure from sparse data sets, 2001.

[70] J.P. Nadal and N. Parga. Non-linear neurons in the low-noise limit: A factorial code maximises information transfer. *Network*, 4:295–312, 1994.

[71] G. Nason. Three-dimensional projection pursuit. *Journal of the Royal Statistical Society, Series C*, 44(4):411–430, 1995.

[72] R.M. Neal. Probabilistic inference using markov chain monte carlo methods. Technical Report CRG-TR-93-1, University of Toronto, September 1993.

[73] J. Nocedal and S.J. Wright. *Numerical Optimization.* Springer-Verlag New York, Inc., 1999.

[74] D. Obradovic and G. Deco. Information maximization and independent component analysis: Is there a difference? *Neural Computation*, 10:2085–2101, 1998.

[75] A. Papoulis. *Probability, Random Variables, and Stochastic Processes.* WCB/McGraw-Hill, 1991.

[76] J. Pearl. Fusion, propagation, and structuring in belief networks. *Artificial Intelligence*, 29:241–288, September 1986.

[77] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufman, 1988.

[78] Judea Pearl and Tom S. Verma. A theory of inferred causation. In James F. Allen, Richard Fikes, and Erik Sandewall, editors, *KR'91: Principles of Knowledge Representation and Reasoning*, pages 441–452, San Mateo, California, 1991. Morgan Kaufmann.

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

[80] J.G. Proakis and D. Manolakis. *Digital Signal Processing: Principles, Algorithms, and Applications.* Prentice Hall, New Jersey, 3 edition, 1996.

[81] M. Rattray and Gleb Basalyga. Scaling laws and local minima in hebbian ica. In S. Becker T.G. Dietterich and Z. Ghahramani Editors, editors, *Advances in Neural Information Processing Systems 14*, Vancouver, Canada, December.

[82] M. Rattray and Gleb Basalyga. Stochastic trapping in a solvable model of on-line Independent Component Analysis. *Neural Computation*, 14:421–435, 2002.

[83] J. Rissanen. Stochastic complexity (with discussion). *Journal of the Royal Statistical Society, Series B*, 49:223–239 and 253–265, 1987.

[84] S. Roberts and R. Everson, editors. *Independent Component Analysis : Principles and Practice*. Cambridge University Press, 2001.

[85] C. Sabatti, L. Rohlin, M.-K. Oh, and J.C. Liao. Co-expression pattern from dna microarray experiments as a tool for operon prediction. *Nucleic Acid Res.*, 30(13):2886–2893, 2002.

[86] B.W. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, New York, 1985.

[87] V.P. Skitovich. On one property of the normal distribution. *Dokl. Akad. Nauk SSSR*, 89:217–219, 1953.

[88] G.C. Tseng, M.-K. Oh, L. Rohlin, J.C. Liao, and W.H. Wong. Issues in cdna microarray analysis: quality filtering, channel normalization, models of variations and assessment of genes effects. *Nucleic Acid Res.*, 29:2549–2557, 2001.

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

[90] D. L. Wallace. Asymptotic approximations to distributions. *Ann. Math. Stat.*, 29:635–654, 1958.