Chapter 2

Independent component analysis and blind source separation

Erkki Oja, Juha Karhunen, Aapo Hyvärinen, Petteri Pajunen, Ricardo Vigário, Harri Valpola, Jaakko Särelä, Ella Bingham, Mika Inki, Antti Honkela, Tapani Raiko, Karthikesh Raju, Alexander Ilin, Rãzvan Cristescu, Simona Mãlãroiu, Kimmo Kiviluoto, Mika Ilmoniemi

2.1 Introduction

Independent Component Analysis (ICA) is a computational technique for revealing hidden factors that underlie sets of measurements or signals. ICA assumes a statistical model whereby the observed multivariate data, typically given as a large database of samples, are assumed to be linear or nonlinear mixtures of some unknown latent variables. The mixing coefficients are also unknown. The latent variables are nongaussian and mutually independent, and they are called the independent components of the observed data. By ICA, these independent components, also called sources or factors, can be found. Thus ICA can be seen as an extension to Principal Component Analysis and Factor Analysis. ICA is a much richer technique, however, capable of finding the sources when these classical methods fail completely.

In many cases, the measurements are given as a set of parallel signals or time series. Typical examples are mixtures of simultaneous sounds or human voices that have been picked up by several microphones, brain signal measurements from multiple EEG sensors, several radio signals arriving at a portable phone, or multiple parallel time series obtained from some industrial process. The term blind source separation is used to characterize this problem.

The technique of ICA is a relatively new invention. It was for the first time introduced in early 1980's in the context of artificial neural networks. In mid-1990's, some highly successful new algorithms were introduced by several research groups, together with impressive demonstrations on problems like the cocktail-party effect, where the individual speech waveforms are found from their mixtures. ICA became one of the exciting new topics both in the field of neural networks, especially unsupervised learning, and more generally in advanced statistics and signal processing.

In our ICA research group, the research stems from some early work on on-line PCA, nonlinear PCA, and separation, that we were involved with in the 80's and early 90's. Since mid-90's, our ICA group grew considerably. That earlier work has been reported in the previous Triennial reports of our laboratory from 1994 to 1999.

In the reporting period 2000 - 2001, our ICA research was extended to several new directions. It is no more possible to report all this work under a single ICA Chapter. The most advanced developments are now presented in their separate Chapters in this report. They are: “Bayesian ensenble learning of generative models” (a project led by prof. Juha Karhunen and Dr. Harri Valpola), “Biomedical data analysis” (a project led by Dr. Ricardo Vigario) and “Computational neuroscience” (a project led by Doc. Aapo Hyvärinen). There is also a separate Chapter for various applications of ICA.

This Chapter covers some theoretical projects that ended during the reporting period. Then, the EU project BLISS is reviewed, and the chapter ends with a description of the ICA2000 international workshop organized by our group.

2.2 Theoretical advances

Local ICA

Juha Karhunen, Simona Mãlãroiu, Mika Ilmoniemi

In standard Independent Component Analysis (ICA), a linear data model is used for a global description of the data. Even though linear ICA yields meaningful results in many cases, it can provide a crude approximation only for general nonlinear data distributions. We have proposed a new structure [1], where local ICA models are used in connection with a suitable grouping algorithm clustering the data. The clustering part is responsible for an overall coarse nonlinear representation of the data, while linear ICA models of each cluster are used for describing local features of the data. The goal is to represent the data better than in linear ICA while avoiding computational difficulties related with nonlinear ICA.

Several data grouping methods were considered, including standard K-means clustering, self-organizing maps, and neural gas. In the journal paper [1] summarizing the results of this research line, connections to existing methods are discussed, and experimental results are given for artificial data and natural images. Furthermore, a general theoretical framework encompassing a large number of methods for representing data is introduced in [1]. These range from global, dense representation methods to local, very sparse coding methods. The proposed local ICA methods lie between these two extremes.

The FastICA algorithm for complex valued signals

Ella Bingham and Aapo Hyvärinen

The purpose of this project was to extend the applicability of the FastICA algorithm [4, 5] to complex-valued signals.

Separation of complex valued signals is a frequently arising problem in signal processing: frequency-domain implementations involving complex valued signals have advantages over time-domain implementations. Especially in the separation of convolutive mixtures it is a common practice to Fourier transform the signals, which results in complex valued signals. It is assumed that the original, complex valued source signals are mutually statistically independent, and the problem is solved by the independent component analysis (ICA) model.

We presented a fast fixed-point type algorithm that is capable of separating complex valued, linearly mixed source signals. The computational efficiency of the algorithm was shown by simulations. Also, the local consistency of the estimator given by the algorithm was proved. The results are presented in detail in [2, 3].

Overlearning in ICA

Jaakko Särelä and Ricardo Vigário

All ICA algorithms face the classical overlearning problem, if the number of data points is too low compared to the dimension of the data. This overlearning problem is studied in [6]. In the extreme case, where there are only as many data points as there are dimensions, the ICA algorithms result in the extreme maximization of their cost functions, totally independent of the data. In case of higher-order ICA algorithms this results in signals that are almost zero everywhere except for a single spike. This overlearning problem is usually quite easy to overcome by collecting more data. Even more efficient way is the reduction of the data dimension by PCA, when the dimension of the data exceeds the number of underlying sources.

If the sources are not independent in time, the problem becomes more severe. This is because the additional information each new data point gives is reduced, thus diminishing the *effective* number of data points. This temporal dependence, with its associated low frequencies, privileges the appearence of *bumps* rather than the spikes. This behaviour is very common to magnetoencephalographic signals (see Chapter 6). The paper [6] presents several preprocessing techniques as well as modifications to the classical ICA algorithms to overcome the problem of bumps.

References

[1] J. Karhunen, S. Mãlãroiu, and M. Ilmoniemi, Local linear independent component analysis based on clustering. *Int. J. Neural Systems*, Vol. 10, No. 6, 2000, pages 439-451.

[2] Ella Bingham and Aapo Hyvärinen. A fast fixed point algorithm for Independent Component Analysis of complex valued signals. *International Journal of Neural Systems*, 10(1):1–8, February 2000.

[3] Aapo Hyvärinen, Juha Karhunen and Erkki Oja. *Independent Component Analysis*, Wiley Interscience, 2001.

[4] Aapo Hyvärinen and Erkki Oja. A fast fixed-point algorithm for Independent Component Analysis. *Neural Computation*, 9:1483–1492, 1997.

[5] Aapo Hyvärinen. Fast and robust fixed-point algorithms for Independent Component Analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, May 1999.

[6] J. Särelä and R. Vigário. The problem of overlearning in high-order ICA approaches: analysis and solutions. Proceedings of the international workshop on artificial neural networks (IWANN'01), (Granada,Spain), pages 818–825, 2001.

2.3 The European joint project BLISS

Juha Karhunen, Harri Valpola, Ricardo Vigario, Erkki Oja, Antti Honkela, Jaakko Särelä

Our laboratory is one of the five participants in a large European joint project on Blind Source Separation and Applications, abbreviated BLISS. The project covers the three year period between June 2000 and June 2003, and its total funding is 1.2 million euros. It belongs to the "Information Society Technologies" programme (1998-2002) funded by the European Community. The other participating institutes and the leaders of the BLISS project there are:

- INESC, Lisbon, Portugal (Prof. Luis Almeida, coordinator);
- INPG (Inst. Nat. Polytechnique de France), Grenoble, France (Profs. Christian Jutten and Dinh-Tuan Pham);
- GMD First (Fraunhofer Institute), Berlin, Germany (Prof. Klaus-Robert Müller);
- McMaster University, Hamilton, Canada (Prof. Simon Haykin; adjunct member, getting its funding from Canada).

In addition, the project has both industrial and scientific advisory board members.

The project divides into two major parts: Theory and Algorithms, and Applications. The first part Theory and Algorithms consists of three subprojects, which are Linear ICA, Nonlinear Separation, and Nonlinear BSS (Blind Source Separation). Our laboratory is involved in the last two subprojects. The second major part Applications has two subprojects, Biomedical Applications and Acoustic Mixtures, and our laboratory participates in the first subproject Biomedical Applications.

The kick-off meeting of the whole project was arranged by our laboratory in context with the ICA2000 workshop held in Espoo in June 2000. The project has official meetings twice a year, with a review meeting after each passed year. The first review meeting was held in Granada, Spain, in June 2001. The reviewers assessed there this project as an excellent one. The third meeting was held in San Diego in December 2001 in context with the ICA2001 workshop. There has also been some visiting researchers between the participating laboratories for deepening practical co-operation. A major event will be the European Meeting on Independent Component Analysis to be held in Vietri sul Mare, Italy, in February 2002. It is largely organized by the participants of the BLISS project, in particular Prof. Erkki Oja, together with Italian researchers. This meeting also serves as a winter school for graduate students.

The research results achieved by our laboratory using the BLISS project funding are described in the chapters "Bayesian ensemble learning for generative models" and "Analysis of independent components in biomedical signals" as well as in the many associated publications. The results have been reported also in the confidential deliverables and reports of the project. More information on the BLISS project is available on its homepage [1].

References

[1] Homepage of the BLISS project `http://www.bliss-project.org/`.

2.4 New book: Independent Component Analysis, by A. Hyvärinen, J. Karhunen, and E. Oja (Wiley, 2001)

The original articles on ICA have been published during the past 20 years in a large number of journals and conference proceedings in the fields of statistics, signal processing, artificial neural networks, information theory, and various application fields. Some edited collections of articles as well as some monographs on ICA, blind source separation, and related subjects have appeared recently. However, while highly useful for their intended readership, these existing texts typically concentrate on some selected aspects of the ICA methods only. In the brief scientific papers and book chapters, mathematical and statistical preliminaries are usually not included, which makes it very hard for a wider audience to gain full understanding of this fairly technical topic. A comprehensive and detailed text-book was missing, which would cover both the mathematical background and principles, algorithmic solutions, and practical applications of the present state-of-the-art of ICA.

The three authors (Dr. Aapo Hyvärinen and professors Juha Karhunen and Erkki Oja) took up the task of writing such a text-book. The ordering of the names reflects the number of chapters that each author wrote, with Aapo Hyvärinen doing the bulk of the work. All the text was carefully read by all the authors, with many rewritings, resulting in the finished manuscript. The project lasted from summer 1999 to autumn 2000, and the book "Independent Component Analysis" [1] came out in spring 2001, published by Wiley-Interscience in the Wiley Series on Adaptive and Learning Systems for Signal Processing, Communications, and Control. The series editor, prof. Simon Haykin, was quite influential in convincing us to write the book and in seeing it through the various stages of publication.

This book provides a comprehensive introduction to ICA as a statistical technique. The emphasis is on the fundamental mathematical principles and basic algorithms. Much of the material is based on the original research conducted in the authors' own research group, which is naturally reflected in the weighting of the different topics. We give a wide coverage especially to those algorithms that are scalable to large problems, having a large number of mixtures and sources available. These will be increasingly used in near future, when ICA moves from the toy problems used in theoretical research into practical real world problems. Respectively, somewhat less emphasis is given to signal processing methods involving convolutive mixtures, delays, and other blind source separation techniques than ICA.

This book serves as a fundamental introduction to ICA. It is expected that the readership will be from a variety of disciplines such as statistics, signal processing, neural networks, cognitive science, information theory, artificial intelligence, mathematics, and engineering. The book is suitable for a graduate level university course on ICA, which is facilitated by the exercise problems and computer assignments given throughout the book, as well as by the lecture slides that are now available, resulting from the course given by J. Karhunen and E. Oja at HUT in autumn 2001.

For more details of the book, see `http://www.cis.hut.fi/projects/ica/book/`.

References

[1] Aapo Hyvärinen, Juha Karhunen and Erkki Oja. *Independent Component Analysis*, Wiley Interscience, 2001.

2.5 ICA2000: the Second International Workshop on Independent Component Analysis and Blind Source Separation

Erkki Oja, Juha Karhunen, Visa Koivunen, Petteri Pajunen, Aapo Hyvärinen, Jukka Iivarinen

Beginning from mid-nineties, Independent Component Analysis (ICA) has become a very popular research topic in the intersection of signal processing and neural networks. The first major scientific meeting exclusively devoted to ICA was the International Workshop on Independent Component Analysis and Blind Source Separation (ICA99), organized in January 1999 in Aussois in the French Alps. The workshop was highly successful, with every major ICA research group represented.

During that meeting it became clear that a continuation of the workshop was needed, and the Finnish group at Helsinki University of Technology undertook the job of organizing the second workshop on ICA and BSS. As for the timing, the general opinion was that an annual meeting may be too frequent but a two year interval is too long. We settled for a compromise of one and a half year interval, also motivated by the fact that June is definitely preferable over January for a conference held in northern Europe.

The Second International Workshop on Independent Component Analysis and Blind Source Separation (ICA2000) was held on June 19 - 22, 2000, on the beautiful Hanasaari island in Espoo. It was a major effort for our ICA group. The General Chairman was Prof. Erkki Oja, the Program Chair Prof. Juha Karhunen, the Local Arrangements Chair Prof. Visa Koivunen, the Publications Chair Prof. Petteri Pajunen, the Publicity Chair Dr. Aapo Hyvärinen, and the Financial Chair Dr. Jukka Iivarinen. Mr. Jaakko Särelä and Mr. Markus Peura were in charge of the graphical design and homepages of the workshop, and a large number of the members of our local ICA research community acted as volunteer assistants in the various conference tasks. The core task, selecting the papers and making up the program, was expertly handled by the International Program Committee, consisting of 26 eminent ICA researchers from Finland, France, Germany, UK, Portugal, USA, Canada, Japan, and P.R.China.

The workshop program contained four invited papers by eminent scientists working on ICA, BSS, or closely related fields, as well as 99 high-quality technical papers. The invited talks were “Source separation: from dusk till dawn” by prof. C. Jutten (Institut National Polytechnique de Grenoble, France), “Multiuser detection: an overview and a new result” by prof. U. Madhow (University of Santa Barbara, USA), “Nature vs. math: interpreting ICA in light of recent work in harmonic analysis” by prof. D. Donoho (Stanford University, USA), and “Independent component analysis of biomedical signals” by prof. T. Sejnowski (Salk Institute, San Diego, USA). The regular sessions ranged from theoretical ones to several applications. The number of registered participants was about 120. This was a well focussed, high level workshop covering all the main aspects of ICA and blind source separation.

Full contents of the papers are included in the Proceedings [1], now also available in CD Rom. The Web page `www.cis.hut.fi/ica2000/` gives some more information, including several photos of the main events.

The two ICA meetings had a succession in December, 2001, when the Third International Workshop on ICA and BSS was held in San Diego, USA. The next ICA workshop, to be held in Nara, Japan, is already at the planning stage.

Figure 2.1: Participants of the ICA workshop on the lawn of Hanasaari Cultural Centre

References

[1] *Proceedings of the Second International Workshop on Independent Component Analysis and Blind Signal Separation*. P. Pajunen and J. Karhunen, Editors. Helsinki University of Technology, Neural Networks Research Centre, 2000.