Chapter 2

Independent component analysis and blind source separation

Erkki Oja, Juha Karhunen, Harri Valpola, Jaakko Särelä, Mika Inki, Antti Honkela, Alexander Ilin, Karthikesh Raju, Tapani Ristaniemi, Ella Bingham

2.1 Introduction

What is Independent Component Analysis and Blind Source Separation? 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. Also other criteria than independence can be used for finding the sources.

Our contributions in ICA research. 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. This earlier work has been reported in the previous Triennial and Biennial reports of our laboratory from 1994 to 2003. A notable achievement from that period was the textbook "Independent Component Analysis" (Wiley, May 2001) by A. Hyvärinen, J. Karhunen, and E. Oja. It has been very well received in the research community; according to the latest publisher's report, over 3900 copies have been sold by August, 2005. The book has been extensively cited in the ICA literature and seems to have evolved into the standard text on the subject worldwide. In 2005, the Japanese translation of the book appeared.

Another tangible contribution has been the public domain FastICA software package (`http://www.cis.hut.fi/projects/ica/fastica/`). This is one of the few most popular ICA algorithms used by the practitioners and a standard benchmark in algorithmic comparisons in ICA literature.

In the reporting period 2004 - 2005, ICA/BSS research stayed as a core project in the laboratory. It was extended to several new directions. This Chapter starts by introducing some theoretical advances on FastICA undertaken during the reporting period. Then, several extensions and applications of ICA and BSS are covered, namely nonlinear ICA and BSS, the Denoising Source Separation (DSS) algorithm, its applications to climate data analysis and telecommunications signals, ICA for image representations, and a latent variable method for analyzing binary data.

2.2 Finite sample behaviour of the FastICA algorithm

In ICA, a set of original source signals are retrieved from their mixtures based on the assumption of their mutual statistical independence. The simplest case for ICA is the instantaneous linear noiseless mixing model. In this case, the mixing process can be expressed as

$$\mathbf{X} = \mathbf{AS}, \tag{2.1}$$

where $\mathbf{X}$ is an $d \times N$ data matrix. Its rows are the observed mixed signals, thus d is the number of mixed signals and N is their length or the number of samples in each signal. Similarly, the unknown $d \times N$ matrix $\mathbf{S}$ includes samples of the original source signals. $\mathbf{A}$ is an unknown regular $d \times d$ mixing matrix. It is assumed square because the number of mixtures and sources can always be made equal in this simple model.

In spite of the success of ICA in solving even large-scale real world problems, some theoretical questions remain partly open. One of the most central questions is the theoretical accuracy of the developed algorithms. Mostly the methods are compared through empirical studies, which may demonstrate the efficacy in various situations. However, the general validity cannot be proven like this. A natural question is, whether there exists some theoretical limit for separation performance, and whether it is possible to reach it.

Many of the algorithms can be shown to converge in theory to the correct solution giving the original sources, under the assumption that the sample size N is infinite. This is unrealistic. For finite data sets, what typically happens is that the sources are not completely unmixed but some traces of the other sources remain in them even after the algorithm has converged. This means that the obtained demixing matrix $\widehat{\mathbf{W}}$ is not exactly the inverse of $\mathbf{A}$, and the matrix of estimated sources $\mathbf{Y} = \widehat{\mathbf{W}}\mathbf{X} = \widehat{\mathbf{W}}\mathbf{AS}$ is only approximately equal to $\mathbf{S}$. A natural measure of error is the deviation of the so-called gain matrix $\mathbf{G} = \widehat{\mathbf{W}}\mathbf{A}$ from the identity matrix, i.e., the variances of its elements.

The well-known lower limit for the variance of a parameter vector in estimation theory is the Cramér-Rao lower bound (CRB). In publications [1, 2], the CRB for the demixing matrix of the FastICA algorithm was derived. The result depends on the score functions of the sources,

$$\psi_k(s) = -\frac{d}{ds}\log p_k(s) = -\frac{p_k'(s)}{p_k(s)} \tag{2.2}$$

where $p_k(s)$ is the probability density function of the k-th source. Let

$$\kappa_k \quad = \quad \mathrm{E}\left[\psi_k^2(s_k)\right]. \tag{2.3}$$

Then, assuming that the correct score function is used as the nonlinearity in the FastICA algorithm, the asymptotic variances of the off-diagonal elements (k,ℓ) of matrix $\mathbf{G}$ for the one-unit and symmetrical FastICA algorithm, respectively, read

$$V_{k\ell}^{1U-opt} \quad = \quad \frac{1}{N}\frac{1}{\kappa_k - 1} \tag{2.4}$$

$$V_{k\ell}^{SYM-opt} \quad = \quad \frac{1}{N}\frac{\kappa_k + \kappa_\ell - 2 + (\kappa_\ell - 1)^2}{(\kappa_k + \kappa_\ell - 2)^2}, \tag{2.5}$$

while the CRB reads

$$\mathrm{CRB}(\mathbf{G}_{k\ell}) = \frac{1}{N}\frac{\kappa_k}{\kappa_k\kappa_\ell - 1}. \tag{2.6}$$

Comparison of these results implies that the algorithm FastICA is nearly statistically efficient in two situations:

(1) One-unit version FastICA with the optimum nonlinearity is asymptotically efficient for $\kappa_k \to \infty$, regardless of the value of κ_ℓ.

(2) Symmetric FastICA is nearly efficient for κ_i lying in a neighborhood of 1^+, provided that all independent components have the same probability distribution function, and the nonlinearity is equal to the joint score function.

The work was continued to find a version of the FastICA that would be asymptotically efficient, i.e. with some choice of nonlinearities would be able to attain the CRB. This work [3] will be reported later.

References

[1] Koldovský, Z., Tichavský, P. and Oja, E.: Cramér-Rao lower bound for linear independent component analysis. *Proc. Int. Conf. on Acoustics, Speech, and Signal Proc. (ICASSP'05)*, March 20 - 23, 2005, Philadelphia, USA (2005).

[2] Tichavský, P., Koldovský, Z. and Oja, E.: Performance analysis of the FastICA algorithm and Cramér-Rao bounds for linear independent component analysis. *IEEE Trans. on Signal Processing 54*, no. 4, April 2006.

[3] Koldovský, Z., Tichavský, P., and Oja, E.: Efficient variant of algorithm FastICA for independent component analysis attaining the Cramér-Rao lower bound. *IEEE Trans. on Neural Networks*, to appear (2006).

2.3 Nonlinear ICA and BSS

Juha Karhunen, Antti Honkela, Alexander Ilin

Recent advances in blind source separation (BSS) and independent component analysis (ICA) for nonlinear mixing models have been reviewed in the invited journal paper [1]. After a general introduction to BSS and ICA, uniqueness and separability issues are discussed in more detail, presenting some new results. A fundamental difficulty in the nonlinear BSS problem and even more so in the nonlinear ICA problem is that they provide non-unique solutions without extra constraints, which are often implemented by using a suitable regularization. In the paper [1], two possible approaches are explored in more detail. The first one is based on structural constraints. Especially, post-nonlinear mixtures are an important special case, where a nonlinearity is applied to linear mixtures. For such mixtures, the ambiguities are essentially the same as for the linear ICA or BSS problems. The second approach uses Bayesian inference methods for estimating the best statistical parameters, under almost unconstrained models in which priors can be easily added. In the later part of the paper [1], various separation techniques proposed for post-nonlinear mixtures and general nonlinear mixtures are reviewed.

Our own research on nonlinear BSS has concentrated on the Bayesian approach which is described in Sec. 4.4. The latest results include the use of kernel PCA to initialize the model for improved accuracy in highly nonlinear problems as well as a variational Bayesian generative model for post-nonlinear ICA.

There exist few comparisons of nonlinear ICA and BSS methods, and their limitations and preferable application domains have been studied only a little. We have experimentally compared two approaches introduced for nonlinear BSS: the Bayesian methods developed at the Neural Network Research Centre (NNRC) of Helsinki University of Technology, and the BSS methods introduced for the special case of post-nonlinear (PNL) mixtures developed at Institut National Polytechnique de Grenoble (INPG) in France. This comparison study took place within the framework of the European joint project BLISS on blind source separation and its applications.

The Bayesian method developed at NNRC for recovering independent sources consists of two phases: Applying the general nonlinear factor analysis (NFA) [3] to obtain Gaussian sources; and their further rotation with a linear ICA technique such as the FastICA algorithm [2]. The compared BSS method, developed at INPG for post-nonlinear mixtures, is based on minimization of the mutual information between the sources. It uses a separating structure consisting of nonlinear and linear stages [4].

Both approaches were applied to the same ICA problems with artificially generated post-nonlinear mixtures of two independent sources. Based on the experimental results, the following conclusions were drawn on the applicability of the INPG and Bayesian NFA+FastICA approaches to post-nonlinear blind source separation problems:

1. The INPG method performs better in classical post-nonlinear mixtures with the same number of sources and observations when all post-nonlinear distortions are invertible.

2. The performance of both methods can be improved by exploiting more mixtures than the number of sources especially in the case of noisy mixtures.

3. The advantage of the Bayesian methods in post-nonlinear BSS problems is that they can separate post-nonlinear mixtures with non-invertible post-nonlinearities

provided that the full mapping is globally invertible. The existing INPG methods cannot do this due to their constrained separation structure.

The results of this comparison study were presented in [5].

References

[1] C. Jutten and J. Karhunen. Advances in Blind Source Separation (BSS) and Independent Component Analysis (ICA) for nonlinear mixtures. *Int. J. of Neural Systems*, 14(5):267-292, 2004.

[2] A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483–1492, 1997.

[3] H. Lappalainen and A. Honkela. Bayesian nonlinear independent component analysis by multi-layer perceptrons. In Mark Girolami, editor, *Advances in Independent Component Analysis*, pages 93–121. Springer-Verlag, Berlin, 2000.

[4] A. Taleb and C. Jutten. Source separation in post-nonlinear mixtures. *IEEE Transactions on Signal Processing*, 47(10):2807–2820, 1999.

[5] A. Ilin, S. Achard, and C. Jutten. Bayesian versus constrained structure approaches for source separation in post-nonlinear mixtures. In *Proc. of the Int. Joint Conf. on Neural Networks (IJCNN 2004)*, pages 2181–2186, Budapest, Hungary, 2004.

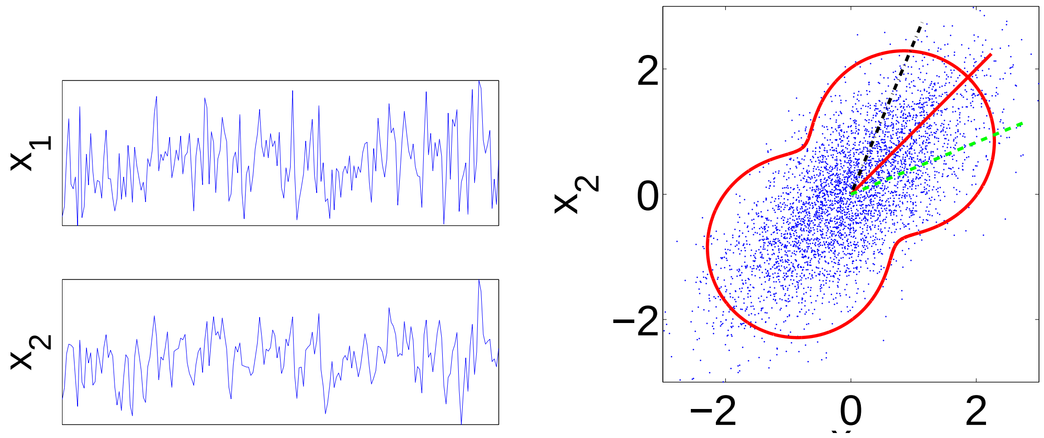
2.4 Denoising source separation

Jaakko Särelä

Denoising source separation (DSS,[1]) is a recently developed framework for linear source separation and feature extraction. In DSS, the algorithms are constructed around denoising operations. With certain types of denoising, DSS realises ICA.

With linear denoising, the algorithm consists of three steps: 1) sphering (whitening) 2) denoising 3) PCA. If the denoising is nonlinear, an iterative procedure has to be used, and the algorithm resembles nonlinear PCA and has the following iteration after presphering: 1) estimation of the sources using the current mapping, 2) denoising of the source estimates, 3) re-estimation of the mapping. The crucial part is the denoising, and available prior or acquired information may be implicitly implemented in it.

As an example, consider the two observations in Fig. 2.1. The correlation structure between the observations becomes apparent, when they are plotted against each others in a scatter-plot. The red curve illustrates the variance of different projections and the red line the direction where the variance is maximised. As it happens, the observations are linear mixtures of two independent sources. The mixing vectors are shown in the scatter-plot using the black and the green line.



(a) (b)

Figure 2.1: *a) Observed signals. b) The scatter-plot of the observed signals.*

As the original sources are independent of each others, a good attempt to recover them is to project the data to the principal components and thus remove any correlations between them. The resulting scatter-plot after normalisation of the variances (sphering or witening) is shown in Fig. 2.2. As illustrated by the red circle, the variance of any projection equals to one. However, the principal directions (y_1 and y_2) do not recover the original sources as shown by the black and the green line. Furthermore, there is no structure left in the scatter-plot.

The scatter-plot loses all the temporal structure the data may have. A good frequency representation is given by the discrete cosine transform (DFT). DFT of the sphered signals is shown in Fig. 2.3a. It seems that there are relatively more low frequencies than high frequencies. One hyphothesis could be that a source with low frequencies exists in the data. This source would become more visible (or denoised) by low-pass filtering. The resulting scatter-plot is shown in Fig. 2.3b.

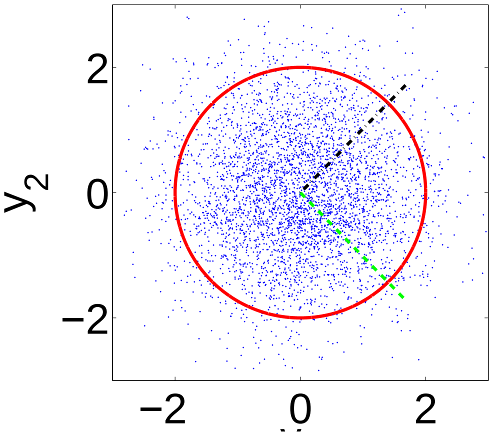


Figure 2.2: *Scatter-plot of the sphered signals.*

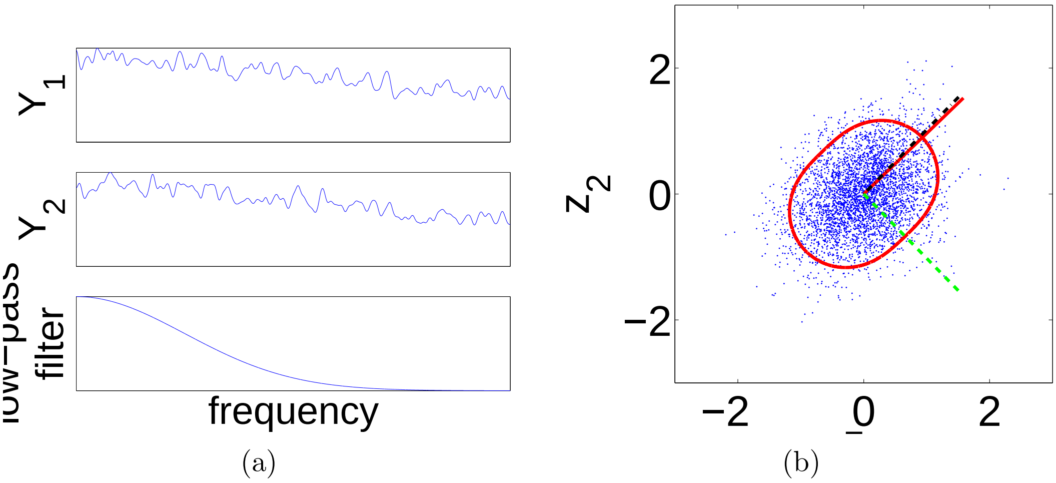


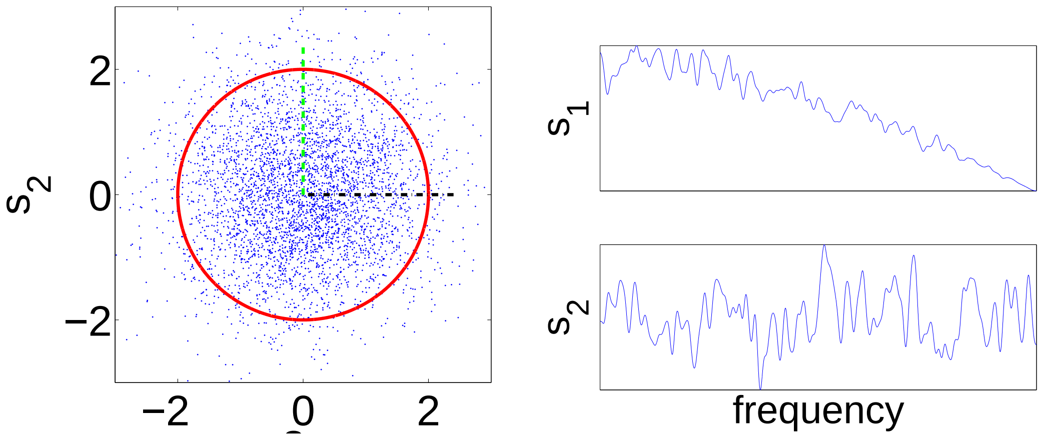
Figure 2.3: *a) Amplitude spectra of the sphered signals. b) The denoised signals.*

Now all directions do not have unit variance, and the maximal variance may be identified by another PCA. The principal directions align with the sphered mixing vectors (black and green line) and the original sources are recovered. The estimated sources and their amplitude spectra are shown in Fig. 2.4.

The DSS framework has been applied in several application fields. In this laboratory, we have applied it, for instance, to blind suppression of various interfering signals appearing in direct sequence CDMA communication systems (Sec. 2.6), to exploratory source separation of climate phenomena (Sec. 2.5) and to neuroinformatics (Ch. 3).

References

[1] J. Särelä and H. Valpola, “Denoising source separation,” *Journal of Machine Learning Research*, vol. 6, pp. 233–272, 2005.



(a) (b)

Figure 2.4: *a) The estimated sources. b) The amplitude spectra of the estimated sources.*

2.5 Climate data analysis with DSS

Alexander Ilin, Harri Valpola, Erkki Oja

One of the main goals of statistical analysis of climate data is to extract physically meaningful patterns of climate variability from highly multivariate weather measurements. The classical technique for defining such dominant patterns is principal component analysis (PCA), or empirical orthogonal functions (EOF) as it is called in climatology (see, e.g., [1]). However, the maximum remaining variance criterion used in PCA can lead to such problems as mixing different physical phenomena in one extracted component [2]. This makes PCA a useful tool for information compression but limits its ability to isolate individual modes of climate variation.

To overcome this problem, rotation of the principal components has proven useful. The classical rotation criteria used in climatology are based on the general concept of "simple structure" which can provide spatially or temporally localizated components [2]. Denoising source separation (DSS) is a tool which can also be used for rotating the principal components. It is particularly efficient when some prior information exists (e.g., the general shape of the time curves of the sources or their frequency contents). For example, in the climate data analysis we might be interested in some phenomena that would be cyclic over a certain period, or exhibit slow changes. Then, exploiting the prior knowledge may significantly help in finding a good representation of the data.

We use the DSS framework for exploratory analysis of the large spatio-temporal dataset provided by the NCEP/NCAR reanalysis project [3]. The data is the reconstruction of the daily weather measurements around the globe for a period of 56 years.

In our first works, we concentrate on slow climate oscillations and analyze three major atmospheric variables: surface temperature, sea level pressure and precipitation. In [4], we show that optimization of the criterion that we term clarity helps find the sources exhibiting the most prominent periodicity in a specific timescale. In the experiments, the components with the most prominent interannual oscillations are clearly related to the well-known El Niño–Southern Oscillation (ENSO) phenomenon. For all three variables, the most prominent component is a good ENSO index (see Fig. 2.5–2.6) and the second component is close to the derivative of the first one.

In [5], we extend the analysis to a more general case where slow components are separated by their frequency contents. The sources found using the frequency-based criterion give a meaningful representation of the slow climate variability as combination of trends, interannual oscillations, the annual cycle and slowly changing seasonal variations.

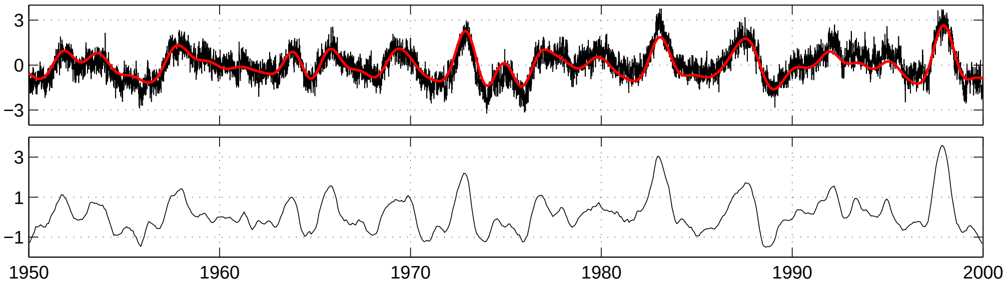


Figure 2.5: The dark curve on the upper plot shows the component with the most prominent interannual oscillations extracted with DSS. The red curve is the found component filtered in the interannual timescale. The lower plot presents the index which is used in climatology to measure the strength of El Niño. The curves have striking resemblance.

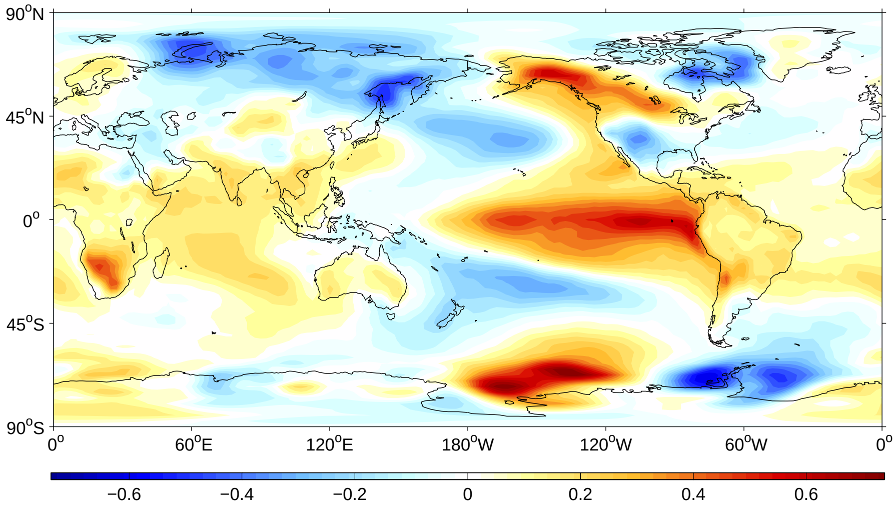


Figure 2.6: The temperature pattern corresponding to the component with the most prominent interannual oscillations. The map tells how strongly the component is expressed in the measurement data. The pattern has many features traditionally associated with El Niño. The scale of the map is in degrees centigrade.

References

[1] H. von Storch, and W. Zwiers. *Statistical Analysis in Climate Research.* Cambridge University Press, Cambridge, U.K, 1999.

[2] M. B. Richman. Rotation of principal components. *Journal of Climatology*, 6:293–335, 1986.

[3] E. Kalnay and coauthors. The NCEP/NCAR 40-year reanalysis project. Bulletin of the American Meteorological Society, 77:437–471, 1996.

[4] A. Ilin, H. Valpola, and E. Oja. Semiblind source separation of climate data detects El Niño as the component with the highest interannual variability. In *Proc. of the Int. Joint Conf. on Neural Networks (IJCNN 2005)*, pages 1722–1727, Montréal, Québec, Canada, 2005.

[5] A. Ilin, and H. Valpola. Frequency-based separation of climate signals. In *Proc. of 9th European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD 2005)*, pages 519–526, Porto, Portugal, 2005.

2.6 ICA and denoising source separation in CDMA communications

Karthikesh Raju, Tapani Ristaniemi, Juha Karhunen,
Jaakko Särelä and Erkki Oja

In wireless communication systems, like mobile phones, an essential issue is division of the common transmission medium among several users. A primary goal is to enable each user of the system to communicate reliably despite the fact that the other users occupy the same resources, possibly simultaneously. As the number of users in the system grows, it becomes necessary to use the common resources as efficiently as possible.

During the last years, various systems based on CDMA (Code Division Multiple Access) techniques [1, 2] have become popular, because they offer several advantages over the more traditional FDMA and TDMA schemes based on the use of non-overlapping frequency or time slots assigned to each user. Their capacity is larger, and it degrades gradually with increasing number of simultaneous users who can be asynchronous. On the other hand, CDMA systems require more advanced signal processing methods, and correct reception of CDMA signals is more difficult because of several disturbing phenomena [1, 2] such as multipath propagation, possibly fading channels, various types of interferences, time delays, and different powers of users.

Direct sequence CDMA data model can be cast in the form of a linear independent component analysis (ICA) or blind source separation (BSS) data model [3]. However, the situation is not completely blind, because there is some prior information available. In particular, the transmitted symbols have a finite number of possible values, and the spreading code of the desired user is known.

In this project, we have applied independent component analysis and denoising source separation (DSS) to blind suppression of various interfering signals appearing in direct sequence CDMA communication systems. The standard choice in communications for suppressing such interfering signals is the well-known RAKE detection method [2]. RAKE utilizes available prior information, but it does not take into account the statistical independence of the interfering and desired signal. On the other hand, ICA utilizes this independence, but it does not make use of the prior information. Hence it is advisable to combine the ICA and RAKE methods for improving the quality of interference cancellation.

In the journal paper [4], various schemes combining ICA and RAKE are introduced and studied for different types of interfering jammer signals under different scenarios. By using ICA as a preprocessing tool before applying the conventional RAKE detector, some improvement in the performance is achieved, depending on the signal-to-interference ratio, signal-to-noise ratio, and other conditions [4]. These studies have been extended to consider multipath propagation and coherent jammers in [5].

All these ICA-RAKE detection methods use the FastICA algorithm [3] for separating the interfering jammer signal and the desired signal. In the case of multipath propagation, it is meaningful to examine other temporal separation methods, too. The results of such a study have been presented in [7].

The paper [6] deals with application of denoising source separation [9] to interference cancellation. This is a semi-blind approach which uses the spreading code of the desired user but does not require training sequences. The results of the DSS-based interference cancellation scheme show improvements over conventional detection.

Work on both uplink and downlink interference cancellation in direct sequence CDMA systems has been summarized in the joint paper [8]. In this paper, an effort is made to present both uplink and downlink methods under a unified framework.

References

[1] S. Verdu, *Multiuser Detection.* Cambridge Univ. Press, 1998.

[2] J. Proakis, *Digital Communications.* McGraw-Hill, 3rd edition, 1995.

[3] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis.* Wiley, 2001, 481+xxii pages.

[4] K. Raju, T. Ristaniemi, J. Karhunen, and E. Oja, Jammer cancellation in DS-CDMA arrays using independent component analysis. *IEEE Trans. on Wireless Communications*, Vol. 5, No. 1, January 2006, pp. 77–82.

[5] K. Raju, T. Ristaniemi, and J. Karhunen, Semi-blind interference suppression on coherent multipath environments. In *Proc. of the First IEEE Int. Symp. of Control, Communications, and Signal Processing (ISCCSP2004)*, Hammamet, Tunisia, March 21-24, 2004, pp. 283–286.

[6] K. Raju and J. Särelä, A denoising source separation based approach to interference cancellation for DS-CDMA array systems. In *Proc. of the 38th Asilomar Conf. on Signals, Systems, and Computers*, Pacific Grove, California, USA, November 7-10, 2004, pp. 1111-1114.

[7] K. Raju and B. Phani Sudheer, Blind source separation for interference cancellation - a comparison of several spatial and temporal statistics based techniques. In *Proc. of the 3rd Workshop on the Internet, Telecommunications, and Signal Processing*, Adelaide, Australia, December 20-22, 2004.

[8] K. Raju, T. Huovinen, and T. Ristaniemi, Blind interference cancellation schemes for DS-CDMA systems. In *Proc. of the IEEE Int. Symp. on Antennas and Propagation and UNSC/URSI National Radio Science Meeting*, Washington, USA, July 3-8, 2005.

[9] J. Särelä and H. Valpola, Denoising source separation. *J. of Machine Learning Research*, Vol. 6, 2005, pp. 233–272.

2.7 ICA for image representations

Mika Inki

Already the earliest adapters of ICA on small image windows noted the similarity of the features to cortical simple cell receptive fields [1, 5]. This can be considered as support for the notion that the primary visual cortex (and early visual system in general) employs a strategy of sparse coding or redundancy reduction. In any case, the features obtained by ICA, and especially their efficiency in image coding and functionality in edge detection, can be argued to be useful when the objective is to build a hierarchical system capable of image analysis or understanding.

However, there are many limitations on the usefulness of the ICA description of images. A basic limitation is that ICA considers the components to be independent, which they are not in any sense with image data. Also, it can be argued that every possible scaling, translation and rotation of every ICA feature should also be in the basis, resulting in very highly overcomplete description, computationally infeasible to estimate. Another computational hindrance is the small window size necessitated by the curse of dimensionality.

We have focused on removing these limitations, and extending the ICA model to better account for image statistics, while comparing it to biological visual systems. We have, for example, examined the dependencies between ICA features in image data [3], built models based on these findings, studied overcomplete models [2, 4], and examined how the features can be extended past the window edges, cf. Figure 2.7.

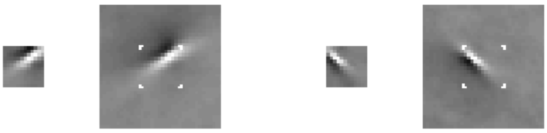


Figure 2.7: A couple of typical ICA features for images and their extensions.

References

[1] A.J. Bell and T.J. Sejnowski. The 'independent components' of natural scenes are edge filters. *Vision Research*, 37:3327–3338, 1997.

[2] A. Hyvärinen and M. Inki. Estimating overcomplete independent component bases for image windows. *Journal of Mathematical Imaging and Vision*, 2002.

[3] M. Inki. A model for analyzing dependencies between two ICA features in natural images. In *Proc. Int. Conf. on Independent Component Analysis and Blind Signal Separation*, Granada, Spain, 2004.

[4] M. Inki. An easily computable eight times overcomplete ICA method for image data. In *Proc. Int. Conf. on Independent Component Analysis and Blind Signal Separation*, Charleston, South Carolina, 2006.

[5] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37:3311–3325, 1997.

2.8 Analyzing 0-1 data

Ella Bingham

A novel probabilistic latent variable method for analyzing 0-1-valued (binary) data was presented. This method, termed as "aspect Bernoulli", was first described in [1]. The method is able to detect and distinguish between two types of 0s in the data: those that are "true absences" and those that are "false absences"; both of these are coded as 0 in the data.

As an example we may consider text documents in which some words are missing because they do not fit the topical content of the data — they are "true absences". Some other words are missing just because the author of the document did not use them although they would nicely fit the topic — these are "false absences" and the document could be augmented with these words. Another application might be black-and-white images in which some pixels are turned to white by an external noise process, resulting in "true" and "false" white pixels. Our method finds a specific latent component that accounts for the "false absences". Figure 2.8 shows results on this.

Similarly, the method can distinguish between two types of 1s: "true presences" and "false presences"; the latter could be extra black pixels in a black-and-white image, for example.

The method can be used in several applications: noise removal in black-and-white images; detection of false data instances in noisy data; and query expansion where topically related words are added into a document.

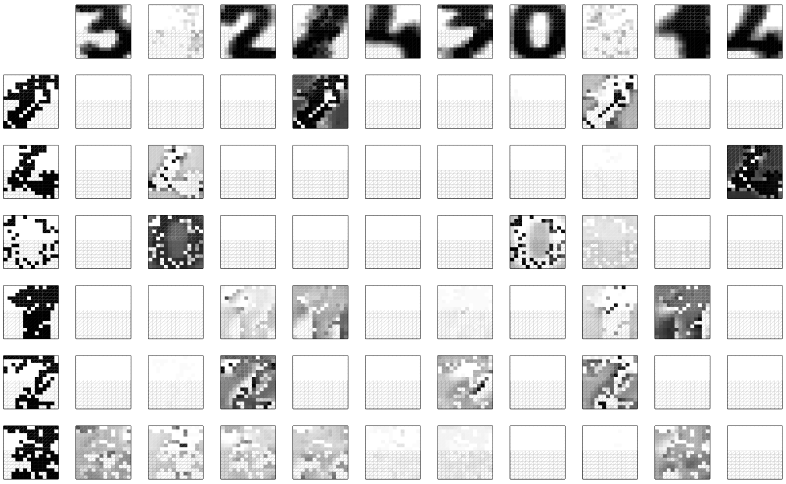


Figure 2.8: Analyzing corrupted black-and-white images. The top row shows the basis images estimated by the aspect Bernoulli model when corrupted images are fed into the algorithm. Examples of observed corrupted images are shown in the first column. The middle rows and columns give the pixel-wise probability that a basis image is responsible for generating a pixel in the observed image. For example, the observed digit "1" in the second row is largely generated by the 4th basis image resembling the digit "1", but the corrupted pixels are generated by the 8th basis image which is almost completely white and accounts for the corruption.

References

[1] Ata Kabán, Ella Bingham and Teemu Hirsimäki. Learning to read between the lines: The aspect Bernoulli model. *Proceedings of the 4th SIAM International Conference on Data Mining*, pp.462–466, 2004.