Chapter 3

Independent component analysis and blind source separation

Erkki Oja, Juha Karhunen, Alexander Ilin, Antti Honkela, Karthikesh Raju, Tomas Ukkonen, Zhirong Yang, Zhijian Yuan

3.1 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 2005. 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 5000 copies had been sold by August, 2007. The book has been extensively cited in the ICA literature and seems to have evolved into the standard text on the subject worldwide. In Google Scholar, the number of hits (in early 2008) is over 2300. In 2005, the Japanese translation of the book appeared (Tokyo Denki University Press), and in 2007, the Chinese translation (Publishing House of Electronics Industry).

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 2006 - 2007, ICA/BSS research stayed as one of the core projects in the laboratory, with the pure ICA theory somewhat waning and being replaced by several new directions. Chapter 3 starts by introducing some theoretical advances on the FastICA algorithm undertaken during the reporting period, followed by a number of extensions of ICA and BSS. The first one is the method of independent subspaces with decoupled dynamics, that can be used to model complex dynamical phenomena. The second extension is related to Canonical Correlation Analysis, and the third one is nonnegative separation by the new Projective Nonnegative Matrix Factorization (P-NMF) principle. An application of ICA to telecommunications is also covered. Then the Denoising Source Separation (DSS) algorithm is applied to climate data analysis. This is an interesting and potentially very useful application that will be under intensive research in the future in the group.

Another way to formulate the BSS problem is Bayesian analysis. This is covered in the separate Chapter 2.

3.2 Convergence and finite-sample behaviour of the FastICA algorithm

Erkki Oja

In Independent Component Analysis, 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

$$\mathbf{X} = \mathbf{AS}, \tag{3.1}$$

where $\mathbf{X}$ is a $d \times N$ 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.

Sometimes 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*. In [1], the FastICA algorithm was analyzed from this point of view. A central factor in the algorithm is a nonlinear function that is the gradient of the ICA cost function. It may be a polynomial, e.g. a cubic function in the case of kurtosis maximization/minimization, but it can be some other suitable nonlinearity as well. According to [1], let us present an example of convergence when the nonlinearity is the third power, and the 2×2 case is considered for the mixing matrix $\mathbf{A}$ in model (3.1).

In the theoretical analysis a linear transformation was made first, so that the correct solution for the separation matrix $\mathbf{W}$ (essentially the inverse of matrix $\mathbf{A}$) is a unit matrix or a variant (permutation and/or sign change). Thus the four matrix elements of $\mathbf{W}$ convergence to zero or to ± 1. The FastICA algorithm boils down to an iteration $w_{t+1} = f(w_t)$ for all the four elements of the separation matrix. The curve in Figure 3.2 shows the iteration function $f(.)$ governing this convergence. It is easy to see that close to the stable points, the convergence is very fast, because the iteration function is very flat.

In practice, however, the assumption of infinite sample size 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 [2], the CRB for the demixing matrix of the FastICA algorithm was derived. The result depends on the score functions of the sources,

$$\psi_k(s) = -\frac{d}{ds}\log p_k(s) = -\frac{p_k'(s)}{p_k(s)} \tag{3.2}$$

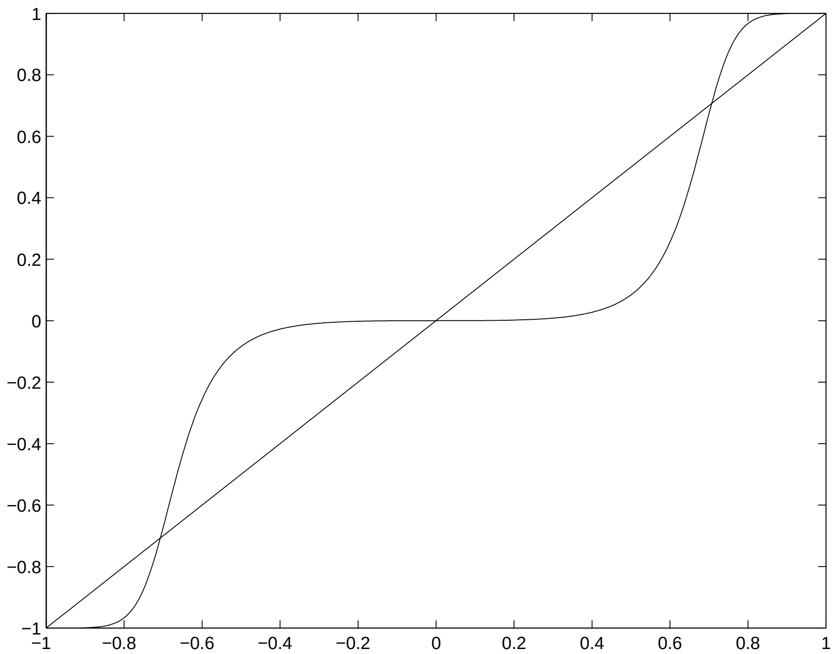

Figure 3.1: Shape of the iteration function for separation matrix elements, kurtosis case

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

$$\kappa_k = \mathrm{E}\left[\psi_k^2(s_k)\right]. \tag{3.3}$$

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

$$V_{k\ell}^{1U-opt} = \frac{1}{N}\frac{1}{\kappa_k - 1} \tag{3.4}$$

$$V_{k\ell}^{SYM-opt} = \frac{1}{N}\frac{\kappa_k + \kappa_\ell - 2 + (\kappa_\ell - 1)^2}{(\kappa_k + \kappa_\ell - 2)^2}, \tag{3.5}$$

while the CRB reads

$$\mathrm{CRB}(\mathbf{G}_{k\ell}) = \frac{1}{N}\frac{\kappa_k}{\kappa_k\kappa_\ell - 1}. \tag{3.6}$$

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. able to attain the CRB. This can be achieved in the orthogonalization stage of the FastICA algorithm: instead of requiring strict orthogonalization, this condition is relaxed to allow small deviations from orthogonality, controlled by a set of free parameters. These parameters can be optimized so that the exact CRB is reached by the new algorithm, given that the correct score functions are used as nonlinearities.

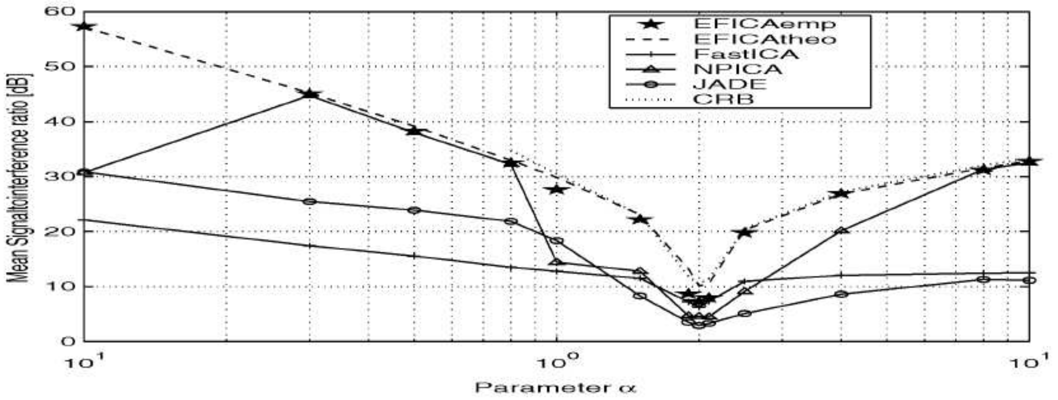

Figure 3.2: The Mean Signal-to Inference Ratio of EFICA, compared to CRB and some other ICA algorithms

The new efficient FastICA algorithm, dubbed EFICA, requires two phases because the score functions have to estimated first. Once they have been estimated, the new approximative orthogonalization scheme is run for a number of steps to reach the optimal solution. Figure 3.2 shows the efficiency of EFICA. To make meaningful comparisons, 13 source signals were artificially generated, each having a generalized gamma density $GG(\alpha)$ (where the value $\alpha = 2$ corresponds to the Gaussian density). The α values ranged from 0.1 to 10 and their places are marked by asterisks in the figure. The Mean Signal-to-Inference Ratio (SIR), both theoretical and experimental, obtained by EFICA is shown in the image (uppermost curve). It is very close to the Cramér-Rao Bound attainable in this situation, and far better than the Mean SIR attained by some other algorithms such as plain FastICA, NPICA, or JADE.

References

[1] Oja, E. and Yuan, Z.: The FastICA algorithm revisited – convergence analysis. *IEEE Trans. on Neural Networks 17*, no. 6, pp. 1370 - 1381 (2006).

[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, pp. 1189 - 1203 (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 17*, no. 5, pp. 1265 - 1277 (2006).

3.3 Independent subspaces with decoupled dynamics

Alexander Ilin

Independent subspace models extend the general source separation problem by allowing groups (subspaces) $\mathbf{s}_k$ of sources:

$$\mathbf{x}(t) = \sum_{k=1}^{K} \mathbf{A}_k \mathbf{s}_k(t)\,. \tag{3.7}$$

The sources within one group $\mathbf{s}_k$ are assumed dependent while signals from different groups are mutually independent. Similarly to classical BSS, subspaces can be separated exploiting non-Gaussianity or temporal structures of the mixed signals. The technique presented in [2] uses a first-order nonlinear model to model the dynamics of each subspace:

$$\mathbf{s}_k(t) = \mathbf{g}_k(\mathbf{s}_k(t-1)) + \mathbf{m}_k(t)\,, \quad k = 1, \ldots, K\,, \tag{3.8}$$

Both the de-mixing transformation and the nonlinearities $\mathbf{g}_k$ governing the dynamics are estimated simulateously by minimizing the mean prediction error of the subspace dynamical models (3.8). The optimization procedure can be implemented using the algorithmic structure of denoising source separation [1].

The algorithm was tested on artificially generated data containing linear mixtures of two independent Lorenz processes with different parameters, a harmonic oscillator and two white Gaussian noise signals (see Fig. 3.3). The algorithm is able to separate the three subspaces using only the information about their dimensionalities.

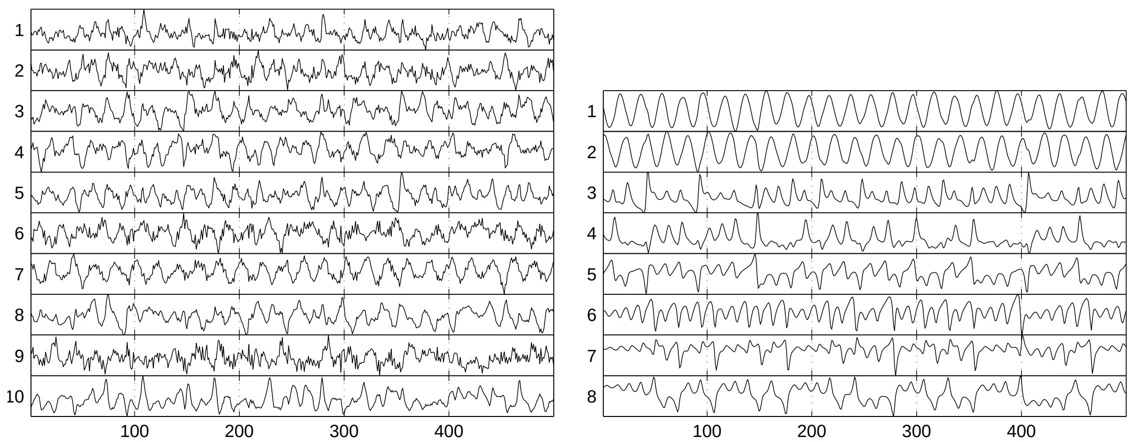

Figure 3.3: Left: Artificially generated linear mixtures of three dynamical processes and white noise signals. Right: Sources extracted by the technique extracting subspaces (signals 1–2, 3–5 and 6–9) with decoupled dynamics.

References

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

[2] A. Ilin. Independent dynamics subspace analysis. In *Proc. of the 14th European Symposium on Artificial Neural Networks (ESANN 2006)*, pp. 345–350, April 2006.

3.4 Extending ICA for two related data sets

Juha Karhunen, Tomas Ukkonen

Standard linear principal component analysis (PCA) [2, 1] and independent component analysis (ICA) [1] are both based on the same type of simple linear latent variable model for the observed data vector $\mathbf{x}(t)$:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) = \sum_{i=1}^{n} s_i(t)\mathbf{a}_i \tag{3.9}$$

In this model, the data vector $\mathbf{x}(t)$ is expressed as a linear transformation of the coefficient vector $\mathbf{s}(t) = [s_1(t), s_2(t), \ldots, s_n(t)]^T$. The column vectors $\mathbf{a}_i$, $i = 1, 2, \ldots, n$, of the transformation matrix $\mathbf{A}$ comprise the basis vectors of PCA or ICA, and the components $s_i(t)$ of the source vector $\mathbf{s}(t)$ are respectively principal or independent components corresponding to the data vector $\mathbf{x}(t)$. For simplicity, we assume that both the data vector $\mathbf{x}(t)$ and the source vector $\mathbf{s}(t)$ are zero mean n-vectors, and that the basis matrix $\mathbf{A}$ is a full-rank constant $n \times n$ matrix.

In PCA, the basis vectors $\mathbf{a}_i$ are required to be mutually orthogonal, and the coefficients $s_i(t)$ to have maximal variances (power) in the expansion (3.9) [2, 1]. While in ICA the basis vectors $\mathbf{a}_i$ are generally non-orthogonal, and the expansion (3.9) is determined under certain ambiguities from the strong but often meaningful condition that the coefficients $s_i(t)$ must be mutually statistically independent or as independent as possible [1].

Canonical correlation analysis (CCA) [2] is a generalization of PCA for two data sets whose data vectors are denoted by $\mathbf{x}$ and $\mathbf{y}$. CCA seeks for the linear combinations of the components of the vectors $\mathbf{x}$ and $\mathbf{y}$ which are maximally correlated. In this work, we have considered a similar expansion as (3.9) for both $\mathbf{x}$ and $\mathbf{y}$:

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \quad \mathbf{y} = \mathbf{B}\mathbf{t} \tag{3.10}$$

We then try to find in a similar manner as in ICA the maximally independent and dependent components from $\mathbf{x}$ and $\mathbf{y}$ by using higher-order statistics. As a result, we get an ICA style counterpart for canonical correlation analysis.

These ideas are introduced in [3], and discussed in more detail in the journal paper [4]. The methods introduced in these papers are somewhat heuristic, buut seem to work adequately both for artificially generated data and in a difficult cryptographic problem. We also consider in these papers practical measures for statistical dependence or independence of two random variables.

References

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

[2] A. Rencher, *Methods of Multivariate Analysis, 2nd ed.* Wiley, 2002.

[3] J. Karhunen and T. Ukkonen. Generalizing independent component analysis for two related data sets. In *Proc. of the IEEE 2006 Int. Conf. on Neural Networks / 2006 IEEE World Congress on Computational Intelligence (IJCNN2006/WCCI2006)*, Vancouver, Canada, July 2006, pp. 1822-1829.

[4] J. Karhunen and T. Ukkonen, Extending ICA for finding jointly dependent components from two related data sets. *Neurocomputing*, Vol. 70, Issues 16-18, October 2007, pp. 2969-2769.

3.5 ICA in CDMA communications

Karthikesh Raju, Tapani Ristaniemi, Juha Karhunen, 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].

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. We have also applied denoising source separation [5] 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.

All the results achieved in this project have been collected and presented in the monograph type doctoral thesis [6].

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

[6] K. Raju, *Blind Source Separation for Interference Cancellation in CDMA Systems.* PhD Thesis, Helsinki Univ. of Technology, 2006. Published as Report D16, Laboratory of Computer and Information Science.

3.6 Non-negative projections

Zhirong Yang, Jorma Laaksonen, Zhijian Yuan, Erkki Oja

Projecting high-dimensional input data into a lower-dimensional subspace is a fundamental research topic in signal processing, machine learning and pattern recognition. Non-negative projections are desirable in many real-world applications where the original data are non-negative, consisting for example of digital images or various spectra. It was pointed out by Lee and Seung [1] that the positivity or non-negativity of a linear expansion is a very powerful constraint, that seems to lead to sparse representations for the data. Their method, *non-negative matrix factorization (NMF)*, minimizes the difference between the data matrix $\mathbf{X}$ and its non-negative decomposition $\mathbf{WH}$. The difference can be measured by the Frobenius matrix norm or the Kullback-Leibler divergence.

Yuan and Oja [2] proposed the *projective non-negative matrix factorization (P-NMF)* method which replaces $\mathbf{H}$ in NMF with $\mathbf{W}^T\mathbf{X}$. This actually combines the objective of principal component analysis (PCA) with the non-negativity constraint. The P-NMF algorithm has been applied to facial image processing [4] using a popular database, FERET [3]. Figure (3.4) visualizes the basis images learned by NMF and P-NMF. The empirical results indicate that P-NMF is able to produce more spatially localized, part-based representations of visual patterns.

Another attractive feature of the NMF and P-NMF methods is that their multiplicative update rules do not involve human-specified parameters such as the learning rate. Thus the analysis results are completely data driven. In [5] we have studied how to construct multiplicative update rules for non-negative projections based on Oja's iterative learning rule. Our method integrates the multiplicative normalization factor into the original additive update rule as an additional term which generally has a roughly opposite direction. As a consequence, the modified additive learning rule can easily be converted to its multiplicative version, which maintains the non-negativity after each iteration. With this technique, almost identical results to P-NMF can be obtained by imposing the non-negativity constraint on linear Hebbian networks.

The derivation of our approach provides a sound interpretation of learning non-negative projection matrices based on iterative multiplicative updates—a kind of Hebbian learning with normalization. A convergence analysis is provided by interpreting the multiplicative

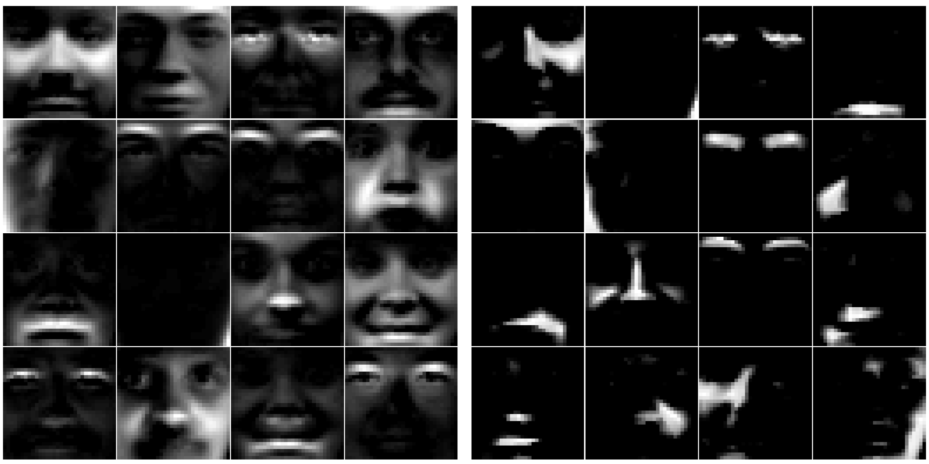

Figure 3.4: NMF (left) and P-NMF (right) bases of 16 dimensions.

updates as a special case of natural gradient learning. Furthermore, our non-negative variant of *linear discriminant analysis (LDA)* can serve as a feature selector. Its kernel extension can reveal an underlying factor in the data and be used as a sample selector.

References

[1] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[2] Zhijian Yuan and Erkki Oja. Projective nonnegative matrix factorization for image compression and feature extraction. In *Proc. of 14th Scandinavian Conference on Image Analysis (SCIA 2005)*, pages 333–342, Joensuu, Finland, June 2005.

[3] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The FERET evaluation methodology for face recognition algorithms. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22:1090–1104, October 2000.

[4] Zhirong Yang, Zhijian Yuan, and Jorma Laaksonen. Projective non-negative matrix factorization with applications to facial image processing. *International Journal on Pattern Recognition and Artificial Intelligence*, 21(8):1353–1362, December 2007.

[5] Zhirong Yang and Jorma Laaksonen. Multiplicative updates for non-negative projections. *Neurocomputing*, 71(1-3):363–373, 2007.

3.7 Climate data analysis with DSS

Alexander Ilin, Harri Valpola, Erkki Oja

An important task for which statistical methods are used in climate research is seeking physically meaningful interpretations of observed climate variability, for example, identification of 'modes' in the observational record. Statistical techniques which are widely used in this task include principal component analysis (PCA) or empirical orthogonal functions (EOFs), extended EOFs, and Hilbert EOFs [1]. Although EOFs have probably been the most popular tool for an efficient representation of climate records, EOF representation may be intuitively meaningless in a meteorological sense [2]. Therefore several techniques of rotated PCA/EOF have been proposed to ensure easier interpretation of the results. The rotation is realized using a linear transformation of principal components such that a suitably chosen criterion of "simple structure" is optimized. The objective is to find a data representation allowing for compact scientific explanation of a variable with a smaller number of principal components. Different assumptions on simplicity yield different rotation techniques.

We extend the concept of rotated PCA by introducing the concept of "interesting structure". In our case, the goal of exploratory analysis is to find signals with some specific structures of interest. They may for example manifest themselves mostly in specific variables, which exhibit prominent variability in a specific timescale etc. An example of such analysis can be extracting clear trends or quasi-oscillations from climate records. The procedure for obtaining suitable rotations of EOFs can be based on the general algorithmic structure of denoising source separation (DSS) [3].

In our initial studies, we tested the effectiveness of the proposed methodology to discover climate phenomena which are well-known in climatology, using very little information about their properties. One of the most prominent results is the extraction of the El Niño–Southern Oscillation phenomenon, using only a very generic assumption of its prominent variability in the interannual timescale (see Figs. 3.5-3.6) [4]. Other prominent signals found in this analysis might correspond to significant climate phenomena as well; for example, the second signal with prominent interannual variability somewhat resembles the derivative of the El Niño index (see Fig. 3.5).

Several other techniques for studying prominent climate variations have been introduced in our papers [4, 5]. Analysis which separates prominent quasi-oscillations in climate records by their frequency contents gives a meaningful representation of the slow climate variability as combination of trends, interannual oscillations, the annual cycle and slowly changing seasonal variations [4]. The technique presented in [5] can be used for studying slow variability present in fast weather fluctuations.

The results of the climate research were presented at the Fifth Conference on Artificial Intelligence Applications to Environmental Science as part of the 87th Annual Meeting of the American Meteorological Society (best student presentation) [6] and at the 10th International Meeting on Statistical Climatology.

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.

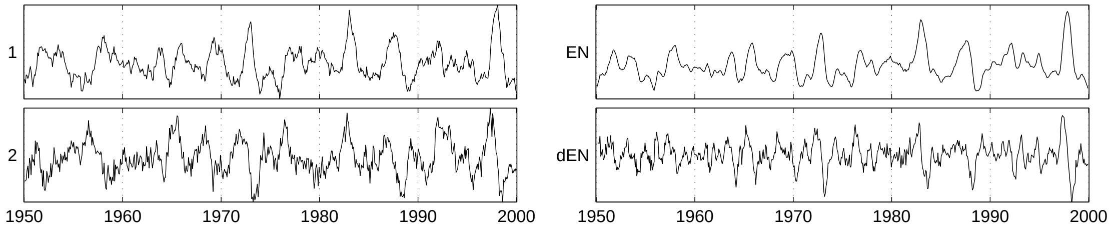

Figure 3.5: Left: The time courses of the two interannual phenomena found in global temperature, air pressure and precipitation data using DSS. Right: The index used in climatology to measure the strength of El Niño (marked as EN) and the derivative of the El Niño index (marked as dEN). The similarity is striking for the upper signals and some common features can be observed in the lower signals.

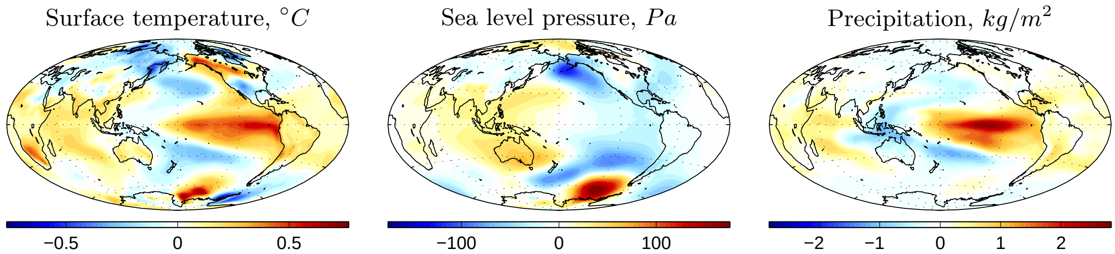

Figure 3.6: Spatial patterns corresponding to the most prominent interannual phenomenon found in climate data. The maps display the regions in which the effect of the phenomenon is most prominent. The maps contain many features traditionally associated with El Niño–Southern Oscillation phenomenon.

[3] J. Särelä and H. Valpola. Denoising source separation. *Journal of Machine Learning Research*, 6:233–272, 2005.

[4] A. Ilin, and H. Valpola, and E. Oja. Exploratory analysis of climate data using source separation methods. *Neural Networks*, Vol. 19, No. 2, pp. 155–167, March 2006.

[5] A. Ilin, and H. Valpola, and E. Oja. Extraction of components with structured variance. In *Proc. of the IEEE World Congress on Computational Intelligence (WCCI 2006)*, pp. 10528–10535, Vancouver, BC, Canada, July 2006.

[6] A. Ilin, and H. Valpola, and E. Oja. Finding interesting climate phenomena by exploratory statistical techniques. In *Proc. of the Fifth Conference on Artificial Intelligence Applications to Environmental Science as part of the 87th Annual Meeting of the American Meteorological Society*, San Antonio, TX, USA, January 2007. Best student presentation.