APPLICATIONS OF NEURAL BLIND SEPARATION TO SIGNAL AND IMAGE PROCESSING

Juha Karhunen, Aapo Hyvärinen, Ricardo Vigário, Jarmo Hurri and Erkki Oja

Helsinki University of Technology, Laboratory of Computer and Information Science
Rakentajanaukio 2 C, FIN-02150 Espoo, Finland
Email: Juha.Karhunen@hut.fi

ABSTRACT

In blind source separation one tries to separate statistically independent unknown source signals from their linear mixtures without knowing the mixing coefficients. Such techniques are currently studied actively both in statistical signal processing and unsupervised neural learning. In this paper, we apply neural blind separation techniques developed in our laboratory to extraction of features from natural images and to separation of medical EEG signals. The new analysis method yields features that describe the underlying data better than for example classical principal component analysis. We briefly discuss difficulties related with real-world applications of blind signal processing, too.

1. INTRODUCTION

Blind signal separation (BSS) and the closely related Independent Component Analysis (ICA) have been studied in statistical signal processing since 1980's. Most of the developed methods are batch type, though adaptive approaches have been considered, too. For references and reviews, see [1, 2, 3]. In these statistical approaches, separation is usually achieved by optimizing some constrast functions that are defined explicitly in terms of cumulants (higher-order moments) of the observed data.

On the other hand, in neural approaches to BSS and ICA, cumulants are typically replaced by suitable nonlinearities in the learning algorithms. The nonlinearities implicitly introduce the higher-order statistics which is necessary for blind separation. The first neural BSS algorithm, discussed in [4], was proposed by Jutten and Herault already in 1985. During the last couple of years, there has been a strong renewed interest in neural realizations of BSS and ICA. Several research groups have independently developed new, more efficient learning algorithms for separation problems. Neural approaches to BSS and ICA are reviewed in the recent tutorial paper [5]. Some new developments can be found in the special sessions arranged recently on this topic [6, 7].

Blind signal separation and ICA can be applied to a wide variety of problems. These include at least:

- Array signal processing; see for example [3].
- Separation of speech sources (Cocktail party problem) [8, 9]. More references can be found in [9].
- Several communications problems, such as multipath propagation in mobile communications, and separation of QAM sources [10, 11].
- Medical signal processing. Examples are electroencephalography (EEG) (separation of brain signals) [12, 13], and separation of ECG (heart) signals.
- Industrial problems, such as fault detection [3, 14].
- Extraction of meaningful features from data. Independent component analysis has been successfully applied at least to image [15, 16] and speech data [17].
- Generally, ICA can be applied to the same problems as standard Principal Component Analysis (PCA) [1]. If the internal representation of the data is not important (for example in technical data compression) or the data are Gaussian, it is easier to use PCA. But if higher-order statistics contain important information and the goal is meaningful representation of the data, ICA generally provides better results than PCA.

The above list is not complete (especially with respect to references) due to space limitations.

Of the neural based algorithms, the most widely applied in various forms are probably the seminal Herault-Jutten algorithm (for example [4, 8, 11]), Bell-Sejnowski algorithm [9, 12, 15, 17], and the fixed-point rules [13, 16] discussed later on in this paper.

2. NEURAL BLIND SOURCE SEPARATION

The basic data model employed both in blind source separation and independent component analysis is as follows [1, 5]. Assume that there exist m zero mean source signals (independent components) $s_1(t), \ldots, s_m(t)$ that are scalar-valued and mutually statistically independent (or as independent as possible) at each index value t. The original sources $s_i(t)$ are unknown, and all that we have are n possibly noisy but different linear mixtures $x_1(t), \ldots, x_n(t)$ of the sources. The mixing coefficients are some unknown constants. In blind source separation, the task is to find the waveforms $\{s_i(t)\}$ of the sources, knowing only the mixtures $x_j(t)$.

Denote by $\mathbf{x}(t) = [x_1(t), \ldots, x_n(t)]^T$ the n-dimensional t:th data (mixture) vector at discrete time (or point) t. The BSS signal model can then be written in the form

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

Here $\mathbf{s}(t) = [s_1(t), \ldots, s_m(t)]^T$ is the source (independent component) vector. $\mathbf{A}$ is a constant full-rank $n \times m$ mixing matrix whose elements are the unknown coefficients of the mixtures. The vectors $\mathbf{a}_i$, $i, \ldots, m$, are the basis vectors of ICA; see [1, 5].

The number of different mixtures n must be at least as large as the number of sources m (unless there is some extra information). Usually m is assumed known, and often $m = n$. Furthermore, each source signal $s_i(t)$ is assumed to be a stationary zero-mean stochastic process. Only one of the sources is allowed to have a Gaussian distribution.

In neural and adaptive BSS, an $m \times n$ separating matrix $\mathbf{B}(t)$ is updated so that the m-vector

$$\mathbf{y}(t) = \mathbf{B}(t)\mathbf{x}(t) \tag{2}$$

becomes an estimate $\mathbf{y}(t) = \hat{\mathbf{s}}(t)$ of the original independent source signals. In neural realizations, $\mathbf{y}(t)$ is the output vector of the network, and the matrix $\mathbf{B}(t)$ is the total weight matrix between the input and output layers. The estimate $\hat{s}_i(t)$ of the i:th source signal may appear in any component $y_j(t)$ of $\mathbf{y}(t)$. The amplitudes of the estimates $y_j(t)$ are typically scaled so that they have unit variance.

In several BSS algorithms, the data vectors $\mathbf{x}(t)$ are preprocessed by whitening (sphering) them: $\mathbf{v}(t) = \mathbf{V}(t)\mathbf{x}(t)$. Here $\mathbf{v}(t)$ denotes the t:th whitened vector, and $\mathbf{V}(t)$ is an $m \times n$ whitening matrix. After prewhitening the subsequent separating matrix, denoted here for clarity by $\mathbf{W}^T(t)$, can be taken orthogonal. The relationship between the whitening and output layers is $\mathbf{y}(t) = \mathbf{W}^T(t)\mathbf{v}(t)$, and the total separating matrix between input and output layers becomes $\mathbf{B}(t) = \mathbf{W}^T(t)\mathbf{V}(t)$.

Various neural algorithms for learning either the separating matrix $\mathbf{B}(t)$ or the matrix $\mathbf{W}(t)$ after prewhitening are reviewed in [5]. Some algorithms are based on heuristic independence conditions, while others try to optimize an information-theoretic criterion or a simpler contrast function leading to independent outputs. We have used the kurtoses $\mathrm{E}\{y_i(t)^4\} - 3[\mathrm{E}\{y_i(t)^2\}]^2$ of the components $y_i(t)$ of the vector $\mathbf{y}(t)$ as a separating criterion, because this approach leads to simple learning algorithms. It can be shown that the source signals or independent components are found from the local maxima of the modulus of the kurtosis for prewhitened data. A generalization of this approach can be found in [22].

3. SEPARATION ALGORITHMS

We have earlier derived simple stochastic gradient type algorithms [5, 18, 19, 21] for minimizing or maximizing the kurtosis criterion. These truly neural algorithms employ nonlinear Hebbian learning, but due to the crude instantaneous estimate of the gradient their accuracy is limited, and convergence speed may be slow. Therefore, it may be difficult to apply them to separation problems where there are more than about ten sources. Another problem is that in certain applications some of the source signals are sub-Gaussian (having a negative kurtosis) while others are super-Gaussian (with positive kurtosis), but the gradient algorithms are directly applicable to either type of sources only. Quite recently, we have developed a new recursive least-squares type neural or adaptive learning algorithm [20] which is more accurate and converges clearly faster, but it has not yet been applied to large-scale practical problems.

However, we have recently introduced fixed-point algorithms [21, 22] which are simple to implement, accurate, and converge fast to the local maxima of the kurtosis criterion. They can be applied to both sub-Gaussian and super-Gaussian sources simultaneously. Therefore, these algorithms are very useful in practical applications. A drawback is that the fixed-point algorithms are not strictly neural and data-adaptive. However, they originate from our earlier neural gradient algorithms, and could be replaced by them in some situations at least.

In the basic generalized fixed-point algorithm [22] the data vectors $\mathbf{x}(t)$ are first whitened using for example standard PCA [5]. Random vectors normalized to unit length are chosen to the initial values of the rows $\mathbf{w}_i$ $(i = 1, \ldots, m)$ of the orthogonal $m \times m$ separating matrix $\mathbf{W}^T$. The key step in the generalized fixed-point algorithm is to compute a new $(k+1)$:th estimate for $\mathbf{w}_i$ using the iteration

$$\mathbf{w}_i^*(k+1) = \mathrm{E}\{\mathbf{v}g(\mathbf{w}_i(k)^T\mathbf{v}) - g'(\mathbf{w}_i(k)^T\mathbf{v})\mathbf{w}_i(k)\}, \quad (3)$$

$$\mathbf{w}_i(k+1) = \mathbf{w}_i^*(k+1) / \parallel \mathbf{w}_i^*(k+1) \parallel . \quad (4)$$

Here E denotes the mathematical expectation. In practice it is replaced by sample mean computed using a large number of whitened vectors $\mathbf{v}(t)$. The function $g(u)$ can be chosen any odd, sufficiently regular nonlinear function, and $g'(u)$ denotes its derivative. The choice $g(u) = u^3$ directly maximizes the kurtosis criterion. In practice, it is often advisable to use a robust nonlinearity that grows less than linearly; a typical choice is $g(u) = \tanh(u)$. This also has a relationship to the kurtosis criterion [18]. For preventing the vectors $\mathbf{w}_i$, $i = 1, \ldots, m$, from converging to the same directions, they are orthogonalized against each other. This can be done either symmetrically or sequentially using a deflation type procedure [21, 22].

It can be proven [22] that $\mathbf{w}_i(k)$ converges (up to the sign) to one of the rows of the separating matrix $\mathbf{W}^T$ under very mild conditions. The convergence of the fixed-point algorithms is cubic, and our experiments show that usually less than 10 iterations provide sufficiently accurate estimates. This means that the fixed-point algorithms are very fast compared to typical gradient-based adaptive blind separation algorithms. Another advantage is that they don't require any learning parameters. They are also much simpler than the currently best known batch algorithm introduced in [1]. In [22], versions that need not prewhitening have been introduced.

4. SEPARATION OF EEG SOURCES

In electroencephalography (EEG), a practical problem is to extract the meaningful brain activity information from measured signals distorted by various artifacts. Typical artifacts consist of eye movements, muscle activity, and mechanical displacements in the measuring apparatus. Traditional methods of artifact canceling are usually based on discarding the portions of EEG measurements that contain high amounts of these disturbances.

It seems that the model (1) describes well some important aspects of the EEG measurement data. The source signals $s_i(t)$ can be divided into roughly mutually independent brain activity signals and artifacts. In [13], we have applied blind source separation (independent component analysis) to practical EEG data. The 23-dimensional data vectors $\mathbf{x}(t)$ are not shown in this paper due to space limitations; see [13]. They were whitened and used in the standard fixed-point algorithm. This algorithm is actually a special case of the generalized algorithm (3) for the cubic nonlinearity $g(u) = u^3$. However, (3) becomes even simpler in this case because the derivative $\mathrm{E}\{g'(\mathbf{w}^T\mathbf{v})\} = 3\mathrm{E}\{(\mathbf{w}^T\mathbf{v})^2\} = 3 \parallel \mathbf{w} \parallel^2 = 3$.

The first 8 source signals $s_1(t), \ldots, s_8(t)$ (independent components ICA1,...,ICA8) found after learning are shown in Fig. 1. The first source signal $s_1(t)$ (ICA1) isolates eye blinking, a mechanical disturbance appears in ICA3, and the electronic adaptation triggered by it in ICA5. The signals ICA2 and ICA8 explain the rest of the eye activity. The utility of ICA/BSS in EEG is very clearly demonstrated in the 4th component ICA4, where a signal not visible in the EEG is separated. A possible interpretation of this source signal is a k-complex associated to initialization of sleeping.

The results are very promising, allowing to isolate typical artifacts from EEG monitoring while keeping the rest of the brain activities untouched. They open up some new vistas for EEG research. Makeig et al. have earlier applied Bell's and Sejnowski's algorithm to EEG data in [12]. However, the features found in [13] seem to be physiologically more significant. Possible explanations are that the fixed-point algorithm is more accurate, and that the EEG sources can be either sub-Gaussian or super-Gaussian.

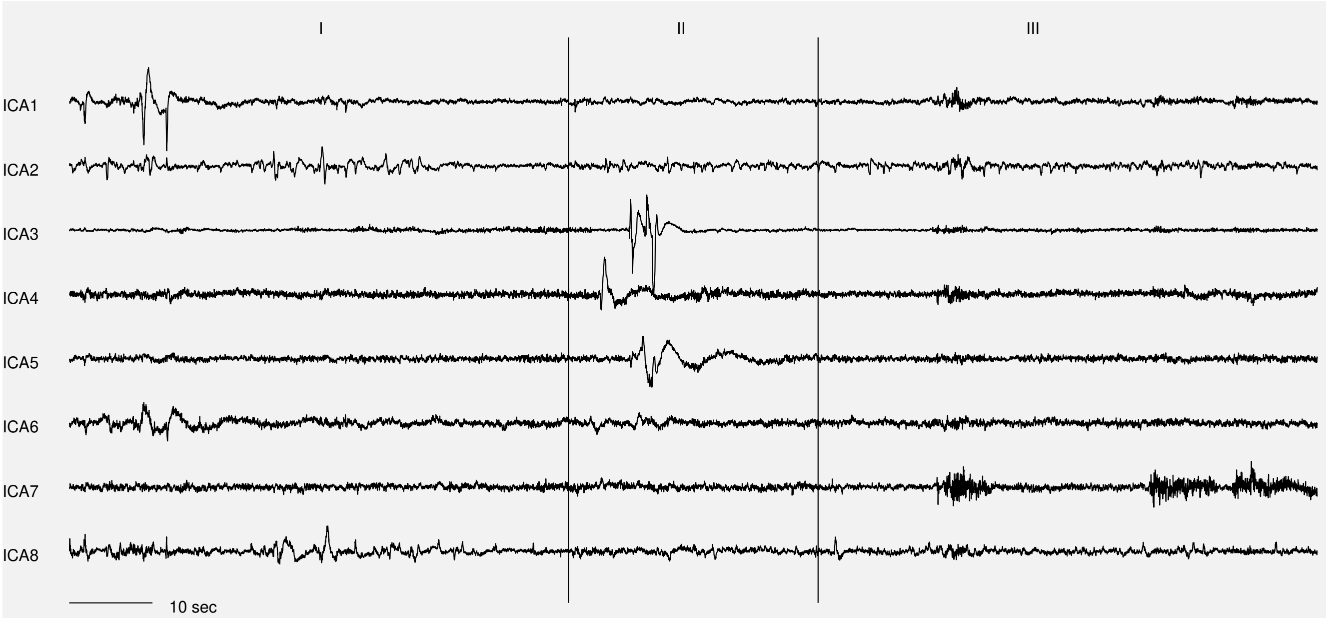

Figure 1: Eight first independent components of the EEG data.

5. IMAGE FEATURE EXTRACTION

As a second application example, consider unsupervised extraction of image features using ICA. This is described in more detail in [16].

The raw data consisted of 15 different images representing various natural objects or scenes. From these images, we randomly picked 10000 12×12 subimages. Each subimage was then separately made zero-mean and normalized to unit length. The obtained 144-dimensional data vectors $\mathbf{x}(t)$ were whitened using standard PCA. The PCA whitening matrix is $\mathbf{V} = \mathbf{D}^{-1/2}\mathbf{E}^T$, where the columns of the matrix $\mathbf{E}$ contain the PCA eigenvectors, and the diagonal matrix $\mathbf{D}$ the respective eigenvalues [5]. The generalized fixed-point algorithm (3) was then applied to the whitened vectors $\mathbf{v}(t)$ using the sigmoidal nonlinearity $g(u) = \tanh(u)$. From the learned vectors $\mathbf{w}_i$, one can compute the estimates of the corresponding basis vectors $\mathbf{a}_i$ of ICA using the formula [5]

$$\hat{\mathbf{a}}_i = \mathbf{E}\mathbf{D}^{1/2}\mathbf{w}_i. \tag{5}$$

Figure 2 shows typical examples of the estimated basis $\hat{\mathbf{a}}_i$ vectors of ICA, represented again as 12×12 subimages scaled suitably. Most of the ICA basis vectors correspond to wavelet type filters that are sensitive to local features and spatial frequencies in the images. However, a part of the estimated basis vectors yield filters that are sensitive to edges and lines of varying thickness in different orientations.

Image feature extraction using ICA has been independently considered in [15] using different algorithms and preprocessing. The results are, however, qualitatively fairly similar than in Fig. 2. It is noteworthy that ICA is a new image analysis method which extracts meaningful features from the input images in a completely unsupervised manner. Contrary to fixed transforms and filter masks, the results depend on the image data. The masks provided by ICA should be in many tasks more useful than the PCA masks which are mainly sensitive to spatial frequencies only.

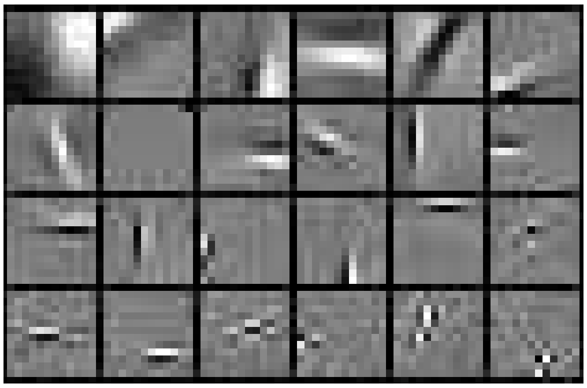

Figure 2: Some ICA basis vectors of natural image data.

6. PROBLEMS AND PROSPECTS

Until recently, blind source separation and independent component analysis have been applied to small dimensional problems where there are a few source signals only. The development of new, computationally efficient algorithms has enabled applications to larger scale problems. Two such applications, showing promising results and the potential of these techniques, have been described in this paper.

In developing practical real-world applications, the basic model (1) is often too simple or the assumptions made on it are not realistic. The model (1) can be extended and/or modified in several ways. We just mention here some possibilities. Some more information can be found in [5, 6, 7].

- It is usually not possible to separate noise from the source signals $s_i(t)$ unless there is some prior information on noise available. If there are more mixtures than sources ($n > m$), the amount of noise can be suppressed using PCA prewhitening [5, 23].

- Either the sources or the mixture coefficient or both can be nonstationary. Adaptive algorithms can be used in principle, but blind separation becomes even more difficult than normally.
- The source signals may have different time delays in each mixture. This problem has been studied by many authors, see e.g. [3, 8, 9].
- The number m of sources is often unknown. The situation where there are more sources than mixtures ($m > n$) is especially problematic. It can be sometimes handled if there is some prior information constraining the form of sources (for example the sources are binary).
- Some work for extending the linear data model to nonlinear ICA and source separation has been done; see [5, 6, 7, 14]. However, there are computational problems, and some additional constraints must be imposed for making the problem tractable. Recently, nonlinear ICA has been applied to detection of motor faults [14].
- In practice, there is often some prior information available, but the number of unknown parameters in (1) is too large for applying classical parameter estimation methods. This prior information should be utilized for getting optimal separation results. Some cases have been discussed in [23].
- The source signals may not be statistically independent. However, some algorithms are able to approximately separate moderately correlated sources in practice.

A lot of work is still required for developing satisfactory BSS and ICA methods for these and other extensions of the basic data model (1). Another important topic is to develop more accurate and faster converging truly neural or adaptive learning algorithms for large-scale problems.

7. REFERENCES

[1] P. Comon, "Independent component analysis - a new concept?," *Signal Processing*, vol. 36, pp. 287-314, 1994.

[2] J.-F. Cardoso and P. Comon, "Independent component analysis, a survey of some algebraic methods," in *Proc. IEEE Int. Symp. on Circuits and Systems*, Atlanta, GA, May 1996, vol. 2, pp. 93-96.

[3] R.-W. Liu, "Blind signal processing: an introduction," in *Proc. IEEE Int. Symp. on Circuits and Systems*, Atlanta, GA, May 1996, vol. 2, pp. 81-84.

[4] C. Jutten and J. Herault, "Blind separation of sources, part I: an adaptive algorithm based on neuromimetic architecture," *Signal Processing*, vol. 24, no. 1, pp. 1-10, July 1991.

[5] J. Karhunen, "Neural approaches to independent component analysis and source separation," in *Proc. 4th European Symp. on Artificial Neural Networks*, Bruges, Belgium, April 1996, pp. 249-266.

[6] Special session on "Blind signal processing - adaptive and neural network approaches," in *Proc. 1996 Int. Conf. on Neural Information Processing*, Hong Kong, September 1996, pp. 1187-1239.

[7] A. Cichocki and A. Back (Eds.), *NIPS'96 Postconference Workshop on Blind Signal Processing and Their Applications,* Snowmass, Colorado, December 1996.

[8] H.-L. Nguyen Thi and C. Jutten, "Blind source separation for convolutive mixtures," *Signal Processing*, vol. 45, pp. 209-229, 1995.

[9] K. Torkkola, "Blind separation of convolved sources based on information maximization," in *Proc. IEEE Workshop on Neural Networks for Signal Processing*, Kyoto, Japan, September 1996, pp. 423-432.

[10] B. Laheld and J.-F. Cardoso, "Adaptive source separation with uniform performance," in *Signal Processing VII: Theories and Applications*, M. Holt et al. (Eds.). Lausanne: EURASIP, 1994, vol. 2, pp. 183-186.

[11] Y. Deville and L. Andry, "Application of blind source separation techniques to multi-tag contactless identification systems," in *Proc. 1995 Int. Symp. on Nonlinear Theory and Applications*, Las Vegas, USA, December 1995, vol. 1, pp. 73-78.

[12] S. Makeig, A. Bell, T.-P. Jung, and T. Sejnowski, "Independent component analysis of electroencephalographic data," in *Advances in Neural Information Processing Systems 8,* D. Touretzky et al. (Eds.). Cambridge, MA: MIT Press, 1996, pp. 145-151.

[13] R. Vigário, A. Hyvärinen, and E. Oja, "ICA fixed-point algorithm in extraction of artifacts from EEG," in *Proc. 1996 IEEE Nordic Signal Processing Symp.*, Espoo, Finland, September 1996, pp. 383-386.

[14] L. Parra, G. Deco, and S. Miesbach, "Statistical independence and novelty detection with information preserving nonlinear maps," *Neural Computation*, vol. 8, 1996, pp. 260-269.

[15] A. Bell and T. Sejnowski, "Edges are the independent components of natural scenes," to appear in *Advances in Neural Information Processing Systems 9* (Proc. NIPS'96, Denver, Colorado, December 1996).

[16] J. Hurri, A. Hyvärinen, J. Karhunen, and E. Oja, "Image feature extraction using independent component analysis," in *Proc. 1996 IEEE Nordic Signal Processing Symp.*, Espoo, Finland, September 1996, pp. 475-478.

[17] A. Bell and T. Sejnowski, "Learning the higher-order structure of natural sound," 1996, to appear in *Network.*

[18] J. Karhunen, L. Wang, and R. Vigário, "Nonlinear PCA type approaches for source separation and independent component analysis," in *Proc. 1995 IEEE Int. Conf. on Neural Networks*, Perth, Australia, November 1995, pp. 995-1000.

[19] E. Oja, J. Karhunen, and A. Hyvärinen, "From neural PCA to neural ICA", in *NIPS'96 Postconference Workshop on Blind Signal Processing and Their Applications*, Snowmass, Colorado, December 1996.

[20] J. Karhunen and P. Pajunen, "Blind source separation using least-squares type adaptive algorithms," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Munich, Germany, April 21-24, 1997.

[21] A. Hyvärinen and E. Oja, "One-unit learning rules for independent component analysis," to appear in *Advances in Neural Information Processing Systems 9* (Proc. NIPS'96, Denver, Colorado, December 1996).

[22] A. Hyvärinen, "A family of fixed-point algorithms for independent component analysis," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Munich, Germany, April 21-24, 1997.

[23] C. Jutten and J.-F. Cardoso, "Separation of sources: really blind?," in *Proc. 1995 Int. Symp. on Nonlinear Theory and Applications*, Las Vegas, USA, December 1995, vol. 1, pp. 79-84.