31 Blind Signal Separation and Independent Component Analysis

Erkki Oja, Juha Karhunen, Jyrki Joutsensalo, Aapo Hyvärinen, Petteri Pajunen, and Ricardo Vigário

Both Principal Component Analysis (PCA) and Independent Component Analysis (ICA) [2] attempt to find a coordinate transformation of a collection of multivariate data, by which the new coordinates or feature dimensions have some desirable properties in terms of data compression and representation. In the case of classical PCA, the new coordinates are uncorrelated and an optimal linear compression is achieved in the minimum mean square sense. In the case of ICA, the new coordinates are statistically independent, which means that a very efficient data representation is possible.

An especially interesting connection of ICA exists to the problem of Blind Signal Separation [1,3]. A mathematical definition is the following: an L-dimensional vector-valued discrete signal $\mathbf{x}_k = [x_k(1), \dots, x_k(L)]^T$ at the discrete time k is assumed to be of the form

$$\mathbf{x}_k = \mathbf{A}\mathbf{s}_k + \mathbf{n}_k = \sum_{i=1}^{M} s_k(i)\mathbf{a}(i) + \mathbf{n}_k. \tag{84}$$

Here $\mathbf{s}_k = [s_k(1), \dots, s_k(M)]^T$ is a source vector consisting of M *unknown* source signals (independent components) $s_k(i)$ $(i = 1, \dots, M)$ at time k. $\mathbf{A} = [\mathbf{a}(1), \dots, \mathbf{a}(M)]$ is a fixed $L \times M$ *unknown* mixing matrix whose columns $\mathbf{a}(i)$ are the basis vectors of ICA, and $\mathbf{n}_k$ denotes possible corrupting additive noise. The noise term $\mathbf{n}_k$ is often omitted from (84), because it is usually impossible to distinguish it from the source signals. Instead of time, k can also stand for the spatial location of a pixel, like in the example of Figs. 47, 48.

The problem is to find the mixing matrix $\mathbf{A}$, when only a sample $\mathbf{x}_k, k = 1, 2, ...$ of the mixtures is available.

The following assumptions are typically made [1]:

1. $\mathbf{A}$ is a constant matrix with full column rank. Thus the number of mixtures L is at least as large as the number of sources M, which is usually assumed to be known in advance. If $M < L$, the data vectors $\mathbf{x}_k$ roughly lie in the M-dimensional subspace spanned by the basis vectors of ICA.

2. The source signals $s_k(i)$ $(i = 1, \dots, M)$ must be *mutually statistically independent* at each time instant k, or as independent as possible. The degree of independence can be measured using suitable constrast functions.

3. Each source signal $s_k(i)$ is a stationary zero-mean stochastic process. Only one of the source signals $s_k(i)$ is allowed to have a Gaussian marginal distribution.

Note that very little prior information is available for the matrix $\mathbf{A}$. Therefore, the strong independence assumptions are required to fix the ICA expansion (84). Even then, only the directions of the ICA basis vectors $\mathbf{a}(i)$, $i = 1, \dots, M$, are defined.

To get a more unique solution, one can normalize the variances of the source signals to unity.
In the techique called *blind source (or signal) separation*, one tries to extract the unknown waveforms $\{s_k(i)\}$, $k = 1, \ldots,$ of the independent source signals in (84) from the data vectors $\mathbf{x}_k$ by a linear transformation

$$\mathbf{y}_k = \mathbf{B}\mathbf{x}_k, \tag{85}$$

where $\mathbf{B}$ is called a separating matrix. The elements of $\mathbf{y}_k$ approximate the source signals $s_k(i)$. Such blind techniques are useful for example in array processing, speech enhancement, and communications. A typical example is the "cocktail party effect": suppose we can record the mixed voices from a party by several microphones. The blind source separation would give the voices of the individual speakers.
In several blind separation algorithms, the data vectors $\mathbf{x}_k$ are preprocessed by whitening (sphering) them, so that their covariance matrix becomes the unit matrix. After whitening, the separating matrix $\mathbf{B}$ can be assumed orthogonal. This auxiliary constraint makes the separating algorithms simpler, and also normalizes the variances of the estimated sources automatically to unity.
A practical difficulty in designing source separation and ICA algorithms is reliable verification of the independence condition. It is impossible to do this directly because the involved probability densities are unknown. Therefore, approximating contrast functions which are maximized by separating matrices have been introduced [2]. As an example, for prewhitened input vectors it can be shown that the relatively simple contrast function based on the fourth-order cumulant or *kurtosis*

$$J_1(\mathbf{y}) = \sum_{i=1}^{M} \mid \mathrm{cum}[y(i)^4] \mid = \sum_{i=1}^{M} \mid [\mathrm{E}\{y(i)^4\} - 3\mathrm{E}^2\{y(i)^2\} \mid \tag{86}$$

is maximized by the separating matrix $\mathbf{B}$ in model (85), if the sign of the (unnormalized) kurtosis $\mathrm{cum}[s(i)^4]$ is the same for all the source signals $s_k(i)$, $i = 1, \ldots, M$.
A 3-layer feedforward network was proposed in [4] for ICA and blind source separation. Each of the 3 layers performs one of the processing tasks required for complete ICA: 1. whitening; 2. separation; and 3. estimation of the mixing matrix. Any of these three tasks can be performed either neurally or conventionally.
For whitening, simplified versions of neural PCA learning rules are convenient. For separation, we can use the *nonlinear PCA rule* [5]:

$$\mathbf{W}_{k+1} = \mathbf{W}_k + \mu_k[\mathbf{x}_k - \mathbf{W}_k\mathbf{g}(\mathbf{y}_k)]\mathbf{g}(\mathbf{y}_k^T). \tag{87}$$

with μ_k the learning rate, $\mathbf{x}_k$ the mixture vectors that are now assumed whitened, and $\mathbf{y}_k = \mathbf{W}_k^T\mathbf{x}_k$ the output vector from a neural layer whose weights are given by matrix $\mathbf{W}_k$. The function $g(.)$ is a suitable nonlinearity, e.g. the hyperbolic tangent function. During learning, the weight matrix $\mathbf{W}_k$ converges to a (transposed) separating matrix [5], and the elements of $\mathbf{y}_k$, or the outputs from the neural layer, tend to the source signals.
A connection of nonlinear PCA to some other statistical and information theoretic criteria, as well as the learning rules, are discussed in another Section of this report.
In 1995, we also developed another so-called *bigradient algorithm* [6], which is applied for learning the orthonormal separating matrix $\mathbf{B}$ as follows:

$$\mathbf{W}_{k+1} = \mathbf{W}_k + \mu_k\mathbf{x}_k\mathbf{g}(\mathbf{y}_k^T) + \gamma_k\mathbf{W}_k(\mathbf{I} - \mathbf{W}_k^T\mathbf{W}_k). \tag{88}$$

Here γ_k is another gain parameter. usually about 0.5 or 1 in practice. Again, the weight matrix $\mathbf{W}_k^T$ tends to the separating matrix $\mathbf{B}$.
Since 1996, new algorithmic development into the ICA and BSS problem has concentrated on the fixed-point learning rules, implemented in the FastICA software package (see the section on one-unit and fixed point ICA algorithms). Also several extensions have been studied recently, like nonlinear mixing models, robust algorithms, and relations with complexity criteria - see the separate section on extensions.
Our ICA / BSS ideas have been applied to a number of artificial and real signals, e.g. to separate 10 speech signals from their mixtures. As an illustrative example, Fig. 47 shows 9 mixtures of 9 natural images. This means that the 9 original images (not shown) have been multiplied pixel-wise by randomly chosen coefficients and added together, to obtain one of the mixtures shown here. Different multiplying coefficients have been used for the 9 different mixtures. The 9-dimensional mixture vectors $\mathbf{x}_k$ in eq. (84) are obtained by collecting the gray levels of pixels in the 9 mixture images at the same pixel location. Thus k is a running index for the pixel location. In this experiment, there was no additive noise in the mixtures. These mixture vectors where whitened by PCA and input to the nonlinear PCA learning rule, eq. (87). The outputs after learning, again collected into images, are shown in Fig. 48. These are quite close to the original images used in forming the mixtures. Note that no information whatsoever was used on the mixing coefficients (elements of matrix $\mathbf{A}$) or the original images in computing these results. The only information the algorithm had were the mixtures of Fig. 47.
More conrete applications are in biomedical signal analysis, financial time series analysis, and feature extraction for digital images. All of these are covered in their separate Sections in this report.

References

[1] J.-F. Cardoso and B. Laheld, "Equivariant adaptive source separation," *IEEE Trans. on Signal Processing*, vol. 44, pp. 3017 - 3030, 1996.

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

[3] 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, 1991.

[4] J. Karhunen, E. Oja, L. Wang, R. Vigario, and J. Joutsensalo, " A class of neural networks for Independent Component Analysis," *IEEE Trans. on Neural Networks 8*, pp. 486 - 504 (1997).

[5] Oja, E., "The nonlinear PCA learning rule in Independent Component Analysis," *Neurocomputing 17*, pp. 25 - 45 (1997).

[6] L. Wang, J. Karhunen, and E. Oja, "A bigradient optimization approach for robust PCA, MCA, and source separation," *Proc. 1995 IEEE Int. Conf. on Neural Networks*, Perth, Australia, November 1995, pp. 1684 - 1689.

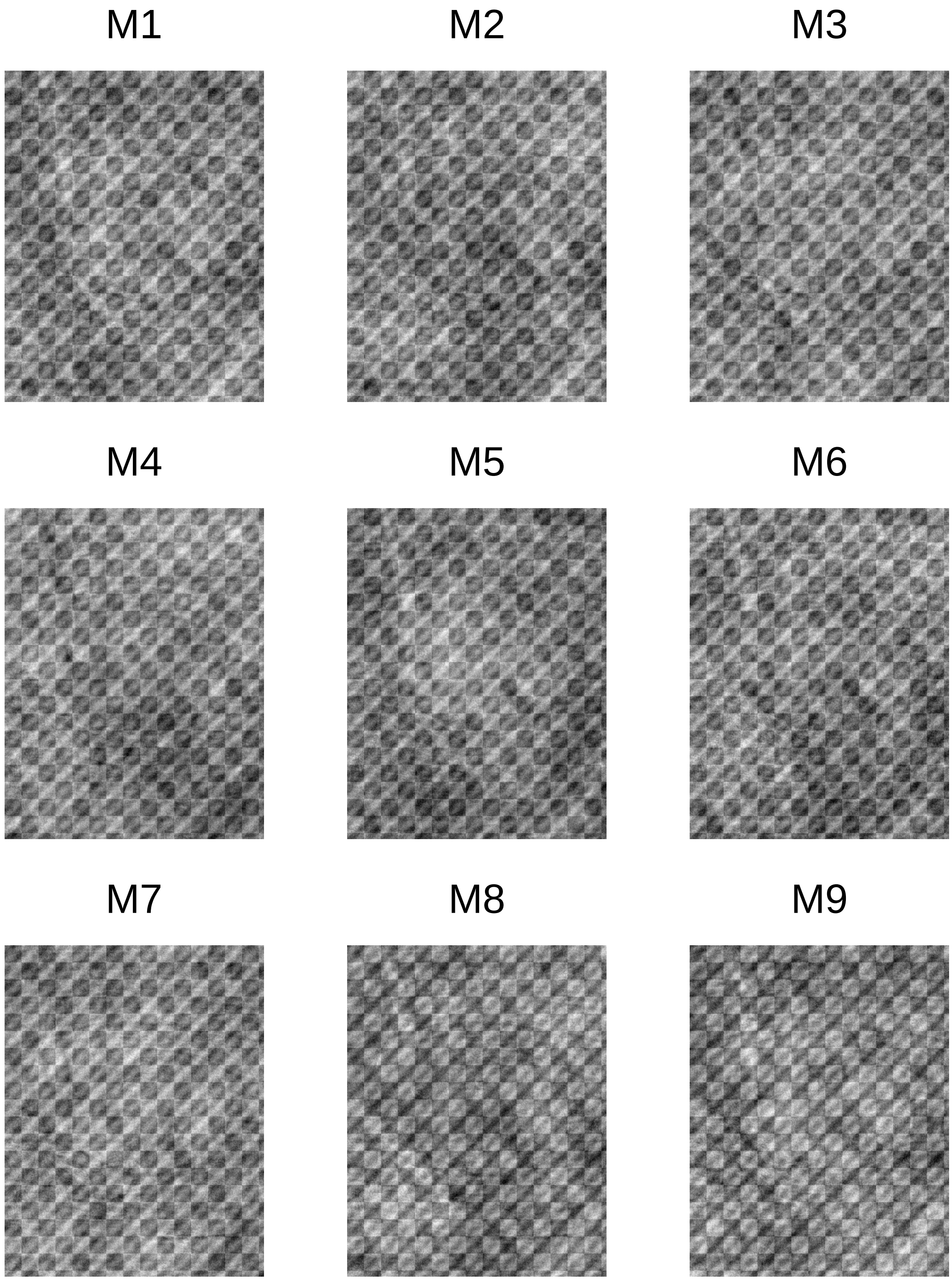

Figure 47: Mixtures of 9 natural images.

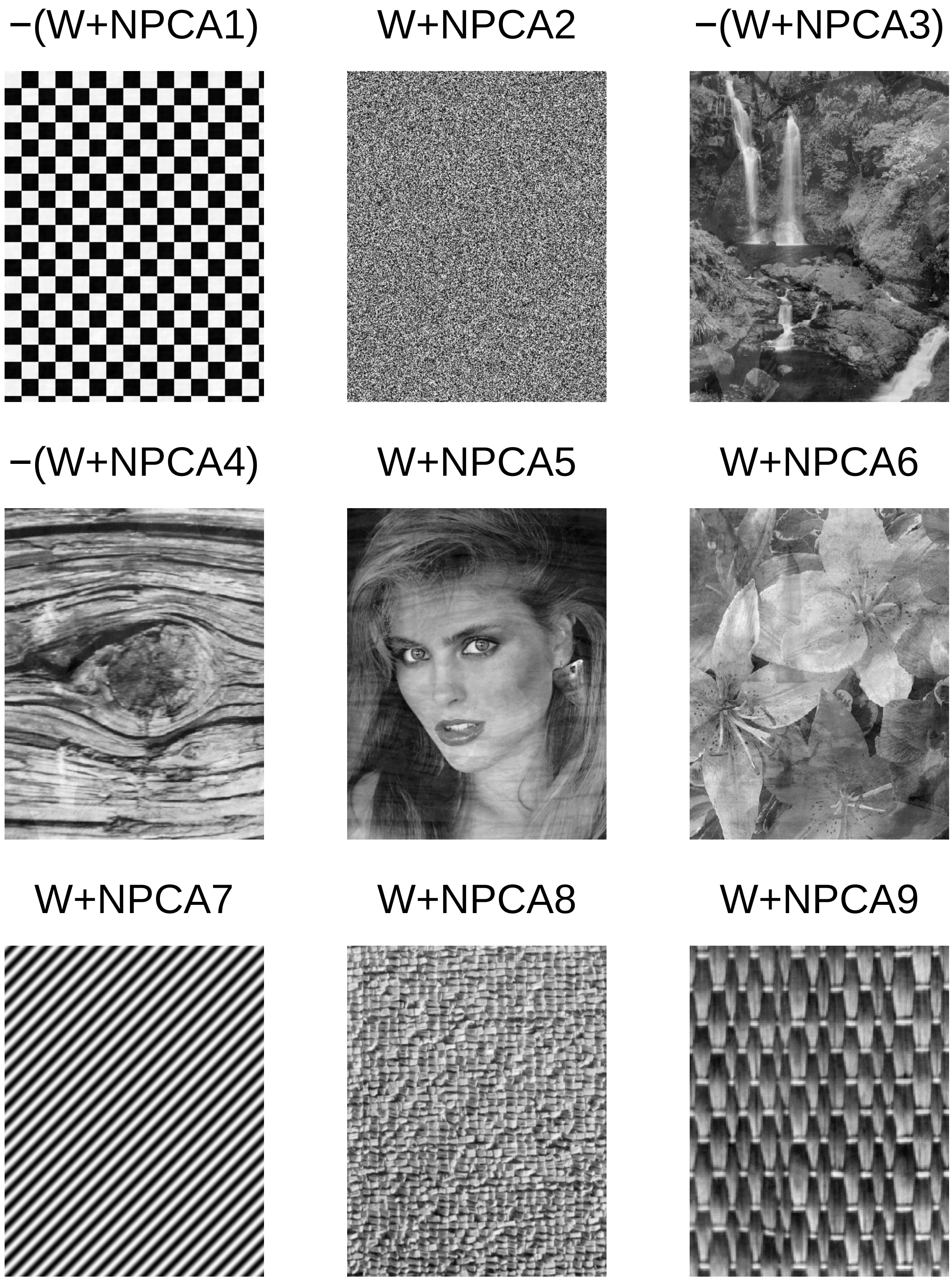

Figure 48: Separated images found by the 3-layer ICA network, using whitening and the nonlinear PCA algorithm.