# DENOISING SOURCE SEPARATION: A NOVEL APPROACH TO ICA AND FEATURE EXTRACTION USING DENOISING AND HEBBIAN LEARNING

Jaakko Särelä
Helsinki University of Technology
Neural Network Research Centre
P.O. Box 5400, FIN-02015 HUT, Finland
Jaakko.Sarela@hut.fi

## 1. INTRODUCTION

The traditional principal component analysis (PCA) has often been suggested for data analysis and feature extraction. PCA can be realised by simple Hebbian learning, e.g., by Oja's rule [5]. PCA explains the largest amount of variance of the data in each of its components. For this reason it tends to extract global features and is not very suitable for representing structures in the data. The structures would be better represented by sparse features (see, e.g., [7] for sparse coding of natural images).

Independent component analysis (ICA) [2] provides a better tool for feature extraction. While in PCA, only uncorrelation is guaranteed between the sources, ICA renders the sources completely independent. The features provided by ICA tend to be sparse. ICA has been frequently applied to computational neuroscience and modelling simple and complex cells in human primary visual cortex (V1) [4].

Many algorithms have proposed for ICA. For our purpose the most interesting one stem from the idea of nonlinear PCA [6]: combination of a batch version of the nonlinear PCA and presphering of the data, lead to a very fast ICA algorithm, FastICA [3]. However, not much has been said about how to choose the nonlinearity.

In this presentation, we discuss a new interpretation to the nonlinearity as denoising. We call this new framework denoising source separation (DSS) [8]. Prior knowledge can be easily incorporated in this denoising.

Denoising corresponds to procedural knowledge while in most approaches to source separation, the algorithms are derived from explicit objective functions or generative models. This corresponds to declarative knowledge. Algorithms are procedural, however. Thus declarative knowledge has to be translated into procedural form, which may result in complex and computationally demanding algorithms. Furthermore, we argue that it is easier to suggest biologically plausible procedural algorithms than to suggest such objective functions or generative models. Some interesting extensions that make DSS more relevant models of human information processing are discussed in the end.

## 2. DENOISING SOURCE SEPARATION

Assume the data $\mathbf{X}$ presphered. Then the DSS algorithm is as follows:

$$\mathbf{s} = \mathbf{w}^T\mathbf{X} \tag{1}$$

$$\mathbf{s}^+ = \mathbf{f}(\mathbf{s}) \tag{2}$$

$$\mathbf{w}^+ = \mathbf{X}\mathbf{s}^{+T} \tag{3}$$

$$\mathbf{w}_{\text{new}} = \frac{\mathbf{w}^+}{||\mathbf{w}^+||}. \tag{4}$$

The first step (1) calculates the current estimate of one source. The second step (2) is the denoising step and it should reflect the structure of a source. The step (3) re-estimates the mixing vector to fit the denoised source estimate. Finally, normalisation (4) stabilises the norm of the mixing vector.

We have shown that with linear denoising, i.e., matrix multiplication $\mathbf{s}^+ = \mathbf{f}(\mathbf{s}) = \mathbf{s}\mathbf{D}$, DSS is equivalent to the classical power method that is performed for a denoised (presphered) data $\mathbf{Z} = \mathbf{X}\mathbf{D}^*$ where $\mathbf{D} = \mathbf{D}^*\mathbf{D}^{*T}$.

Consider such a separation scheme using linear denoising through an example: two sources, shown in Fig. 1a, are mixed together (scatter-plot shown in Fig. 1b, and presphered data in Fig. 1c). Note that though the presphering renders the orthogonal projections uncorrelated, it does not identify the sources, but some mixing still exists.

The first source seem to vary more slowly than the other one. Hence, a reasonable denoising scheme would be to low-pass filter the data. The result is shown in Fig. 1d. The maximal variance direction now identifies the slower source. Actually, any denoising would have sufficed that makes the eigenvalues of the two sources different.

In the following table, we list some used denoising functions and the function of the corresponding DSS algorithm. More denoising functions and methods for speeding up the convergence can be found in the DSS papers [8, 10].

| $f(\mathbf{s})$ | algorithm |
|---|---|
| LTI-filtering | temporal ICA |
| $s^3$ | kurtosis-based ICA |
| $\mathbf{s} - \tanh(\mathbf{s})$ | robust ICA |

## 3. BIOLOGICALLY PLAUSIBLE DSS

The DSS algorithm given above is a batch algorithm. Simply, by changing the denoising scheme to be real-time, DSS becomes an online feature extractor. This is useful in adaptation to changing environments.

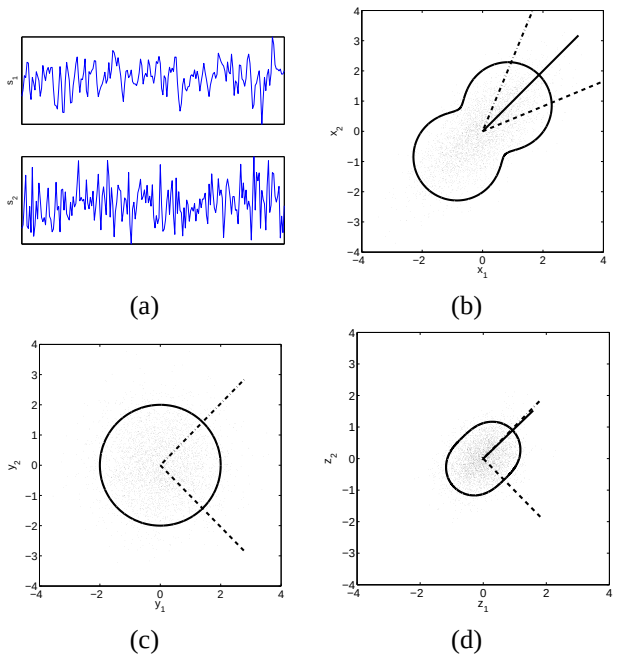


**Fig. 1**. *(a) Original source. (b) Scatter-plot of the mixtures. (c) Sphered data. (d) Data denoised using lp-filtering* $\mathbf{Z} = \mathbf{X}\mathbf{D}^*$. *The dashed lines depict the mixing vectors and the solid lines the largest eigenvector. The curves denote the standard deviation of the projection of the data.*

In this abstract, we only considered orthogonal projections from the data. However, nonorthogonal projections resulting from over-complete representations provide some clear advantages, especially in sparse codes [1], and may be found useful in the DSS framework as well.

Linear features provide only a limited representation of data. Even building of hierarchies becomes futile, because hierarchies of linear models stay linear. In DSS, nonlinear hierarchical networks may be achieved by outputting the denoised source estimate (2) to the next level of hierarchy, instead of the linear source estimate (1).

Many neuroscientists have proposed that feedback connections are crucial for contextual and attentive mechanisms. In DSS such connections, temporal, top-down or lateral, can be used for denoising and hence for feature extraction. For instance, Valpola [9] extracted complex-cell-like features from a static image using lateral context as the guiding principle.

A synthesis of the above extensions leads to powerful hierarchical nonlinear feature extractor (see Fig. 2). With temporal context, such a network becomes similar to the slow feature analysis (SFA) [11]. Each layer on the hierarchical DSS would consist of two phases: 1) nonlinear feature expansion, realised for instance by over-complete DSS and 2) context guided learning of features by lateral, top-down or temporal feedback.

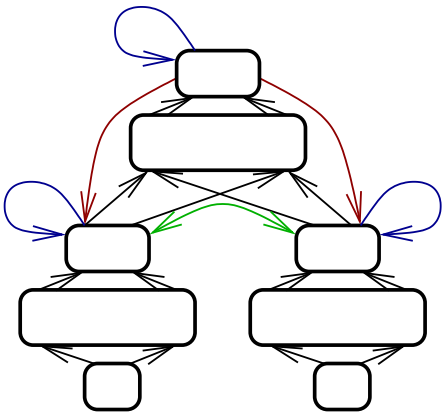

**Fig. 2**. *Feature extraction using DSS and contextual denoising.*

## 4. REFERENCES


[1] P. Földiák. Forming sparse representations by local anti-hebbian learning. *Biol. Cybernetics*, 64:165 – 170, 1990.

[2] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent component analysis*. Wiley, 2001.

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

[4] A. Hyvärinen, P. O. Hoyer, and J. Hurri. Extensions of ica as models of natural images and visual processing. In *Proc of the 4th Int. Conf. on ICA and Signal Separation (ICA2003)*, pages 963 – 974, Nara, Japan, 2003.

[5] E. Oja. A simplified neuron model as a principal component analyzer. *J. of Mathematical Biology*, 15:267–273, 1982.

[6] E. Oja, H. Ogawa, and J. Wangviwattana. Learning in nonlinear constrained Hebbian networks. In T. Kohonen et al., editor, *Artificial Neural Networks, Proc. ICANN'91*, pages 385–390, Espoo, Finland, 1991. North-Holland, Amsterdam.

[7] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

[8] J. Särelä and H. Valpola. Denoising source separation. *Journal of Machine Learning Research*, 2005. In press, a revision available at `http://lib.hut.fi/Diss/2004/isbn9512273438/article5.pdf`.

[9] H. Valpola. Behaviourally meaningful representations from normalisation and context-guided denoising. Technical report, AILab, Dept. of Inf. Tech., University of Zurich, 2004. Avail. at Cogprints: `http://cogprints.ecs.soton.ac.uk/archive/00003633/`.

[10] H. Valpola and J. Särelä. Accurate, fast and stable denoising source separation algorithms. In *Proc of the 5th Int. Conf. on ICA and Signal Separation (ICA2004)*, pages 64 – 71, Granada, Spain, 2004.

[11] L. Wiskott and T. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 14:715 – 770, 2002.