Journal of Machine Learning Research ? (?) ??-?? l Submitted 03/04; Published ??/??

Denoising Source Separation

Jaakko Särelä JAAKKO.SARELA@HUT.FI
Neural Networks Research Centre
Helsinki University of Technology
P.O.Box 5400, FI-02015 HUT, Espoo, FINLAND

Harri Valpola HARRI.VALPOLA@HUT.FI
Artificial Intelligence Laboratory
University of Zurich
Andreasstrasse 15, 8050 Zurich, Switzerland
or *Neural Networks Research Centre*
Helsinki University of Technology
P.O.Box 5400, FI-02015 HUT, Espoo, FINLAND

Editor: ????

Abstract

A new algorithmic framework called denoising source separation (DSS) is introduced. The main benefit of this framework is that it allows for easy development of new source separation algorithms which are optimised for specific problems. In this framework, source separation algorithms are constucted around denoising procedures. The resulting algorithms can range from almost blind to highly specialised source separation algorithms. Both simple linear and more complex nonlinear or adaptive denoising schemes are considered. Some existing independent component analysis algorithms are reinterpreted within DSS framework and new, robust blind source separation algorithms are suggested. Although DSS algorithms need not be explicitly based on objective functions, there is often an implicit objective function that is optimised. The exact relation between the denoising procedure and the objective function is derived and a useful approximation of the objective function is presented. In the experimental section, various DSS schemes are applied extensively to artificial data, to real magnetoencephalograms and to simulated CDMA mobile network signals. Finally, various extensions to the proposed DSS algorithms are considered. These include nonlinear observation mappings, hierarchical models and overcomplete, nonorthogonal feature spaces. With these extensions, DSS appears to have relevance to many existing models of neural information processing.

Keywords: blind source separation, BSS, prior information, denoising, denoising source separation, DSS, independent component analysis, ICA, magnetoencephalograms, MEG, CDMA

1. Introduction

Over the recent years, source separation of linearly mixed signals has attracted a wide range of researchers. The focus of this research has been on developing algorithms that make minimal assumptions on the underlying process, thus approaching blind source separation

(BSS). Independent component analysis (ICA) (Hyvärinen et al., 2001b) clearly follows this tradition. This blind approach certainly has its assets, giving the algorithms a wide range of possible applications. ICA has been a valuable tool, in particular, in testing the hypotheses of the target field of research (*c.f.,* Vigário et al., 2000).

Nearly always, however, there is further information due to the experimental setup, other design specifications or cumulated knowledge due to scientific research. For example in biomedical signal analysis (*c.f.,* Gazzaniga, 2000, Rangayyan, 2002), careful design of experimental setups provides us with presumed signal characteristics. In man-made technology, such as a CDMA mobile system (*c.f.,* Viterbi, 1995), the transmitted signals are even more restricted.

The Bayesian approach provides a sound framework for including prior information into inferences about the signals. This has been used, for instance, by Knuth (1998), Valpola and Karhunen (2002), Särelä et al. (2001). However, the algorithms are not always simple or computationally efficient. In fast point estimate algorithms, prior information has been used for denoising the results given by ICA (Vigneron et al., 2003). Additional information has also well been used to extract signals corresponding to some specific feature of the measured phenomenon (for a review of biomedical applications, see Rangayyan, 2002).

In this paper, we introduce denoising source separation (DSS), a new framework for incorporating prior knowledge into the separation process. DSS is developed by generalising the principles introduced by Valpola and Pajunen (2000). We argue that it is often easy and practical to express the available knowledge in terms of denoising.

Some previous works (*c.f.,* Hyvärinen et al., 1999, Valpola and Pajunen, 2000) have acknowledged that it is possible to interpret some parts of ICA algorithms as denoising. In this paper, we show that it is actually possible to construct the source separation algorithms around the denoising methods themselves.

Denoising corresponds to procedural knowledge. This differs from most approaches to ICA where the algorithms are derived from explicit objective functions leading to declarative knowledge. We also derive the exact relation between the objective function and the corresponding denoising. This makes it possible to mix the two types of information and also provides a new interpretation to some of the existing ICA algorithms.

Additionally, we review a method to speed up convergence and discuss the applicability of DSS for exploratory source separation when no detailed information of the sources is available.

The paper is organised as follows: In Sec. 2, we introduce the principles of denoising source separation in a linear framework and proceed to nonlinear denoising. In Sec. 3, some practical denoising functions are discussed. In Sec. 4, a useful approximation for the objective function of DSS is derived and its uses are discussed. The use of the algorithmic framework is demonstrated in Sec. 5 in experiments with artificial and real-world data. Finally, in Sec. 6, we discuss extensions to DSS framework and their connections to models of neural information processing.

2. Source separation by denoising

Consider a linear instantaneous mixing of sources:

$$\mathbf{X} = \mathbf{A}\mathbf{S} + \boldsymbol{\nu}, \tag{1}$$

where

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_M \end{bmatrix}, \qquad \mathbf{S} = \begin{bmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \\ \vdots \\ \mathbf{s}_N \end{bmatrix}. \tag{2}$$

The source matrix $\mathbf{S}$ consists of N sources. Each individual source $\mathbf{s}_i$ consists of T samples, that is, $\mathbf{s}_i = [s_i(1) \dots s_i(t) \dots s_i(T)]$. Note that in order to simplify the notation throughout the paper, we have defined each source to be a row vector instead of the more traditional column vector. The symbol t often stands for time, but other possibilities include, *e.g.*, space. For the rest of the paper, we refer to t as time, for convenience. The observations $\mathbf{X}$ consist of M mixtures of the sources, that is, $\mathbf{x}_i = [x_i(1) \dots x_i(t) \dots x_i(T)]$. Usually it is assumed that $M \geq N$. The linear mapping $\mathbf{A} = [\mathbf{a}_1\, \mathbf{a}_2 \cdots \mathbf{a}_M]^T$ consists of the mixing vectors $\mathbf{a}_i = [a_{i1}\, a_{i2} \dots a_{iN}]^T$, and is usually called mixing matrix. In the model, there is some Gaussian noise $\boldsymbol{\nu}$, too.

If the sources are assumed Gaussian, this is a general, linear factor analysis model with rotational invariance. There are several ways to fix the rotation, *i.e.*, to separate the original sources, $\mathbf{S}$. Some approaches assume structure for the mixing matrix. If no structure is assumed, the solution to this problem is usually called blind source separation (BSS). Note that this approach is not really blind, since one always needs some information to be able to fix the rotation. One such information is the non-Gaussianity of the sources, which leads to the recently popular ICA methods (*c.f.,* Hyvärinen et al., 2001b). Other properties may include their temporal structure as in Belouchrani et al. (1997), Ziehe and Müller (1998).

In this section we introduce a general framework for using denoising algorithms for source separation. We show how source separation algorithms can be constructed around methods for removing noise from the estimated sources. We first analyse the situation in linear denoising and then consider the nonlinear case.

2.1 Linear denoising

Many ICA algorithms preprocess the data by removing the mean and normalising the covariance to unit matrix, *i.e.*, $\mathbf{X}\mathbf{X}^T/T = \mathbf{I}$. This is referred to as sphering, whitening or decorrelation and its result is that any signal obtained by projecting the sphered data on any unit vector has zero mean and unit variance. Furthermore, orthogonal projections yield uncorrelated signals. Often sphering is combined with reducing the dimension of the data by selecting a principal subspace which contains most of the energy of the original data. In ICA, the motivation for sphering is that with at most as many sources as observations, the original independent signals can be recovered by an orthogonal rotation $\mathbf{W}$ of the sphered signals, provided that the noise $\boldsymbol{\nu}$ is negligible. If there is substantial noise present, unbiased estimate for $\mathbf{S}$ may not be achieved by an orthogonal $\mathbf{W}$. Even in these cases, orthogonal $\mathbf{W}$ usually offers a good approximation. For the rest of the paper, the data $\mathbf{X}$ is assumed to be spherical or sphered and we mainly consider orthogonal rotations $\mathbf{W}$.

Let us consider the effect of modifying the sphered data by a linear operation, for instance by low-pass filtering the sphered signals. In general, low-pass filtering decreases the energy of the signals and the remaining energy depends on their frequency content. This means that projecting the data on a vector with unit length no longer yields a signal

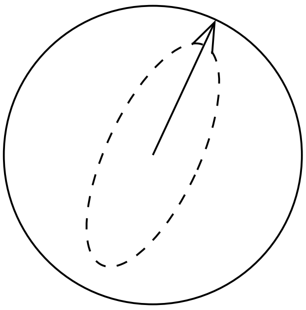


Figure 1: *Sphering renders the variance of all unit projections equal and the projections can be depicted by a circle (solid line). After denoising (dashed line), the variance depends on the direction of the projection, and the signal or signals of interest can be identified by PCA. The arrow points to the direction of the first principal component.*

with unit variance and, more importantly, not all projections result in the same variance. After low-pass filtering, it is therefore possible to identify the signals having higher than average proportion of low frequencies by principal component analysis (PCA).

If signals of interest are characterised as having low-frequency components, low-pass filtering can be regarded as denoising. It is thus possible to identify the following steps in the above signal separation: sphering, denoising and PCA. These are illustrated in Fig. 1.

For now we consider linear denoising. This means that denoising can be mathematically expressed as matrix multiplication. Then for denoising matrix $\mathbf{D}$ the denoised data $\mathbf{Z}$ becomes:

$$\mathbf{Z} = \mathbf{X}\mathbf{D}\,. \tag{3}$$

Note that $\mathbf{D}$ operates on each signal $\mathbf{x}_i$ separately, *i.e.*, denoising is defined for one-dimensional signals. Furthermore, the denoising is performed over time, *i.e.*, $\mathbf{D}$ is $T \times T$ -matrix.

The first principal component of the denoised data $\mathbf{Z}$, can be computed by the *classical power method* (see any elementary matrix textbook):

$$\mathbf{w}^+ = \mathbf{Z}\mathbf{Z}^T\mathbf{w} \tag{4}$$

$$\mathbf{w}_{\text{new}} = \frac{\mathbf{w}^+}{||\mathbf{w}^+||}, \tag{5}$$

where $\mathbf{w}_{\text{new}}$, a column vector, means the new estimate of the principal direction. This power method finds the maximum eigenvector[1] of the covariance matrix of the denoised data. Thus its objective function is

$$g_{\text{lin}}(\mathbf{w}) = \mathbf{w}^T\mathbf{Z}\mathbf{Z}^T\mathbf{w}, \tag{6}$$

[1] The maximum eigenvector corresponds to the maximum eigenvalue.

subject to the constraint $\mathbf{w}^T\mathbf{w} = 1$. Note that $g_{\text{lin}}(\mathbf{w})$ is a scalar function of a vector argument.

Let us now substitute the denoising (3) into the power method (4):

$$\mathbf{w}^+ = \mathbf{Z}\mathbf{Z}^T\mathbf{w}^T = \mathbf{X}\mathbf{D}\mathbf{D}^T\mathbf{X}^T\mathbf{w}. \tag{7}$$

Further, let us denote

$$\mathbf{D}^* = \mathbf{D}\mathbf{D}^T. \tag{8}$$

Then the classical power method applied to filtered data can be reformulated as follows:

$$\mathbf{s} = \mathbf{w}^T\mathbf{X} \tag{9}$$

$$\mathbf{s}^+ = \mathbf{s}\mathbf{D}^* \tag{10}$$

$$\mathbf{w}^+ = \mathbf{X}\mathbf{s}^{+T} \tag{11}$$

$$\mathbf{w}_{\text{new}} = \frac{\mathbf{w}^+}{||\mathbf{w}^+||}, \tag{12}$$

where $\mathbf{s}$ is used to denote the current estimate of the signal corresponding to the eigenvector estimate $\mathbf{w}$. From now on, we shall refer to this signal as source. Mathematically, algorithm (9)–(12) is equivalent to the classical power method applied to the filtered data (3)–(5). In the classical version, the denoising $\mathbf{D}$ was applied to the whole data, but in Eq. (10) the denoising $\mathbf{D}^*$ is applied to the current source estimate $\mathbf{s}$, instead. Equations (11) and (12) compute the weight vector which yields a new source which is closest to the denoised $\mathbf{s}^+$ in the least-mean-squares (LMS) sense. We call Eqs. (9)–(12) the *linear DSS algorithm.* The corresponding objective function, starting from Eq. (6), can be written as

$$g_{\text{lin}}(\mathbf{s}) = \mathbf{w}^T\mathbf{Z}\mathbf{Z}^T\mathbf{w} = \mathbf{s}\mathbf{D}^*\mathbf{s}^T, \tag{13}$$

In order to study further the relation between $\mathbf{D}^*$ and $\mathbf{D}$, let us now assume that $\mathbf{D}^*$ is given and try to find a corresponding $\mathbf{D}$ that satisfies Eq. (8). Since $\mathbf{D}\mathbf{D}^T$ is always symmetric, a necessary condition for the existence of a corresponding $\mathbf{D}$ is symmetry of $\mathbf{D}^*$. In this case it means symmetry in time since $\mathbf{D}^*$ is a $T \times T$ matrix. Symmetric matrices have an eigenvalue decomposition

$$\mathbf{D}^* = \mathbf{V}\boldsymbol{\Lambda}^*\mathbf{V}^T, \tag{14}$$

where $\boldsymbol{\Lambda}^*$ is diagonal and $\mathbf{V}$ is orthonormal[2]. Since $\mathbf{D} = \mathbf{V}\boldsymbol{\Lambda}^{*1/2}$ satisfies Eq. (8), symmetry of $\mathbf{D}^*$ is both a necessary and sufficient condition for the existence of a corresponding $\mathbf{D}$. Power method applied to the denoised data $\mathbf{Z}$ always converges (assuming that the largest eigenvalue is not degenerate). The above analysis shows that the same holds for linear DSS provided that denoising is symmetric in time.

The eigenvalue decomposition (14) shows that any denoising in linear DSS can be implemented as an orthonormal rotation followed by a point-wise adjustment of the samples and rotation to the original space. A good example of this is presented by linear time-invariant

[2] We consider here eigenvalue decompositions containing real matrices only, thus the transpose.

(LTI) filtering. In such denoising, $\mathbf{V}$ corresponds for example to the Fourier transform[3] or discrete cosine transform (DCT). After the transform, the signal is filtered using the diagonal matrix $\mathbf{\Lambda}^*$, *i.e.*, by a point-wise adjustment of the frequency bins. Finally the signal is inverse transformed using $\mathbf{V}^T$. In the case of LTI-filtering, the denoising characteristics are manifested only in the diagonal matrix, while the transforming matrix $\mathbf{V}$ portrays a constant rotation. When this is the case, the algorithm can be further simplified by imposing the transformation on the sphered data, $\mathbf{X}$. Then the iteration can be performed in the transformed basis. This trick has been exploited in the first experiment of Sec. 5.2.

When the whole data is denoised by $\mathbf{D}$ and power method is used, it does not matter if the inverse transformation is applied. In fact, any orthonormal rotation $\mathbf{U}$ can be applied to the data without changing the covariance. This means that $\mathbf{D}$ corresponding to $\mathbf{D}^*$ is not unique and all orthonormal rotations of $\mathbf{D}$ satisfy $(\mathbf{DU})(\mathbf{DU})^T = \mathbf{D}(\mathbf{UU}^T)\mathbf{D}^T = \mathbf{D}^*$, too. If $\mathbf{U}$ is chosen to be the inverse transform $\mathbf{V}^T$, the denoising $\mathbf{D}$ is

$$\mathbf{D} = \mathbf{V}\mathbf{\Lambda}^{*1/2}\mathbf{V}^T = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T, \tag{15}$$

where the diagonal matrix has been denoted by $\mathbf{\Lambda} = \mathbf{\Lambda}^{*1/2}$. Note that a denoising $\mathbf{D}$ is usually meaningful only if $\mathbf{\Lambda}$ is real and positive. The denoising $\mathbf{D}^*$ used in linear DSS should therefore be positive definite in addition to being symmetric.

In some practical cases the diagonal elements of $\mathbf{\Lambda}$ have binary values which means that $\mathbf{D} = \mathbf{DD}^T = \mathbf{D}^*$ and the denoisings used in the different approaches are exactly the same. Examples of such denoisings are presented in Sec. 3.

2.2 Nonlinear noise reduction

In general, denoising is not restricted to linear operations. Median filtering is a clear example of nonlinear denoising which cannot be implemented as mere matrix multiplication. Another example of nonlinear denoising is encountered when the denoising is tuned adaptively to improving estimates of the source characteristics as the iteration progresses. A good review on nonlinear filtering is given by Kuosmanen and Astola (1997).

One common way to develop nonlinear algorithms, such as ICA, from linear algorithms, such as PCA, is to replace the quadratic criterion (6) by a criterion which contains other than second-order moments. However, we argue that it is often easier and more practical to simply replace Eq. (10) by a nonlinear denoising step

$$\mathbf{s}^+ = \mathbf{f}(\mathbf{s}). \tag{16}$$

The function $\mathbf{f}(\mathbf{s})$ denotes the result of denoising, *i.e.*, both $\mathbf{s}$ and $\mathbf{f}(\mathbf{s})$ are row vectors of the same length. In the linear case $\mathbf{f}(\mathbf{s}) = \mathbf{sD}^*$, but in general, almost any type of denoising procedure can be applied. When more than one source is estimated, it may be desirable to use the information in all sources $\mathbf{S}$ for denoising any particular source $\mathbf{s}_i$. This leads to the following denoising function: $\mathbf{s}_i^+ = \mathbf{f}_i(\mathbf{S})$.

Denoising is useful as such and therefore there is a wide literature of sophisticated denoising methods to choose from (*c.f.,* Anderson and Moore, 1979). Moreover, one usually

[3] Note that the eigenvalue decomposition (14) contains real rotations instead of complex, but Fourier transform is usually seen as a complex tranformation. To keep the theory simple, we consider real Fourier transform where the corresponding sine and cosine terms have been separated in different elements.

has some knowledge about the signals of interest and thus possesses the information needed for denoising. In fact, quite often the signals extracted by BSS techniques would be post-processed to reduce noise in any case (*c.f.,* Vigneron et al., 2003). In the DSS framework, the available denoising methods can be directly applied to source separation, producing better results than purely blind techniques. There are also very general noise reduction techniques such as wavelet denoising (Donoho et al., 1995, Vetterli and Kovacevic, 1995) or median filtering (Kuosmanen and Astola, 1997) which can be applied in exploratory data analysis. The DSS framework thus suggests new algorithms ranging from BSS to highly specialised applications.

The *nonlinear DSS algorithms* (9), (16), (11), (12) can be implemented without any reference to an optimisation criterion $g(\mathbf{w})$. However, to justify the algorithm, we shall next establish the relation between the denoising function $\mathbf{f}(\mathbf{s})$, and the objective function $g(\mathbf{w})$ that is implicitly optimised. The following Lagrange equation holds at the optimum:

$$\nabla_{\mathbf{w}}[g_{\mathbf{s}}(\mathbf{s}) - \boldsymbol{\lambda}^T \mathbf{h}(\mathbf{w})] = 0, \tag{17}$$

where $g_{\mathbf{s}}(\mathbf{s}) = g(\mathbf{w}^T\mathbf{X})$ denotes the optimisation criterion as a function of the source estimate $\mathbf{s}$. The row vectors $\boldsymbol{\lambda}$ and $\mathbf{h}$ denote the Lagrange multipliers and the corresponding constraints under which the optimisation is performed, respectively. In this case, the constraint is that $\mathbf{w}$ has a unit norm, *i.e.*, $h(\mathbf{w}) = \mathbf{w}^T\mathbf{w} - 1 = 0$, and it thus follows

$$\mathbf{X}\nabla_{\mathbf{s}} g_{\mathbf{s}}(\mathbf{s})^T - 2\lambda\mathbf{w} = 0. \tag{18}$$

This results in the following fixed point:

$$\mathbf{w}_g = \frac{\mathbf{X}\nabla_s g_{\mathbf{s}}^T(\mathbf{s})}{||\mathbf{X}\nabla_s g_{\mathbf{s}}^T(\mathbf{s})||}. \tag{19}$$

This should coincide with the fixed point of the nonlinear DSS presented by Eqs. (9), (16), (11), (12):

$$\mathbf{w}_{\mathbf{f}} = \frac{\mathbf{X}\mathbf{f}^T(\mathbf{s})}{||\mathbf{X}\mathbf{f}^T(\mathbf{s})||}. \tag{20}$$

One possible solution is obviously $\mathbf{f}(\mathbf{s}) = \nabla_{\mathbf{s}} g_{\mathbf{s}}(\mathbf{s})$ but, more generally, the relation between denoising, $\mathbf{f}$, and the optimisation criterion, g, is

$$\mathbf{f}(\mathbf{s}) = \alpha(\mathbf{s})\nabla_{\mathbf{s}} g_{\mathbf{s}}(\mathbf{s}) + \beta(\mathbf{s})\mathbf{s}, \tag{21}$$

where $\alpha(\mathbf{s})$ and $\beta(\mathbf{s})$ are scalar valued functions. The scalar $\alpha(\mathbf{s})$ disappears in the normalisation of Eq. (20). Furthermore, the fixed point of DSS is not altered by addition of any multiple of $\mathbf{w}$ to the right side of Eq. (11). Since for sphered data it always holds that $\mathbf{w} \propto \mathbf{X}\mathbf{s}^T$, the term $\beta(\mathbf{s})\mathbf{s}$ in Eq. (21) does not change the fixed point (20) either.

Selecting $\beta(\mathbf{s}) = 1$ yields an intuitive interpretation for the relation between $\mathbf{f}(\mathbf{s})$ and $g_{\mathbf{s}}(\mathbf{s})$: denoising by $\mathbf{f}$ modifies the source estimate $\mathbf{s}$ in the direction where the optimisation criterion g grows maximally. In general, $g_{\mathbf{s}}(\mathbf{s})$ can be seen as a signal-to-noise ratio (SNR) where signal and noise are defined implicitly by the denoising function $\mathbf{f}(\mathbf{s})$.

An arbitrary function $\mathbf{f}(\mathbf{s})$ does not necessarily have a corresponding objective function $g_{\mathbf{s}}(\mathbf{s})$ which would satisfy Eq. (21). Derivation of $\mathbf{f}(\mathbf{s})$ from $g_{\mathbf{s}}(\mathbf{s})$ via Eq. (21) always results a denoising with which DSS converges as long as the step size is kept in control by suitable choices of $\alpha(\mathbf{s})$ and $\beta(\mathbf{s})$. Conversely, if $\mathbf{f}(\mathbf{s})$ leads to DSS which does not converge for any choice of $\alpha(\mathbf{s})$ and $\beta(\mathbf{s})$, there cannot be $g_{\mathbf{s}}(\mathbf{s})$ satisfying Eq. (21). In practice, however, as long as $\mathbf{f}(\mathbf{s})$ is a reasonable denoising function, each iteration step improves some particular function $g_{\mathbf{s}}(\mathbf{s})$ which can be interpreted as a measure of SNR[4]. Consequently, DSS iterations using the denoising $\mathbf{f}(\mathbf{s})$ converge even if it may be difficult to write down the corresponding $g_{\mathbf{s}}(\mathbf{s})$.

One last point can be made before we proceed from the basic DSS to some useful extensions. Let us assume that the nonlinear denoising (16) operates point-wise, *i.e.*, the denoised signal at time t depends only on the original signal at time t. Then the nonlinear DSS algorithm is actually equivalent to the nonlinear PCA (Karhunen and Joutsensalo, 1994). In general, however, DSS is not restricted to time independent denoising.

2.3 Deflation

The classical power method has two common extensions: deflation and spectral shift. They are readily available for the linear DSS since it is equivalent to the power method applied to filtered data via Eq. (8). We shall now generalise the deflation to nonlinear DSS algorithms. Spectral shift will be discussed in Section 4.3.

The power method (4)–(5) estimates only the most powerful source in terms of the eigenvalues of $\mathbf{Z}\mathbf{Z}^T$. Deflational method is a procedure which allows one to estimate several sources by iteratively applying power method several times. The convergence to previously found sources is prevented simply by restricting the separating vectors to be orthogonal to each other, for instance by $\mathbf{w}_{\mathrm{orth}} = \mathbf{w} - \mathbf{W}^T\mathbf{W}\mathbf{w}$ (Luenberger, 1969), where $\mathbf{W} = [\mathbf{w}_1\, \mathbf{w}_2 \cdots \mathbf{w}_N]^T$ contains the already estimated projections. As mentioned, the deflational method is readily available for the linear DSS algorithms.

It turns out that deflation is directly applicable to nonlinear DSS as well. Additional constraints $h(\mathbf{w}) = \mathbf{w}^T\mathbf{w}_i^T = 0$ in Eq. (17) give rise to additive terms $\lambda_i \mathbf{w}_i$ in Eq. (18). This shows that the denoising procedure itself is not affected by the orthogonalisation of $\mathbf{w}$ and that consecutive runs of the algorithm optimise the same $g(\mathbf{w})$ as the first run but under the constraint of orthogonality to the previously extracted components.

Note that in this deflation scheme, it is possible to use different kinds of denoising procedures when the sources differ in characteristics. This will be discussed in more detail in Sec. 4.2.

If more than one sources are estimated simultaneously, the symmetric orthogonalisation methods proposed for symmetric FastICA (Hyvärinen, 1999) can be used.

3. Denoising functions in practice

DSS is a framework for designing source separation algorithms. The idea is that the algorithms differ mainly in the denoising function $\mathbf{f}(\mathbf{s})$ while the other parts of the algorithm remain mostly the same. In this section, we discuss denoising functions ranging from simple but powerful linear ones to sophisticated nonlinear ones with the goal of inspiring others to

[4] It is possible to show that $g_{\mathbf{s}}(\mathbf{s})$ exists if and only if DSS using $\mathbf{f}(\mathbf{s})$ converges for any additional constraint on $\mathbf{s}$.

try out their own denoising methods. The range of applicability of the examples spans from cases where the knowledge about the signals is relatively specific to almost blind source separation where very little is assumed about the signal characteristics. Many of the denoising functions discussed in this section are applied in experiments in Section 5.

Before proceeding to examples of denoising functions, we note that it is usually not crucial for the denoising to be very exact. Otherwise DSS would not be very useful because one would only get what is asked from the algorithm in terms of the denoising function. Fortunately, this is not the case: Assuming that the signals are recoverable by linear projections from the observations, it is enough for the denoising function $\mathbf{f}(\mathbf{s})$ to remove more noise than signal (*c.f.,* Hyvärinen et al., 2001b, Theorem 8.1). This is because the reestimation steps (11) and (12) constrain the source $\mathbf{s}$ to the subspace spanned by the data. Even if the denoising discards parts of the signal, reestimation steps restore them.

In practice, the observations contain noise which does not fully disappear by any linear projection and then the quality of the separated signals depends on the accuracy of denoising. If there is no detailed knowledge about characteristics of the signals to start with, it is useful to bootstrap the denoising functions. This can be achieved by starting with relatively general signal characteristics and then tuning the denoising functions based on analyses of the structure in the noisy signals extracted in the first phase. In fact, some of the nonlinear DSS algorithms can be regarded as linear DSS algorithms where a linear denoising function is adapted to the sources, leading to nonlinear denoising.

3.1 Detailed linear denoising functions

In this section we consider several detailed, simple, but powerful, linear denoising schemes. We introduce the denoisings using the denoising matrix, $\mathbf{D}^*$ when feasible. We consider effective implementation of the denoisings as well.

3.1.1 On/off-denoising

Consider designed experiments, *e.g.*, in fields of psychophysics or biomedicine. It is usual to control them by having periods of activity and non-activity. In such experiments the denoising can be simply implemented by

$$\mathbf{D}^* = \mathrm{diag}(\mathbf{b}), \tag{22}$$

where $\mathbf{D}^*$ refers to the linear denoising matrix in Eq. (10) and

$$\mathbf{b} = \begin{cases} 1, \text{for the active parts} \\ 0, \text{for the inactive parts} \end{cases} \tag{23}$$

This amounts to multiplying the source estimate $\mathbf{s}$ by a binary mask[5], where ones represent the active parts and zeroes the non-active parts. Notice that this masking procedure actually satisfies $\mathbf{D}^* = \mathbf{D}^*\mathbf{D}^{*T}$. This means that DSS is equivalent to the power method even with exactly the same filtering. In practice this DSS algorithm could be implemented by PCA, applied to the active parts of the data while the sphering stage would still involve the whole data.

[5] By masking we refer to point-wise multiplication of a signal or a transformation of a signal.

3.1.2 Denoising based on the frequency content

If, on the other hand, signals are characterised by having certain frequency components, one can transform the source estimate by DCT, mask the spectrum, *e.g.*, with a binary mask, and inverse transform to obtain the denoised signal:

$$\mathbf{D}^* = \mathbf{V}\mathbf{\Lambda}^*\mathbf{V}^T, \tag{24}$$

where $\mathbf{V}$ is the transform, $\mathbf{\Lambda}^*$ is the matrix with the mask on its diagonal, and $\mathbf{V}^T$ is the inverse transform. Again, a computational implementation of the algorithm needs not resort to matrix multiplications and it is possible to implement DSS by applying PCA on selected parts of the transformed data.

3.1.3 Spectrogram denoising

Often a signal is well characterised by what frequencies occur at what times. This is evident, *e.g.*, in oscillatory activity in the brain where oscillations often occur in bursts. An example of source separation in such data is studied in Sec. 5.2. The time-frequency behaviour can be described by calculating discrete cosine transform (DCT) in short windows in time. This results in a combined time and frequency representation, spectrogram, where the masking can be applied.

There is a known dilemma in the calculation of the spectrogram: detailed description of the frequency content does not allow detailed information of the activity in time and vice versa. In other words, large amount of different frequency bins T_f will result in small amount of time locations T_t. Wavelet transforms (Donoho et al., 1995, Vetterli and Kovacevic, 1995) have been suggested to overcome this problem. There an adaptive or predefined basis, different from the pure sinusoids used in Fourier transform or DCT, is used to divide the resources of time and frequency behaviour optimally in some sense.

Here we apply a related overcomplete-basis approach. Instead of having just one spectrogram, we use several time-frequency analyses with different T_t's and T_f's. Then the new estimate of the projection $\mathbf{w}^+$ is achieved by summing the new estimates $\mathbf{w}_i^+$ of each of the time-frequency analyses: $\mathbf{w}^+ = \sum_i \mathbf{w}_i^+$.

3.1.4 Denoising of quasiperiodic signals

As a final example of denoising based on detailed source characteristics, consider Fig. 2a. There a source estimate $\mathbf{s}$ has been reached. The apparent quasiperiodic structure of the signal can be used to perform DSS to get a better estimate. The denoising proceeds as follows:

1. Estimate the locations of the peaks of the current source estimate $\mathbf{s}$ (Fig. 2b).
2. Chop each period from peak to peak.
3. Dilate each period to a fixed length L (linearly or nonlinearly).
4. Average the dilated periods (Fig. 2c).
5. Let the denoised source estimate $\mathbf{s}^+$ be a signal where each period has been replaced by the averaged period dilated back into the original length (Fig. 2d).

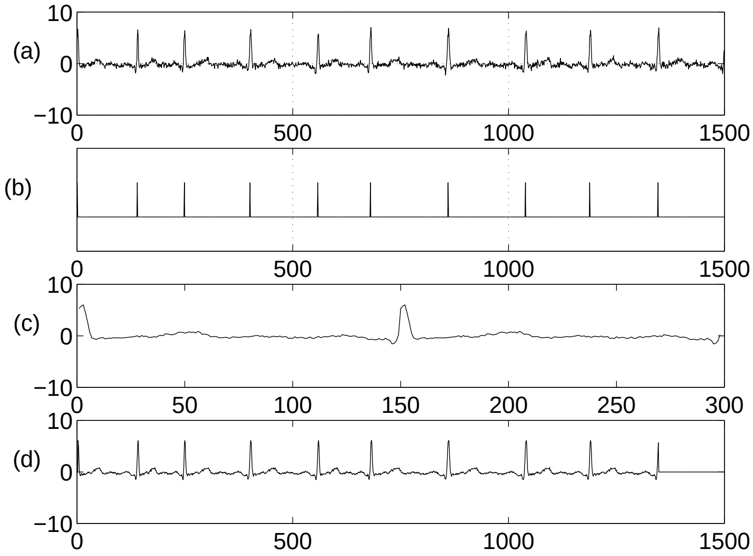


Figure 2: *a) Current source estimate* $\mathbf{s}$ *of a quasiperiodic signal b) Peak estimates c) Average signal* s_{ave} *d) Denoised source estimate* $\mathbf{s}^+$.

The denoised signal $\mathbf{s}^+$ in Fig 2d show significantly better SNR compared to the original source estimate $\mathbf{s}$, in Fig. 2a.

This averaging is a form of linear denoising since it can be implemented as matrix multiplication. Furthermore, it presents another case in addition to the binary masking, where DSS is equivalent to the power method even with exactly the same filtering. It would not be easy to see from the denoising matrix $\mathbf{D}^*$ itself that $\mathbf{D}^* = \mathbf{D}^*\mathbf{D}^{*T}$. However, this becomes evident should one consider the averaging of source estimate $\mathbf{s}^+$ (Fig. 2d) that is already averaged.

Note that there are cases where chopping from peak to peak does not guarantee the best result. This is especially true when the periods do not span the whole section from peak to peak, but there are parts where the response is silent. Then there is need to estimate the lengths of the periods separately.

3.2 Denoising based on estimated signal variance

In the previous section, several denoising schemes were introduced. In all of them, the details of the denoising were assumed to be known. It is as well possible to estimate the denoising specifications from the data. This makes the denoising nonlinear or adaptive. In this section we consider a particular ICA algorithm in the DSS framework, suggesting modifications which improve separation results and robustness.

3.2.1 Kurtosis based ICA

Consider one of the best known BSS approaches, ICA by optimisation of the sample kurtosis of the sources. The objective function is then $g_{\mathbf{s}}(\mathbf{s}) = \sum s^4(t)/T - 3\left(\sum s^2(t)/T\right)^2$. Since the source variance has been fixed to unity, we can simply use $g_{\mathbf{s}}(\mathbf{s}) = \sum s^4(t)/T$ and derive the function $\mathbf{f}(\mathbf{s})$ via Eq. (21). This yields $\nabla_{\mathbf{s}} g_{\mathbf{s}}(\mathbf{s}) = 4/T\,\mathbf{s}^3$, where $\mathbf{s}^3 = [s^3(1)\, s^3(2)\, \ldots]$. Selecting $\alpha(\mathbf{s}) = T/4$ and $\beta(\mathbf{s}) = 0$ in Eq. (21) then result in

$$\mathbf{f}(\mathbf{s}) = \mathbf{s}^3\,. \tag{25}$$

This implements an ICA algorithm with nonlinear denoising. So far, we have not referred to denoising, but a closer examination of Eq. (25) reveals that one can, in fact, interpret $\mathbf{s}^3$ as being $\mathbf{s}$ masked by $\mathbf{s}^2$, the latter being a somewhat naïve estimate of signal energy and thus relating to SNR.

Kurtosis as an objective function is notorious for being prone to overfitting and producing very spiky source estimates (Särelä and Vigário, 2003, Hyvärinen, 1998). For illustration of this consider Fig. 3. There one iteration of DSS using kurtosis based denoising is shown. Assume that via some means source estimate shown in Fig. 3a has been reached. The source seems to contain increased activity in three portions (around time instances 1000, 2300 and 6000). It as well contains a peak roughly at time instance 4700. The signal variance estimate, *i.e.*, the mask is shown in Fig. 3b. While it has boosted somewhat the broad activity compared to the silent parts, the magnification of the peak is far greater. Thus the denoised source estimate $\mathbf{s}^+$ (Fig 3c) has nearly nothing else than the peak. The new source estimate $\mathbf{s}_{\text{new}}$, based on the new projection $\mathbf{w}_{\text{new}}$, is a clear spike having little left of the broad activity.

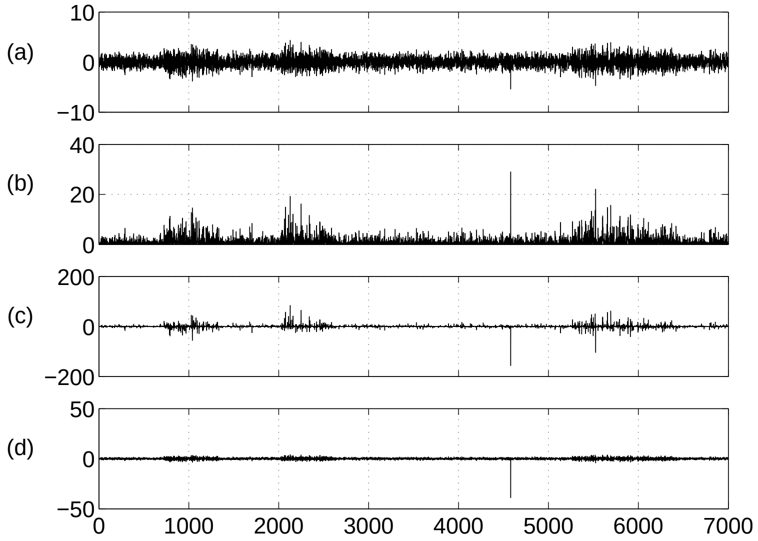


Figure 3: *a) Source estimate* $\mathbf{s}$ *b) Mask* $s^2(t)$ *c) Denoised source estimate* $\mathbf{s}^+ = \mathbf{f}(\mathbf{s}) = \mathbf{s}^3$ *d) Source estimate corresponding to the reestimated* $\mathbf{w}_{\text{new}}$.

The denoising interpretation suggests that the failure to extract the broad activity is due to a poor estimate of SNR.

3.2.2 Better estimate for the signal variance

Let us now consider a related but better founded estimate. Assuming that $\mathbf{s}$ is composed of Gaussian noise with a constant variance σ_n^2 and Gaussian signal with non-stationary variance $\sigma_s^2(t)$, the maximum-a-posteriori (MAP) estimate of the signal is

$$s^+(t) = s(t)\frac{\sigma_s^2(t)}{\sigma_{\text{tot}}^2(t)}\,, \tag{26}$$

where $\sigma_{\text{tot}}^2(t) = \sigma_s^2(t) + \sigma_n^2(t)$ is the total variance of the observation.

The kurtosis based DSS (25) can be acquired from this MAP estimate if the signal variance is assumed to be far smaller than the total variance. In that case it is reasonable to assume σ_{tot}^2 to be constant and $\sigma_s^2(t)$ can be estimated by $s^2(t) - \sigma_n^2$. Subtraction of σ_n^2 does not affect the fixed points as it can be embedded in the term $\beta(\mathbf{s}) = -\sigma_n^2/\sigma_{\text{tot}}$ in Eq. (21) as will be argued in Sec. 4.3. Likewise, division by $\sigma_{\text{tot}}^2(t)$ is absorbed by $\alpha(\mathbf{s})$.

Comparison of Eq. (26) and Eq. (25) immediately suggests improvements to the kurtosis based DSS. For instance, it is clear that if $s^2(t)$ is large enough, it is not reasonable to assume that $\sigma_s^2(t)$ is small compared to $\sigma_n^2(t)$. Instead, the mask should saturate for large $s^2(t)$. This already improves robustness against outliers and alleviates the tendency to produce spiky source estimates.

We suggest the following improvements over kurtosis based denoising function (25):

1. The estimates of signal variance and total variance are based on several observations. The rationale of smoothing is the assumption of smoothness of the signal variance. In practice this can be achieved by low-pass filtering the time, frequency or time-frequency description of $s^2(t)$ yielding the approximation of total variance.

2. The noise variance is likewise estimated from data. It should be some kind of soft minimum of the estimated total variances because the estimate can be expected to have random fluctuations. We suggest the following formula:

$$\sigma_n^2 = C\left(\exp\{E[\log \sigma_n^2 + \sigma_{\text{tot}}^2(t)]\} - \sigma_n^2\right)\,. \tag{27}$$

   The noise variance σ_n^2 appears on both sides of the equation, but at the right-hand side, it appears only to prevent rare small values of σ_{tot}^2 from spoiling the estimate. Hence, we used the previously estimated value on the right-hand side. The constant C is tuned such that the formula gives a consistent estimate of the noise variance if the source estimate is, in fact, nothing but Gaussian noise.

3. The signal variance should be close to the estimate of the total variance minus the estimate of the noise variance. Since a variance cannot be negative and the estimate of the total variance has fluctuations, we use a formula which yields zero only when the total variance is zero but which asymptotically approaches $\sigma_{\text{tot}}^2(t) - \sigma_n^2$ for large values of the total variance:

$$\sigma_s^2(t) = \sqrt{\sigma_{\text{tot}}^4(t) + \sigma_n^4} - \sigma_n^2\,. \tag{28}$$

As an illustration of these improvements consider Fig. 4 where one iteration of DSS using the MAP estimate is shown. The first two subplots (Fig. 4a and b) are identical to the ones using kurtosis based denoising. In Fig. 4c, the energy estimate is smoothed using low-pass filtering. Note that the broad activity has been magnified when compared to the spike around time instance 4700. The noise level σ_n^2, calculated using Eq. (27), is shown in dashed line. Corresponding masking (Fig. 4d) results in a denoised source estimate using Eq. (26), shown in Fig. 4e. Finally, the new source estimate $\mathbf{s}_{\text{new}}$ is shown after five iterations of DSS in Fig. 4f. DSS using the MAP-based denoising has clearly removed a considerable amount of background noise as well as the lonely spike.

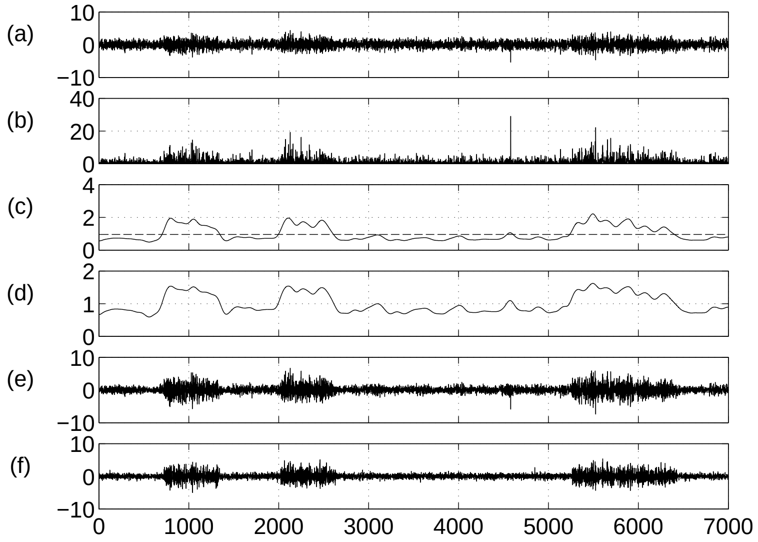


Figure 4: *a) Source estimate* $\mathbf{s}$ *b)* $s^2(t)$*c) Smoothed total energy with the noise level in dashed line d) Denoising mask e) Denoised source estimate* $\mathbf{s}^+$ *f) Source estimate after five iterations of DSS.*

The exact details of these improvements are not crucial, but we wanted to show that the denoising interpretation of Eq. (25) can carry us quite far. The above estimates plugged into Eq. (26) yield a DSS algorithm which is far more robust against overfitting, does not produce the spiky signal estimates and in general yields signals with better SNRs than kurtosis.

Despite the merits of the DSS algorithm described above, there is still one problem with it. While the extracted signals have excellent SNR, they do not necessarily correspond to independent sources, *i.e.*, the sources may remain mixed. This is because there is nothing in the denoising which could discard other sources. Assume, for instance, that two sources have clear-cut and non-overlapping times of strong activity ($\sigma_s^2(t) \gg 0$) and remain silent for most of the time ($\sigma_s^2(t) = 0$). Suppose that one source is present for some time at the beginning of the data and another at the end. If the current source estimate is a mixture of both, the mask will have values close to one at the beginning and at the end of the signal.

Denoising can thus clean the noise from the signal estimate, but it cannot decide between the two sources.

In this respect, kurtosis actually works better than DSS based on the above improvements. This is because the mask never saturates and small differences in the strengths of the relative contributions of two original sources in the current source estimate will be amplified. The problem only occurs in the saturated regime of the mask and we therefore suggest a simple modification of the MAP estimate (26):

$$\mathbf{f}_t(\mathbf{s}) = s(t)\frac{\sigma_s^{2\mu}(t)}{\sigma_{\text{tot}}^2(t)}\,, \tag{29}$$

where μ is a constant slightly greater or equal to one. Note that this modification is usually needed at the beginning of the iterations only. Once the source estimate is dominated by one of the original sources and the contributions of the other sources fall closer to the noise level, the values of the mask are smaller for the other original sources possibly still present in the estimated source.

Another approach is based on the finding that the orthogonalisation of the projection vectors $\mathbf{W}$ cancels only the linear correlation between different sources. Higher-order correlations may still exist. For instance, the variances of different sources can be correlated. Schwartz and Simoncelli (2001) have suggested that variances may be decorrelated by a divisive procedure, in contrast to the orthogonalisation of $\mathbf{W}$, a subtractive procedure. Then it is necessary to estimate explicitly the correlation between one source and the other sources in the current source estimate. The estimate of the total variance can be based in $\sigma_{\text{tot}}^2(t) = \sigma_s^2(t) + \sigma_n^2(t) + \sigma_{\text{others}}^2(t)$, where $\sigma_{\text{others}}^2(t)$ stands for the estimate of total leakage of variance from the other sources. This approach has been further pursued by Valpola and Särelä (2004).

The problems related to kurtosis are well known and several other improved nonlinear functions $\mathbf{f}(\mathbf{s})$ have been proposed. However, some aspects of the above denoising, especially smoothing of the total-variance estimate $s^2(t)$, have not been suggested previously although they arise quite naturally from the denoising interpretation.

3.2.3 Tanh-nonlinearity interpreted as saturated energy estimate

A popular replacement of the kurtosis-based nonlinearity (25) is the hyperbolic tangent $\tanh(\mathbf{s})$ operating point-wise for the sources. It is generally considered to be more robust against overfitted and spiky source estimates than kurtosis. By selecting $\alpha(\mathbf{s}) = -1$ and $\beta(\mathbf{s}) = 1$, we arrive at

$$\mathbf{f}_t(\mathbf{s}) = s(t) - \tanh[s(t)] = s(t)\left(1 - \frac{\tanh[s(t)]}{s(t)}\right)\,. \tag{30}$$

Now the term multiplying $s(t)$ can be interpreted as a mask related to SNR. Unlike the naïve mask $s^2(t)$ resulting from kurtosis, the tanh-based mask (30) saturates, though not very fast.

The energy based mask (29) with the improvements considered above offers a new intepretation for the robustness of the tanh-mask. Parameter values $\sigma_n^2 = 1$ and $\mu = 1.08$ give an excellent fit between the masks as shown in Fig. 5. The advantages of the denoising

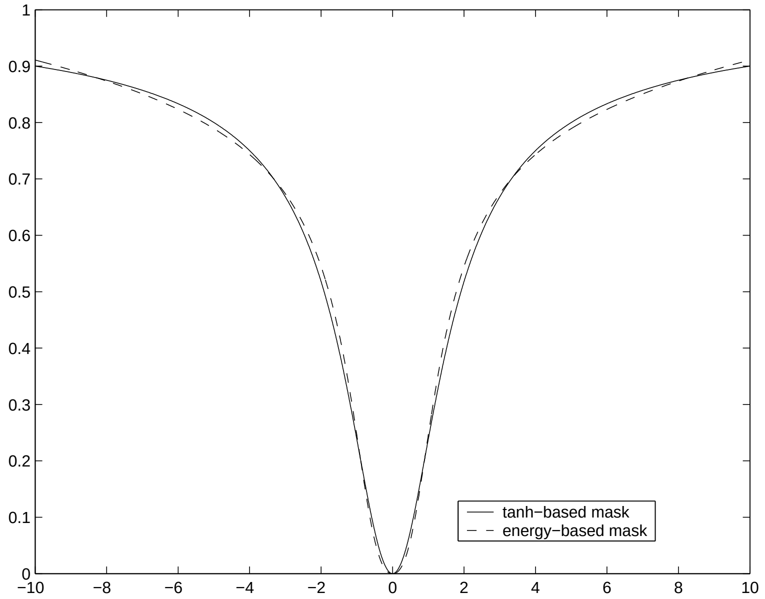


Figure 5: *The tanh-based denoising mask* $1 - \tanh(s)/s$ *is shown together with the energy-based denoising mask proposed here. The parameters in the proposed mask were* $\sigma_n^2 = 1$ *and* $\mu = 1.08$. *We have scaled the proposed mask to match the scale of the tanh-based mask.*

we propose are that σ_n^2 can be tuned to the source estimate, μ can be controlled during the iterations and the estimate of the signal energy can be smoothed. These features contribute to the resistance against overfitting and spiky source estimates.

3.3 Other denoising functions

There are cases where the system specification itself suggests some denoising schemes. One such case is described in Sec. 5.6. Another example is source separation with a microphone array combined with speech recognition. Many speech recognition systems rely on generative models which can be readily used to denoise the speech signals.

Often it would be useful to be able to separate the sources online, *i.e.*, in real time. Since there exists online sphering algorithms (*c.f.,* Douglas and Cichocki, 1997, Oja, 1992), real time DSS can be considered as well. One simple case of online denoising is presented by MA filters. However, such online filters typically are not symmetric and thus have no definite objective function (see Sec. 4.1) resulting in potentially unstable DSS. Consider, for example, a case of two harmonic oscillatory sources. It has a rotational invariance in a space defined by the corresponding sine-cosine pair. Batch DSS algorithms would converge to some particular rotation, but non-symmetric on-line denoising by $\mathbf{f}(s(t)) = s(t-1)$ would not converge at all. Thus, in the case of online DSS, denoising would be best kept symmetric.

Sometimes the sources can be grouped to form interesting subspaces. This could happen, *e.g.*, when all the sources are not independent of each others, but there exists anyway subspaces that are mutually independent. Some form of subspace rules can be used to guide the extraction of interesting subspaces in DSS. It is possible to further relax the independence criterion at the borders of the subspaces. This can be achieved by incorporating a neighbourhood denoising rule in DSS, resulting in a topographic ordering of the sources. One such topographic rule was used in topographic ICA (Hyvärinen et al., 2001a).

It is possible to combine various denoising functions when the sources are characterised by more than one type of structure. Note that the combination order might be crucial for the outcome. This is simply because, in general, $\mathbf{f}_i\left(\mathbf{f}_j(\mathbf{s})\right) \neq \mathbf{f}_j\left(\mathbf{f}_i(\mathbf{s})\right)$ where $\mathbf{f}_i$ and $\mathbf{f}_j$ present two different linear or nonlinear denoisings. As an example, consider the combination of the linear on/off-mask (22) and (23), and the nonlinear energy based mask (29): the noise estimation becomes significantly more accurate when the on/off-masking is performed only after the nonlinear denoising.

Finally, a source might be almost completely known. Then it is possible to apply a detailed matched filter to estimate the mixing coefficients or the noise level. Detailed matched filters have been used in Sec. 5.1 to get an upper limit of the SNRs of the source estimates.

4. Approximation for the objective function

The virtue of DSS framework is that it allows one to develop procedural source separation algorithms without referring to an exact objective function. However, in many cases an approximation of the objective function is nevertheless useful. In this section, we propose such an approximation and discuss its uses, including monitoring and acceleration of convergence as well as analysis of separation results.

4.1 Derivation of the approximation

As shown in Sec. 2.1, Eq. (13), the objective function corresponding to linear denoising $\mathbf{f}(\mathbf{s}) = \mathbf{s}D^*$ is $g_\mathbf{s}(\mathbf{s}) = \mathbf{s}D^*\mathbf{s}^T$, given that D^* is a symmetric matrix[6]. This can be written as $g_\mathbf{s}(\mathbf{s}) = \mathbf{s}\,\mathbf{f}_{\text{lin}}^T(\mathbf{s})$. This formula is exact for linear DSS and we propose it as an approximation $\hat{g}_\mathbf{s}$ for the objective function for nonlinear DSS as well:

$$\hat{g}_\mathbf{s}(\mathbf{s}) = \mathbf{s}\,\mathbf{f}^T(\mathbf{s})\,. \tag{31}$$

There is, however, an important caveat to be made. Note that Eq. (21) includes the scalar functions $\alpha(\mathbf{s})$ and $\beta(\mathbf{s})$. This means that functionally equivalent DSS algorithms can be implemented with slightly different denoising functions $\mathbf{f}(\mathbf{s})$ and while they would converge exactly to the same results, the approximation (31) might yield completely different values. In fact, by tuning $\alpha(\mathbf{s})$, $\beta(\mathbf{s})$ or both, the approximation $\hat{g}_\mathbf{s}(\mathbf{s})$ could be made to yield any desired function $h(\mathbf{s})$ which needs have no correspondance to the true $g_\mathbf{s}(\mathbf{s})$.

Due to $\alpha(\mathbf{s})$ and $\beta(\mathbf{s})$, it seems virtually impossible to write down a simple approximation of $g_\mathbf{s}(\mathbf{s})$ that could not go wrong with a malevolent choice of $\mathbf{f}(\mathbf{s})$. In the following, however, we argue that Eq. (31) is in most cases a good approximation and it is usually easy to check

[6] For unsymmetric denoising $\mathbf{D}^*$ there is no objective function that is optimised.

whether it behaves as desired—yields values which are monotonic in SNR. If it doesn't, $\alpha(s)$ and $\beta(s)$ can be easily tuned to correct this.

Let us first check what would be the DSS algorithm minimising $\hat{g}_{\mathbf{s}}(\mathbf{s})$. Obviously, the approximation is good if the algorithm turns out to use a denoising similar to $\mathbf{f}(\mathbf{s})$. Substituting $\mathbf{s} = \mathbf{w}^T\mathbf{X}$ in Eq. (31) and deriving the DSS algorithm corresponding to $\hat{g}_{\mathbf{s}}$ similarly as in Eqs. (17)–(18) results in:

$$\mathbf{w}^+ = \mathbf{X}[\mathbf{f}^T(\mathbf{s}) + \mathbf{J}^T(\mathbf{s})\mathbf{s}^T], \tag{32}$$

where $\mathbf{J}$ is the Jacobian of $\mathbf{f}$. This should conform with the corresponding steps in the nonlinear DSS (16) and (11) which uses $\mathbf{f}(\mathbf{s})$ for denoising. For this to be ture, the two terms in the square brackets should have the same form, *i.e.*, $\mathbf{f}(\mathbf{s}) \propto \mathbf{s}\,\mathbf{J}(\mathbf{s})$.

As expected, in the linear case the two algorithms are exactly the same because the Jacobian is a constant matrix and $\mathbf{f}(\mathbf{s}) = \mathbf{s}\mathbf{J}$. The denoised sources are also proportional to $\mathbf{s}\,\mathbf{J}(\mathbf{s})$ in some special nonlinear cases, for instance, when $\mathbf{f}(\mathbf{s}) = \mathbf{s}^n$.

As an example of how Eq. (31) can fail to approximate the true objective function, consider the masking based denoisings discussed in Section 3 where denoising is implemented by multiplying the source point-wise by a mask. This means that according to Eq. (31), $g(\mathbf{s})$ will be a sum of $s^2(t)$ weighted by the values of the mask. If the mask is constant w.r.t. $\mathbf{s}$, denoising is linear and Eq. (31) is an exact formula, but let us assume that the mask is computed based on the current source estimate $\mathbf{s}$.

In some cases it may be useful to normalise the mask and this could be implemented in several ways. Some possiblities that may come to mind are to normalise the maximum value or the sum of squared values of the mask. While this type of normalisation has no effect on the behaviour of DSS, it can render the approximation (31) useless. This is because a maximally flat mask usually corresponds to a source with a low SNR. However, after normalisation, the sum of values in the mask would be greatest for a maximally flat mask and this tends to produce high values of the approximation of $g(\mathbf{s})$ conflicting the low SNR.

As a simple example, consider the mask to be $m(t) = s^2(t)$. This corresponds to the kurtosis-based denoising (25). Now the sum of squared values of the mask is $\sum s^4(t)$, but so is $\mathbf{s}\mathbf{f}^T(\mathbf{s})$. If the mask were normalised by dividing by the sum of squares, the approximation (31) would always yield a constant value of one, totally independent of $\mathbf{s}$.

A better way of normalising a mask is to normalise the sum of the values. Then Eq. (31) should always yield approximately the same value if the mask and source estimate are unrelated, but the value would be greater for cases where the magnitude of the source is correlated with the value of the mask. This is usually a sign of a structured source and consequently a high SNR.

4.2 Negentropy ordering

The approximation (31) can be readily used for monitoring the convergence of DSS algorithms. It is always easy to use it for ordering the sources based on their SNR if several sources are estimated using DSS with the same $\mathbf{f}(\mathbf{s})$. However, simple ordering based on Eq. (31) is not possible if different denoising functions are used for different sources.

In these cases it is useful to order the source estimates by their negentropy which is a normalised measure of structure in the signal. Differential entropy H of a random variable is a measure of disorder and is dependent on the variance of the variable. Negentropy is a normalised quantity measuring the difference between the differential entropy of the component and a Gaussian component with the same variance. Negentropy is zero for the Gaussian distribution and non-negative for all distributions since among the distributions with a given variance, the Gaussian distribution has the highest entropy.

Calculation of the differential entropy assumes the distribution to be known. Usually this is not the case and the estimation of the distributions is often difficult and computationally demanding. Following Hyvärinen (1998), we approximate the negentropy $N(\mathbf{s})$ by

$$N(\mathbf{s}) = H(\boldsymbol{\nu}) - H(\mathbf{s}) \approx \eta_g[\hat{g}_{\mathbf{s}}(\mathbf{s}) - \hat{g}_{\mathbf{s}}(\boldsymbol{\nu})]^2, \tag{33}$$

where $\boldsymbol{\nu}$ is a normally distributed variable. The reasoning behind Eq. (33) is that $\hat{g}_{\mathbf{s}}(\mathbf{s})$ carries information about the distribution of $\mathbf{s}$. If $\hat{g}_{\mathbf{s}}(\mathbf{s})$ equals $\hat{g}_{\mathbf{s}}(\boldsymbol{\nu})$, there is no evidence of the negentropy to be greater than zero, so this is when $N(\mathbf{s})$ should be minimised. A Taylor series expansion of $N(\mathbf{s})$ w.r.t. $\hat{g}_{\mathbf{s}}(\mathbf{s})$ around $\hat{g}_{\mathbf{s}}(\boldsymbol{\nu})$ yields the approximation (33) as the first non-zero term.

Comparison of signals extracted with different optimisation criteria presumes that the weighting constants η_g should be known. We propose that η_g can be calibrated by generating a signal with a known, nonzero negentropy. Negentropy ordering is most useful for signals which have a relatively poor SNR—the signals with a good SNR will most likely be selected in any case. Therefore we choose our calibration signal to have SNR of 0 dB, *i.e.*, it contains equal amounts of signal and noise in terms of energy: $\mathbf{s}_s = (\boldsymbol{\nu} + \mathbf{s}_{\text{opt}})/\sqrt{2}$, where $\mathbf{s}_{\text{opt}}$ is a pure signal having no noise. It obeys fully the signal model implicitly defined by the corresponding denoising function $\mathbf{f}$. Since $\mathbf{s}_{\text{opt}}$ and $\boldsymbol{\nu}$ are uncorrelated, $\mathbf{s}_s$ has unit variance. The entropy of $\boldsymbol{\nu}/\sqrt{2}$ is

$$H(\boldsymbol{\nu}/\sqrt{2}) = H(\boldsymbol{\nu}) + \log 1/\sqrt{2} = H(\boldsymbol{\nu}) - 1/2 \log 2\,. \tag{34}$$

Since the entropy can only increase by adding a second, independent signal $\mathbf{s}_{\text{opt}}$, $H(\mathbf{s}_{\mathbf{s}}) \geq H(\boldsymbol{\nu}) - 1/2 \log 2$. It thus holds $N(\mathbf{s}_s) = H(\boldsymbol{\nu}) - H(\mathbf{s}_s) \leq 1/2 \log 2$. One can usually expect that $\mathbf{s}_{\text{opt}}$ has a lot of structure, *i.e.*, its entropy is low. Then its addition to $\boldsymbol{\nu}/\sqrt{2}$ does not significantly increase the entropy. It is therefore often reasonable to approximate

$$N(\mathbf{s}_s) \approx 1/2 \log 2 = 1/2\,\text{bit}, \tag{35}$$

where we chose base-2 logarithm yielding bits. Depending on $\mathbf{s}_{\text{opt}}$, it may also be possible to compute the negentropy of $N(\mathbf{s}_s)$ exactly. This can then be used instead of the approximation (35).

The coefficients η_g in Eq. (33) can now be solved by requiring that the approximation (33) yields Eq. (35) for $\mathbf{s}_s$. This results in

$$\eta_g = \frac{1}{2(\hat{g}(\mathbf{s}_s) - \hat{g}(\boldsymbol{\nu}))^2}\text{bit} \tag{36}$$

and finally, substitution of the approximation of the objective function (31) and Eq. (36) into Eq. (33) yields the calibrated approximation of the negentropy:

$$N(\mathbf{s}) \approx \frac{\left[\mathbf{s}\,\mathbf{f}^T(\mathbf{s}) - \boldsymbol{\nu}\,\mathbf{f}^T(\boldsymbol{\nu})\right]^2}{2\left[\mathbf{s}_s\,\mathbf{f}^T(\mathbf{s}_s) - \boldsymbol{\nu}\,\mathbf{f}^T(\boldsymbol{\nu})\right]^2}\,\text{bit}. \tag{37}$$

4.3 Spectral shift

In the classical power method, the convergence speed depends on the ratio of the largest eigenvalues, $|\lambda_1/\lambda_2|$, where $|\lambda_1| > |\lambda_2|$. If this ratio is close to unity, the matrix multiplication (4) does not promote the largest eigenvalue effectively compared to the second largest eigenvalue.

The convergence speed in such cases can be increased by so called spectral shift[7] which modifies the eigenvalues without changing the fixed points. At the fixed point of the linear DSS,

$$\lambda \mathbf{w} = \mathbf{X}\mathbf{f}_{\mathrm{lin}}^T(\mathbf{s})\,. \tag{38}$$

Then it also holds that $(\lambda + \lambda_\beta)\mathbf{w} = \mathbf{X}\mathbf{f}_{\mathrm{lin}}(\mathbf{s})^T + \lambda_\beta \mathbf{w}$ for any λ_β. The additional term simply adds λ_β to all eigenvalues. The spectral shift modifies the ratio of two largest eigenvalues[8] which becomes $|(\lambda_1 + \lambda_\beta)/(\lambda_2 + \lambda_\beta)| > |\lambda_1/\lambda_2|$, provided that λ_β is negative but not much smaller than $-\lambda_2$. This can greatly increase the convergence speed of the classical power method.

However, for very negative λ_β, some eigenvalues will become negative. In fact, if λ_β is small enough, the absolute value of the originally smallest eigenvalue will exceed that of the originally largest eigenvalue. Iterations of linear DSS will then minimise the eigenvalue rather than maximise it. It does not seem very useful to minimise $g(\mathbf{s})$, a function that measures the SNR of the sources. But as we saw with negentropy and the approximation (33) of it, values $g(\mathbf{s}) < g(\boldsymbol{\nu})$ are, in fact, indicative of signal. A reasonable selection for λ_β is thus $-\boldsymbol{\nu}\mathbf{f}_{\mathrm{lin}}^T(\boldsymbol{\nu})$ which leads linear DSS to extremise $g(\mathbf{s}) - g(\boldsymbol{\nu})$ or, equivalently, to maximise the negentropy approximation (33).

A side effect for this type of spectral shift is that the estimate $\mathbf{w}$ of the principal direction changes its sign at each iteration if the eigenvalue is negative. This needs to be kept in mind when determining the convergence of DSS.

Above, the spectral shift has been applied to the eigenvalues of the matrix $\mathbf{Z}\mathbf{Z}^T$. However, in DSS the spectral shift can be embedded in the denoising of the source using $\beta(\mathbf{s})$ according to Eq. (21) because $\beta(\mathbf{s}) = \lambda_\beta/T$. From now on, we use $\beta(\mathbf{s})$ to implement the spectral shift.

Unlike in linear DSS, the approximation (31) may not accurately represent the objective function in nonlinear DSS. Consequently, nonlinear DSS does not necessarily optimise the eigenvalues λ defined by Eq.(38). As we argued before, Eq. (31) nevertheless offers a good approximation $\hat{g}_{\mathbf{s}}(\mathbf{s})$. Hence we suggest a spectral shift

$$\beta = -\hat{g}_{\mathbf{s}}(\boldsymbol{\nu})/T \tag{39}$$

for nonlinear DSS, too. One way to improve the efficiency of this approach is to try to scale the denoising such that a Gaussian noise signal always has a similar contribution in the denoised signal. For example, if the denoising is implemented by masking the source signal, the contribution of a fixed amount of Gaussian noise to the denoised source signal can be equalised by normalising the sum of the masking components.

[7] The set of the eigenvalues is often called eigenvalue spectrum.

[8] Since the denoising operation presumably preserves some of the signal and noise, it is reasonable to assume that all eigenvalues are originally positive.

It is not necessary to base the spectral shift on a global approximation of $g_{\mathbf{s}}(\boldsymbol{\nu})$. An alternative is to linearise $\mathbf{f}(\mathbf{s})$ around the current source estimate $\mathbf{s}$ and use this to compute $\beta(\mathbf{s})$ as follows:

$$\hat{\mathbf{f}}(\boldsymbol{\nu}) = \mathbf{f}(\mathbf{s}) + (\boldsymbol{\nu} - \mathbf{s})\mathbf{J}(\mathbf{s}) \tag{40}$$

$$\begin{aligned} \beta(\mathbf{s}) &= -\hat{g}_{\mathbf{s}}(\boldsymbol{\nu})/T = -\boldsymbol{\nu}[\mathbf{f}(\mathbf{s}) + (\boldsymbol{\nu} - \mathbf{s})\mathbf{J}(\mathbf{s})]^T/T \\ &= -\operatorname{tr}\mathbf{J}(\mathbf{s})/T \end{aligned} \tag{41}$$

The last step follows from the fact that the elements of $\boldsymbol{\nu}$ are mutually uncorrelated and have zero mean and unit variance. If denoising is instantaneous, $\mathbf{f}(\mathbf{s}) = [f_1(s(1))\, f_2(s(2))\, \ldots]$, the shift can be written as $\beta = -\sum_t f'_t(s(t))/T$. This is the spectral shift used in FastICA (Hyvärinen, 1999), but it has been justified as an approximation to Newton's method and our analysis thus provides a novel interpretation.

In general, iterations converge faster with the FastICA-type spectral shift (41) than with the fixed shift (39) but the latter has the benefit that no gradients need to be computed. This is important when the denoising is defined by a complex nonlinear procedure such as median filtering.

A well known example where the spectral shift by the eigenvalue of a Gaussian signal is useful is the mixture of both super- and sub-Gaussian distributions. DSS algorithm designed for super-Gaussian distributions would lead to $\lambda > \lambda_G$ for super-Gaussian and $\lambda < \lambda_G$ for sub-Gaussian distributions, λ_G being the eigenvalue of the Gaussian signal. By shifting the eigenvalue spectrum by $-\lambda_G$, the most non-Gaussian distributions will result in the largest absolute eigenvalues regardless of whether the distribution is super- or sub-Gaussian. By using the spectral shift it is therefore possible to extract both super- and sub-Gaussian distributions with a denoising scheme which is designed for one type of distributions only.

Consider for instance $\mathbf{f}(\mathbf{s}) = \tanh\mathbf{s}$ which can be used as denoising for sub-Gaussian signal while, as we saw before, $\mathbf{s} - \tanh\mathbf{s} = -(\tanh\mathbf{s} - \mathbf{s})$ is a suitable denoising for super-Gaussian signals. This shows that depending on the choice of β, DSS can find either sub-Gaussian ($\beta = 0$) or super-Gaussian ($\beta = -1$) sources. With the FastICA spectral shift (41), β will always lie in the range $-1 < \beta \leq \tanh^2 1 - 1 \approx -0.42$. In general, β will be closer to -1 for super-Gaussian sources which shows that FastICA is able to adapt its spectral shift to the source distribution.

None of the above methods always work for nonlinear DSS. Sometimes the spectral shift turns out to be either too modest or strong, leading to slow convergence or lack of convergence, respectively. For this reason, we suggest a simple stabilisation rule: instead of updating $\mathbf{w}$ into $\mathbf{w}_{\text{new}}$ defined by (12), it is updated into

$$\mathbf{w}_{\text{adapted}} = \operatorname{orth}(\mathbf{w} + \gamma\Delta\mathbf{w}) \tag{42}$$

$$\Delta\mathbf{w} = \mathbf{w}_{\text{new}} - \mathbf{w}\,, \tag{43}$$

where γ is the step size and sthe orthogonalisation has been added in case several sources are to be extracted. Originally $\gamma = 1$, but if the conseqtive steps are taken in nearly opposite directions, *i.e.*, the angle between $\Delta\mathbf{w}$ and $\Delta\mathbf{w}_{\text{old}}$ is greater than 179°, then $\gamma = 0.5$ for the rest of the iterations. A stabilised version of FastICA has been proposed by Hyvärinen (1999) as well and procedure similar to the one above has been used. The different speedup techniques considered above, and some additional ones, are studied further in Valpola and Särelä (2004).

Sometimes there are several signals with similar large eigenvalues. It may then be impossible to use spectral shift to accelerate their separation significantly because of small eigenvalues that would assume very negative values exceeding the signal eigenvalues in magnitude. In that case, it may be beneficial to first separate the subspace of the signals with large eigenvalues from the smaller ones. Spectral shift will then be useful in the signal subspace.

4.4 Detection of overfitting

In exploratory data analysis DSS is very useful for giving a better insight to the data using a linear factor model. However, it is possible that DSS extracts structures that are not actually present in the data but are generated by the denoising function, *i.e.*, the results may be due to overfitting.

Overfitting in ICA has been extensively studied by Särelä and Vigário (2003). It was observed that it typically results in signals that are mostly inactive, except for a single spike. In DSS the outlook of the overfitted results depends on the denoising criterion. The results of exploratory DSS should thus be treated with a healthy amount of scepticism.

To detect an overfitted result, one should know how it looks like. As a first approximation, DSS can be performed with same amount of i.i.d Gaussian data. Then all the results present cases of overfitting. Even better characterisation of the overfitting results can be obtained by mimicking the actual data characteristics as well as possible. In that case it is important to make sure that the structure assumed by the signal model has been broken. Both the Gaussian overfitting test and the more advanced test are used throughout the experiments in Secs. 5.2–5.5.

Note that in addition to visual test, the methods described above provide us with a quantitative measure as well. Using the negentropy approximation (37), we can set a threshold under which the sources are very likely overfits and do not carry much real structure. In the simple case of linear DSS, the negentropy can be approximated easily using the corresponding eigenvalue.

5. Experiments

In this section we demostrate the separation capabilities of the algorithms presented earlier. First, in Sec. 5.1, we separate artificial signals with different DSS schemes, some of which can be implemented by FastICA (1998), Hyvärinen (1999). Furthermore, we compare the results to one standard ICA algorithm, JADE (1999), Cardoso (1999). In Secs. 5.2–5.5 linear and nonlinear DSS algorithms are applied extensively in the study of magnetoencephalograms (MEG). Finally, in Sec. 5.6, recovery of CDMA signals is demonstrated. In each experiment after the case of artificial sources, we first discuss the nature of the expected underlying sources. Then we describe this knowledge in the form of denoising.

5.1 Artificial signals

Artificial signals were mixed to compare different DSS schemes and JADE (Cardoso, 1999). Ten mixtures of the five sources were produced and independent white noise was added with different SNRs ranging from nearly noiseless mixtures of 50dB to -10dB, a very noisy case. The original sources and the mixtures are shown in Figs. 6a and 6b respectively. The mixtures shown have SNR of 50 dB.

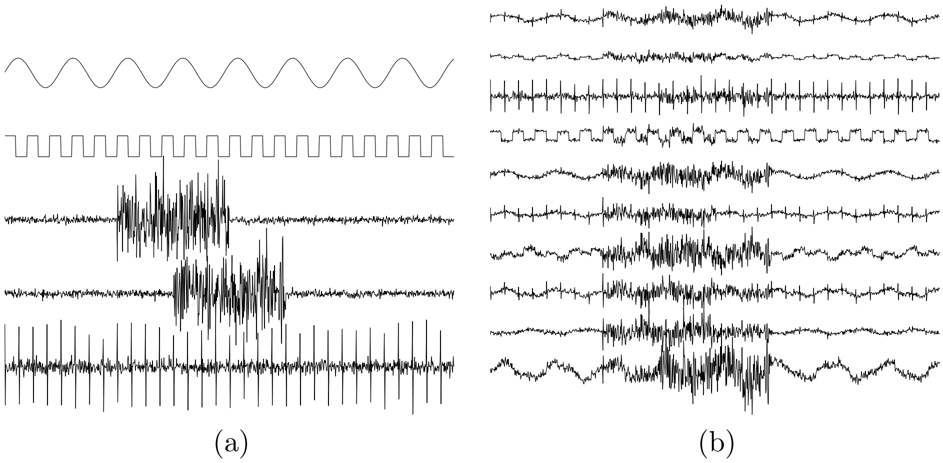


Figure 6: *(a) Five artificial signals with simple frequency content (signals #1 and #2), simple on/off non-stationarity in time domain (signals #3 and #4) or quasi-periodicity (signal #5). (b) Ten mixtures of the signals in (a).*

5.1.1 Linear denoising

In this section, we show how the simple linear denoising schemes described in Sec. 3.1 can be used to separate the artificial sources. These schemes require prior knowledge about the source characteristics.

The base frequencies of the first two signals were assumed to be known. Thus two band-pass filtering masks were constructed around these base frequencies. The third and fourth source estimates were known to have periods of activity and non-activity. Third was known to be active in the second quadrant and the fourth a definite period in the latter half. They were denoised using binary masks in time domain. Finally, the fifth source had a known quasi-periodic repetition rate and was denoised using the averaging procedure described in Sec. 3.1.4 and Fig. 2. Since all the five denoisings are linear, five separate filtered data sets were produced and PCA was used to recover the principal components. The separation results are described in Sec. 5.1.3 together with the results of other DSS schemes and JADE.

5.1.2 Nonlinear exploratory denoising

In this section, we describe an exploratory source separation of the artificial signals. One author of this paper gave the mixtures to the other author whose task was to separate the original signals. The author did not receive any additional information, so he was forced to apply a blind approach. He chose to use the masking procedure based on the instantaneous energy estimate, described in Sec. 3.2. To enable the separation of both sub- and super-Gaussian sources in the MAP-based signal-variance-estimate denoising, he used the spectral shift (39). To ensure convergence, he used the 179-rule to control the step size γ (42). Finally, he did not smooth $s^2(t)$ but used it directly as the estimate of the total instantaneous variance $\sigma^2_{\mathrm{tot}}(t)$.

Based on the separation results of the variance-based DSS, he further devised specific masks for each of the source. He chose to denoise the first source in frequency domain with a strict band-pass filter around the main frequency. The author decided to denoise the second source by a simple denoising function $\mathbf{f}(\mathbf{s}) = \mathrm{sign}(\mathbf{s})$. This makes quite an accurate signal model though it neglects the behaviour of the source in time. The third and fourth signal seemed to have periods of activity and non-activity. He found an estimate for the active periods by inspecting the instantaneous variance estimates $\mathbf{s}^2$, and devised simple binary masks. The last signal seemed to consist of alternating positive and negative peaks with fixed inter-peak-interval as well as some additive Gaussian noise. The signal model was tuned to model the peaks only.

5.1.3 Separation results

In this section, we compare the separation results of the linear denoising (Sec. 5.1.1), variance-based denoising and adapted denoising (Sec 5.1.2) to other DSS algorithms. In particular, we compare to the popular denoising schemes $\mathbf{f}(\mathbf{s}) = \mathbf{s}^3$ and $\mathbf{f}(\mathbf{s}) = \tanh(\mathbf{s})$, suggested for use with FastICA (1998). We compare to JADE (Cardoso, 1999) as well. During sphering in JADE, the number of dimensions were either reduced ($n = 5$) or all the ten dimensions were kept ($n = 10$).

We restrained from using deflation in all the different DSS schemes to avoid suffering from cumulative errors in separation of the first sources. Instead one source was extracted with each of the masks several times using different initial vector $\mathbf{w}$ until five sufficiently different source estimates were reached (see Himberg and Hyvärinen, 2003, Meinecke et al., 2002, for further possibilities along these lines). Deflation was only used if no estimate could be found for all the 5 sources. This was often the case for poor SNR under 0dB.

To get some idea of statistical significance of the results, each algorithm was used to separate the sources ten times with the same mixtures, but different measurement noises. The average SNRs of the sources are depicted in Fig. 7. The straight line above all the DSS schemes represents the optimal separation. It is achieved by calculating the unmixing matrix explicitly using the true sources.

With outstanding SNR (> 20 dB), linear DSS together with JADE and kurtosis-based DSS seem to perform worst, while the other, nonlinear DSS approaches: tanh-based, sophisticated variance estimate and the adapted one seem to perform better. The gap between these groups is more that two standard deviations of the 10 runs, making the difference statistically significant. In practice the difference in performance probably does not matter.

With moderate SNRs (between 0 and 20 dB), all algorithms perform quite alike. With poor SNR (< 0 dB), the upper group consist of the linear and adapted DSS as well as the

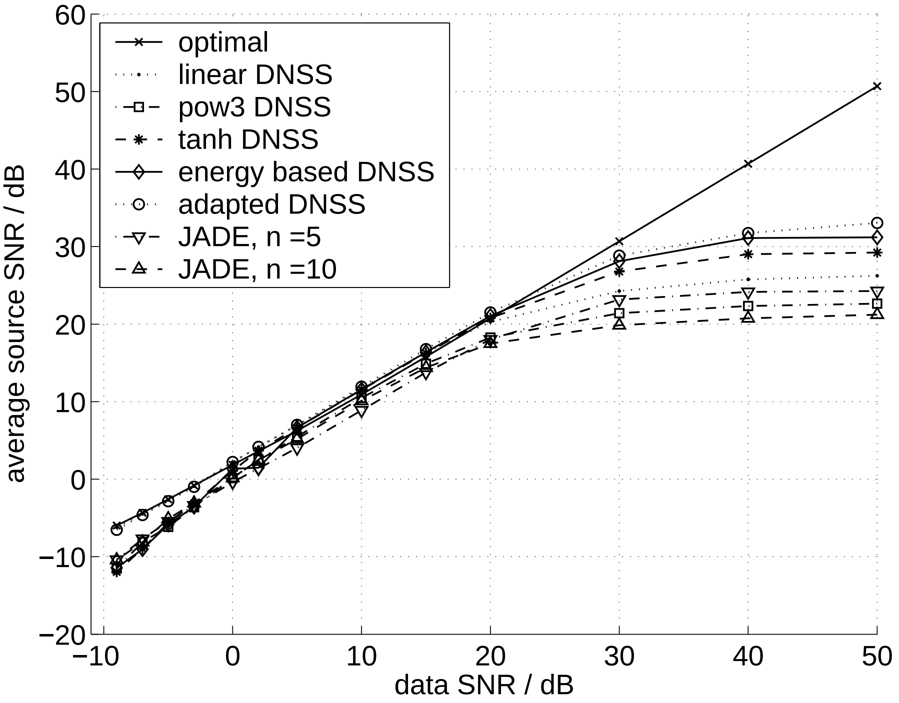


Figure 7: *Average SNRs for the estimated sources averaged over 10 runs.*

optimal one and the lower group consists of the blind approaches. This seems reasonable, since it makes sense to rely more on prior knowledge when the data is very noisy.

5.2 Exploratory source separation in rhythmic MEG data

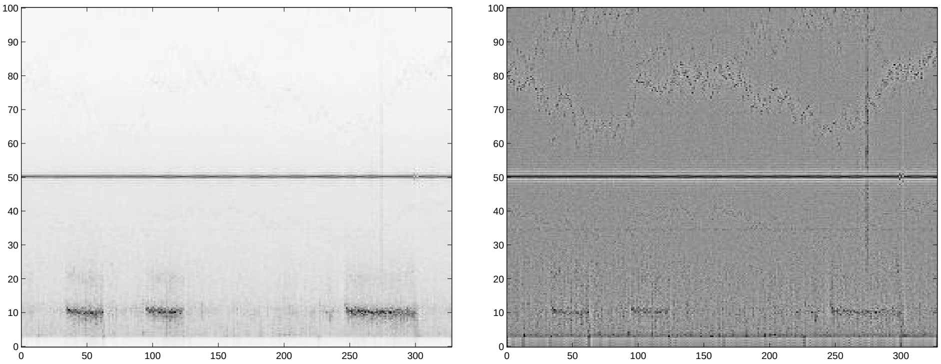
In biomedical research it is usual to design detailed experimental frameworks to examine interesting phenomena. Hence it offers a nice field of application for both blind and specialised DSS schemes. In the following we test the developed algorithms in signal analysis of magnetoencephalograms (MEG, Hämäläinen et al., 1993). MEG is a completely non-invasive brain imaging technique measuring the magnetic fields on scalp caused by syncronous activity in the cortex.

Since the early EEG and MEG recordings, cortical electromagnetic rhythms have played an important role in clinical research, *e.g.*, in detection of various brain disorders, and in studies of development and aging. It is believed that the spontaneous rhythms, in different parts of the brain, form a kind of resting state that allows for quicker responses to stimuli by those specific areas. For example deprivation of visual stimuli by closing one's eyes induces so called α-rhythm on the visual cortex, characterised by a strong 8–13 Hz frequency component. For a more comprehensive discussion regarding EEG and MEG, and their spontaneous rhythms, *c.f.,* Niedermeyer and Lopes da Silva (1993), Hämäläinen et al. (1993).

In this paper, we examine an MEG experiment where the subject is asked to relax by closing her eyes (producing α-rhythm). There is also a control state where the subject has her eyes open. The data has been sampled with $f_s = 200$ Hz, and there are $T = 65536$ time samples giving total of more than 300 seconds of measurement. The magnetic fields are measured using a 122-channel MEG device. Some source separation results of this data have been reported by Särelä et al. (2001). Prior to any analysis, the data is high-pass filtered with cut-off frequency of 1 Hz, to get rid of the dominating very low frequencies.

5.2.1 Denoising in rhythmic MEG

Examination of the average spectrogram in Fig. 8a reveals clear structures indicating the existence of several, presumably distinct, phenomena. The burst-like activity around 10 Hz and the steady activity at 50 Hz dominate the data, but there seem to be some weaker phenomena as well, *e.g.*, on higher frequencies than 50 Hz. To amplify these, we not only sphere the data spatially but temporally as well. This temporal decorrelation actually makes the separation harder, but enables the finding of the weaker phenomena. The normalised and filtered spectrogram is shown in Fig. 8b.



(a) (b)

Figure 8: *(a) Averaged spectrogram of all the 122 MEG channels. (b) Frequency normalised spectrogram.*

The spectrogram data seems well suited for demonstrating the exploratory data analysis use of DSS. As some of the sources seem to have quite steady frequency content in time, but others changing in time, we used two different time-frequency-analyses as described in Sec. 3.1.3 with lengths of the spectra $T_f = 1$ and $T_f = 256$. The first spectrogram is then actually the original frequency-normalised and filtered data with only time information.

We apply the several noise reduction principles based on the estimated variance of the signal and the noise discussed in Sec. 3.2. Specifically, the power spectrogram of the source estimate is smoothed over time and frequency using 2-D convolution with Gaussian windows. The standard deviations of the Gaussian windows were $\sigma_t = 8/\pi$ and $\sigma_f = 8/\pi$. After this, the instantaneous estimate of the source variance is found using Eq. (28). Then we get the denoised source estimate using Eq. (29) together with the spectral shift (39). Initially we have set $\mu = 1.3$. This is then decreased by 0.1 every time DSS has converged, until $\mu < 1$ is reached. Finally, the new projection vector is calculated using the stabilised version (42), (43) with the 179-rule in order to ensure convergence.

5.2.2 Separation results

The separated signals, depicted in Fig. 9, include several interesting sources. Due to poor contrast in Fig. 9, we show enhanced and smoothed spectrograms of selected interesting, but low contrast, components (#1, #2, #3 and #18) in Fig. 10. In the rest of this paper, we always show the enhanced spectrograms of extracted components. First of all, there exist several sources with α-activity (#1, #4 and #7 for example). The second and 6th source are clearly related to the power-line. The fourth source depicts an interesting signal caused probably by some anomaly in either the measuring device itself or its physical surroundings. In source #18, there is another, presumably artefactual source, composed of at least two steady frequencies around 70 Hz.

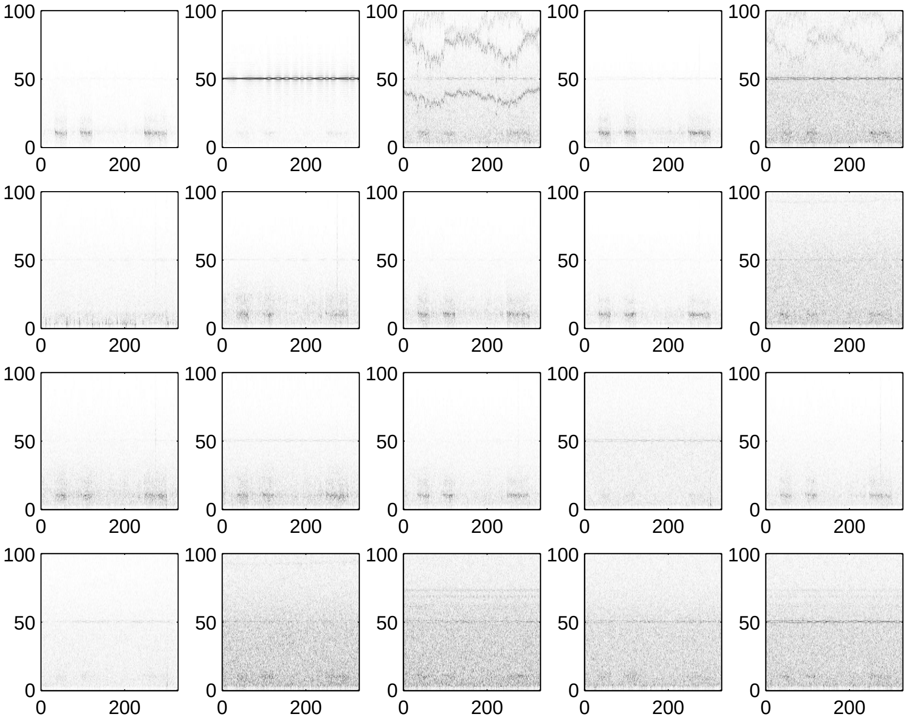


Figure 9: *Spectrograms of the extracted components (comps. 1–5 on the topmost row)*

The DSS approach described above seems to be reliable and fast: the temporal decorrelation of the data enabled the finding of very weak sources and yet we found several clear α-sources as well. Valpola and Särelä (2004) have further studied the convergence speed, reliability and stability of DSS with various speedup methods, such as the spectral shift used in FastICA. Convergence speed exceeding standard FastICA by 50 % was reported.

Though quite a clear separation of the sources was achieved, some cross-talk between the signals remains. We now turn to more specific masks by taking advantage of the structures

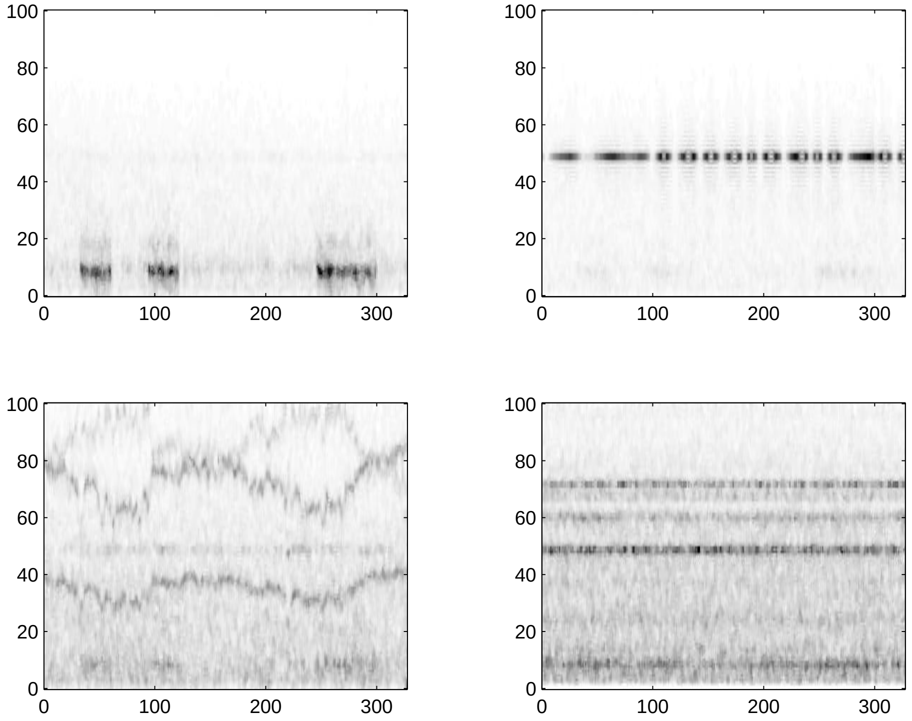


Figure 10: *Enhanced and smoothed spectrograms of the selected components (correspond to sources #1, #2, #3 and #18 in Fig. 9)*

uncovered by variance-based masking. First we take a look at the α-subspace. Then, in Sec. 5.4 the anomalous signals (bottom row of the Fig. 10) are inspected further. Finally, in Sec. 5.5 we show that with specific knowledge it is possible to find even very weak phenomena in MEG data using DSS.

5.3 Adaptive extraction of the α-subspace

Exploratory source separation revealed several sources with significant activity on the α-band (*e.g.*, #3 and #6). But in previous section, only general noise reduction principles were used. In this section, we further tune the masks to achieve maximal signal to noise ratio in the α-subspace.

5.3.1 Denoising of the α-sources

The α-sources have a characteristic frequency around 10 Hz. Furthermore, some of the sources (see especially #3 in Fig. 9) have waveforms differing considerably from a pure sinusoid. This is manifested in the significant activity around 20 Hz as well. Another noticable characteristics for the α-sources is that they appear in bursts in time, specifically being active, when the subject has her eyes closed.

These two characteristics (10 Hz and 20 Hz frequency content and on/off activity) are directly exploited to make a band-pass filter in the spectrogram (see Sec. 3.1). Since this denoising is completely linear, we have chosen to use the PCA type of DSS, as described by Eqs. (3), (4) and (5).

5.3.2 Separation results

We calculated all the 122 principal components related to the α-masked data. Spectrograms of the components having the 20 largest eigenvalues are shown in Fig. 11. They are in descending order according to corresponding eigenvalues.

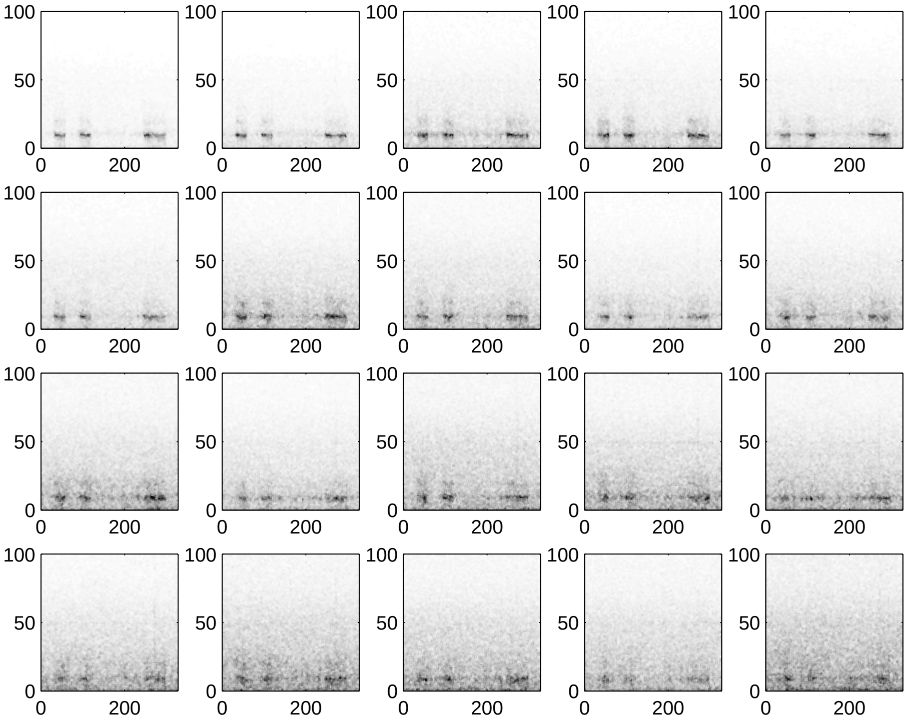


Figure 11: *Spectrograms of the components of the extracted α-subspace (comps. 1–5 on the topmost row).*

The data seem to contain more α-components than previously found with ICA or decorrelation types of BSS (Vigário and Matilainen, 2004). Additionally, there are components (#4 and #11 for example) with considerable activity on the secondary band around 20 Hz.

The tuned α-mask is quite a strict one, and there is reason to believe that part of the found α-related components are definitely overfitted results (see Sec. 4.4). To get an estimate of the number of α-related sources we can trust, we mimicked the $1/f$-type spectrum typical for MEG data. The maximum eigenvalue reached from this data with the α-mask was very close to the eigenvalue of the 18th α-source, the 17th and 16th being rather close as well.

Thus it is not reasonable to assume that there exists much more than 15 α-components in the data.

5.3.3 Rotation of the α-subspace

There is a wide literature concerning the localisation of MEG and EEG sources (*c.f.,* Hämäläinen et al., 1993, Niedermeyer and Lopes da Silva, 1993). Though we do not intend to go into details here, we note that there is little reason to believe that the linear α-mask used above would actually rotate the α-subspace in physiologically meaningful components. It probably only effectively separates between the noise and the signal subspace.

To actually find a meaningful separation, we need an additional criterion for the rotation. Recently, ICA has been suggested for this task by Vigário et al. (2000), Jung et al. (2000), Tang and Pearlmutter (2003). It can find a meaningful rotation only in the case where the sources have non-Gaussian distributions. This might easily not be the case in burst-like activity of α-sources (see Vigário et al., 2000, for an example). Furthermore, the separation is not guaranteed even if the distributions are non-Gaussian. For instance, the rotational ambiguity of a sine and a cosine remains even if they are modulated with envelopes, if the envelopes are similar enough.

Another possible approach, more in lines with the philosophy of this paper, is to find properties that the sources might differ in. One interesting possibility is to consider further rotation of the α-subspace by the mutual strengths of the base frequence (10 Hz) and first harmonics (20 Hz) of the α-activity.

We propose to achieve the rotation by another linear DSS scheme in the resphered α-subspace of 16 α-related components, where only the active portions in time have been preserved. The denoising is based on band-pass filtering around 20 Hz. From the results shown in Fig. 12 it can be seen that the first components have the highest amount of 20 Hz compared to the 10 Hz. Similarly, the last components are ones having the least amount of 20 Hz.

5.4 Adaptive extraction of artefacts

Exploratory DSS separated some presumable artefacts as well. In Fig. 9, the 5th component has a curious wandering frequency around 30–40 Hz and some higher harmonics. Another interesting phenomenon is seen in the 14th component, with its steady frequencies around 60 Hz. In this section we adaptively maximise SNR of these signals. We as well check whether some weaker, related signals come forward when the masks are adapted.

5.4.1 Denoising of the steady frequency components

The steady frequencies can be denoised with detailed linear band-pass filter tuned to the dominating frequencies. We extracted four components using band-pass filtering around 68–69 Hz. The mask was adapted to the dominating frequency of each of the four sources and four new components were estimated with each of the masks.

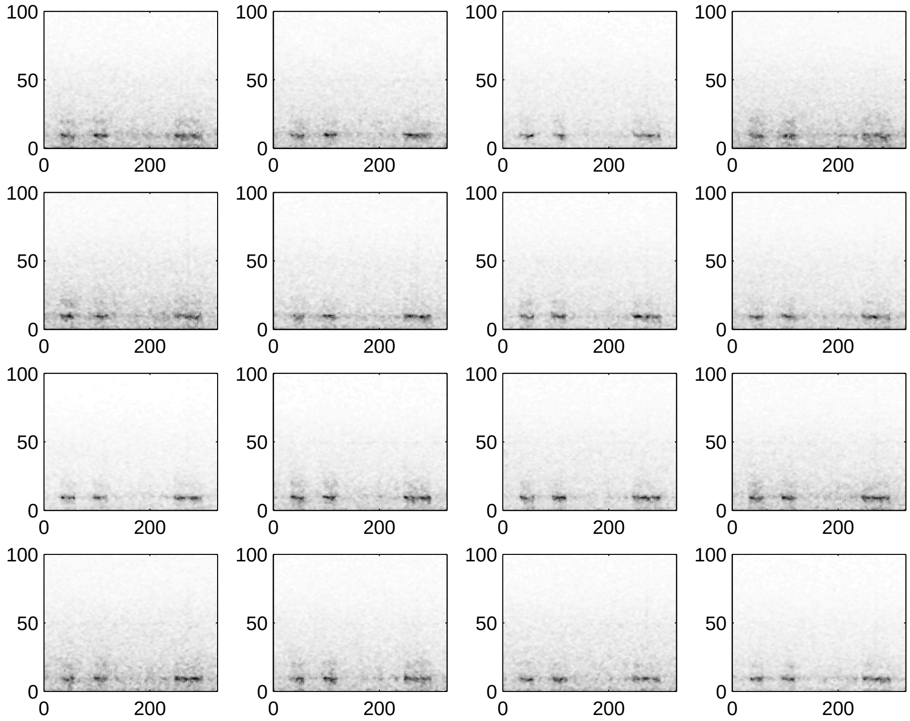


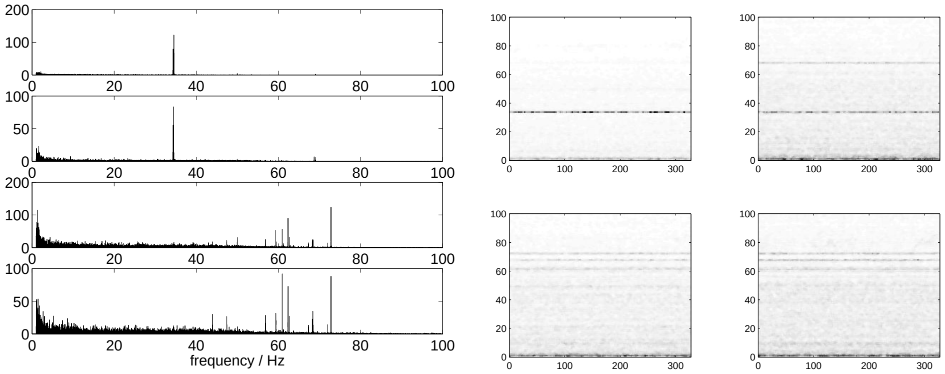
Figure 12: *Spectrograms of the components of the rotated α-subspace (comps. 1–4 on the topmost row).*

5.4.2 Separation results

Using the adaptive procedure described above, 16 components were reached. Some of them are highly mutually correlated because their corresponding masks are similar. The data seemed to have a total of four different components uncorrelated to each others. The spectra and the spectrograms of the different components found are shown in Fig. 13. The first two are quite similar, having steady frequency around 34 Hz. Likewise, the last two signals have similar frequency content having several steady higher frequencies, mainly on band 60–75 Hz.

5.4.3 Denoising of the wandering frequency components

For the wandering signal (bottom left in Fig. 10), we adaptively tune a mask in the time-frequency-space so that it takes into account the slow drifting of the base frequency. The very clear 2nd and 3rd harmonics are used to aid the estimation of the base frequency. Note that the third harmonic surpasses the Nyquist frequency of $f_s/2 = 100$ Hz at certain locations and causes an aliasing effect.

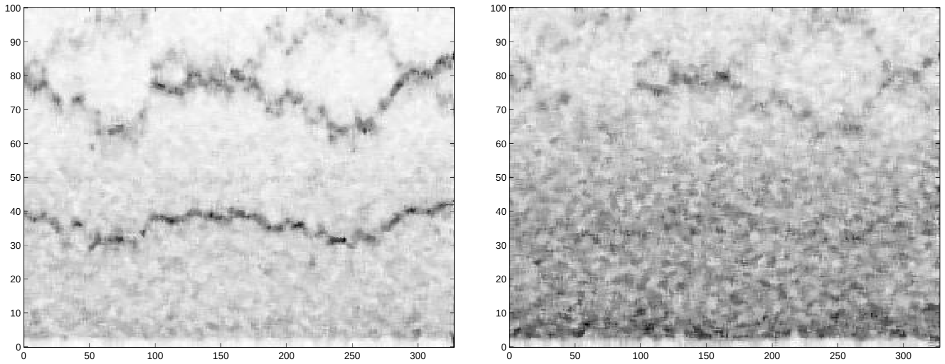


(a)   (b)

Figure 13: *(a) Power spectra of the components having several steady frequencies. (b) Smoothed spectrograms of the corresponding components (comps. 1–2 on the topmost row).*

5.4.4 Separation results

Using the DSS procedure described above, we extracted several signals having wandering frequency around 30–40 Hz and higher harmonics. Two of these are shown in Fig. 14. As the tuned mask is a very narrow one, it can see similar structure in pure Gaussian data already. Comparison of the corresponding eigenvalues revealed that all the other extracted wandering components, expect for the two shown, are caused by overfitting. The base frequency of the second source is not clearly visible but this appears to be caused by greater noise variance on the frequencies compared to the higher frequencies where the harmonics are.



(a)   (b)

Figure 14: *(a) Enhanced spectrogram of one artefact having wandering frequency around 30–40 Hz and harmonics. (b) Enhanced spectrogram of another similar component.*

5.5 Adaptive extraction of cardiac subspace in MEG

Cardiac activity causes magnetic fields as well. Sometimes these are strongly reflected in MEG and can pose a serious problem for the signal analysis of the neural phenomena of interest. In this data, however, the cardiac signals are not visible to the naked eye. Thus, we want to demonstrate the capability of DSS to extract some very weak cardiac signals, using detailed prior information in an adaptive manner.

5.5.1 Denoising of the cardiac subspace

A clear QRS complex can be extracted from the MEG data using standard BSS methods, such as kurtosis- or tanh-based denoising. Due to its sparse nature, this QRS signal can be used to estimate the places of the heart beats. With the places known, we can guide further search using the averaging DSS, as described in Sec. 3.1. Every now and then, we reestimate the QRS onsets needed for the averaging DSS.

When the estimation of the QRS locations has been stabilised, a subspace that compose of signals having activity phase-locked to the QRS complexes can be extracted.

5.5.2 Separation results

Figure 15 depicts five signals averaged around the QRS complexes, found using the procedure above[9]. The first signal presents a very clear QRS complex, whereas the second one contains the small P and the T waves. An interesting phenomenon is found in the third signal: there is a clear peak at the QRS onset, which is followed by a slow attenuation phase. We presume that it originates from some kind of relaxing state.

Two other heart related signals were also extracted. They both show a clear deflection during the QRS compex, but have as well significant activity elsewhere. These two signals might present a case of overfitting, contemplated in Sec. 4.4. To test this hypothesis, we performed DSS using the same procedure, but for reversed data. This should break the underlying repetitive structure and resulting signals should be pure overfits. The results are shown in Fig. 16. The eigenvalues corresponding to the QRS-complex and the second signal having the P and T waves are approximately 10 times higher than the principal eigenvalue of the reversed data. Thus they clearly exhibit some real structure in the data, as already expected. The eigenvalues corresponding to the last three signals are comparable to the principal eigenvalue of the reversed data, the two largest being somewhat greater. It is reasonable to expect that all the three carry some real structure as there is a nonzero correlation between the first two signals having the main cardiac responses and the overfitted component corresponding to the largest eigenvalue from the reversed data. In the three other signals, there probably occurs some overfitting as well, since the signals have similar structures as the last two signals of the actual subspace experiment shown in Fig. 15.

It is worth noticing that even the strongest component of the cardiac subspace is rather weakly present in the original data. The other components of the subspace are hardly detectable without advanced methods beyond blind source separation. This clearly demonstrates the power that DSS can provide for an exploring researcher.

[9] For clarity, two identical cycles of averaged heart beats are always shown.

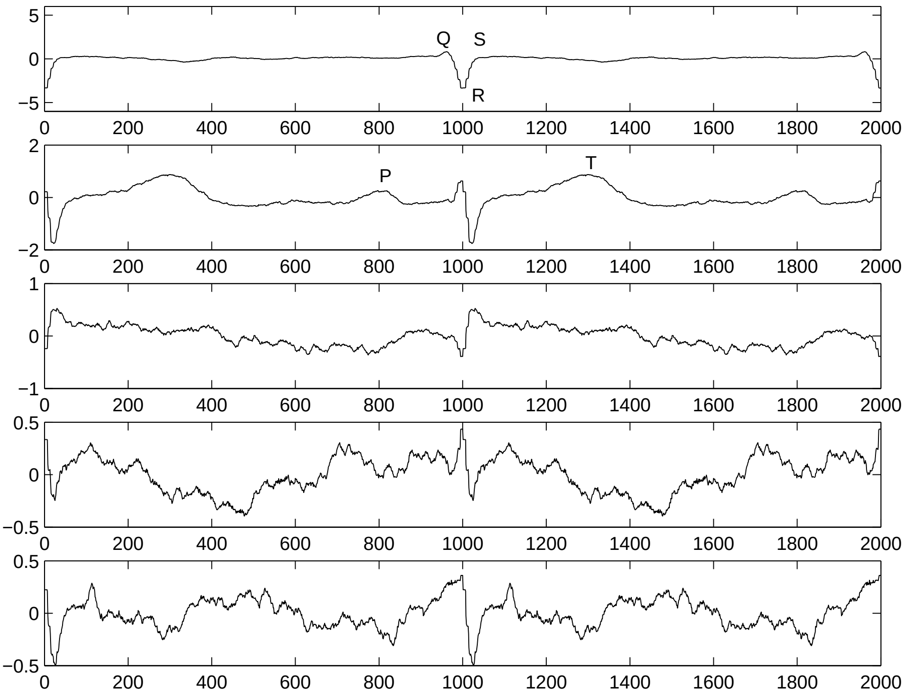


Figure 15: *Averages of three heart related signals and presumably two overfitting results.*

5.6 Signal recovery in CDMA

Mobile systems constitute another important signal processing application area, in addition to biomedical signal processing. There are several ways to allow multiple users to use the same communication channel, one being modulation scheme called code-division-multiple-access (CDMA, Viterbi, 1995). In this section we consider bit-stream recovery in a simplified simulation of a CDMA network.

In CDMA each user has a unique signature quasiorthogonal to the signatures of the other users. The user codes each complex bit[10] which he sends using this signature. This coded bit stream is transmitted through the communication channel, where it is mixed with the signals of the other transmitters. The mixture is infected with some noise as well, due to multi-path propagation, doppler shifts, interfering signals, etc.

To recover the sent bit stream, receiver decodes the signal with the known signature. Ideally then, the result would be ones and zeros repeated number of times corresponding to the signature length. In practice, noise and other interefering signals cause variation and the bits are usually extracted by majority voting.

If there are multiple paths a particular bit stream is sent to the receiver or the transmitter and receiver have multiple antennas, so called RAKE procedure can be used: The path coefficients are estimated based on the so called pilot bit streams that are fixed known bit streams and sent frequently by the transmitter. Different bit streams are then summed

[10] Here a scheme called QAM is used. There two bits are packed into one complex bit by making a 90° phase shift in the other bit.

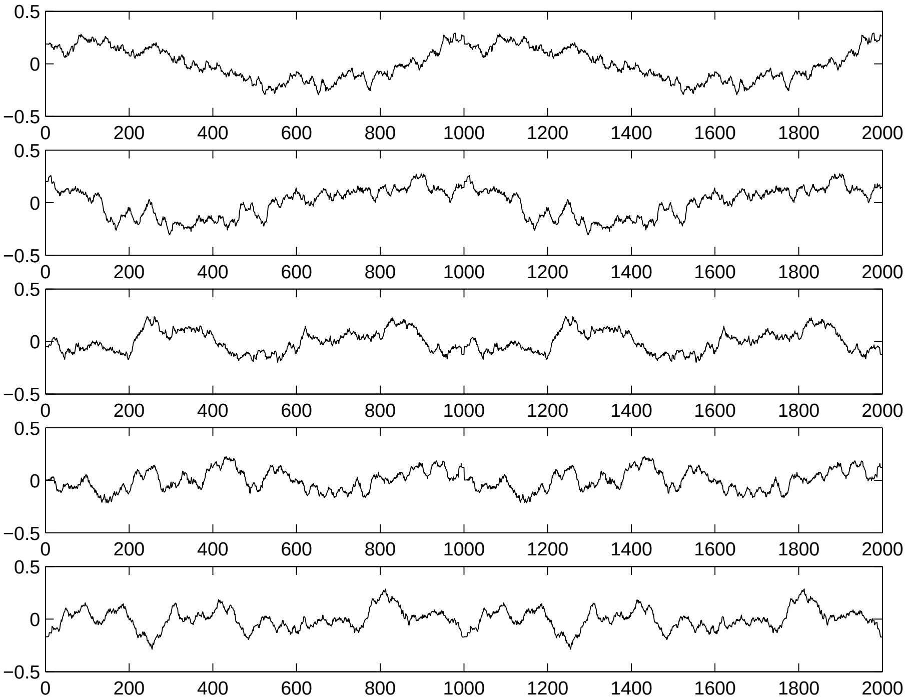


Figure 16: *Averages of five signals from the cardiac control experiment, showing clear overfittings.*

together before the majority voting. In RAKE-ICA (Raju and Ristaniemi, 2002), ICA is used to blindly separate the desired signal from the interference of other users and noise. This yields better results in the majority voting.

5.6.1 Denoising of CDMA signals

We know that the original bit stream should consist of repeated coding signatures convoluted by the original complex bits. First the bit stream is decoded using a standard detection algorithm. The denoised signal is then the recoding of the decoded bit stream.

This DSS approach is nonlinear. If the original bit-stream estimate is very inaccurate, *e.g.*, due to serious interference of other users or external noise, the nonlinear approach might get stuck in deficient local minimum. To prevent this, we first initialise by running a simpler, linear DSS. There we only exploit the fact that the signal should consist of repetitions of the signature multiplied by a complex number. The nonlinearity of the denoising is gradually increased in the first iterations.

5.6.2 Separation results

We sent 100 blocks of 200 complex bits. The sent bits were mixed using the streams of 15 other users. For simplicity we set all the path delays to zero. The signal-to-noise-ratio

(SNR) varied from -10 to 15 dB. The length of the spreading signature was 31. The mixtures were measured using three antennas. We did not consider multi-path propagation.

Figure 17 sums up the results of CDMA experiments. The comparison to the RAKE algorithm shows that DSS performs better in all situations except in the highest SNR, where RAKE is slightly better. Note that RAKE needs the pilot bits to estimate the mixing while DSS does not need them. The better performance of DSS for low SNR is explained by the fact that DSS in effect actively cancels disturbing signals while RAKE ignores them.

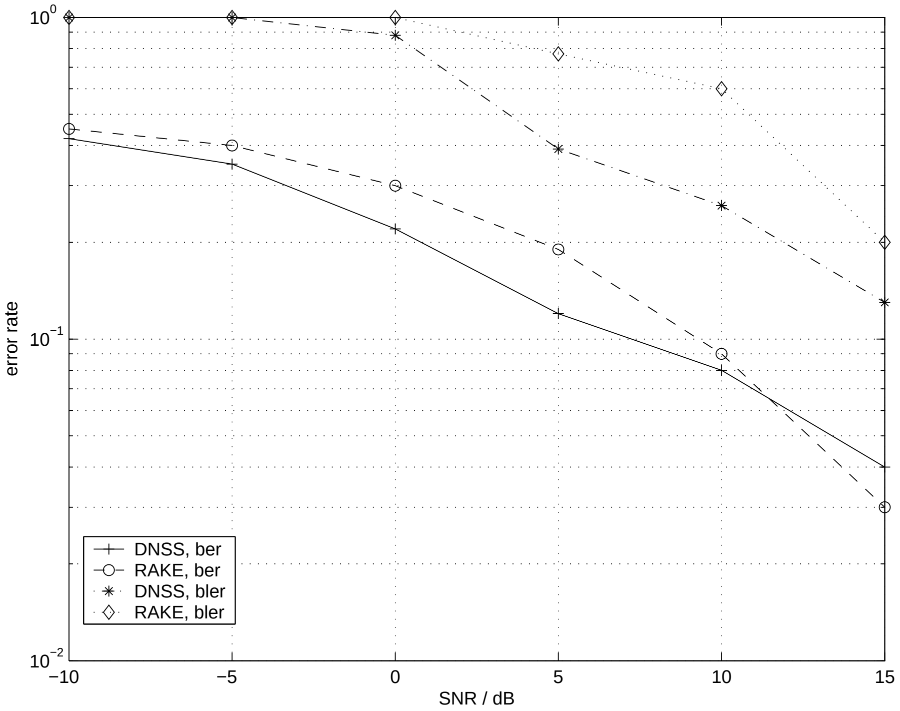


Figure 17: *Bit- and block-error rates for different SNRs for DSS and RAKE.*

CDMA bit streams consist of known headers that are necessary for standard CDMA techniques to estimate several properties of the transmission channel. In DSS framework, these become useless and can be replaced by proper signal, thus increasing the channel capacity. In addition, bits defined by the actual data such as error-correcting or check bits allow an even better denoising of the desired stream. Furthermore, it is possible to take multipath propagation into account using several delayed versions of the received signal. This should then result in a kind of averaging denoising when proper delay is used analogous to the multi-resolution spectrogram DSS described in Sec. 3.1.3. In case of moving transmitters and receivers, DSS may exploit the Doppler effect.

6. Discussion

In this paper, we developed several DSS algorithms. Moreover, DSS offers a promising framework for developing additional extensions. In this section, we first summarise the extensions that have already been mentioned in previous sections and then discuss some auxiliary extensions, as well.

We discussed online learning strategy in Sec. 3.3, where we noted that asymmetric online denoising may lead to noncorvergent DSS algorithms. However, symmetric denoising procedures performing similar functions may easily be generated.

We also noted that the masking based on the instantaneous variance in Sec. 3.2 may have problems in separating the actual sources, though it effectively separates the noise subspace from the signal subspace. We proposed a simple modification to magnify small differences between the variance estimates of different sources. Furthermore, we noted that a better founded alternative is to consider explicitly the leakage of variance between the signals. Then the variances of the signals can be decorrelated using similar techniques as suggested by Schwartz and Simoncelli (2001). This idea has been pursued further in the DSS framework (Valpola and Särelä, 2004), making the variance-based masking a very powerful approach to source separation. Furthermore, the variance-based mask saturates on large values. This reduces the tendency to suffer from outliers. However, data values that differ utterly from other data points probably carry no interesting information at all. Even more robustness would then be achieved if the mask would start to decrease on large enough values.

In this paper, we usually considered the sources to have one-dimensional structure, which is used to implement the denoising. We already applied successfully two-dimensional denoising techniques for the spectrograms. Furthermore, it was mentioned in Sec. 2 that the index t of different samples $\mathbf{s}(t)$ might refer as well to space as to time. In space it becomes natural to apply filtering in 2D or even in 3D. For example, the astrophysical ICA (Funaro et al., 2003) would clearly benefit from multidimensional filtering.

Source separation is not the only application of ICA-like algorithms. Another, important field of application is feature extraction. ICA has been used for example in extraction of features from natural images, similar to those that are found in the primary visual cortex (Olshausen and Field, 1996). It is reasonable to consider DSS extensions that have been suggested in the field of feature extraction as well. For instance, until now we have only considered extraction of multiple components by forcing the projections to be orthogonal. However, nonorthogonal projections resulting from overcomplete representations provide some clear advantages, especially in sparse codes (Földiák, 1990), and may be found useful in DSS framework as well.

Throughout this paper, we have considered linear mapping from the sources to the observations but nonlinear mappings can be used, too. One such approach is slow feature analysis (SFA, Wiskott and Sejnowski, 2002) where the observations are first expanded nonlinearly and sphered. The expanded data is then high-pass filtered and projections minimising the variance are estimated. Due to the nonlinear expansion, it is possible to stack several layers of SFA on top of each others to extract higher-level slowly changing features, resulting in hierarchical SFA.

Interestingly, SFA is directly related to DSS. Instead of minimising the variance after high-pass filtering as in SFA, it is also possible to maximise the variance after low-pass filtering. SFA is thus equivalent to DSS with nonlinear data expansion and low-pass filtering as denoising. This is similar to earlier proposals, *e.g.*, by Földiák (1991).

There are several possibilities for the nonlinear feature expansion in hierarchical DSS. For instance kernel PCA (Schölkopf et al., 1998), sparse coding or liquid state machines (Maass et al., 2002) can be used.

Parga and Rolls (1998) proposed that recurrent activity in cortical circuits might mediate the low-pass filtering. Considering this activity as contextual input suggests that context could be used for denoising more generally, even when the context does not change slowly. Such contextual information has been considered, *e.g.*, by (Becker and Hinton, 1992, Deco and Schürmann, 2000). In particular, Deco and Schürmann (2000) have shown that top-down bias combined with local lateral competition accounts for many of the characteristics of covert attention. In their model, attention emerges because the influence of the representations activated at the higher levels propagates downward the hierarchy. Deco and Rolls (2004) showed that it is possible to learn the features needed for the model but top-down weights were only used for attention. Interpreting the top-down bias as denoising suggests that the same mechanism could account both for learning features and for emergent attention.

The hierarchical DSS can be used in a fully supervised setting by fixing the activations of the topmost layer to target outputs. Supervised learning often suffers from slow learning in deep hierarchies because the way information is represented gradually changes in the hierarchy. It is therefore difficult to use the information about the target output for learning the layers close to the inputs. The benefit of hierarchical DSS is that learning on lower levels is not only dependent on the information propagated from the target output because the context includes lateral or delayed information from the inputs. In this approach, the mode of learning shifts smoothly from mostly unsupervised learning to mostly supervised learning from the input layer towards the output layer. A similar mixture of supervised and unsupervised learning has been suggested by Körding and König (2001).

7. Conclusion

The work in linear source separation has concentrated on blind approaches to fix the rotational ambiguity left by the factor analysis model. Usually, however, there would be additional information to find the rotation either more efficiently or more accurately. In this paper we developed an algorithmic framework called denoising source separation (DSS). We showed that denoising can be used for source separation and that the results are often better than with blind approaches. The better the denoising is, the better the results are. Furthermore, many blind source separation techniques can be interpreted as DSS algorithms using very general denoising principles. In particular, we showed that FastICA is a special case of DSS which also implies that DSS can be computationally very efficient.

The main benefit of DSS framework is that it allows for easy development of new source separation algorithms which are optimised for the specific problem at hand. There is a wide literature on signal denoising to choose from and in some cases denoising would be used for post-processing in any case. All the tools needed for DSS are then readily available.

In the experimental section, we demonstrated DSS in various source separation tasks. We showed how denoising can be adapted to the observed characteristics of signals extracted with denoising based on vague knowledge. From MEG signals, we were able to extract very accurately subspaces such as the α-subspace or the very weak components of the cardiac subspace. DSS also proved to be able to recover CDMA signals better than the standard RAKE technique under poor SNR.

Finally, we discussed potential extensions of DSS. It appears that DSS offers a sound basis for developing hierarchical, nonlinear feature extraction methods and the connections to cortical models of attention and perception suggest a promising starting point for future work.

8. Acknowledgements

This work is funded by the Academy of Finland, under the project New information processing principles, and by European Commission, under the project ADAPT (IST-2001-37137).

We would like to show gratitude to Dr. Ricardo Vigário for the fruitful discussions concerning the method in general as well as the MEG experiments in detail and Dr. Aapo Hyvärinen for the method itself and its connections to ICA. We would like to thank as well Mr. Karthikesh Raju for his suggestions and help concerning the CDMA experiments. Finally, we would like to thank prof. Erkki Oja for his comments on the draft version of this manuscript.

References

B. D. Anderson and J. B. Moore. *Optimal filtering.* Prentice-Hall, 1979.

S. Becker and G. E. Hinton. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355:161 – 163, 1992.

A. Belouchrani, K. Abed Meraim, J.-F. Cardoso, and E. Moulines. A blind source separation technique based on second order statistics. *IEEE Trans. on S.P.*, 45(2):434–44, 1997.

J.-F. Cardoso. High-order contrasts for independent component analysis. *Neural computation*, 11(1):157 – 192, 1999.

G. Deco and E. T. Rolls. A neurodynamical cortical model of visual attention and invariant object recognition. *Vision research*, 44:621 – 642, 2004.

G. Deco and B. Schürmann. A hierarchical neural system with attentional top-down enhancement of the spatial resolution for object recognition. *Vision research*, 40:2845 – 2859, 2000.

D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage: asymptopia? *Journal of the Royal Statistical Society ser. B*, 57:301–337, 1995.

S. C. Douglas and A. Cichocki. Neural networks for blind decorrelation of signals. *IEEE Trans. Signal Processing*, 45(11):2829 – 2842, 1997.

FastICA. The FastICA MATLAB package. 1998. Available at `http://www.cis.hut.fi/projects/ica/fastica/`.

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

P. Földiák. Learning invariance from transformation sequences. *Neural Computation*, 3: 194–200, 1991.

M. Funaro, E. Oja, and H. Valpola. Independent component analysis for artefact separation in astrophysical images. *Neural networks*, 16(3 – 4):469 – 478, 2003.

M. S. Gazzaniga, editor. *The New Cognitive Neurosciences.* A Bradford book/MIT Press, 2nd edition, 2000.

M. Hämäläinen, R. Hari, R. Ilmoniemi, J. Knuutila, and O. V. Lounasmaa. Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews of Modern Physics*, 65:413–497, 1993.

J. Himberg and A. Hyvärinen. Icasso: software for investigating the reliability of ica estimates by clustering and visualization. In *Proc. 2003 IEEE workshop on neural networks for signal processing (NNSP'2003)*, pages 259–268, Toulouse, France, 2003.

A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. on Neural Networks*, 10(3):626–634, 1999.

A. Hyvärinen. New approximations of differential entropy for independent component analysis and projection pursuit. In *Advances in Neural Information Processing 10 (Proc. NIPS'98)*, pages 273–279. MIT Press, 1998.

A. Hyvärinen, P. Hoyer, and M. Inki. Topographic independent component analysis. *Neural Computation*, 13(7):1525–1558, 2001a.

A. Hyvärinen, P. Hoyer, and E. Oja. Sparse code shrinkage: Denoising by maximum likelihood estimation. In *Advances in Neural Information Processing Systems 11 (Proc. NIPS'98)*, 1999.

A. Hyvärinen, J. Karhunen, and E. Oja. *Independent component analysis.* Wiley, 2001b.

JADE. The JADE MATLAB package. 1999. Available at `http://www.tsi.enst.fr/icacentral/Algos/cardoso/`.

T.-P. Jung, S. Makeig, T.-W. Lee, M. McKeown, G. Brown, A. Bell, and T. J. Sejnowski. Independent component analysis of biomedical signals. In *Proc. Int. Workshop on Independent Component Analysis and Blind Separation of Signals (ICA'00)*, pages 633 – 644, Helsinki, Finland, 2000.

J. Karhunen and J. Joutsensalo. Representation and separation of signals using nonlinear PCA type learning. *Neural Networks*, 7(1):113–127, 1994.

K. H. Knuth. Bayesian source separation and localization. In A. Mohammad-Djafari, editor, *SPIE'98 Proceedings: Bayesian Inference for Inverse Problems*, pages 147–158, San Diego, USA, 1998.

K.P. Körding and P. König. Neurons with two sites of synaptic integration learn invariant representations. *Neural Computation*, 13:2823 – 2849, 2001.

P. Kuosmanen and J. T. Astola. *Fundamentals of nonlinear digital filtering.* CRC press, 1997.

D. G. Luenberger. *Optimization by Vector Space Methods.* John Wiley & Sons, 1969.

W. Maass, T. Natschläger, and H. Markram. Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural Computation*, 14(11):2531 – 2560, 2002.

F. Meinecke, A. Ziehe, M. Kawanabe, and K.-R. Müller. A resampling approach to estimate the stability of one- and multidimensional independent components. *IEEE Trans. Biom. Eng.*, 49(12):1514 – 1525, 2002.

E. Niedermeyer and F. Lopes da Silva, editors. *Electroencephalography. Basic principles, clinical applications, and related fields.* Baltimore: Williams & Wilkins, 1993.

E. Oja. Principal components, minor components, and linear neural networks. *Neural Networks*, 5:927–935, 1992.

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.

N. Parga and E. T. Rolls. Transform invariant recognition by association in a recurrent network. *Neural Computation*, 10(6):1507 – 1525, 1998.

K. Raju and T. Ristaniemi. ICA-RAKE switching for jammer cancellation in DS-CDMA array systems. In *Proc. of the IEEE Int. Symposium on Spread Spectrum Techniques and Applications (ISSSTA)*, Prague, September 2002. to appear.

R. M. Rangayyan. *Biomedical signal analysis: A case-study approach.* IEEE Press Series in Biomedical Engineering, 2002.

J. Särelä, H. Valpola, R. Vigário, and E. Oja. Dynamical factor analysis of rhythmic magnetoencephalographic activity. In *Proc. Int. Conf. on Independent Component Analysis and Signal Separation (ICA2001)*, pages 451–456, San Diego, USA, 2001.

J. Särelä and R. Vigário. Overlearning in marginal distribution-based ICA: analysis and solutions. *Journal of machine learning research*, 4 (Dec):1447–1469, 2003.

B. Schölkopf, S. Mika, A. Smola, Gunnar Rätsch, and K.-R. Müller. Kernel PCA pattern reconstruction via approximate pre-images. In *Proc. 8th Int. Conf. on Artificial neural networks (ICANN'98)*, pages 147 – 152, Skövde, 1998.

O. Schwartz and E. P. Simoncelli. Natural signal statistics and sensory gain control. *Nature Neuroscience*, 4(8):819 – 825, 2001.

A. C. Tang and B. A. Pearlmutter. *Independent components of magnetoencephalography: localization*, chapter Blind Decomposition of Multimodal Evoked Responses and DC Fields, pages 129 – 162. The MIT Press, 2003.

H. Valpola and J. Karhunen. An unsupervised ensemble learning method for nonlinear dynamic state-space models. *Neural Computation*, 14:2647 – 2692, 2002.

H. Valpola and P. Pajunen. Fast algorithms for Bayesian independent component analysis. In *Proceedings of the second international workshop on independent component analysis and blind signal separation, ICA'00*, pages 233–238, Espoo, Finland, 2000.

H. Valpola and J. Särelä. Accurate, fast and stable denoising source separation algorithms. *Submitted to a conference*, 2004.

M. Vetterli and J. Kovacevic. *Wavelets and subband coding.* Prentice-Hall, 1995.

R. Vigário and J. Matilainen. Personal communication. 2004.

R. Vigário, J. Särelä, V. Jousmäki, M. Hämäläinen, and E. Oja. Independent component approach to the analysis of EEG and MEG recordings. *IEEE transactions on biomedical engineering*, 47(5):589–593, 2000.

V. Vigneron, A. Paraschiv-Ionescu, A. Azancot, O. Sibony, and C. Jutten. Fetal electrocardiogram extraction based on non-stationary ICA and wavelet denoising. In *Proceedings of ISSPA 2003*, Paris (France), July 2003.

A. J. Viterbi. *CDMA : Principles of Spread Spectrum Communication.* Wireless Info Networks Series. Addison-Wesley, 1995.

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

A. Ziehe and K.-R. Müller. TDSEP — an effective algorithm for blind separation using time structure. In *Proc. int. conf. at neural networks (ICANN'98)*, pages 675–680, Skövde, Sweden, 1998.