Accurate, Fast and Stable Denoising Source Separation Algorithms

Harri Valpola[1,2⋆] and Jaakko Särelä[2⋆⋆]

[1] Artificial Intelligence Laboratory, University of Zurich
Andreasstrasse 15, 8050 Zurich, Switzerland
[2] Neural Networks Research Centre, Helsinki University of Technology
P.O.Box 5400, FI-02015 HUT, Espoo, Finland
{harri.valpola,jaakko.sarela}@hut.fi

Abstract. Denoising source separation is a recently introduced framework for building source separation algorithms around denoising procedures. Two developments are reported here. First, a new scheme for accelerating and stabilising convergence by controlling step sizes is introduced. Second, a novel signal-variance based denoising function is proposed. Estimates of variances of different source are whitened which actively promotes separation of sources. Experiments with artificial data and real magnetoencephalograms demonstrate that the developed algorithms are accurate, fast and stable.

1 Introduction

In denoising source separation (DSS) framework [1], separation algorithms are built around a denoising function. This makes it easy to tailor source separation algorithms for the task at hand. Good denoisings usually result in fast and accurate algorithms. Furthermore, explicit objective function is not needed, in contrast to most existing source separation algorithms.

Here we report further developments of two aspects. First, we introduce a new method for stabilising and accelerating convergence which is inspired by predictive controllers. Second, we develop further the signal-variance-based denoising principles. The resulting algorithms yield good results in terms of signal-to-noise ratio (SNR) and exhibit fast and stable convergence.

2 Source separation by denoising

Consider a linear instantaneous mixing of sources:

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

⋆ Funded by the European Commission, under the project ADAPT (IST-2001-37173) and by the Academy of Finland, under the project New information processing principles.
⋆⋆ Funded by the Academy of Finland.

C.G. Puntonet and A. Prieto (Eds.): ICA 2004, LNCS 3195, pp. 65–72, 2004.

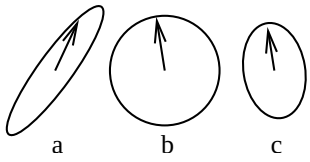


Fig. 1. a) Original data set , b) after sphering and c) after denoising. After these steps, the projection yielding the best signal-to-noise ratio, denoted by arrow, can be obtained by simple correlation-based learning.

where the $N \times T$ matrix $\mathbf{S}$ are the sources, the $M \times T$ matrix $\mathbf{X}$ are the observations and there is noise $\boldsymbol{\nu}$. If the sources are assumed Gaussian, this is a general, linear factor analysis model with rotational invariance.

DSS, as many other computationally efficient ICA algorithms, resorts to sphering. In the case of DSS, the main reason is that after sphering, denoising combined with simple correlation based estimation akin to Hebbian learning (on-line) or power method (batch) is able to retrieve the signal with the highest SNR. Here SNR is implicitly defined by the denoising. The effect of sphering and subsequent denoising is depicted in Fig. 1.

Assuming that $\mathbf{X}$ is already sphered and $\mathbf{f}(\mathbf{s})$ is the denoising procedure, a simple DSS algorithm can be written as follows:

$$\mathbf{s} = \mathbf{w}^T \mathbf{X} \tag{2}$$

$$\mathbf{s}^+ = \mathbf{f}(\mathbf{s}) \tag{3}$$

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

$$\mathbf{w}_{\text{new}} = \text{orth}(\mathbf{w}^+)\,, \tag{5}$$

where $\mathbf{s}$ is the source estimate (a row vector), $\mathbf{s}^+$ is the denoised source estimate, $\mathbf{w}$ is the previous weight vector (a column vector), $\mathbf{w}^+$ is the new weight vector before and $\mathbf{w}_{\text{new}}$ after orthonormalisation (e.g., deflatory or symmetric orthogonalisation as in FastICA [2]).

Note that if $\mathbf{X}$ were not sphered and no denoising were applied, i.e., $\mathbf{f}(\mathbf{s}) = \mathbf{s}$, the above equations would describe the power method for computing the principal eigenvector. When $\mathbf{X}$ is sphered, all eigenvalues are equal to one and without denoising the solution is degenerate, i.e., any unit vector $\mathbf{w}$ is a fixed point of the iterations. This shows that for sphered $\mathbf{X}$, even the slightest denoising $\mathbf{f}(\mathbf{s})$ can determine the convergence point.

If, for instance, $\mathbf{f}(\mathbf{s})$ is chosen to be low-pass filtering, implicitly signals are assumed to have relatively more low frequencies than noise and the above iteration converges to the signal which has the most low-frequency components. On the other hand, if $\mathbf{f}(\mathbf{s})$ is a shrinkage function, suppressing small components of $\mathbf{s}$ while leaving large components relatively untouched, signals are implicitly assumed to have heavy tails and thus super-Gaussian distributions.

It is possible to begin with an objective function $g(\mathbf{s})$ in which case the denoising can be chosen[1] to be the gradient: $\mathbf{f}(\mathbf{s}) = \nabla g(\mathbf{s})$. In practice, denoising functions can easily be designed without explicitly starting from objective functions. They often work exceedingly well and good denoisings result in fast and accurate algorithms.

3 Accelerating and stabilising convergence by spectral shift and adaptation of learning rate

If the denoising function is not able to reduce noise significantly more than signal, the basic DSS iterations (2)–(5) may converge slowly. This is closely related to the fact that power method converges slowly if the largest eigenvalue is only slightly larger than the next largest. Consequently, convergence in DSS can be accelerated in a very similar manner as in power method.

A well-known speedup for power method is spectral shift. It is based on modifying an iteration of the form $\mathbf{w}^+ = \mathbf{A}\mathbf{w}$ into $\mathbf{w}^+ = \mathbf{A}\mathbf{w} + \beta\mathbf{w}$. In the original iteration, it holds $\mathbf{w}^+ = \lambda\mathbf{w}$ at the fixed points and consequently $\mathbf{w}^+ = (\lambda + \beta)\mathbf{w}$ after the modification. The fixed points remain the same but the eigenvalues λ are shifted by β, hence the name spectral shift.

If all eigenvalues are large and their differences are small, convergence can be greatly accelerated by using β which is negative and whose absolute value is close to the second largest eigenvalue. On the other hand, power method converges to the eigenvector that corresponds to the eigenvalue having the largest absolute value. This means that instead of finding the principal component, the minor component is obtained with negative enough β.

In DSS, (3) can be modified into

$$\mathbf{s}^+ = \alpha(\mathbf{s})\mathbf{f}(\mathbf{s}) + \beta(\mathbf{s})\mathbf{s} \tag{6}$$

without changing the fixed points as long as $\alpha(\mathbf{s})$ and $\beta(\mathbf{s})$ are scalar functions. Since $\alpha(\mathbf{s})$ only scales the source estimate, from now on we assume $\alpha(\mathbf{s}) = 1$.

In DSS, $\mathbf{s}^+\mathbf{s}^T/T$ plays the role of the eigenvalue [1]. Since Gaussian signals are the least desirable ones in source separation, a reasonable choice for β is the one that shifts the eigenvalue of Gaussian signals to zero:

$$\beta = E\{\mathbf{f}(\boldsymbol{\nu})\boldsymbol{\nu}^T/T\}\,, \tag{7}$$

where $\boldsymbol{\nu}$ is a normally distributed signal.

It is interesting to note that the fixed-point equation of FastICA [2] can be interpreted within this framework although normally the speedup used in FastICA is justified as an approximation to Netwon's method. In [1], it was shown that if $\beta(\mathbf{s})$ is based on a linearisation of $\mathbf{f}(\mathbf{s})$ around the current source estimate $\mathbf{s}$, the spectral shift (7) will be

$$\beta(\mathbf{s}) = -\operatorname{tr}\mathbf{J}(\mathbf{s})/T\,, \tag{8}$$

[1] There is some freedom in this choice because there are several denoising functions which have the same convergence points. They are given in (6).

which is identical to the one used in FastICA. Here $\mathbf{J}(\mathbf{s})$ is the Jacobian of $\mathbf{f}(\mathbf{s})$. Interpreting the speedup as a spectral shift corresponding to Gaussian noise gives an intuitive explanation to why FastICA is able to extract both super- and sub-Gaussian signals with the same nonlinearity: power-method-like iterations converge to the eigenvector whose eigenvalue has the largest magnitude. The sign of the eigenvalue is different depending on whether the component is super- or sub-Gaussian but the magnitude increases when moving away from Gaussian signal whose eigenvalue has been shifted to zero.

In general, iterations converge faster with the FastICA-type spectral shift (8) than with the global Gaussian approximation (7) 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.

Neither of the spectral shifts, (7) or (8), always results in stable or fast convergence. Sometimes the spectral shift is too large, which due to the nonlinearity of denoising typically leads to oscillatory behaviour: the iteration oscillates between two weight values. Some other times the spectral shift is too modest leading to slow convergence characterised by small changes of $\mathbf{w}$ in the same direction during several iterations.

For this reason, we have suggested a simple stabilisation rule [1]: instead of updating $\mathbf{w}$ into $\mathbf{w}_{\text{new}}$ defined by (5), it is updated into

$$\mathbf{w}_{\text{adapted}} = \text{orth}(\mathbf{w} + \gamma \Delta\mathbf{w}) \tag{9}$$

$$\Delta\mathbf{w} = \mathbf{w}_{\text{new}} - \mathbf{w}\,, \tag{10}$$

where γ is the step size. Originally $\gamma = 1$, but if the consecutive 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. There exist a stabilised version of FastICA as well [2] and a similar procedure has been used in practice.

The above modification is able to stabilise convergence in case of oscillations but sometimes the spectral shift is too small and then an increase in step size would be appropriate, i.e., $\gamma > 1$. We propose a simple rule for adapting γ which is inspired by predictive controllers used in robotics: a simple but slow and possibly unstable reactive controller is used for teaching a new, predictive controller. Usually stable and rapid convergence are difficult to achieve simultaneously, but in this setup the new controller can be both faster and stabler.

Translated in our problem, the old slow and unstable controller is the weight modification rule which proposes a modification of weight according to (10). The new controller is implemented by (9), i.e., it modifies the step size. The new controller tries to do immediately what the old controller would do in the future. The step at the previous time instant was apparently optimal if the step proposed at this time instant is orthogonal with it. If not, γ should have been different and, assuming that the optimal γ is constant, the gamma used at this time step should be

$$\gamma_{\text{new}} = \gamma_{\text{old}} + \Delta\mathbf{w}_{\text{old}}^T \Delta\mathbf{w} / ||\Delta\mathbf{w}_{\text{old}}||^2\,. \tag{11}$$

As it does not seem productive to take steps in the direction opposite from what is suggested by $\Delta\mathbf{w}$ or to take extremely short steps, we require that $\gamma \geq 0.5$.

The above adaptation of γ has turned out to be very useful and it can both stabilise and accelerate convergence. According to (11), γ keeps increasing as long as the steps are taken to the same direction and decreases if they are taken backwards.

4 Denoising based on estimated signal variance

Several denoising procedures based on masking the source estimate were proposed in [1]. The basic idea is to multiply the source estimate by a positive envelope, a mask which has low values when SNR is low and vice versa. Depending on how the mask is computed, several types of prior information about the sources can be used for separation.

A simple and well-founded mask can be obtained from the maximum-a-posteriori (MAP) estimate. Assuming that the signals are Gaussian with changing variance $\sigma_s^2(t)$ (for related methods, see, e.g., [3]) and additive Gaussian noise σ_n^2, the MAP estimate of the signal is

$$s^+(t) = s(t)\frac{\sigma_s^2(t)}{\sigma_{\text{tot}}^2(t)}\,, \tag{12}$$

where $\sigma_{\text{tot}}^2(t) = \sigma_s^2(t) + \sigma_n^2(t)$ is the total variance of the observation. Masking then boils down to estimating $\sigma_s^2(t)$ and $\sigma_{\text{tot}}^2(t)$ from the observations.

A naïve estimate of the signal variance is $\sigma_s^2(t) \approx s^2(t)$. It can be improved by low-pass filtering in time, e.g., by convolving with a Gaussian kernel. Simple estimation of the baseline noise-level σ_n^2 was suggested in [1] resulting in a simple DSS algorithm. However, from the viewpoint of the estimated signal, other signals should be treated as noise. DSS algorithm using the above approximation separates easily the signal subspace from noise but the separation in the signal subspace is slow and may even fail. In [1], this was solved by using $\sigma_s^{2\mu}(t)$ with $\mu > 1$ in (12). This way the mask does not saturate so quickly for large signal variances, giving competitive edge to the source which is strongest. A close connection to the familiar tanh-nonlinearity was shown: $\mathbf{f}(\mathbf{s}) = \mathbf{s} - \tanh \mathbf{s}$ has the same fixed points as $\mathbf{f}(\mathbf{s}) = \tanh \mathbf{s}$ but the former can be interpreted as $\mathbf{s}$ masked by a slowly saturating envelope.

In this paper, we propose a new and better founded solution to the separation problem. One can simply whiten the estimated total variance $\sigma_{\text{tot}}(t)$ by a symmetric whitening matrix. This bares resemblance to proposals of the role of divisive normalisation on cortex [4] and to the classical ICA-method called JADE [5]. Whitening naturally requires that all sources are estimated simultaneously and deflation approach is thus not applicable. The total variance is obtained by smoothing $s^2(t)$ as described above. We obtain $\sigma_s^2(t)$ by taking the positive part of the whitened $\sigma_{\text{tot}}^2(t)$. Whitening here includes removing the mean. Separation by (12) is accelerated significantly because the differences between the envelopes of source estimates are actively emphasised.

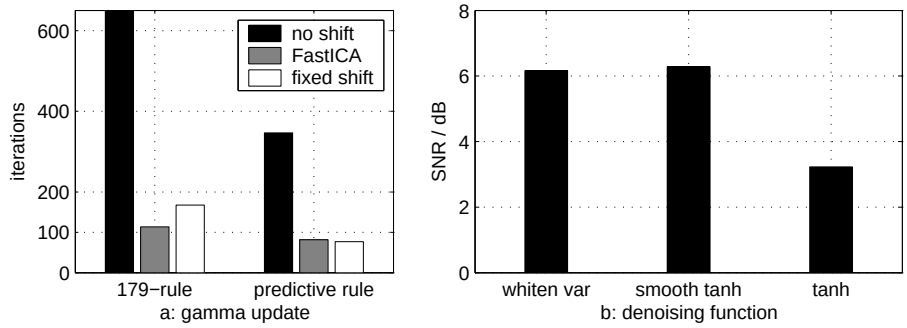


Fig. 2. Speedup tests. a) Effects of spectral shift and step-size adaptation on convergence speed. The leftmost bar not fully shown. b) Average SNRs for different denoising functions: variance whitening and tanh with and without smoothing.

5 Experiments

In this section, we show that the developed algorithms are fast, stable, accurate and produce meaningful results. First, in Sec. 5.1, we demonstrate the different spectral shifts and step-size adaptation. Then the accuracy of different denoising algorithms is tested with artificial data (Sec. 5.2). Finally, we demonstrate the separation capability and convergence speed of the variance-based-denoising in real MEG data (Sec. 5.3).

5.1 Speedup comparison

In Sec. 3, we reviewed two spectral shifts that can accelerate convergence in DSS algorithms. Later in the section, we proposed two additional methods to adapt these spectral shifts to increase stability. In this section, we compare these adaptive-spectral-shift methods together with the stability improvements in deflatory separation. The data consists of $M = 50$ channels and $T = 8192$ time samples of rhythmic magnetoencephalograms (MEG) [6, 1]. The data was preprocessed as in [1] to enhance weak phenomena. Simple $\mathbf{f}(\mathbf{s}) = \mathbf{s} - \tanh \mathbf{s}$ was used as the denoising function. DSS was run to extract 30 components from this data and average number of iterations was calculated. To be fair for all the methods, each of them was run until convergence, where the angle between old and new projection vectors ($\mathbf{w}$ and $\mathbf{w}_{\text{new}}$) was less than 0.0001°. We then measured the number of iterations that had taken $\mathbf{w}$ within 0.1° of the final solution.

The results are shown in Fig. 2a. Both types of spectral shift and γ adaptation always accelerated convergence. Convergence without any speedups took on average more than 1500 iterations. Without γ adaptation, the FastICA-type scheme (8) converged faster on average than the fixed-shift scheme (7), but γ adaptation reversed the situation. Standard FastICA used about 50% more iterations than the best method.

5.2 Comparison of denoising functions

We next compare DSS schemes based on source-variance estimates to the classical tanh-based approach in symmetrical separation of artificial signals. The signals were generated as follows. First, six signals were generated by modulating Gaussian noise with slowly changing envelope. Then the signals were divided into two subspaces, three signals in each. In each of the subspaces, the signals were modulated by another envelope common to all the signals in the subspace. The common envelopes of the subspaces were stronger than the individual envelopes of the sources. Finally, the unit-variance signals were mixed linearly (with $M = N$). Mixing coefficients were sampled from normal distribution and Gaussian noise with variance $\sigma_\nu^2 = 0.09$ was added.

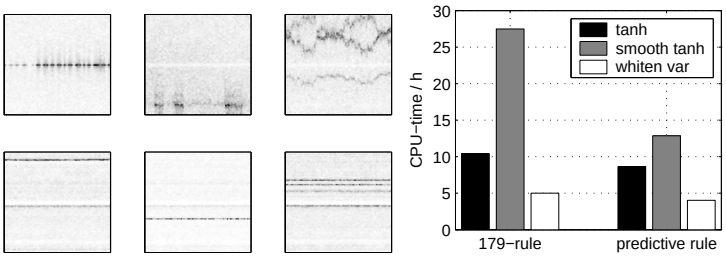
One hundred different data sets were generated and DSS was used to separate the sources with three different denoising functions. Two methods were based on smoothed estimate of source variance. Either the whitening scheme described in Sec. 4 or tanh-based scheme were used in order to promote separation, the tanh-mask being $1 - \tanh[\sigma_{\text{tot}}(t)]/\sigma_{\text{tot}}(t)$. If $\sigma_{\text{tot}}^2(t) = s^2(t)$, this reduces to the popular tanh-nonlinearity. With these methods, spectral shift was computed by assuming that the mask does not significantly depend on any individual source value, i.e. $-\beta$ equals to the average of elements of the mask. The third method was the popular tanh-nonlinearity with FastICA-type spectral shift. The step size was adapted by the 179-rule.

As before, the algorithms were run until convergence. The average SNRs of the separation over the one hundred runs are shown in Fig. 2b. Smoothing the variance estimate clearly improves the SNR with tanh-nonlinearity. Variance whitening achieved comparable SNR but used significantly less iterations.

5.3 MEG signal separation

Finally, we used the DSS algorithms and acceleration methods studied in the previous sections to separate sources from rhythmic MEG data. The whole data set ($M = 122$ and $T = 65536$) was used and 30 components were extracted using the same denoising functions as in the previous section. Both the 179-rule and the predictive rule (11) were tested. The number of iterations was taken to be the limit where the projection vector $\mathbf{w}$ of the slowest converging component reaches 0.1° of the final projection. Enhanced spectrograms of some interesting components extracted by the variance-whitening DSS are depicted in Fig. 3a.

Tanh-nonlinearity with smoothed variance estimate extracted similar components, but the usual tanh-nonlinearity without smoothing seemed to have trouble in finding the weak steady frequencies shown in the bottom row of Fig. 3a. The processing times of different denoising functions and different step size adaptations are shown in Fig. 3b. Since the computational complexity of one iteration depends on the denoising function, the total CPU-time is reported. Compared to the variance-whitening DSS, the tanh-nonlinearities used more than two times more processing time, independent of the step-size adaptation. Compared to the 179-rule, the adaptive γ reduced the total processing time by 20–50 %, depending



(a) (b)

Fig. 3. a) Spectrograms of some of the sources separated using variance whitening. Time on the horizontal and frequency on the vertical axis. b) Used processing time for different denoising functions and step sizes.

on the denoising function. Tanh-nonlinearity with smoothed variance estimate used a fixed spectral shift and benefitted more from adaptation of γ than the plain tanh-nonlinearity with FastICA-type spectral shift.

6 Conclusion

DSS framework offers a sound basis for developing simple but efficient and accurate source separation algorithms. We proposed a method for stabilising and accelerating convergence and showed that convergence is faster than with FastICA. Additional benefit is that gradient of the nonlinearity is not needed. We also proposed a new denoising procedure which was justified as the MAP-estimate of signals with changing variance. Denoising which makes use of non-stationarity of variance was shown to yield better results than the popular tanh-nonlinearity as measured by SNR in the artificially generated data. The variance-whitening DSS also extracted cleaner signals in MEG data, while the tanh-nonlinearity had difficulties with some weak but clear phenomena. Whitening the estimated variances of different sources significantly improved convergence.

References

1. Särelä, J., Valpola, H.: Denoising source separation. Submitted to a journal (2004) Available at Cogprints `http://cogprints.ecs.soton.ac.uk/archive/00003493/`.
2. Hyvärinen, A.: Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans. on Neural Networks **10** (1999) 626–634
3. Matsuoka, K., Ohya, M., Kawamoto, M.: A neural net for blind separation of nonstationary signals. Neural Networks **8** (1995) 411–419
4. Schwartz, O., Simoncelli, E.P.: Natural signal statistics and sensory gain control. Nature Neuroscience **4** (2001) 819 – 825
5. Cardoso, J.F.: High-order contrasts for independent component analysis. Neural computation **11** (1999) 157 – 192
6. Särelä, J., Valpola, H., Vigário, R., Oja, E.: Dynamical factor analysis of rhythmic magnetoencephalographic activity. In: Proc. Int. Conf. on Independent Component Analysis and Signal Separation (ICA2001), San Diego, USA (2001) 451–456