Time-Scale Approach to Blind Source Separation

M.G. Jafari, and J.A. Chambers
Department of Electronic & Electrical Engineering
University of Bath, Bath BA2 7AY, U.K.
E-mail: m.g.jafari@bath.ac.uk

ABSTRACT

A new approach to blind source separation (BSS) in the wavelet domain is introduced. The technique improves the speed of convergence of the natural gradient algorithm (NGA), and overcomes the problem of having to select the non-linearities required to separate mixed sub- and super-Gaussian signals. The distribution of the wavelet coefficients of certain natural source signals is modeled by a Laplacian density, and therefore in the time-scale domain the problem of selecting an appropriate activation function is overcome. Experimental results show the validity of this method.

1 Introduction

The recovery of a number of source signals from observations which contain only mixtures of these signals is the essence of blind source separation. One of the assumptions at the heart of BSS is that at most one source has a Gaussian distribution. In practice, however, the performance of BSS algorithms improves as the probability density functions (pdfs) of the sources become less Gaussian, a phenomenon that has been observed when mapping certain signals from the time domain to the frequency domain. In this paper we address the BSS problem in the wavelet domain, and make use of an image processing result to obtain a model for the sample distribution of the wavelet coefficients of the sources. The proposed method allows the separation of mixtures of both sub- and super-Gaussian signals without having to use different non-linearities, as is shown by computer simulation.

2 Problem statement

When n real sources are mixed by a time-invariant instantaneous channel, and no noise is present, the m observed signals are given by [1]

$$\mathbf{x}(k) = \mathbf{A}\mathbf{s}(k) \tag{1}$$

where $\mathbf{x}(k) \in \mathbf{R}^m$ is the vector of observed signals, and $\mathbf{s}(k) \in \mathbf{R}^n$ is the vector of source signals, assumed to be zero-mean and mutually independent. $\mathbf{A} \in \mathbf{R}^{mxn}$ is an unknown, full column rank, mixing matrix, and typically it is assumed that there are at least as many sensors as sources, that is $m \geq n$, and that at most one source has Gaussian distribution. The sources are recovered using the following linear separating system

$$\mathbf{y}(k) = \mathbf{W}(k)\mathbf{x}(k) = \mathbf{W}(k)\mathbf{A}\mathbf{s}(k) \tag{2}$$

where $\mathbf{y}(k) \in \mathbf{R}^n$ estimates $\mathbf{s}(k)$, $\mathbf{W}(k) \in \mathbf{R}^{nxm}$ is the separating matrix, and the product $\mathbf{P}(k) = \mathbf{W}(k)\mathbf{A} \in \mathbf{R}^{nxn}$ is known as the global mixing-separating matrix. The performance of a BSS method can be assessed by plotting the performance index (PI), which is a measure of the closeness between $\mathbf{W}(k)$ and the pseudo-inverse of the mixing matrix, taking into account the scaling and ordering ambiguities.

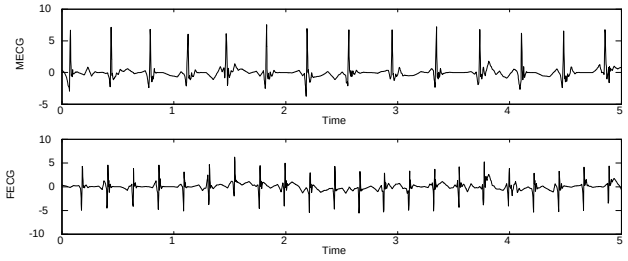

Figure 1: Maternal and foetal electrocardiogram (ECG) components extracted with the JADE algorithm, and de-noised with WT.

3 The Wavelet transform

The wavelet transform (WT) maps a signal from the time domain to the time-scale domain. Discrete wavelets are defined as

$$\psi_{j,q}(k) = 2^{-j/2}\psi\left(2^{-j}k - q\right) \tag{3}$$

where $j, q \in \mathbf{Z}$. The resulting functions form a set of discrete wavelet basis functions [2], and the wavelet transform of a signal $x(k)$ is given by the inner product of

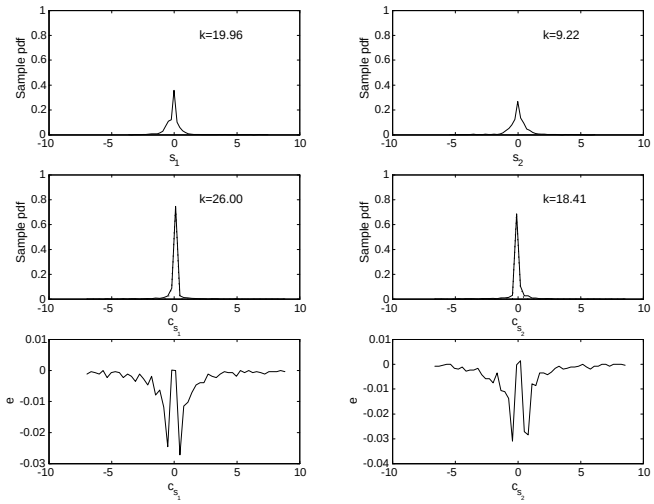

Figure 2: Sample pdfs for the sources shown in Fig. 1 (upper plots), sample pdfs of corresponding wavelet coefficients (middle plots, solid lines), fitted distributions (middle plots, dotted lines) and residual error (lower plots).

the signal with each wavelet:

$$c_{j,q} = \left(x\left(k\right), \psi_{j,q}\left(k\right)\right) \tag{4}$$

where $c_{j,k}$ denotes the transform coefficients, and $(\cdot,\cdot)$ represent the inner product. The expression in equation (4), together with the linearity property of the inner product, leads to the following result, useful when assessing the performance of BSS methods operating in the wavelet domain. When the mixing matrix $\mathbf{A}$ is time-invariant, the wavelet transform of the i-th observed signal $x_i\left(k\right) = \sum_{l=1}^{n} a_{il} s_l\left(k\right)$ is given by

$$\mathcal{W}\left\{x_i\left(k\right)\right\} = \sum_{l=1}^{n} a_{il}\left(s_i\left(k\right), \psi_{j,q}\left(k\right)\right) \tag{5}$$

Thus, the wavelet transform of (1) in matrix form is

$$\mathbf{c}_x = \mathbf{A}\mathbf{c}_s \tag{6}$$

where $\mathbf{c}_x = \left[\left(x_1\left(k\right), \psi_{j,q}\left(k\right)\right), \ldots, \left(x_n\left(k\right), \psi_{j,q}\left(k\right)\right)\right]$ and $\mathbf{c}_s = \left[\left(s_1\left(k\right), \psi_{j,q}\left(k\right)\right), \ldots, \left(s_m\left(k\right), \psi_{j,q}\left(k\right)\right)\right]$ are, respectively, the vectors of wavelet transformed sensors and sources. It follows that the sources estimated in the wavelet domain $\mathbf{y}_{\mathcal{W}}\left(k\right)$ are given by

$$\mathbf{y}_{\mathcal{W}}\left(k\right) = \mathbf{W}\left(k\right)\mathbf{c}_x = \mathbf{W}\left(k\right)\mathbf{A}\mathbf{c}_s \tag{7}$$

Hence, PI remains a meaningful performance measure for BSS algorithms operating in the wavelet domain. The wavelet coefficients of natural images have been reported to have highly non-Gaussian statistics, which can be modeled using a Laplacian pdf of the form [3]

$$f_{s,p}\left(c\right) = e^{-|c/s|^p}/N\left(s,p\right) \tag{8}$$

where $N\left(s,p\right) = 2s\Gamma\left(1/p\right)/p$, and $\Gamma\left(l\right) = \int_0^\infty t^{l-1}e^{-t}dt$, is the Gamma function. Expressions for the variance σ^2, and kurtosis k of the distribution are given in [3]. Frequency and time-frequency approaches to BSS have been motivated by the observation that certain signals are less Gaussian in the frequency domain than they are in the time domain. To demonstrate the validity of the model (8), applied to the distribution of the wavelet coefficients of one-dimensional (1-D) natural signals, we use the least squares curve fitting method, to fit the Laplacian pdf in (8) to the sample pdf of the wavelet domain representation of the signal. The sample pdfs of the sources in Fig. 1, and fitted sample pdfs of their wavelet coefficients are illustrated in Fig. 2 (middle plots). The lower plots show the residual error, given by $e = f_{s,p}\left(c_{s_i}\right) - q\left(c_{s_i}\right)$, where $f_{s,p}\left(c_{s_i}\right)$ is the fitted pdf for the wavelet coefficients of the i-th source, and $q\left(c_{s_i}\right)$ is their true sample pdf. Clearly, the Laplacian distribution models the statistics of the wavelet coefficients very closely. The kurtoses of the sources in both domains are also compared in Fig. 2, indicating that the signals are less Gaussian in the wavelet domain than in the time domain.

4 Natural gradient algorithm

The NGA algorithm update equation is given by

$$\mathbf{W}(k+1) = \mathbf{W}(k) + \eta(k)[\mathbf{I} - \mathbf{f}(\mathbf{y}(k))\mathbf{y}^T(k)]\mathbf{W}(k) \tag{9}$$

where $\mathbf{f}(\ \mathbf{y}(k))$ is an odd non-linear function of the output, called the activation function, whose choice depends on the statistics of the sources, and $\eta(k)$ is a positive adaptive learning parameter defined in [1].

5 Time-scale approach

The time-scale approach is as follows:

1. the mixtures are divided into blocks of N samples, their wavelet transform is evaluated, and hard-thresholding is applied.
2. NGA separates sequentially the transformed signals in the wavelet domain, leading to $\mathbf{y}_{\mathcal{W}_N}\left(k\right)$
3. the inverse wavelet transform gives N samples of the estimated sources $\mathbf{y}\left(k\right)$. Steps 1-3 are then repeated for the next block of data.

Reducing the noise level is expected to improve the performance of NGA because invariably true measurements are noisy, while the algorithm is derived on the assumption that the sources are mixed in the absence of noise. Although not satisfying this hypothesis fully, the mixtures obtained after de-noising are better suited for processing by NGA than prior to noise removal. NGA requires a priori knowledge about the statistics of the sources, and different non-linearities are selected for the separation of sub- and super-Gaussian sources.

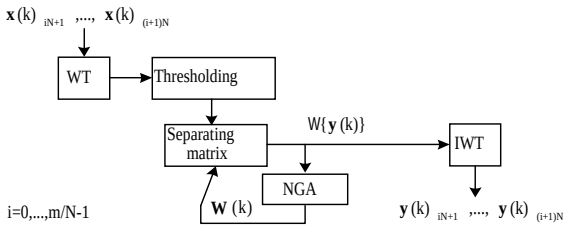

Figure 3: BSS in the wavelet domain.

Also, the algorithm may fail when mixtures of both sub- and super-Gaussian signals are observed. To address these difficulties, time-varying non-linearities can be employed, appropriately selected for each channel according to the statistics of the output [1]. The most remarkable property of the time-scale approach is that the problem of switching between activation functions is overcome, because the wavelet coefficients of the sources can be modeled by a Laplacian pdf, and therefore the activation function need not change when the sources are sub-Gaussian.

5.1 Convergence of algorithm

The increase in source kurtosis in the wavelet domain can be expressed mathematically as

$$\kappa_1^{\mathcal{W}} > \kappa_1 \qquad \text{and} \qquad \kappa_2^{\mathcal{W}} > \kappa_2 \tag{10}$$

where κ_i and $\kappa_i^{\mathcal{W}}$ are, respectively, the kurtosis of the i-th signal and of its wavelet coefficients. Post-multiplying (9) by the mixing matrix $\mathbf{A}$, we obtain an expression for the global mixing-separating system

$$\mathbf{P}(k+1) = \mathbf{P}(k) + \eta(k)\left[\mathbf{I} - \mathbf{F}_y(k)\right]\mathbf{P}(k) \tag{11}$$

where $\mathbf{F}_y(k) = \mathbf{f}(\mathbf{y}(k))\,\mathbf{y}^T(k)$. Characterisation of the transient behaviour of this system is typically a very challenging task [4], due to the cross-coupling of the elements of $\mathbf{P}(k)$. Convergence speed depends on the second term on the right-hand side of (11). To a first approximation, an increase in $(\mathbf{I} - \mathbf{F}_y(k))$ results effectively in the algorithm taking a larger step in the descent direction, which is desirable during initial convergence when the filter parameters are away from their optimal values. Conversely, convergence of the mean of the algorithm is ensured when $\lim_{k\longrightarrow\infty} E\{\mathbf{F}_y(k)\} = \mathbf{I}$. Thus, some growth in the diagonal elements of $-\mathbf{F}_y(k)$ will generally increase the convergence speed of the algorithm, as will a more rapid decay of the off-diagonal elements. As in [5], we approximate the activation function for super-Gaussian sources $\tanh(y_i(k))$ with the Maclaurin's series. However, since we seek to express (11) in terms of the kurtoses of the sources, the series is truncated at degree 3. Thus, ignoring the time index for convenience, the non-linearity is $\tanh(y_i) \simeq y_i - \frac{1}{3}y_i^3$. Then from (2), and $\kappa_i = E\{s_i^4\} - 3E\{s_i^2\}^2$, considering only the diagonal elements of $\mathbf{F}_y(k)$, and applying the statistical expectation operator, we have

$$\begin{aligned} E\{f_1 y_1\} &\simeq \underbrace{E\left\{\left(p_{12}^2 + p_{11}^2\right)\left(1 - p_{11}^2 - p_{12}^2\right)\right\}}_{\mathbf{B}_1} \\ &\quad -\frac{1}{3}\left(\underbrace{E\{p_{11}^4\}}_{\alpha}\kappa_1 + \underbrace{E\{p_{12}^4\}}_{\gamma}\kappa_2\right) \\ E\{f_2 y_2\} &\simeq \underbrace{E\left\{\left(p_{22}^2 + p_{21}^2\right)\left(1 - p_{21}^2 - p_{22}^2\right)\right\}}_{\mathbf{B}_2} \\ &\quad -\frac{1}{3}\left(\underbrace{E\{p_{21}^4\}}_{\delta}\kappa_1 + \underbrace{E\{p_{22}^4\}}_{\zeta}\kappa_2\right) \end{aligned} \tag{12}$$

where the expectation is with respect to the elements of $\mathbf{W}(k)$ and the sources and, as in [4], it has been assumed that the elements of $\mathbf{W}(k)$ are independent of the sources. At time $k = 0$ we have

$$\begin{aligned} E\{f_1(0)\,y_1(0)\} &\simeq \mathbf{B}_1(0) - \frac{1}{3}\left(\alpha(0)\,\kappa_1 + \gamma(0)\,\kappa_2\right) \\ E\{f_2(0)\,y_2(0)\} &\simeq \mathbf{B}_2(0) - \frac{1}{3}\left(\delta(0)\,\kappa_1 + \zeta(0)\,\kappa_2\right) \end{aligned} \tag{13}$$

Separation in the time-scale domain leads to

$$\mathbf{P}^{\mathcal{W}}(k+1) = \mathbf{P}^{\mathcal{W}}(k) + \mu(k)\left[\mathbf{I} - \mathbf{F}_y^{\mathcal{W}}(k)\right]\mathbf{P}^{\mathcal{W}}(k) \tag{14}$$

Generally, $\mathbf{P}(k) \neq \mathbf{P}^{\mathcal{W}}(k)$, because typically (11) and (14) will have different dynamical characteristics. At time $k = 0$, however, assuming that $\mathbf{W}(0)$ is the same in both domains, (13) becomes

$$\begin{aligned} E\{f_{\mathcal{W}1}(0)\,y_{\mathcal{W}1}(0)\} &\simeq \mathbf{B}_1(0) - \frac{1}{3}\left(\alpha(0)\,\kappa_1^{\mathcal{W}} + \gamma(0)\,\kappa_2^{\mathcal{W}}\right) \\ E\{f_{\mathcal{W}2}(0)\,y_{\mathcal{W}2}(0)\} &\simeq \mathbf{B}_2(0) - \frac{1}{3}\left(\delta(0)\,\kappa_1^{\mathcal{W}} + \zeta(0)\,\kappa_2^{\mathcal{W}}\right) \end{aligned} \tag{15}$$

From (10), considering $E\{-\mathbf{F}_y(0)\}$ and $E\{-\mathbf{F}_y^{\mathcal{W}}(0)\}$, it can be shown that

$$\begin{aligned} &-\mathbf{B}_1(0) + \frac{1}{3}\left(\alpha(0)\,\kappa_1^{\mathcal{W}} + \gamma(0)\,\kappa_2^{\mathcal{W}}\right) > \\ &\quad -\mathbf{B}_1(0) + \frac{1}{3}\left(\alpha(0)\,\kappa_1 + \gamma(0)\,\kappa_2\right) \end{aligned} \tag{16}$$

and similarly

$$\begin{aligned} &-\mathbf{B}_2(0) + \frac{1}{3}\left(\delta(0)\,\kappa_1^{\mathcal{W}} + \zeta(0)\,\kappa_2^{\mathcal{W}}\right) > \\ &\quad -\mathbf{B}_2(0) + \frac{1}{3}\left(\delta(0)\,\kappa_1 + \zeta(0)\,\kappa_2\right) \end{aligned} \tag{17}$$

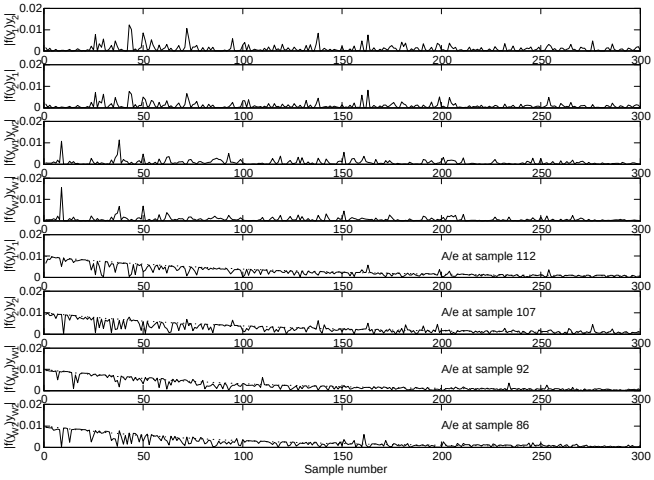

Figure 4: Average behaviour of the elements of $\eta(k)(\mathbf{I}-\mathbf{F}_y(k))$ and $\mu(k)\left(\mathbf{I}-\mathbf{F}_y^{\mathcal{W}}(k)\right)$: off-diagonal (upper four plots), diagonal elements (lower four plots, solid lines) fitted exponential envelopes (dotted lines), and comparison of convergence of the two algorithms.

Thus,

$$diag\left(\mathbf{I}-\mathbf{F}_y^{\mathcal{W}}(0)\right) > diag\left(\mathbf{I}-\mathbf{F}_y(0)\right) \qquad (18)$$

In general, assuming that at time k the matrix $\mathbf{P}(k)$ has the same value in both domains, (18) becomes

$$diag\left(\mathbf{I}-\mathbf{F}_y^{\mathcal{W}}(k)\right) > diag\left(\mathbf{I}-\mathbf{F}_y(k)\right) \qquad (19)$$

In the above analysis, the off-diagonal terms of $(\mathbf{I}-\mathbf{F}_y(k))$ and $\left(\mathbf{I}-\mathbf{F}_y^{\mathcal{W}}(k)\right)$ have been ignored because during initial convergence the diagonal elements are large and dominate algorithm performance. This is illustrated in Fig. 4, which shows the evolution of the elements of the matrices $\eta(k)(\mathbf{I}-\mathbf{F}_y(k))$ and $\mu(k)\left(\mathbf{I}-\mathbf{F}_y^{\mathcal{W}}(k)\right)$, averaged over 30 independent trials, where the contributions of the step-size parameters have been taken into account since, due to their self-adaptive nature, they play a role in the behaviour of the algorithm. Exponential envelopes are fitted to the waveforms to compare the rate of convergence of the two algorithms, and indicate that NGA has faster convergence speed when operating in the wavelet domain.

6 Simulations

The sources in Fig. 1 are mixed by a stationary channel, zero mean Gaussian noise at $10dB$ SNR is added, and the non-linearity is $f_i(y_i(k)) = \tanh(y_i(k))$. The signals are separated, in 30 independent trials, with conventional NGA, and NGA in the wavelet domain. Since PI in the time-scale domain is a valid performance measure, the performance indices obtained with the two

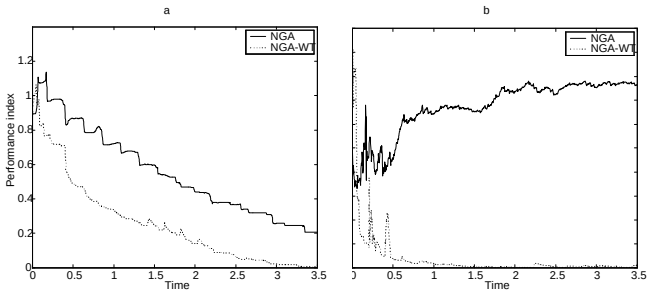

Figure 5: PIs obtained with NGA and NGA in the wavelet domain when the sensors are mixtures of a) two super-Gaussian sources, and b) one super-Gaussian and two sub-Gaussian sources.

methods are compared in Fig. 5a. It shows that the wavelet domain approach results in faster convergence speed than time domain NGA. Following convergence, a lower PI is also obtained. Improved algorithm performance is largely due to the sample pdf of the sources being closer to a Gaussian distribution than the pdf of their wavelet coefficients. Fig. 5b shows the average performance of the two methods when the foetal ECG and two sub-Gaussian sources are mixed by a time-invariant instantaneous channel, and the activation function is $f_i(y_i(k)) = \tanh(6 * y_i(k))$. The results clearly show that when operating in the time domain NGA diverges, thus failing to separate the sources. Conversely, when separation is carried out in the time-scale domain the algorithm converges quite quickly and the performance index remains low thereafter.

7 Conclusions

NGA operating in the wavelet domain results in higher convergence speed than when it separates in the time domain. Furthermore, this approach separates mixtures of sub- and super-Gaussian signals without the need to switch between different non-linearities.

References

[1] S. Amari and A. Cichocki, "Adaptive blind signal processing - neural network approaches," *Proceedings of the IEEE*, pp. 2026–2048, 1998.

[2] I. Daubechies, *Ten lectures on wavelets*. SIAM, 1992.

[3] R. Buccigrossi and E. Simoncelli, "Image compresion via joint statistical characterization in the wavelet domain," tech. rep., Univ. of Pennsylvania, 1997.

[4] S. Douglas and A. Cichocki, "Convergence analysis of local algorithms for blind decorrelation," in *ANIPS*, pp. 2–7, 1996.

[5] S. Amari, T. Chen, and A. Cichocki, "Stability analysis of adaptive blind source separation," *Neural Networks*, pp. 1345–1351, 1997.