Third IEEE Signal Processing Workshop on Signal Processing Advances in Wireless Communications,Taoyuan, Taiwan, March 20-23, 2001

Unbiased Blind Separation Using the Threshold Nonlinearity

Heinz Mathis
Signal and Information Processing Lab.
Swiss Federal Institute of Technology
CH - 8092 Zurich, Switzerland
e-mail: mathis@isi.ee.ethz.ch

Marcel Joho
Signal and Information Processing Lab.
Swiss Federal Institute of Technology
CH - 8092 Zurich, Switzerland
e-mail: joho@isi.ee.ethz.ch

Abstract — **The stability analysis of blind signal separation algorithms using the threshold nonlinearity is extended to discrete distributions, such as used in digital communication systems. Conditions for the threshold parameter are derived. An extension to the standard algorithm mitigates separation coefficient bias, which is introduced by additive noise at the sensors. Simulations quantify the improvement in terms of steady-state interchannel interference.**

I. Introduction

The blind separation task of a linear instantaneous mixture of signals is a problem that is frequently solved by the natural gradient update equation (see e.g. [1])

$$\mathbf{W}_{t+1} = \mathbf{W}_t + \mu \left(\mathbf{I} - \mathbf{g}(\mathbf{u})\mathbf{u}^T\right)\mathbf{W}_t \tag{1}$$

where $\mathbf{W}$ is the separation matrix used to unwind the mixing process given by the linear mixing matrix $\mathbf{A}$, so that the recovered signals are

$$\mathbf{u} = \mathbf{W}\mathbf{x} = \mathbf{W}(\mathbf{A}\mathbf{s}+\mathbf{n}). \tag{2}$$

$\mathbf{s} = [s_1, \ldots, s_{M_S}]^T$, $\mathbf{x} = [x_1, \ldots, x_{M_S}]^T$, and $\mathbf{n} = [n_1, \ldots, n_{M_S}]^T$ denote the vectors of the source signals, the mixed signals and the noise, respectively. $\mathbf{g}(\mathbf{u})$ is the vector containing the output elements after the nonlinearity. In this paper, we assume the same nonlinearity for all signals, thus $\mathbf{g}(\mathbf{u}) = [g(u_1), \ldots, g(u_{M_S})]^T$. We also assume $\mathbf{A}$ to be square, i.e., we have the same number of sensors as sources. The shape of the nonlinearity $g(.)$, which is an essential part of this type of algorithm, depends on the distribution of the source signals. For communication signals, the threshold nonlinearity

$$g(u_i) = \begin{cases} 0, & |u_i| < \vartheta \\ a\,\mathrm{sign}(u_i), & |u_i| \geq \vartheta \end{cases} \tag{3}$$

has been demonstrated to work well in terms of convergence speed of the separation task [2]. For unit-variance output signal the scaling condition

$$\int_{-\infty}^{\infty} p_{U_i}(u_i) g(u_i) u_i \, du_i = 1 \tag{4}$$

if $p_{U_i}(.)$ is the output distribution with unit variance $\sigma_{U_i}^2 = 1$, must be satisfied. For the threshold nonlinearity given by (3), this means

$$a = \frac{1}{2\int_{\vartheta}^{\infty} p_{U_i}(u_i) u_i \, du_i}. \tag{5}$$

II. Stability analysis for discrete distributions

While computer simulations have shown the convergence of the separation algorithm (1) using the threshold nonlinearity [2], its stability has only been proven explicitly for continuous distributions [3]. In the following, this stability analysis is extended to discrete distributions. The stability condition for the threshold nonlinearity is [3]

$$\frac{p_{U_i}(\vartheta)}{\int_{\vartheta}^{\infty} p_{U_i}(u_i) u_i \, du_i} > 1, \qquad i = 1 \ldots M_S. \tag{6}$$

with M_S denoting the number of sources. Whereas the integral in the denominator of (6) can be written as a sum for discrete distributions

$$\int_{\vartheta}^{\infty} p_{U_i}(u_i) u_i \, du_i = \sum_{k,\, A_k \geq \vartheta} \Pr(u_i = A_k) A_k \tag{7}$$

the evaluation of a probability density, as appearing in the numerator of (6), needs a closer look. Close to an equilibrium point we may model the output distribution as a convolution of the discrete probability model of the sources by some mixing noise distribution, which is Gaussian distributed. The probability density at a certain constellation point is therefore the discrete probability of that point multiplied by the mode of the Gaussian kernel $\frac{1}{\sqrt{2\pi}\sigma_N}$, with σ_N^2 being the variance of the mixing noise. In other words, the discrete-level distribution is convolved with the probability density function (pdf) of a Gaussian noise signal. The resulting pdf for a 4-PAM (Pulse Amplitude Modulation) signal with -20 dB residual mixing noise is depicted in Fig. 1. The stability regions are thus dependent on the mixing noise. Figs. 2 and 3 show the stable regions as derived from the evaluation of Eq. (6) for BPSK (Binary Phase Shift Keying) and 4-PAM, respectively. It is interesting to note that in addition to the region around the outer symbols, which looks similar for BPSK and 4-PAM, there is a further stable region around the inner symbols in the case of 4-PAM.

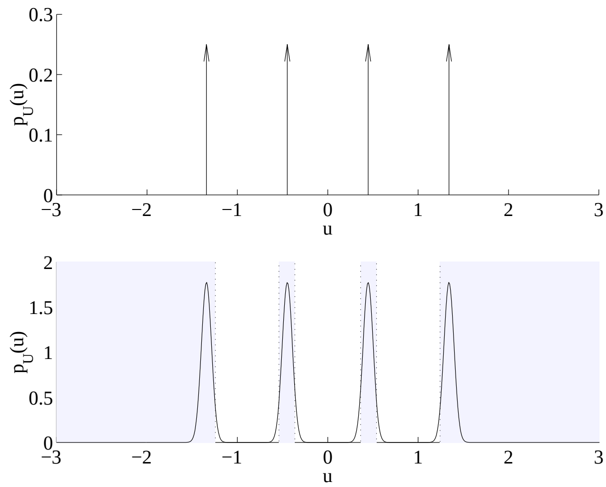

Fig. 1: Top: discrete distribution of 4-PAM signal with unit variance. Bottom: pdf of 4-PAM signal with additive Gaussian noise, SNR$=25$ dB. The shaded regions indicate the stable region of the threshold parameter ϑ as derived from Fig. 3.

It becomes apparent that for a stable update equation for BPSK signals, the threshold ϑ has to be in the neighborhood of the symbol amplitude, otherwise the algorithm becomes unstable. A closer look at (6) reveals that the mixing noise keeps the algorithm stable through a finite pdf in the neighborhood of the symbol amplitude. In other words, if the threshold ϑ is chosen too far away from the symbol amplitude, more mixing noise is needed to satisfy (6). For BPSK, the threshold ϑ should therefore be chosen directly at the symbol amplitude $A_1 = 1$. For this choice, with probability 0.5 the signal will be larger (smaller) than the threshold, enforcing a choice of the scaling factor $a = 2$ in order to satisfy the scaling condition (4). For all choices of the threshold ϑ smaller than $A_1 = 1$ and low residual mixing, a scaling factor of $a = 1$ is needed. For larger threshold values, the gain gets impractically high due to (4). For M-PAM signals with $M > 2$, stable algorithms can be obtained by setting the threshold to the outermost symbol amplitude

$$\vartheta = \sqrt{\frac{3(M-1)}{M+1}}. \tag{8}$$

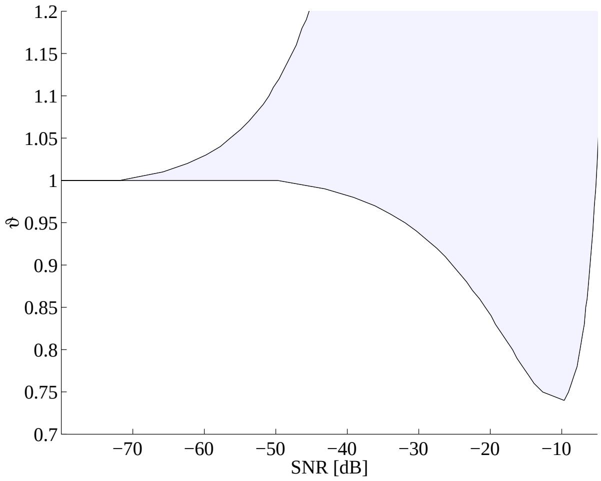

Fig. 2: Stable region (shaded) for noisy BPSK signals and the threshold nonlinearity.

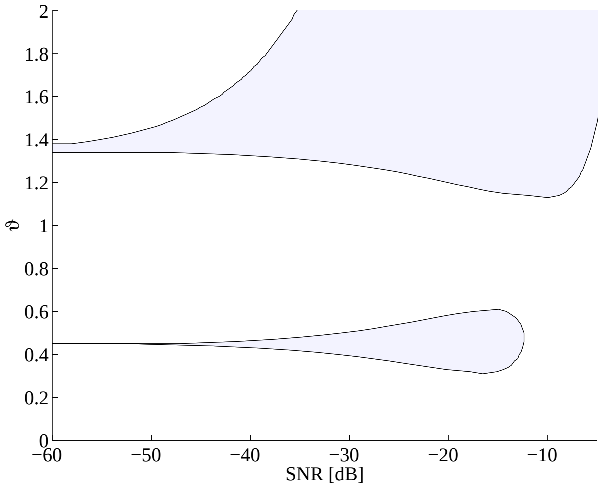

Fig. 3: Stable regions (shaded) for noisy 4-PAM signals and the threshold nonlinearity.

The corresponding gain is

$$a = M\sqrt{\frac{M+1}{3(M-1)}}. \tag{9}$$

III. Unbiased blind separation

Algorithms of the form given by (1) lead to a biased solution, if additive noise is present at the sensors. Particularly in communication environments this is often the case. A bias is generated, because by making the source signals independent through a linear combination of the input signals—essentially following a zero-forcing criterion—the noise signals get correlated at the output, introducing dependency between the sensor signals. A combined learning process involving unsupervised learning for the separation and supervised learning for noise reduction was presented in [4]. The lack of a noise reference in practice, however, makes this approach unapplicable to most common problems. It is possible to devise an update equation with an additional term in the update equation, which involves either some expectation of the signal derivatives [5] or their stochastic versions [6]

$$\mathbf{W}_{t+1} = \mathbf{W}_t + \mu\left(\mathbf{I} - \mathbf{g}(\mathbf{u})\mathbf{u}^T + \mathbf{B}\mathbf{W}_t\mathbf{R}_N\mathbf{W}_t^T\right)\mathbf{W}_t \tag{10}$$

where $\mathbf{B}$ is a diagonal matrix with entries

$$b_{ii} = E\left\{\frac{dg(u_i)}{du_i}\right\} \tag{11}$$

and $\mathbf{R}_N$ is the covariance matrix of the noise contribution. Eq. (10) can be seen by defining an unbiased estimate of the source signal as

$$\hat{\mathbf{u}} = \mathbf{W}\mathbf{A}\mathbf{s}. \tag{12}$$

The recovered signals can then be written as

$$\mathbf{u} = \hat{\mathbf{u}} + \mathbf{W}\mathbf{n}. \tag{13}$$

The expectation of the input-output product of the nonlinearity can be expressed as

$$E\left\{\mathbf{g}(\mathbf{u})\mathbf{u}^T\right\} = E\left\{\mathbf{g}(\hat{\mathbf{u}}+\mathbf{W}\mathbf{n})(\hat{\mathbf{u}}+\mathbf{W}\mathbf{n})^T\right\}. \tag{14}$$

A first-order truncated Taylor series expansion of the nonlinearity around $\hat{\mathbf{u}}$ yields

$$\mathbf{g}(\hat{\mathbf{u}}+\mathbf{W}\mathbf{n}) = \mathbf{g}(\hat{\mathbf{u}}) + \operatorname{diag}\left(\mathbf{g}'(\hat{\mathbf{u}})\right)\mathbf{W}\mathbf{n} \tag{15}$$

where $\operatorname{diag}(\mathbf{g}'(\hat{\mathbf{u}}))$ is a diagonal matrix with the elements $g'(u_i)$ sitting on the diagonal. Inserted into (14) this results in

$$\begin{aligned} E\left\{\mathbf{g}(\mathbf{u})\mathbf{u}^T\right\} =& E\left\{\mathbf{g}(\hat{\mathbf{u}})\hat{\mathbf{u}}^T\right\} + E\left\{\operatorname{diag}\left(\mathbf{g}'(\hat{\mathbf{u}})\right)\mathbf{W}\mathbf{n}\hat{\mathbf{u}}^T\right\} \\ &+ E\left\{\mathbf{g}(\hat{\mathbf{u}})\mathbf{n}^T\mathbf{W}^T\right\} + E\left\{\operatorname{diag}\left(\mathbf{g}'(\hat{\mathbf{u}})\right)\mathbf{W}\mathbf{n}\mathbf{n}^T\mathbf{W}^T\right\}. \end{aligned} \tag{16}$$

Since the noiseless estimate is uncorrelated to the noise, the two middle terms of the RHS of Eq. (16) are zero, hence

$$E\left\{\mathbf{g}(\mathbf{u})\mathbf{u}^T\right\} = E\left\{\mathbf{g}(\hat{\mathbf{u}})\hat{\mathbf{u}}^T\right\} + E\left\{\operatorname{diag}\left(\mathbf{g}'(\hat{\mathbf{u}})\right)\mathbf{W}\mathbf{R}_N\mathbf{W}^T\right\}. \tag{17}$$

The second term of the RHS of Eq. (17) is now identified as the bias term and has to be subtracted in the original update equation, leading to (10).

Although the threshold nonlinearity is non-differentiable, its expectation can be expressed by integration over a dirac impulse

$$\begin{aligned} E\left\{g'(u_i)\right\} &= \int_{-\infty}^{\infty} p_{U_i}(u_i) g'(u_i) du_i \\ &= \int_{-\infty}^{\infty} p_{U_i}(u_i) a(\delta(u_i+\vartheta) + \delta(u_i-\vartheta)) du_i \\ &= 2a \cdot p_{U_i}(\vartheta). \end{aligned} \tag{18}$$

In the following we assume equal noise power σ_N^2 at each of the sensors, but uncorrelated noise signals, so that the sensor noise vector is described by $\mathcal{N}(0, \sigma_N^2 \cdot \mathbf{I})$, or by $\mathbf{R}_N = \sigma_N^2 \cdot \mathbf{I}$. This is a reasonable assumption, as very often noise is of thermal origin, therefore given by temparature and noise figure and as such equal but mutually uncorrelated for all the channels. Furthermore, the noise power σ_N^2 is presumed to be known, be that from theoretical calculations of thermal noise or by estimating it, e.g., using minor component analysis in an overdetermined separation case [7].

For identical distributions of all signals, Eq. (10) can be simplified to

$$\mathbf{W}_{t+1} = \mathbf{W}_t + \mu\left(\mathbf{I} - \mathbf{g}(\mathbf{u})\mathbf{u}^T + \sigma_N^2 b \mathbf{W}_t \mathbf{W}_t^T\right)\mathbf{W}_t \tag{19}$$

where

$$b = E\left\{g'(u)\right\}. \tag{20}$$

For the uniform distribution, which is a good approximation for M-ary distributions where M is high, with unit variance, implying that the threshold function is properly scaled according to [2]

$$a = \frac{2\sqrt{3}}{3-\vartheta^2} \tag{21}$$

we get

$$b = E\left\{g'(u)\right\} = \frac{2}{3-\vartheta^2}. \tag{22}$$

If the sources have discrete distributions rather than continuous ones, the update equation (19) is not optimal. Since the noise carried through to the outputs determines the probability density at the threshold level ϑ, the term $p_{U_i}(\vartheta)$ in (18) depends on the separation matrix. For an M-ary signalling scheme (e.g. M-PAM) we can write for the probability density at the ith output

$$p_{U_i}(\vartheta) = \frac{1}{M}\frac{1}{\sqrt{2\pi}\sigma_N}\frac{1}{\sqrt{\sum_{k=1}^{M_S} w_{ik}^2}} \tag{23}$$

with w_{ik} being the i,kth element of the separation matrix $\mathbf{W}$, describing the path from the kth sensor to the ith output. For M-PAM signals, using (9) and (23) in (18) and the update equation (10) we get

$$\mathbf{W}_{t+1} = \mathbf{W}_t + \mu\Bigg(\mathbf{I} - \mathbf{g}(\mathbf{u})\mathbf{u}^T + \sqrt{\frac{2(M+1)}{3\pi(M-1)}}\sigma_N\left(\operatorname{diag}\left(\mathbf{W}_t\mathbf{W}_t^T\right)\right)^{-\frac{1}{2}}\mathbf{W}_t\mathbf{W}_t^T\Bigg)\mathbf{W}_t \tag{24}$$

where $\operatorname{diag}\left(\mathbf{W}_t\mathbf{W}_t^T\right)$ means the matrix $\mathbf{W}_t\mathbf{W}_t^T$ with suppressed off-diagonal terms and may also be written by the use of the Hadamard or Schur product: $\operatorname{diag}\left(\mathbf{W}_t\mathbf{W}_t^T\right) = \mathbf{I} \circ (\mathbf{W}_t\mathbf{W}_t^T)$.

IV. MMSE vs. zero-forcing solution

Very often in data communications we are not interested in the solution of $\mathbf{W}$ that directly inverts $\mathbf{A}$—the so-called zero-forcing solution—due to problems associated with noise enhancement at frequencies close to zeros of the system transfer function. In terms of signal purity—the essence of low bit-error rates—we do not care where unwanted contributions to the signal comes from; signals from other channels or thermal noise. This is of course only the case if channels are not jointly detected. For single-channel detection, the proper criterion to choose is the minimum mean squared error (MMSE). If we have a zero-forcing solution $\mathbf{W}_{\mathrm{ZF}}$ we can, by looking at the MMSE solution [7]

$$\mathbf{W}_{\mathrm{MMSE}} = \mathbf{A}^T(\mathbf{A}^T\mathbf{A}^T + \sigma_n^2\mathbf{I})^{-1} \tag{25}$$

reformulate the MMSE solution in terms of the zero-forcing solution. To this end we note that the zero-forcing solution is the inverse of the system matrix but for some permutation

$$\mathbf{W}_{\mathrm{MMSE}} = \mathbf{P}\mathbf{A}^{-1}. \tag{26}$$

Using (26) in (25) leads to

$$\mathbf{W}_{\mathrm{MMSE}} = \mathbf{P}^T\mathbf{W}_{\mathrm{ZF}}^{-T}(\mathbf{W}_{\mathrm{ZF}}^{-1}\mathbf{P}\mathbf{P}^T\mathbf{W}_{\mathrm{ZF}}^{-T} + \sigma_n^2\mathbf{I})^{-1}. \tag{27}$$

Of course, $\mathbf{P}\mathbf{P}^T = \mathbf{I}$, and by premultiplying the solution in (27) by $\mathbf{P}$ we do not challenge its validity, so we get

$$\mathbf{W}_{\mathrm{MMSE}} = \mathbf{W}_{\mathrm{ZF}}^{-T}(\mathbf{W}_{\mathrm{ZF}}^{-1}\mathbf{W}_{\mathrm{ZF}}^{-T} + \sigma_n^2\mathbf{I})^{-1}. \tag{28}$$

V. Simulation

In the following, results of computer simulations of the blind separation using the bias-removal method suggested above are shown. Some important parameters influencing the performance were taken from [6], such as the number of sources and sensors $M_\mathrm{s} = 3$, the mixing matrix

$$\mathbf{A} = \begin{bmatrix} 0.4 & 1.0 & 0.7 \\ 0.6 & 0.5 & 0.5 \\ 0.3 & 0.7 & 0.2 \end{bmatrix} \tag{29}$$

and the condition $\mathbf{W}_0\mathbf{W}_0^T = 0.25 \cdot \mathbf{I}$ (implying that $\mathbf{W}_0$ is a scaled orthogonal matrix) for the one hundred trials with a different initial separation matrix. In the first experiment three uniformly distributed source signals were mixed, and noise was added at the sensors with $\sigma_N^2 = 0.01$. The noise level was assumed to be known to the algorithm. The mixed noisy signals were then separated using the threshold nonlinearity and the update equation (1). The step size was adjusted without noise to obtain an interchannel interference level of $-35\,\mathrm{dB}$ and then fixed to $\mu = 0.00032$ for the other simulations. The performance measure used in the plots is calculated as a function of the global system matrix $\mathbf{P} = [p_{ik}]$

$$J_{\mathrm{ICI}}(\mathbf{P}) = \frac{1}{M_\mathrm{s}}\left(\sum_{i=1}^{M_\mathrm{s}}\frac{\sum_{k=1}^{M_\mathrm{s}} p_{ik}^2}{\max_k p_{ik}^2}\right) - 1 \tag{30}$$

and expresses average interchannel interference. Fig. 4 reveals the convergence improvement of the modified algorithm compared to the standard algorithm without bias removal. Still better results were obtained for binary-distributed signals. Three binary-distributed source signals were mixed using the same mixing matrix as above. To show clearer differences between the algorithms, the noise was increased by 5 dB, resulting in $\sigma_N^2 = 0.0315$. Fig. 5 shows that the modified algorithm is in fact capable of completely removing any bias, albeit at a lower convergence speed. Again, step sizes were chosen equal ($\mu = 0.0018$) for all three cases. It was also observed that the modified algorithm with certain noise levels (e.g. $\sigma_N^2 = 0.01$) consistently outperformed the standard algorithm with no noise. This surprising effect is due to an increased stability region (see Fig. 2) for lower SNRs. The additive noise has then a positive dithering effect. With other nonlinearities (e.g. $g(u) = u^3$) or other distributions (e.g. uniform distribution), this effect cannot be observed. It is only the special arrangement of high derivative of the nonlinearity at the level of spikes in the pdf which benefits from additional noise.

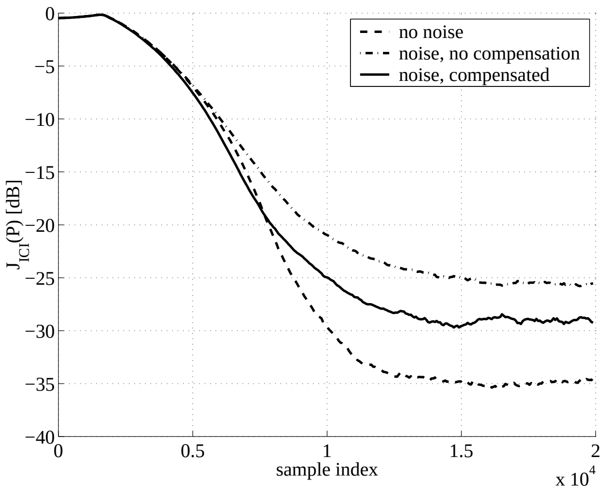

Fig. 4: Separation convergence of bias removal algorithm for uniform distributions.

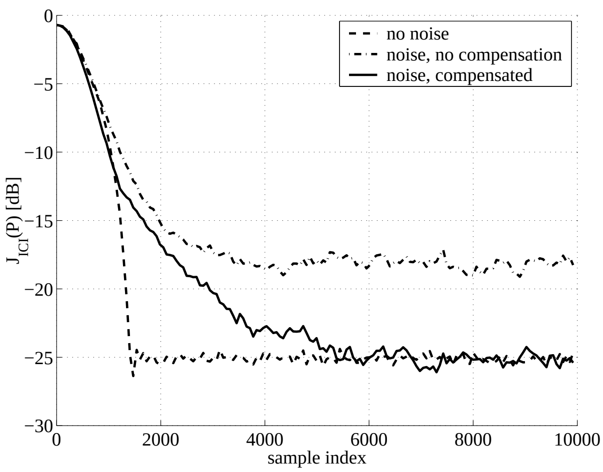

Fig. 5: Separation convergence of bias removal algorithm for binary distributions.

VI. Conclusion

A stability analysis of blind separation algorithms for discrete signals has shown that the threshold nonlinearity is able to separate sub-Gaussian signals with discrete distributions. Extensions reducing the biasing effects, which result from additive sensor noise, have been introduced and verified by simulation. Under certain circumstances, additive noise might even improve convergence properties, if bias removal techniques are properly applied. From an unbiased separation solution, which satisfies a zero-forcing criterion, an MMSE solution can be readily obtained by simple matrix operations. The methods for bias removal shown in this paper can easily be extended to complex quadrature signals. [2] gives some hints as to how the update equations have to be modified.

Acknowledgment

The authors would like to thank Scott C. Douglas for fruitful discussions on the subject of stability analysis.

References

[1] S. C. Douglas and S.-I. Amari, "Natural-gradient adaptation," in *Unsupervised Adaptive Filtering, Volume I: Blind Source Separation*, S. Haykin, Ed. 2000, pp. 13–61, John Wiley & Sons.

[2] H. Mathis, M. Joho, and G. S. Moschytz, "A simple threshold nonlinearity for blind signal separation," in *IEEE International Symposium on Circuits and Systems*, Geneva, Switzerland, May 28–31, 2000, vol. IV, pp. 489–492.

[3] H. Mathis, T. P. von Hoff, and M. Joho, "Blind separation of mixed-kurtosis signals using an adaptive threshold nonlinearity," in *Proc. International Conference on Independent Component Analysis and Blind Signal Separation*, Helsinki, Finland, June 19–22, 2000, pp. 221–226.

[4] A. Cichocki, W. Kasprzak, and S. Amari, "Adaptive approach to blind source separation with cancellation of additive and convolutional noise," in *International Conference on Signal Processing*, Beijing, China, September 1996, pp. 412–415.

[5] A. Cichocki, S. C. Douglas, and A. Amari, "Robust techniques for independent component analysis (ICA) with noisy data," *Neurocomputing*, vol. 22, pp. 113–129, November 1998.

[6] S. C. Douglas, A. Cichocki, and S. Amari, "Bias removal technique for blind source separation with noisy measurements," *Electronics Letters*, vol. 34, no. 14, pp. 1379–1380, July 1998.

[7] M. Joho, H. Mathis, and R. H. Lambert, "Overdetermined blind source separation: Using more sensors than source signals in a noisy mixture," in *Proc. International Conference on Independent Component Analysis and Blind Signal Separation*, Helsinki, Finland, June 19–22, 2000, pp. 81–86.