A Class of Frobenius Norm-Based Algorithms Using Penalty Term and Natural Gradient for Blind Signal Separation

Uttachai Manmontri and Patrick A. Naylor, *Member, IEEE*

***Abstract*—We consider the blind signal separation (BSS) problem of instantaneous mixtures using penalty term and natural gradient. A class of Frobenius norm-based algorithms consisting of the offline/block processing (BP), online processing (OP) algorithms, and their normalized versions is proposed for separating nonstationary and nonwhite signals. The BP and OP algorithms, respectively, suitable for blind separation with offline and online data, are derived by using the nonstationarity and nonwhiteness of signals and the natural gradient method in conjunction with an appropriate penalty term. Associated with almost all algorithms employing a gradient method is a gradient noise problem. We thus develop, from BP and OP, their normalized versions in which the update of an unknown demixing matrix is based on the minimal disturbance principle. We show that the resulting updates are in the same direction as those of the original algorithms but with a scaling factor whose upper bound is unity. Algorithms using the nonstationarity and nonwhiteness properties have been proposed before but, due to the use of logarithms in their derivation, they are not capable of separating signals that are not persistently active and require regularization parameters to mitigate the problem. In this paper, the superior performance of the proposed algorithms to the previously proposed logarithm-based algorithms with and without regularization when separating nonpersistently active source signals is presented through some illustrative numerical experiments.**

***Index Terms*—Blind signal separation (BSS), natural gradient methods, penalty term, second-order statistics.**

I. Introduction

BLIND signal separation (BSS) of both instantaneous and convolutive mixtures has received considerable attention over the last two decades in the fields of signal processing and communications [1], [2].

The classical instantaneous BSS mixing and demixing models of equal numbers of source signals and observed signals can be described in their basic forms by

$$\mathbf{x}(n) = \mathbf{A}\mathbf{s}(n) \qquad (1)$$

$$\mathbf{y}(n) = \mathbf{W}\mathbf{x}(n) \qquad (2)$$

where $\mathbf{s}(n) = [s_1(n), s_2(n), \ldots, s_N(n)]^T$ is the source signals vector, $\mathbf{x}(n) = [x_1(n), x_2(n), \ldots, x_N(n)]^T$ is the observed signals vector, $\mathbf{y}(n) = [y_1(n), y_2(n), \ldots, y_N(n)]^T$ is an estimate of the source signals vector called the output signals vector, $\mathbf{A}$ is an $N \times N$ unknown mixing matrix, $\mathbf{W}$ is a corresponding demixing matrix to be computed, the superscript T denotes transposition, and n is the sample index.

By the establishment of second-order statistical relations from these signals, instantaneous BSS can be solved up to arbitrary scaling and permutations of the source signals [3], [4].

In this paper, we consider the instantaneous BSS problem and propose a class of Frobenius norm-based algorithms for separating signals within the framework of second-order statistics. The proposed algorithms utilize the nonstationarity and nonwhiteness properties of signals together with the penalty term and the natural gradient method.

The use of nonstationarity and nonwhiteness properties has been proposed before in [5] for instantaneous BSS and in [2] for convolutive BSS. The technique of [5], which exploits a whitening process followed by the Jacobi-like technique [6] to seek an orthogonal matrix that jointly diagonalizes a set of whitened signals correlation matrices at different time intervals and several time lags (see also e.g., [1], [7], and [8] for the approach employing only the nonwhiteness property), distorts the structure of the desired demixing matrix in the presence of additive noise [9]. This structural distortion cannot be resolved by any postwhitening process and leads to nonorthogonal search techniques subsequently presented in [4], [10], and [11].

The technique of [2] introduces a generalization of the logarithm-based cost function originally proposed in [12] to a set of novel correlation matrices and employs natural gradient method to seek the desired demixing matrix. It can be shown by using the Oppenheim's inequality [13] that the proposed cost function will attain the minimum at the desired demixing matrix [2]. Although the logarithm-based algorithm [2] employs a systematic search technique through the natural gradient adaptation, its proposed cost function which utilizes the logarithmic function inevitably requires the correlation matrices to be strictly positive definite. As a consequence, all source signals are required to be persistently active with variances always greater than zero so that the logarithm-based cost function will be differentiable. This problem, as suggested in [14], raises the need for regularization parameters to take care of nonpersistently active signals when performing the logarithm-based algorithm. The regularization parameters can, however, lead to biased solutions when they increase robustness to nonpersistently active signals. This

Manuscript received November 8, 2007; revised May 5, 2008. Published July 16, 2008 (projected). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Gael Richard.

U. Manmontri was with the Electrical and Electronic Engineering Department, Imperial College London, London SW7 2AZ, U.K. He is now with the Thailand Institute of Scientific and Technological Research, Pathumthani 12120, Thailand (e-mail: uttachai@tistr.or.th).

P. A. Naylor is with the Electrical and Electronic Engineering Department, Imperial College London, London SW7 2AZ, U.K. (e-mail: p.naylor@imperial.ac.uk).

Digital Object Identifier 10.1109/TASL.2008.2001388

1558-7916/$25.00 © 2008 IEEE

Authorized licensed use limited to: Imperial College London. Downloaded on January 4, 2010 at 08:09 from IEEE Xplore. Restrictions apply.

is due to the fact that the basic feature of the regularization is a compromise between the fidelity to data and the fidelity to prior information about the solution [14].[1] To avoid the biased solutions, we therefore propose the use of Frobenius norm together with the nonstationarity and the nonwhiteness properties of signals to form the cost function and find the solution of the problem using the natural gradient method. By the use of Frobenius norm, the proposed algorithms, without the need of regularization, will be capable of separating signals that are not persistently active. In addition, we also derive the efficient normalized versions of the proposed algorithms to cope with the problem of gradient noise amplification.

We employ the following assumptions in derivations of our proposed algorithms.

A1: $\mathbf{A}$ is a full rank square matrix.

A2: Source signals are a zero mean, nonstationary and nonwhite process with

1) $E[s_i(n)] = 0, \forall i = 1, 2, \ldots, N$;
2) $E[s_i^2(n')] \neq E[s_i^2(n'')], \exists i$ and $\exists n' \neq n''$;
3) $E[s_i(n)s_i(n-\tau)] \neq 0, \exists i$ and $\exists \tau \neq 0$;
also, each source signal is uncorrelated with
4) $E[s_i(n)s_j(n-\tau)] = E[s_i(n)]E[s_j(n-\tau)], \forall 1 \leq i \neq j \leq N$ and $\forall \tau = 0, 1, \ldots, \Gamma$;

where $E[\cdot]$ denotes the statistical expectation operator, $\forall$ denotes *for all*, $\exists$ denotes *for some*, and Γ is the maximum nonzero time lag of interest.

A1 ensures that all source signals are observed in the form of $\mathbf{x}(n)$ by the rank of $\mathbf{A}$ and makes a solution to the problem feasible. A2 is a key assumption that forms the joint diagonalization criterion to be used in our proposed algorithms.

The remainder of the paper is outlined as follows. We propose offline, online Frobenius norm-based algorithms and their normalized versions in Sections II–IV, respectively. In Section II, the offline/block processing (BP) algorithm is derived by employing a set of local time-average correlation matrices. The prerequisites for using second-order statistics to separate signals and the existence of global minimum in the neighborhood of a penalized region are also given. The online processing (OP) algorithm is presented in Section III and is derived in a similar fashion using a set of current time-average correlation matrices. Normalized versions of the BP and OP algorithms, based on the minimal disturbance principle, are presented in Section IV. Computational complexity of all proposed algorithms is presented in Section V. Illustrative numerical results are shown in Section VI and we conclude in Section VII.

The main contribution of this paper is to present the BP, OP algorithms and their normalized versions together with supporting analysis. Earlier work with these algorithms was given in [15] and [16]. We focus on real-valued data, although the extension to the complex-valued data is straightforward.

II. Offline/Block Processing (BP) Algorithm

In this section, we derive an offline algorithm using the Frobenius norm. The main reason to use the Frobenius norm rather than the logarithmic function is that it does not require the source signals to be persistently active. However, the Frobenius norm-based approach, unlike the logarithm-based approach, requires an appropriate penalty term to prevent the algorithms from converging to trivial solutions. Also, as we employ the natural gradient to seek a demixing matrix instead of the ordinary gradient, it is thus worth noting that the BSS algorithm that employs the natural gradient adaptation [17], or equivalently, the relative gradient adaptation [18] to seek a demixing matrix offers preferable convergence properties when compared to its ordinary gradient counterpart [2], [14], [17], [18].[2]

A. Second-Order Statistical Relations Using Local Time-Average Correlation Matrices

The BP algorithm employs both the nonstationarity and the nonwhiteness properties through a set of blocks of nonstationary but locally stationary observed signals. We require an estimator of the correlation matrix for nonstationary source signals. Let $\mathbf{R}_{\mathbf{x}}^{(k,\tau)}$ denote an estimate of the correlation matrix in the kth block, $k = 1, 2, \ldots, K$, and at the time lag τ, $\tau = 0, 1 \ldots, \Gamma$, where Γ is the maximum nonzero time lag of interest. We refer to this estimate as the local time-average correlation matrix of the observed signals

$$\mathbf{R}_{\mathbf{x}}^{(k,\tau)} = \frac{1}{L_k}\mathbf{X}_k(n)\mathbf{X}_k^T(n-\tau) \tag{3}$$

where L_k is the length of the kth block and $\mathbf{X}_k(n)$ is an $N \times L_k$ matrix of stationary samples in the kth block.

The block of stationary samples can be expressed in expanded form as

$$\mathbf{X}_k(n) = [\mathbf{x}(L_{k-1} - D_k + 1), \ldots, \mathbf{x}(L_{k-1} - D_k + L_k)] \tag{4}$$

where D_k is the overlap factor with $0 \leq D_k < L_k$, $D_1 = 0$ and $L_0 = 0$.

Unlike other block methods (see, e.g., [19], [2]), an overlap factor is proposed here to facilitate more flexibility in sectioning the signals. To obtain a set of these locally stationary blocks, it is essential that the parameters used in (4) are chosen properly. These parameters depend on the nature of source signals and, in practice, some *a priori* knowledge of the source signals is required to choose appropriate values. For nonstationary and nonwhite signals such as speech signals, the length of the block is chosen using the average duration of phonemes [19] and the nonzero time lag is chosen to be a small number over which speech exhibits high correlation [2]. Using (3), we write the second-order statistical relations

$$\mathbf{R}_{\mathbf{x}}^{(k,\tau)} = \mathbf{A}\Lambda_{\mathbf{s}}^{(k,\tau)}\mathbf{A}^T \tag{5}$$
$$\mathbf{R}_{\mathbf{y}}^{(k,\tau)} = \mathbf{W}\mathbf{R}_{\mathbf{x}}^{(k,\tau)}\mathbf{W}^T \tag{6}$$

where $\Lambda_{\mathbf{s}}^{(k,\tau)}$, $\mathbf{R}_{\mathbf{x}}^{(k,\tau)}$, and $\mathbf{R}_{\mathbf{y}}^{(k,\tau)}$ are, respectively, the local time-average correlation matrices of the source signals, the observed signals and the output signals, all at (k,τ). It can be seen from (5) and (6) that the blind separation of nonstationary and nonwhite signals becomes one of finding $\mathbf{W}$ that jointly diagonalizes a set of local time-average correlation matrices $\mathbf{R}_{\mathbf{x}}^{(k,\tau)}$.

[1] The use of the logarithmic function and a set of positive definite matrices along with Oppenheim's inequality, which together form the logarithm-based algorithm, as well as the use of regularization will be presented in Section IV.

[2] The superior performance of the natural gradient to the ordinary gradient are presented both theoretically and experimentally in [17] and references therein.

Authorized licensed use limited to: Imperial College London. Downloaded on January 4, 2010 at 08:09 from IEEE Xplore. Restrictions apply.

B. Prerequisites for Using Local Time-Average Correlation Matrices

Before proceeding to the derivation of the BP algorithm, it is useful to study what level of nonstationarity and nonwhiteness is needed in the source signals for blind separation to be feasible. The following prerequisites give *necessary* requirements for source signals to be separated through a set of local time-average correlation matrices.

Prerequisites 1: Let $\rho_{s_1}^{(k,\tau)}, \rho_{s_2}^{(k,\tau)}, \ldots, \rho_{s_N}^{(k,\tau)}$ be corresponding diagonal entries of $\Lambda_{\mathbf{s}}^{(k,\tau)}$,

1) *nonstationarity:* $\mathbf{p}_{s_i}^{(\tau)} = [\rho_{s_i}^{(1,\tau)}, \rho_{s_i}^{(2,\tau)}, \ldots, \rho_{s_i}^{(K,\tau)}]$ and $\mathbf{p}_{s_j}^{(\tau)} = [\rho_{s_j}^{(1,\tau)}, \rho_{s_j}^{(2,\tau)}, \ldots, \rho_{s_j}^{(K,\tau)}]$ must be linearly independent, $\exists\tau = 0, 1, \ldots, \Gamma, \forall 1 \leq i \neq j \leq N$;
2) *nonwhiteness:* $\tilde{\mathbf{p}}_{s_i}^{(k)} = [\rho_{s_i}^{(k,0)}, \rho_{s_i}^{(k,1)}, \ldots, \rho_{s_i}^{(k,\Gamma)}]$ and $\tilde{\mathbf{p}}_{s_j}^{(k)} = [\rho_{s_j}^{(k,0)}, \rho_{s_j}^{(k,1)}, \ldots, \rho_{s_j}^{(k,\Gamma)}]$ must be linearly independent, $\exists k = 1, 2, \ldots, K, \forall 1 \leq i \neq j \leq N$.

Proof: See Appendix A. □

According to the proof of necessity of the Prerequisites 1, it can be seen that, at least, $\mathbf{p}_{s_i}^{(\tau)}$ has to be linearly independent of $\mathbf{p}_{s_j}^{(\tau)}$ when the nonstationarity property is the only property that is used in blind separation of signals and that, when the nonwhiteness property is the only used property, at least, $\tilde{\mathbf{p}}_{s_i}^{(k)}$ and $\tilde{\mathbf{p}}_{s_j}^{(k)}$ have to be linearly independent in order to allow the separation of signals.

C. Derivation of the BP Algorithm

We note that no closed-form solution exists for solving the relations of (5) and (6). Therefore, we form a cost function that can be minimized iteratively. It is more advantageous to use natural gradient adaptation rather than the ordinary gradient adaptation for this problem [17, Theorem 6]. To obtain a more compact form of natural gradients, we employ symmetric matrices and rewrite (5) and (6) as

$$\underline{\mathbf{R}}_{\mathbf{x}}^{(k,\tau)} = \mathbf{A}\underline{\Lambda}_{\mathbf{s}}^{(k,\tau)}\mathbf{A}^T \tag{7}$$

$$\underline{\mathbf{R}}_{\mathbf{y}}^{(k,\tau)} = \mathbf{W}\underline{\mathbf{R}}_{\mathbf{x}}^{(k,\tau)}\mathbf{W}^T \tag{8}$$

where $\underline{\Lambda}_{\mathbf{s}}^{(k,\tau)}$, $\underline{\mathbf{R}}_{\mathbf{x}}^{(k,\tau)}$, and $\underline{\mathbf{R}}_{\mathbf{y}}^{(k,\tau)}$ are, respectively, the symmetric parts of $\Lambda_{\mathbf{s}}^{(k,\tau)}$, $\mathbf{R}_{\mathbf{x}}^{(k,\tau)}$, and $\mathbf{R}_{\mathbf{y}}^{(k,\tau)}$ defined by, e.g., $\underline{\mathbf{R}}_{\mathbf{x}}^{(k,\tau)} = \mathbf{R}_{\mathbf{x}}^{(k,\tau)} + \mathbf{R}_{\mathbf{x}}^{(k,\tau)T}$.

To cope with nonpersistently active signals, we approach the joint diagonalization problem using the Frobenius norm [20]. The joint diagonalization cost function is given by

$$J_{\text{JD}} = \sum_{k=1}^{K}\sum_{\tau=0}^{\Gamma}\beta^{(k,\tau)}J_{\text{JD}}^{(k,\tau)}(\mathbf{W}) \tag{9}$$

where $\beta^{(k,\tau)}$ is a positive weight satisfying $\sum_{k=1}^{K}\sum_{\tau=0}^{\Gamma}\beta^{(k,\tau)} = 1$ and is generally set to $1/K(\Gamma+1)$. The joint diagonalization component is

$$J_{\text{JD}}^{(k,\tau)} = \left\|\text{off}\left(\underline{\mathbf{R}}_{\mathbf{y}}^{(k,\tau)}\right)\right\|_F^2 \tag{10}$$

where off$(\cdot)$ is the operator that returns a matrix with all its diagonal entries being set to zero and $\|\cdot\|_F$ denotes the Frobenius norm. According to [17, Theorem 6], the natural gradient and the ordinary gradient of $J_{\text{JD}}^{(k,\tau)}$ are related by

$$\tilde{\nabla}_{\mathbf{W}}J_{\text{JD}}^{(k,\tau)} = \nabla_{\mathbf{W}}J_{\text{JD}}^{(k,\tau)}\mathbf{W}^T\mathbf{W} \tag{11}$$

where $\tilde{\nabla}_{\mathbf{W}}$ and $\nabla_{\mathbf{W}}$ are, respectively, the natural gradient and the ordinary gradient operators with respect to $\mathbf{W}$. Using matrix differentiation and the relation given in (11), the natural gradient of $J_{\text{JD}}^{(k,\tau)}$ is given by

$$\tilde{\nabla}_{\mathbf{W}}J_{\text{JD}}^{(k,\tau)} = 4\,\text{off}\left(\underline{\mathbf{R}}_{\mathbf{y}}^{(k,\tau)}\right)\underline{\mathbf{R}}_{\mathbf{y}}^{(k,\tau)}\mathbf{W}. \tag{12}$$

Next, we need to examine all possible stationary points of J_{JD} in order to determine an appropriate penalty term. For convenience of presentation, we introduce the following definition.

Definition 1: Let $\mathcal{W}_1, \mathcal{W}_2, \ldots, \mathcal{W}_N$ and $\mathcal{W}_N^*$ be sets of matrices $\mathbf{W} = [w_{ij}]$ defined by $\mathcal{W}_l = \{\mathbf{W} : \mathbf{WA} = \mathbf{D}_l\mathbf{P}\}, l = 1, 2, \ldots, N$, $\mathcal{W}_N^* = \{\mathbf{W} : \mathbf{WA} = \mathbf{D}_N\mathbf{P}$ and $w_{ij} = 1, \forall\, 1 \leq i = j \leq N\}$ where $\mathbf{D}_l$ is an $N \times N$ diagonal matrix with l indicating the number of nonzero elements in the diagonal entries and $\mathbf{P}$ is an $N \times N$ permutation matrix.

The set $\mathcal{W}_N$ is a set of all desired demixing matrices whose existence is ensured by A1. The sets $\mathcal{W}_l$, $l = 1, 2, \ldots, N-1$, which also exist by A1, indicate that $\mathbf{W}$ consists of $N - l$ zero rows, each of which results from replacing all elements in the row of $\mathbf{W} \in \mathcal{W}_N$ with zeros. The special set $\mathcal{W}_N^*$ defined in a similar fashion consists of matrices that have the same properties as the set $\mathcal{W}_N$ but with all diagonal entries being unity. We note here that only the sets $\mathcal{W}_N$ and $\mathcal{W}_N^*$ give the desired demixing matrix and that all other sets give trivial solutions. We show that $\mathcal{W}_N^*$ is a subset of $\mathcal{W}_N$ in Proposition 1.

Proposition 1: 1) If $\mathbf{W} \in \mathcal{W}_N$, then (a) any permutation of the rows of $\mathbf{W}$ is a matrix that belongs to $\mathcal{W}_N$ and (b) any nonzero scaling of the rows of $\mathbf{W}$ is a matrix that belongs to $\mathcal{W}_N$; 2) $\mathcal{W}_N^*$ is a subset of $\mathcal{W}_N$.

Proof: See Appendix B. □

Based on Definition 1, we give the following lemma which provides the existence of all stationary points of J_{JD}.

Lemma 1: $\mathbf{W}$ is a stationary point of J_{JD} *iff* $\mathbf{W} = \mathbf{0}$, where $\mathbf{0}$ is a zero matrix, or $\mathbf{W} \in \mathcal{W}_l$, $l = 1, 2, \ldots, N$.

Proof: See Appendix C. □

It is anticipated that the desired demixing matrix should be found when J_{JD} attains its minimum. However, by using a small perturbation (for details, see Proof of Theorem 1), it can be shown that the stationary points of J_{JD} are all minima. Since only a full rank $\mathbf{W} \in \mathcal{W}_N$, not l-rank $\mathbf{W} \in \mathcal{W}_l$, $l = 1, 2, \ldots, N-1$, is the desired demixing matrix, we thus need to preserve the full rank property of $\mathbf{W}$ during its iterative search process by confining the search to a region. As Proposition 1 suggests that $\mathcal{W}_N^*$ is a subset of $\mathcal{W}_N$, we thus exploit the penalty function given by

$$J_C = \|\text{diag}\,(\mathbf{W} - \mathbf{I})\|_F^2 \tag{13}$$

where diag $(\cdot)$ is the operator that returns a matrix with all its off-diagonal entries being zero. By differentiating J_C with

Authorized licensed use limited to: Imperial College London. Downloaded on January 4, 2010 at 08:09 from IEEE Xplore. Restrictions apply.

respect to $\mathbf{W}$ and using the relation given in Theorem 6 of [17], the natural gradient of J_C is given by

$$\tilde{\nabla}_{\mathbf{W}} J_C = 2\mathrm{diag}(\mathbf{W} - \mathbf{I})\mathbf{W}^T\mathbf{W}. \tag{14}$$

We then combine J_{JD} and J_C and form an unconstrained cost function J_{BP} to be minimized as

$$J_{\mathrm{BP}} = J_{\mathrm{JD}} + \lambda_C J_C \tag{15}$$

where $\lambda_C < 1$ is a small positive penalty factor aiming to reduce the significance of J_C.

Let $\delta\mathbf{W}$ be the change of the demixing matrix from iteration m to $m+1$

$$\delta\mathbf{W}(m) = \mathbf{W}(m+1) - \mathbf{W}(m). \tag{16}$$

The BP update using the natural gradient descent adaptation is of the form

$$\delta\mathbf{W}(m) = -\mu\left(\sum_{k=1}^{K}\sum_{\tau=0}^{\Gamma}\beta^{(k,\tau)}\tilde{\nabla}_{\mathbf{W}} J_{\mathrm{JD}}^{(k,\tau)}(m) + \lambda_C\tilde{\nabla}_{\mathbf{W}} J_C(m)\right) \tag{17}$$

where μ is a positive step-size, $\tilde{\nabla}_{\mathbf{W}} J_{\mathrm{JD}}^{(k,\tau)}(m)$ and $\tilde{\nabla}_{\mathbf{W}} J_C(m)$ can be, respectively, obtained from (12) and (14) by replacing $\mathbf{W}$ including those in $\underline{\mathbf{R}}_{\mathbf{y}}^{(k,\tau)}$ with $\mathbf{W}(m)$.

It can be seen that (17) violates the uniform performance property [18] due to $\tilde{\nabla}_{\mathbf{W}} J_C(m)$. However, we can regard BP as an algorithm that exhibits quasi-uniform performance since λ_C is always much less than unity and therefore considerably reduces the effect of $\tilde{\nabla}_{\mathbf{W}} J_C$. Also, we point out that (17) can be used without restriction and is thereby capable of performing BSS even when some source signals are not persistently active.

As an effect of J_C on J_{JD}, we show that the global minimum of J_{BP} exists and corresponds to the desired demixing matrix.

Theorem 1: In the neighborhood of a matrix whose diagonal entries are all equal to unity, $\mathbf{W}$ is a stationary point of J_{BP} *iff* $\mathbf{W} \in \mathcal{W}_N^*$. In this case, J_{BP} attains the global minimum.

Proof: See Appendix D. □

It can be seen that all minima of J_{JD} except $\mathbf{W} \in \mathcal{W}_N^*$ vanish by adding J_C to J_{JD}. Based on Theorem 1, the convergence of BP to the desired demixing matrix can be achieved by initializing $\mathbf{W}$ with any matrix whose diagonal entries are equal to unity so as to prevent the algorithm from converging to an undesired but possible minimum induced by J_C. Also, since the proof relies on a small perturbation $\delta\mathbf{W}$, it is necessary that an adequately small μ and λ_C are used. The use of λ_C as a penalty factor can be found in [21]–[23] as well as [24] (see [25] and [26] for details).

III. Online Processing (OP) Algorithm

For applications operating on a sample-by-sample basis, it is necessary to devise an algorithm that can adapt itself to the most current given sample. The estimation of the correlation matrix using a local time-average correlation matrix does not suit this purpose as it does not represent the current second-order statistics of signals. In this section, we propose the OP algorithm derived using a set of current time-average correlation matrices.

A. Second-Order Statistical Relations Using Current Time-Average Correlation Matrices

By making use of assumptions A1 and A2, the online joint diagonalization problem can be established through the following second-order statistical relations

$$\underline{\hat{\mathbf{R}}}_{\mathbf{x}}^{(\tau)}(n) = \mathbf{A}\underline{\hat{\boldsymbol{\Lambda}}}_{\mathbf{s}}^{(\tau)}(n)\mathbf{A}^T \tag{18}$$

$$\underline{\hat{\mathbf{R}}}_{\mathbf{y}}^{(\tau)}(n) = \mathbf{W}\underline{\hat{\mathbf{R}}}_{\mathbf{x}}^{(\tau)}(n)\mathbf{W}^T \tag{19}$$

where $\underline{\hat{\boldsymbol{\Lambda}}}_{\mathbf{s}}^{(\tau)}(n)$, $\underline{\hat{\mathbf{R}}}_{\mathbf{x}}^{(\tau)}(n)$ and $\underline{\hat{\mathbf{R}}}_{\mathbf{y}}^{(\tau)}(n)$ are, respectively, the symmetric parts of a *current estimate* of the correlation matrix of the source signals, the observed signals and the output signals, all with a time lag τ.

To obtain the current estimate of the observed signals correlation matrix $\hat{\mathbf{R}}_{\mathbf{x}}^{(\tau)}(n)$, we generalize the estimator in [27] to incorporate the nonzero time lag and propose the following *current time-average correlation matrix*:

$$\hat{\mathbf{R}}_{\mathbf{x}}^{(\tau)}(n) = \alpha\hat{\mathbf{R}}_{\mathbf{x}}^{(\tau)}(n-1) + (1-\alpha)\mathbf{x}(n)\mathbf{x}^T(n-\tau) \tag{20}$$

where α is a forgetting factor with $0 < \alpha < 1$.

For stationary but nonwhite signals, each sample has equal importance in calculating the current time-average correlation matrix. In that case, (20) is thus replaced by

$$\hat{\mathbf{R}}_{\mathbf{x}}^{(\tau)}(n) = \frac{n-1}{n}\hat{\mathbf{R}}_{\mathbf{x}}^{(\tau)}(n-1) + \frac{1}{n}\mathbf{x}(n)\mathbf{x}^T(n-\tau). \tag{21}$$

B. Prerequisites for Using Current Time-Average Correlation Matrices

The online signal separation can be achieved using a set of current time-average correlation matrices, provided that the following prerequisites are satisfied.

Prerequistites 2: Let $\hat{\rho}_{s_1}^{(\tau)}(n), \hat{\rho}_{s_2}^{(\tau)}(n), \ldots, \hat{\rho}_{s_N}^{(\tau)}(n)$ be corresponding diagonal entries of $\hat{\Lambda}_{\mathbf{s}}^{(\tau)}(n)$:

1) *nonstationarity:* $\mathbf{p}_{s_i}^{(\tau)} = [\hat{\rho}_{s_i}^{(\tau)}(1), \hat{\rho}_{s_i}^{(\tau)}(2), \ldots, \hat{\rho}_{s_i}^{(\tau)}(n), \ldots]$ and $\mathbf{p}_{s_j}^{(\tau)} = [\hat{\rho}_{s_j}^{(\tau)}(1), \hat{\rho}_{s_j}^{(\tau)}(2), \ldots, \hat{\rho}_{s_j}^{(\tau)}(n), \ldots]$ must be linearly independent, $\exists\tau = 0, 1, \ldots, \Gamma, \forall\, 1 \le i \ne j \le N$;

2) *nonwhiteness:* $\tilde{\mathbf{p}}_{s_i}(n) = [\hat{\rho}_{s_i}^{(0)}(n), \hat{\rho}_{s_i}^{(1)}(n), \ldots, \hat{\rho}_{s_i}^{(\Gamma)}(n)]$ and $\tilde{\mathbf{p}}_{s_j}(n) = [\rho_{s_j}^{(0)}(n), \rho_{s_j}^{(1)}(n), \ldots, \rho_{s_j}^{(\Gamma)}(n)]$ must be linearly independent, $\forall 1 \le i \ne j \le N$.

Proof: The proof is similar to that for Prerequisites 1. □

It is worth pointing out that the Prerequisite 2 1) is the underlying concept that allows the algorithms proposed in [12], [19] to separate nonstationary source signals and the Prerequisite 2 2) is the key insight that allows source signals that are stationary but nonwhite to be separated in an online fashion. In addition, a sufficient condition for separating nonstationary signals is given in [12] by using the Prerequisite 2 1) together with the logarithm function.

C. Derivation of the OP Algorithm

To find a demixing matrix using any online gradient-based algorithm, it is necessary that variations in the nonstationarity of the source signals are sufficiently slow and a number of samples has to be adequately large so as to allow the algorithm to attain

Authorized licensed use limited to: Imperial College London. Downloaded on January 4, 2010 at 08:09 from IEEE Xplore. Restrictions apply.

the minimum. This means that source signals have to keep exciting (1), but some of them, however, need not be persistently active.

Following the derivation of BP, the OP cost function can be expressed as

$$J_{\text{OP}}(n) = \sum_{\tau=0}^{\Gamma} \beta^{(\tau)} J_{\text{JD}}^{(\tau)}(n) + \lambda_C J_C(n) \tag{22}$$

where $J_{\text{JD}}^{(\tau)}(n) = \|\text{off}(\hat{\underline{\mathbf{R}}}_{\mathbf{y}}^{(\tau)}(n))\|_F^2$ is the joint diagonalization component, $J_C(n) = \|\text{diag}\,(\mathbf{W}(n) - \mathbf{I})\|_F^2$ is the penalty function, $\beta^{(\tau)}$ is a positive weight that satisfies $\sum_{\tau=0}^{\Gamma} \beta^{(\tau)} = 1$ and, in the general case, $\beta^{(\tau)}$ is set to $(1/\Gamma + 1)$, λ_C is a small positive penalty factor, $J_{\text{JD}}^{(\tau)}$ is the joint diagonalization component at time lag τ, and J_C is the penalty function. By use of matrix differentiation and [17, Theorem 6], the corresponding natural gradients are

$$\tilde{\nabla}_{\mathbf{W}} J_{\text{JD}}^{(\tau)}(n) = 4\,\text{off}\left(\hat{\underline{\mathbf{R}}}_{\mathbf{y}}^{(\tau)}(n)\right) \hat{\underline{\mathbf{R}}}_{\mathbf{y}}^{(\tau)}(n)\mathbf{W}(n) \tag{23}$$

$$\tilde{\nabla}_{\mathbf{W}} J_C(n) = 2\,\text{diag}\,(\mathbf{W}(n) - \mathbf{I})\mathbf{W}^T(n)\mathbf{W}(n). \tag{24}$$

Consequently, the change $\delta\mathbf{W}(n) = \mathbf{W}(n+1) - \mathbf{W}(n)$ or equivalently the demixing matrix update of OP is of the form

$$\delta\mathbf{W}(n) = -\mu\left(\sum_{\tau=0}^{\Gamma} \beta^{(\tau)} \tilde{\nabla}_{\mathbf{W}} J_{\text{JD}}^{(\tau)}(n) + \lambda_C \tilde{\nabla}_{\mathbf{W}} J_C(n)\right) \tag{25}$$

where μ is a positive step-size.

The OP algorithm, like BP, requires $\mathbf{W}$ to be initialized in such a way that all its diagonal entries are unity so as to prevent the algorithm from being trapped by local minima induced by J_C. As the difference between OP and BP is in the estimation of correlation matrices, the OP algorithm, at any given n, thus has the global minimum in the neighborhood of the penalized region that corresponds to the desired demixing matrix.

IV. Normalized Versions of the Algorithms

In BSS, the concept underlying most gradient-based algorithms is that the demixing matrix is adjusted with an appropriate step-size in the direction of the computed gradient. Excessively small and large gradients can result in gradient noise problems ranging from slow convergence to divergence of the algorithm. In this section, we develop, from BP and OP, their corresponding normalized algorithms in which the demixing matrix update is based on the minimal disturbance principle [28]. A common theme is that the update of an unknown in the adaptive structure should be disturbed in a minimal fashion.

A. Normalized BP (NBP) Algorithm

At iteration index m, we can rewrite (15) as

$$J_{\text{BP}}(m) = \sum_{k=1}^{K}\sum_{\tau=0}^{\Gamma} \beta^{(k,\tau)} J_{\text{JD}}^{(k,\tau)}(m) + \lambda_C J_C(m) \tag{26}$$

where $J_{\text{BP}}(m)$, $J_{\text{JD}}^{(k,\tau)}(m)$, and $J_C(m)$ are short forms of $J_{\text{BP}}(\mathbf{W}(m))$, $J_{\text{JD}}^{(k,\tau)}(\mathbf{W}(m))$, and $J_C(\mathbf{W}(m))$, respectively. Based on (26), the corresponding update $\delta\mathbf{W}(m)$ in (17) can be written by using a set of components comprising $J_{\text{JD}}^{(k,\tau)}(m)$ and $J_C(m)$ as

$$\delta\mathbf{W}(m) = \sum_{k=1}^{K}\sum_{\tau=0}^{\Gamma} \beta^{(k,\tau)} \delta\mathbf{W}_{\text{JD}}^{(k,\tau)}(m) + \lambda_C \delta\mathbf{W}_C(m) \tag{27}$$

where $\delta\mathbf{W}_{\text{JD}}^{(k,\tau)}(m)$ is an update based on $J_{\text{JD}}^{(k,\tau)}(m)$, and $\delta\mathbf{W}_C(m)$ is an update based on $J_C(m)$. In light of the minimal disturbance principle, every component of $\delta\mathbf{W}(m)$ in (27) should disturb the separation system in a minimal fashion. Let us consider an update $\delta\mathbf{W}_{\text{JD}}^{(k,\tau)}(m)$ based on a particular given $J_{\text{JD}}^{(k,\tau)}(m)$ and form the following constrained minimal disturbance problem

$$\begin{aligned} &\text{minimize } \frac{1}{2}\left\|\delta\mathbf{W}_{\text{JD}}^{(k,\tau)}(m)\right\|_F^2 \\ &\text{subject to } J_{\text{JD}}^{(k,\tau)}(m+1) = 0. \end{aligned} \tag{28}$$

To allow (28) to be differentiable with respect to $\delta\mathbf{W}_{\text{JD}}^{(k,\tau)}(m)$, we estimate $J_{\text{JD}}^{(k,\tau)}(m+1)$ using the Taylor series expansion.

Let $g(\mathbf{X})$ be a function that is differentiable with respect to an $N \times N$ matrix $\mathbf{X}$. The Taylor series expansion of a function g at $\mathbf{X} + \delta\mathbf{X}$ is given by [29]

$$g(\mathbf{X} + \delta\mathbf{X}) = g(\mathbf{X}) + \langle \nabla_{\mathbf{X}} g(\mathbf{X}) | \delta\mathbf{X}\rangle + O(\delta\mathbf{X}^2) \tag{29}$$

where $\delta\mathbf{X}$ is a small change of $\mathbf{X}$, $O(\cdot)$ denotes higher order expansions about $\mathbf{X}$, and $\langle \nabla_{\mathbf{X}} g(\mathbf{X}) | \delta\mathbf{X}\rangle = \text{Trace}\,(\nabla_{\mathbf{X}} g(\mathbf{X})^T \delta\mathbf{X})$.

For a sufficiently small $\delta\mathbf{X}$, we can neglect its higher order terms and estimate $g(\mathbf{X} + \delta\mathbf{X})$ using its first-order. The Taylor series expansion in (29) thus becomes

$$g(\mathbf{X} + \delta\mathbf{X}) \approx g(\mathbf{X}) + \langle \nabla_{\mathbf{X}} g(\mathbf{X}) | \delta\mathbf{X}\rangle. \tag{30}$$

In the above approximation, the steepest direction of a function g at $\mathbf{X}$ is given by the ordinary gradient $\nabla_{\mathbf{X}} g(\mathbf{X})$. However, it is pointed out in [17] that, when the parameter space considered is not Euclidean, for example, the space of nonorthogonal matrices in BSS, the steepest direction is represented by the natural gradient instead of the ordinary gradient. This is due to the fact that there is no orthogonal linear coordinate in the non-Euclidean space. As the demixing matrix $\mathbf{W}$ is not assumed to be orthogonal, it is more general to modify (30) by using the natural gradient, which gives

$$g(\mathbf{X} + \delta\mathbf{X}) \approx g(\mathbf{X}) + \langle \tilde{\nabla}_{\mathbf{X}} g(\mathbf{X}) | \delta\mathbf{X}\rangle. \tag{31}$$

Next, we apply the above approximation of Taylor series together with the natural gradient to estimate $J_{\text{JD}}^{(k,\tau)}(m+1)$. We begin by viewing the update $\delta\mathbf{W}_{\text{JD}}^{(k,\tau)}(m)$ as a small change of a demixing matrix that causes the change between $J_{\text{JD}}^{(k,\tau)}(m)$ and $J_{\text{JD}}^{(k,\tau)}(m+1)$. Since $\delta W_{\text{JD}}^{(k,\tau)}(m)$ is sufficiently small, its higher order terms in Taylor series expansion are neglected. The

Authorized licensed use limited to: Imperial College London. Downloaded on January 4, 2010 at 08:09 from IEEE Xplore. Restrictions apply.

Taylor series expansion of $J_{\mathrm{JD}}^{(k,\tau)}(m+1)$ is approximately given by

$$J_{\mathrm{JD}}^{(k,\tau)}(m+1) = J_{\mathrm{JD}}^{(k,\tau)}(m) + \left\langle \tilde{\nabla}_{\mathbf{W}} J_{\mathrm{JD}}^{(k,\tau)}(m) \middle| \delta\mathbf{W}_{\mathrm{JD}}^{(k,\tau)}(m) \right\rangle. \tag{32}$$

Following the method of Lagrange multipliers, we first replace $J_{\mathrm{JD}}^{(k,\tau)}(m+1)$ with (32) and then convert (28) to the following unconstrained problem:

$$\begin{aligned} J(m) = \frac{1}{2}\left\|\delta\mathbf{W}_{\mathrm{JD}}^{(k,\tau)}(m)\right\|_F^2 &+ \lambda_L \left(J_{\mathrm{JD}}^{(k,\tau)}(m) \right. \\ &\left. + \left\langle \tilde{\nabla}_{\mathbf{W}} J_{\mathrm{JD}}^{(k,\tau)}(m) \middle| \delta\mathbf{W}_{\mathrm{JD}}^{(k,\tau)}(m) \right\rangle \right) \end{aligned} \tag{33}$$

where $J(m)$ is the cost function of an unknown $\delta\mathbf{W}_{\mathrm{JD}}^{(k,\tau)}(m)$ and λ_L is the Lagrange multiplier.

Given $\underline{\mathbf{R}}_{\mathbf{x}}^{(k,\tau)}$ and $\mathbf{W}(m)$, we obtain the first-order conditions of (33) as

$$\begin{aligned} \tilde{\nabla}_{\delta\mathbf{W}_{\mathrm{JD}}^{(k,\tau)}} J(m) &= \delta\mathbf{W}_{\mathrm{JD}}^{(k,\tau)}(m) + \lambda_L \tilde{\nabla}_{\mathbf{W}} J_{\mathrm{JD}}^{(k,\tau)}(m) \\ &= \mathbf{0} \end{aligned} \tag{34}$$

$$\begin{aligned} \frac{\partial J(m)}{\partial \lambda_L} &= J_{\mathrm{JD}}^{(k,\tau)}(m) \\ &\quad + \left\langle \tilde{\nabla}_{\mathbf{W}} J_{\mathrm{JD}}^{(k,\tau)}(m) \middle| \delta\mathbf{W}_{\mathrm{JD}}^{(k,\tau)}(m) \right\rangle \\ &= 0 \end{aligned} \tag{35}$$

and solve for λ_L by substituting $\delta\mathbf{W}_{\mathrm{JD}}^{(k,\tau)}(m)$ from (34) into (35) giving

$$\lambda_L = \frac{J_{\mathrm{JD}}^{(k,\tau)}(m)}{\left\|\tilde{\nabla}_{\mathbf{W}} J_{\mathrm{JD}}^{(k,\tau)}(m)\right\|_F^2}. \tag{36}$$

Replacing λ_L in (34) with (36), we obtain the following optimal component:

$$\delta\mathbf{W}_{\mathrm{JD}}^{(k,\tau)}(m) = -J_{\mathrm{JD}}^{(k,\tau)}(m) \frac{\tilde{\nabla}_{\mathbf{W}} J_{\mathrm{JD}}^{(k,\tau)}(m)}{\left\|\tilde{\nabla}_{\mathbf{W}} J_{\mathrm{JD}}^{(k,\tau)}(m)\right\|_F^2}. \tag{37}$$

A key feature of (37) is that the update $\delta\mathbf{W}_{\mathrm{JD}}^{(k,\tau)}(m)$ obtained from the use of minimal disturbance principle mitigates the gradient noise by normalizing the natural gradient of $J_{\mathrm{JD}}^{(k,\tau)}(m)$ with its squared Frobenius norm. In particular, we can interpret (37) as an update $\delta\mathbf{W}_{\mathrm{JD}}^{(k,\tau)}(m)$ that moves towards the minimum of $J_{\mathrm{JD}}^{(k,\tau)}(m)$ in the natural gradient descent direction with distance proportional to $J_{\mathrm{JD}}^{(k,\tau)}(m)$ and then vanishes at the minimum. To simplify (37), we use (10) and (12) at iteration m and expand the scalar terms in (37) to obtain

$$\delta\mathbf{W}_{\mathrm{JD}}^{(k,\tau)}(m) = -\frac{\left\|\mathrm{off}\left(\underline{\mathbf{R}}_{\mathbf{y}}^{(k,\tau)}(m)\right)\right\|_F^2 \tilde{\nabla}_{\mathbf{W}} J_{\mathrm{JD}}^{(k,\tau)}(m)}{16\left\|\mathrm{off}\left(\underline{\mathbf{R}}_{\mathbf{y}}^{(k,\tau)}(m)\right)\underline{\mathbf{R}}_{\mathbf{y}}^{(k,\tau)}(m)\mathbf{W}(m)\right\|_F^2}. \tag{38}$$

By applying the inequality property of the squared Frobenius norm [20] to the squared Frobenius norm terms in (38), we obtain $(\|\mathrm{off}(\underline{\mathbf{R}}_{\mathbf{y}}^{(k,\tau)}(m))\|_F^2 / \|\mathrm{off}\,(\underline{\mathbf{R}}_{\mathbf{y}}^{(k,\tau)}(m))\underline{\mathbf{R}}_{\mathbf{y}}^{(k,\tau)}(m)\mathbf{W}(m)\|_F^2) \geq \|\mathrm{off}(\underline{\mathbf{R}}_{\mathbf{y}}^{(k,\tau)}(m))\|_F^2 / \|\mathrm{off}(\underline{\mathbf{R}}_{\mathbf{y}}^{(k,\tau)}(m))\|_F^2 \|\underline{\mathbf{R}}_{\mathbf{y}}^{(k,\tau)}(m)\mathbf{W}(m)\|_F^2$. Using this inequality and expanding the remaining $\tilde{\nabla}_{\mathbf{W}} J_{\mathrm{JD}}^{(k,\tau)}(m)$, (38) is approximately simplified to

$$\delta\mathbf{W}_{\mathrm{JD}}^{(k,\tau)}(m) \approx -\frac{\mathrm{off}(\underline{\mathbf{R}}_{\mathbf{y}}^{(k,\tau)}(m))\underline{\mathbf{R}}_{\mathbf{y}}^{(k,\tau)}(m)\mathbf{W}(m)}{4\left\|\underline{\mathbf{R}}_{\mathbf{y}}^{(k,\tau)}(m)\mathbf{W}(m)\right\|_F^2}. \tag{39}$$

We note that the squared Frobenius norm of (39) is always less than that of (38) due to the squared Frobenius norm inequality. Therefore, $\delta\mathbf{W}_{\mathrm{JD}}^{(k,\tau)}(m)$ in (39) still obeys the principle of minimal disturbance. Following the above methodology, we similarly obtain

$$\begin{aligned} \delta\mathbf{W}_C(m) &= -J_C(m)\frac{\tilde{\nabla}_{\mathbf{W}} J_C(m)}{\|\tilde{\nabla}_{\mathbf{W}} J_C(m)\|_F^2} \\ &\approx -\frac{\mathrm{diag}(\mathbf{W}(m)-\mathbf{I})\mathbf{W}^T(m)\mathbf{W}(m)}{2\|\mathbf{W}^T(m)\mathbf{W}(m)\|_F^2}. \end{aligned} \tag{40}$$

In order to control the rate of convergence, we introduce a scaling factor $\bar{\mu}$ to (27) and write

$$\delta\mathbf{W}(m) = \bar{\mu}\left(\sum_{k=1}^{K}\sum_{\tau=0}^{\Gamma}\beta^{(k,\tau)}\delta\mathbf{W}_{\mathrm{JD}}^{(k,\tau)}(m) + \lambda_C\delta\mathbf{W}_C(m)\right) \tag{41}$$

where $0 < \bar{\mu} \leq 1$ with 1 being an upper bound that still keeps the squared Frobenius norm of $\delta\mathbf{W}(m)$ in accordance with the minimal disturbance principle; $\delta\mathbf{W}_{\mathrm{JD}}^{(k,\tau)}(m)$ and $\delta\mathbf{W}_C(m)$ can be obtained from their simplified form in (39) and (40), respectively.

B. Normalized OP (NOP) Algorithm

The normalized OP (NOP) algorithm utilizes a set of current time-average correlation matrices and can be derived in a similar style as NBP. We provide without derivation the update $\delta\mathbf{W}(n)$ of the NOP algorithm which takes the form

$$\delta\mathbf{W}(n) = \bar{\mu}\left(\sum_{\tau=0}^{\Gamma}\beta^{(\tau)}\delta\mathbf{W}_{\mathrm{JD}}^{(\tau)}(n) + \lambda_C\delta\mathbf{W}_C(n)\right) \tag{42}$$

where $\delta\mathbf{W}_{\mathrm{JD}}^{(\tau)}(n)$ and $\delta\mathbf{W}_C(n)$ is, respectively, given in their simplified form by

$$\delta\mathbf{W}_{\mathrm{JD}}^{(\tau)}(n) = -\frac{\mathrm{off}\left(\underline{\mathbf{R}}_{\mathbf{y}}^{(\tau)}(n)\right)\underline{\mathbf{R}}_{\mathbf{y}}^{(\tau)}(n)\mathbf{W}(n)}{4\left\|\underline{\mathbf{R}}_{\mathbf{y}}^{(\tau)}(n)\mathbf{W}(n)\right\|_F^2} \tag{43}$$

$$\delta\mathbf{W}_C(n) = -\frac{\mathrm{diag}\,(\mathbf{W}(n)-\mathbf{I})\mathbf{W}^T(n)\mathbf{W}(n)}{2\|\mathbf{W}^T(n)\mathbf{W}(n)\|_F^2}. \tag{44}$$

Authorized licensed use limited to: Imperial College London. Downloaded on January 4, 2010 at 08:09 from IEEE Xplore. Restrictions apply.

TABLE I
OPERATIONS NEEDED TO COMPUTE $\delta\mathbf{W}$ OF BP, OP, NBP, AND NOP

Algorithms	Operations per Iteration/Sample
BP	$(4K\Gamma + 4K + 2)N^3 - (3K\Gamma + 3K - 1)N^2 + N$
OP	$(4\Gamma + 6)N^3 - (3\Gamma + 2)N^2 + N$
NBP	$(4K\Gamma + 4K + 2)N^3 - 2(K\Gamma + K - 1)N^2 + N$
NOP	$(4\Gamma + 6)N^3 - 2\Gamma N^2 + N$

C. Discussion

In NBP and NOP, normalization is achieved using terms that do not involve the criterion of the original algorithms. It is seen from (39) and (43) that the normalization terms turn out to be dependent on $\mathbf{W}$ which is different from relevant previous work. In [4], [21], [22], and [24], the use of squared Frobenius norm of a correlation matrix as a normalization term is given without derivation. We note that the squared Frobenius norm of a correlation matrix, when used in the algorithm, although similar in some degree to the normalized LMS (NLMS) algorithm in the sense that it is independent of the unknown, i.e., $\mathbf{W}$, does not mitigate the gradient noise problem and also destroys the desirable property of having unity upper bound on the scaling factor.

The only requirement to use the proposed normalized algorithms is that all source signals cannot be simultaneously inactive to avoid a zero normalization term. The proposed normalized algorithms exhibit fast convergence and robustness to gradient noise. Lastly, it is shown by numerical experiments in Section VI that improved performance, when compared to the non-normalized versions, is achieved.

V. Computational Complexity

The computational complexity of the proposed algorithms is given in terms of operations, which include the number of multiplications/divisions and the number of additions/subtractions. The multiplication of two $N \times N$ matrices, the multiplication of two $N \times N$ matrices with one having only off-diagonal entries and the multiplication of two $N \times N$ matrices with one having only diagonal entries require $2N^3 - N^2$, $2N^3 - 3N^2$ and N^2, respectively. The computational complexity of the matrix operators off$(\cdot)$ and $\mathrm{diag}(\cdot)$ is neglected for all algorithms due to their implementation. The operations requirements for computing $\delta\mathbf{W}$ of BP, OP, NBP, and NOP are summarized in Table I.

The table shows that the numbers of operations required to compute $\delta\mathbf{W}$ for NBP and NOP are more than those used by BP and OP. Specifically, NBP and NOP employ the normalization terms achieved at the additional cost of $(K\Gamma + K + 1)N^2$ and $(\Gamma + 2)N^2$ operations, respectively.

VI. Numerical Experiments

To evaluate the performance of the BSS algorithms, the closeness of the global matrix $\mathbf{C} = [c_{ij}] = \mathbf{WA}$ to $\mathbf{DP}$ is measured. We employ the performance index $(\mathcal{P})$ defined as

$$\mathcal{P} = \frac{1}{2N(N-1)} \left(\sum_{i=1}^{N} \left(\sum_{j=1}^{N} \frac{c_{ij}^2}{\max_l(c_{il}^2)} - 1 \right) + \sum_{j=1}^{N} \left(\sum_{i=1}^{N} \frac{c_{ij}^2}{\max_l(c_{lj}^2)} - 1 \right) \right)$$

where $\max_l(c_{il}^2)$ and $\max_l(c_{lj}^2)$ are, respectively, the maximum value of c_{il}^2 and c_{lj}^2 for $1 \leq i, j, l \leq N$. Accordingly, smaller values of $\mathcal{P}$ indicate better performance. A slightly different performance index can be found in [30].

We test the proposed Frobenius norm-based algorithms, both offline and online, by comparing them with the previously proposed logarithm-based algorithms with and without regularization. The experiments are divided into two parts depending on the types of algorithms: offline algorithms and online algorithms.

A. Offline Algorithms

In this section, we aim to compare BP and NBP, which are classified as the Frobenius norm-based algorithm, with the logarithm-based algorithm proposed in [2]. We consider the instantaneous BSS problem and exploit only the nonstationarity property of signals as these do not require the novel matrix formulation in [2]. Since all correlation matrices obtained from the nonstationarity property are symmetric, we thus do not need to employ their symmetric parts. The BP and NBP algorithms with $\Gamma = 0$ are compared with a modified version of the logarithm-based algorithm in [2] or, in short, the log algorithm, which employs the cost function and update, respectively, given by

$$J_L = \sum_{k=1}^{K} \beta^{(k,0)} \log\det \, \mathrm{diag}\left(\mathbf{R}_{\mathbf{y}}^{(k,0)}\right) - \log\det \mathbf{R}_{\mathbf{y}}^{(k,0)} \tag{45}$$

$$\delta\mathbf{W}(m) = -2\,\mu \sum_{k=1}^{K} \beta^{(k,0)} \mathrm{diag}^{-1}\left(\mathbf{R}_{\mathbf{y}}^{(k,0)}(m)\right) \mathrm{off}\left(\mathbf{R}_{\mathbf{y}}^{(k,0)}(m)\right) \mathbf{W}(m) \tag{46}$$

where $\det$ denotes the determinant of a matrix and $\mathrm{diag}^{-1}(\mathbf{R}_{\mathbf{y}}^{(k,0)}) = (\mathrm{diag}(\mathbf{R}_{\mathbf{y}}^{(k,0)}))^{-1}$.

This offers, as close as possible, a like-for-like comparison by removing from [2] the features of convolutive mixing and exploitation of the nonwhiteness property. We also note that (45) is based on the logarithmic function in addition to Oppenheim's inequality and, as a consequence, requires $\mathbf{R}_{\mathbf{y}}^{(k,0)}$ to be *positive definite* rather than nonnegative definite so that it will exist and can be differentiable. This result bounds the performance of the algorithm when separating signals with variances close to zero. To overcome this problem, Aichner *et al.* [14] propose the regularization strategy by adding to the diagonal elements of $\mathbf{R}_{\mathbf{y}}^{(k,0)}$ with a constant regularization parameter δ_{reg}, i.e.,

$$\rho_{y_i}^{(k,0)} \approx \rho_{y_i}^{(k,0)} + \delta_{\mathrm{reg}} \tag{47}$$

where $\rho_{y_i}^{(k,0)}$, $i = 1, 2, \ldots, N$ is corresponding diagonal elements of $\mathbf{R}_{\mathbf{y}}^{(k,0)}$ and δ_{reg} is a positive constant.

Alternatively, a dynamical regularization is also introduced in [14] by computing δ_{reg} using

$$\delta_{\mathrm{reg}} = \delta_{\max} \, e^{-\rho_{y_i}^{(k,0)}/\delta_0} \tag{48}$$

where $\delta_{\max}$ and δ_0 are positive constants.

For convenience, the log algorithm and its regularized versions using a constant and a dynamical regularization pa-

Authorized licensed use limited to: Imperial College London. Downloaded on January 4, 2010 at 08:09 from IEEE Xplore. Restrictions apply.

rameters are, respectively, referred to as LOG, LOG-CR and LOG-DR.

We perform separation using the proposed algorithms and the log algorithms with and without regularization under three different cases, i.e., two source signals having high variance, two source signals with one having low variance and two source signals in a noisy environment.

The two source signals are generated by autoregressive models of order two (AR2)

$$s_1(n) = -1.34s_1(n-1) - 0.76s_1(n-2) + c_1\nu_1(n)$$
$$s_2(n) = -1.24s_2(n-1) - 0.65s_2(n-2) + c_2\nu_2(n)$$

where c_1 and c_2 are positive constants that control the nonstationarity property of the source signals, $\nu_1(n)$ and $\nu_2(n)$ are generated randomly by a normal distribution process.

By changing the values of c_1 and c_2, the signals s_1 and s_2 become nonstationary through the change of their variances. Therefore, we divide s_1 and s_2 into two blocks using two sets of c_1 and c_2. This method not only demonstrates a simple nonstationarity property of the signals but also offers a set of two correlation matrices for the implementation of the generalized eigen decomposition (GED) method [31], [32], which separates the signals by using two correlation matrices.

Since a set of two output signals correlation matrices $(K = 2)$ will be used in all experiments, we are thus able to employ GED as a reference that gives a closed-form solution. All algorithms are tested on three cases, and $\mathcal{P}$ is averaged over 500 independent trials of the above nonstationary source signals, each with 5000 samples. A penalty factor λ_C is set to 0.001 for BP and NBP in all experiments. The full rank mixing matrix with all its elements randomly drawn from a normally distributed random process is given by

$$\mathbf{A} = \begin{bmatrix} 0.19 & 1.72 \\ -1.26 & -0.76 \end{bmatrix}.$$

Case 1 Two Source Signals Both With a High Variance: We set $c_1 = c_2 = 1$ for $n = [1, 2500]$ and $c_1 = 0.75$, $c_2 = 0.5$ for $n = [2501, 5000]$. Fig. 1 shows a faster convergence rate of LOG, which is due to its steeper cost function. All algorithms converge to the result obtained from GED though with different rate. It should be mentioned that, apart from NBP whose upper bound is unity, the upper bounds of all other algorithms have not been given explicitly, and thus their step-sizes used in the experiment have to be chosen empirically to be as large as possible but for which each algorithm still converges.

Case 2 Two Source Signals With one Having a low Variance: We study the effect of gradient noise amplification induced by a low variance signal. Unlike $\tilde{\nabla}_{\mathbf{W}} J_{\mathrm{JD}}^{(k,0)}$, the natural gradient $\tilde{\nabla}_{\mathbf{W}} J_L^{(k,0)}$ requires an inverse of a matrix, meaning that the source signals need to be active throughout the observation. Moreover, a small value of any elements of such a matrix will be sufficient to cause an excessively large natural gradient leading to the performance degradation of the algorithm. In this experiment, c_2 is decreased to 0.1 for $n = [2501, 5000]$ and the other parameters including those for $n = [1, 2500]$ remain similar to *Case 1.*

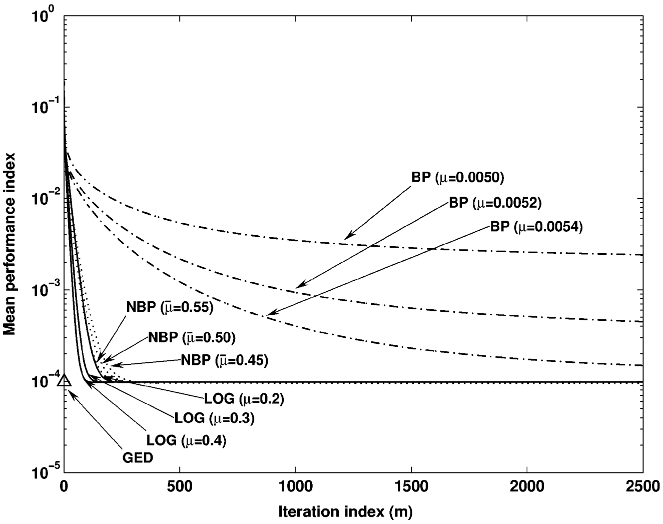


Fig. 1. Mean performance indices of BP $(\Gamma = 0)$, NBP $(\Gamma = 0)$, LOG, and GED obtained from the mixtures of two nonstationary AR2 source signals, both with a high variance.

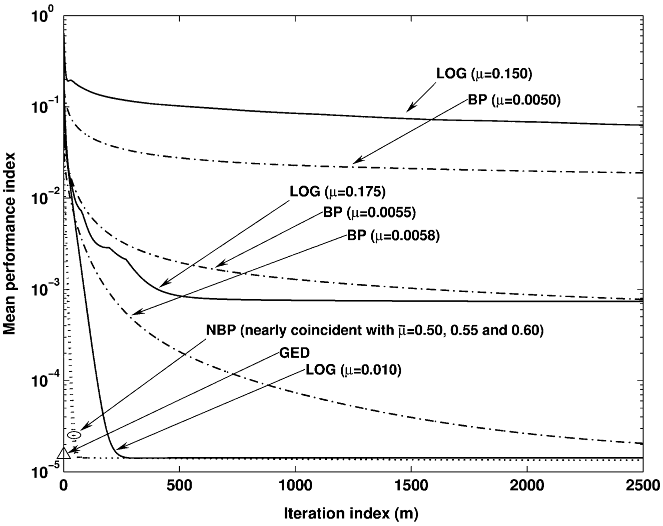


Fig. 2. Mean performance indices of BP $(\Gamma = 0)$, NBP $(\Gamma = 0)$, LOG, and GED obtained from the mixtures of two nonstationary AR2 source signals with one source signal having a low variance for a certain period of time.

In Fig. 2, the best performance is obtained from NBP. The performance of LOG considerably deteriorates when compared to *Case 1* due to the gradient noise amplification caused by the low variance signal. To mitigate this effect, we need to decrease the step-size of LOG. The better performance, however, is achieved at the expense of slower convergence rate.

Fig. 3 shows the performance of BP, NBP, LOG-CR, LOG-DR, and GED when c_2 is reduced to zero for $n = [2501, 5000]$ and the other parameters remain the same including those for $n = [1, 2500]$. In this case, the log algorithm cannot compute $\tilde{\nabla}_{\mathbf{W}} J_L^{(k,0)}$ after a few iterations and fails to separate the signals. To improve the algorithm in such a situation, a constant regularization and a dynamical regularization are needed to prevent the divergence of the algorithm.

Authorized licensed use limited to: Imperial College London. Downloaded on January 4, 2010 at 08:09 from IEEE Xplore. Restrictions apply.

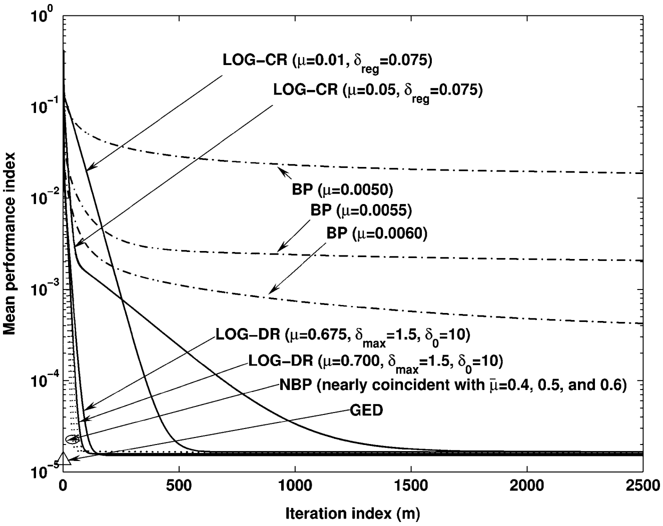


Fig. 3. Mean performance indices of BP $(\Gamma = 0)$, NBP $(\Gamma = 0)$, LOG, LOG-CR, LOG-DR, and GED obtained from the mixtures of two nonstationary AR2 source signals with one source signal having zero variance for a certain period of time.

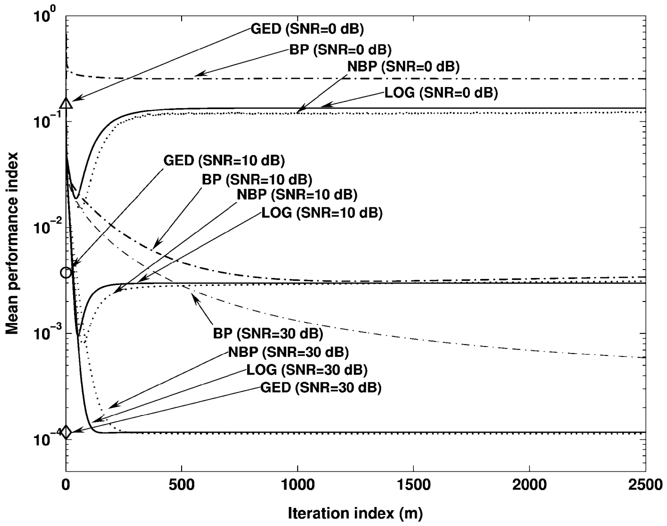


Fig. 4. Mean performance indices of BP $(\Gamma = 0)$, NBP $(\Gamma = 0)$, LOG, and GED obtained from the mixtures of two nonstationary AR2 source signals in noisy environments.

The LOG-CR algorithm $(\delta_{\text{reg}} = 0.075)$ has solved the invertibility problem of $\mathbf{R}_{\mathbf{y}}^{(k,0)}$ by adding to all its diagonal elements a positive constant δ_{reg} enabling the algorithm to converge to the result obtained from GED. A more efficient technique for solving the invertibility problem of $\mathbf{R}_{\mathbf{y}}^{(k,0)}$ is to employ a dynamical regularization. It can be seen that LOG-DR provides performance comparable with NBP. This comparable performance of LOG-DR and NBP is nearly as good as that obtained in *Case 1*. In addition, we investigate the computational complexity of the NBP and LOG-DR algorithms. It is found that NBP $(\Gamma = 0)$ requires $(4K+2)N^3 - 2(K-1)N^2 + N$ operations, which is more than $2KN^3 - 2KN^2 + KN$ operations required by LOG and $2KN^3 - 2KN^2 + (K+1)N$ operations required by LOG-CR. However, only LOG-DR provides performance comparable with NBP. We thus need to investigate further the computational complexity of the dynamical regularization parameters used in LOG-DR. The dynamical regularization exploits the exponential function, which can be computed using [33]

$$e^{-\rho_{y_i}^{(k,0)}/\delta_0} \approx \left(1 - \frac{\rho_{y_i}^{(k,0)}/\delta_0}{M}\right)^M, \quad i = 1, 2, \ldots, N. \quad (49)$$

By use of (49), the LOG-DR algorithm requires $(M+4)N$ additional operations and amounts totally to $2KN^3 - 2KN^2 + (K+M+4)N$ operations. We see that both NBP and LOG-DR have the complexity of $O(N^3)$, but the complexity of LOG-DR also depends on the choice of M. In this experiment $(N = K = 2)$, the computational complexity of NBP will be more than LOG-DR when M is set to be less than 23. However, small values of M are not sufficient to estimate (49) accurately, especially when the absolute value of the exponent term $\rho_{y_i}^{(k,0)}/\delta_0$ is large. It can therefore be said that NBP is more preferable for the BSS problem when N, K, and Γ are small, whereas LOG-DR is more computationally efficient when the problem becomes more complicated and the absolute value of the exponent terms used to compute in the dynamical regularization are small as this requires smaller values of M.

Case 3 Noisy Environments: We investigate the performance of algorithms when performing BSS in noisy environments. The instantaneous mixing model in the presence of additive noise is defined by

$$\mathbf{x}(n) = \mathbf{A}\mathbf{s}(n) + \mathbf{v}(n) \quad (50)$$

where $\mathbf{v}(n) = [v_1(n), v_2(n), \ldots, v_N(n)]^T$ is an $N \times 1$ additive noise vector, whose elements are drawn randomly from a normally distributed random process.

The source signals parameters are set as in *Case 1*. Fig. 4 shows the mean $\mathcal{P}$ of all algorithms converging to those obtained from GED at various levels of signal-to-noise ratio (SNR). The result also indicates the deterioration in performance of all algorithms in the presence of noise with comparable performance between NBP and LOG.

B. Online Algorithms

In the last but more realistic experiment, the OP and NOP algorithms, which employ both the nonstationarity and nonwhiteness properties, and the stochastic relative gradient (SRG) algorithm [19], which employs the nonstationarity property only, are compared. Additionally, we also modify the log algorithm to suit online application by using the following update:

$$\delta\mathbf{W}(n) = -2\,\mu\,\text{diag}^{-1}(\hat{\mathbf{R}}_{\mathbf{y}}(n))\text{off}(\hat{\mathbf{R}}_{\mathbf{y}}(n))\mathbf{W}(n) \quad (51)$$

where $\hat{\mathbf{R}}_{\mathbf{y}}(n) = \mathbf{y}(n)\mathbf{y}^T(n)$ is an instantaneous estimate of the output signals correlation matrix which produces a positive definite matrix required by the use of logarithmic function and Oppenheim's inequality.

In order to separate nonpersistently active signals which, in this experiment, are speech signals, the above online log algorithm needs to be used together with a dynamical regulariza-

Authorized licensed use limited to: Imperial College London. Downloaded on January 4, 2010 at 08:09 from IEEE Xplore. Restrictions apply.

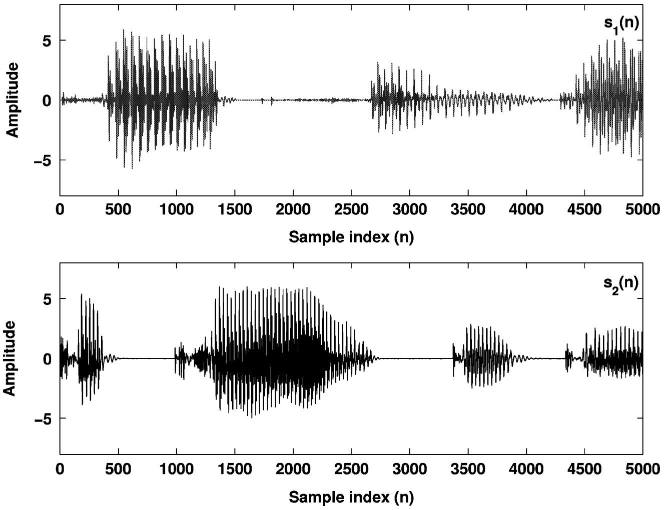


Fig. 5. Speech signals.

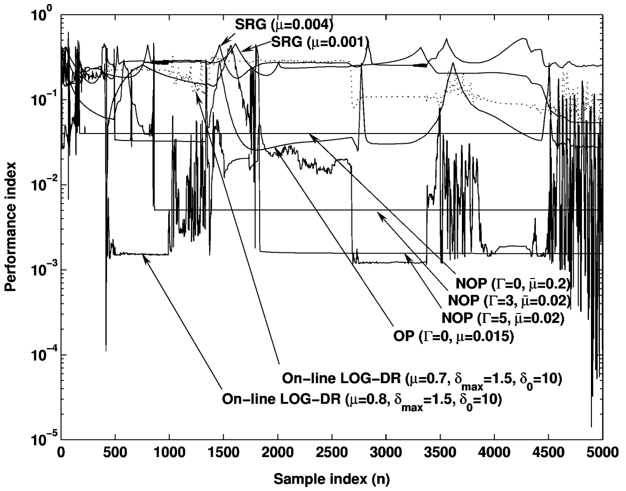


Fig. 6. Performance index of OP, NOP, online LOG-DR, and SRG obtained from the mixtures of two speech signals.

tion and is referred to as online LOG-DR. The speech signals.[3] are shown in Fig. 5 and the mixing matrix whose elements are randomly drawn from a normally distributed random process is given by

$$\mathbf{A} = \begin{bmatrix} 1.98 & 0.72 \\ -0.40 & 0.52 \end{bmatrix}.$$

We set the forgetting factor $\alpha = 0.99$ for OP, NOP, and SRG, and $\lambda_C = 0.01$ for OP and NOP. The step-size μ, the scaling factor $\bar{\mu}$, and the regularization parameters $\delta_{\max}$ and δ_0 are shown in the figure.

In Fig. 6, we can see that all NOP cases $(\Gamma = 0, 3, 5)$ give the results that are more robust to gradient noise than OP, SRG, and online LOG-DR, which is due to the use of normalization term that solves the gradient noise amplification problem. The OP, SRG, and online LOG-DR algorithms all suffer from excessively large gradients during the separation causing high values of performance index. Also, it can be seen that, as Γ is increased, NOP tends to converge more slowly. This can be explained by the fact that, as Γ is increased, NOP has to deal with more correlation matrices, which causes a slower convergence rate. However, the better performance of NOP, as indicated by its lower values of performance index, is achieved when Γ is increased because more information about the nonwhiteness property is included. Inevitably, this better performance of NOP is achieved at the cost of higher computational complexity. It is thus worth comparing the computational complexity of NOP to that of online LOG-DR. Like LOG-DR, the online LOG-DR algorithm also requires sufficiently large M to compute accurately the exponential function in (49) particularly when the absolute values of its exponent are large. By using the complexity in Table I and the previous section, NOP is more computationally efficient than online LOG-DR in the case where N and Γ are small, whereas online LOG-DR is more suitable for solving the problem that has large N and Γ and that employs large absolute values of the exponent in the dynamical regularization.

VII. Conclusion

A class of Frobenius norm-based algorithms using penalty term and natural gradient for blind signal separation which consists of the offline/block processing (BP), the online processing (OP) algorithms, and their normalized versions has been developed in this paper by using the nonstationarity and the nonwhiteness properties. We show that the global minimum corresponding to the desired demixing matrix exists in the region defined by the penalty function. Associated with both BP and OP is the gradient noise problem. We thus develop a normalized version of the BP and OP algorithms so as to mitigate gradient noise. By applying the minimal disturbance principle to the BSS problem, we obtain the normalized BP (NBP) and the normalized OP (NOP) algorithms which exhibit fast convergence and robustness to gradient noise.

The proposed Frobenius norm-based algorithms have been compared with the logarithm-based algorithms. All proposed offline algorithms, when separating persistently active signals, converge to the desired solution, though having a slower convergence rate than the logarithm-based algorithm. The superior performance to the logarithm-based algorithm with and without regularizations is achieved by NBP when separating nonpersistently active signals, which has confirmed the capability of the proposed normalized algorithm to solve the gradient noise problem. When separating speech signals using online algorithms, the NOP algorithm performs better than the online logarithm-based algorithm with dynamical regularization and the stochastic relative gradient algorithm as it employs the normalized term to cope with gradient noise amplification.

The proposed normalized algorithms, although they require higher computational complexity than the logarithm-based algorithm without regularization, are more computationally efficient than the logarithm-based algorithm with dynamical regularization, which requires high complexity to compute its exponential function, especially for the BSS problem having small numbers of signals, blocks, and nonzero time lag.

Future research will focus on applying the proposed algorithms to solve convolutive BSS problem in the frequency domain.

[3] Available: http://www.bsp.brain.riken.jp/ICALAB/ICALABSignalProc/benchmarks/Speech4.mat.

Authorized licensed use limited to: Imperial College London. Downloaded on January 4, 2010 at 08:09 from IEEE Xplore. Restrictions apply.

Appendix

A. *Proof of Prerequisites 1*

To prove the necessity of the Prerequisites 1, we employ the Jacobi rotation to transform an unknown mixing matrix $\mathbf{A}$ and a source signals vector $\mathbf{s}(n)$ into a different unknown mixing matrix $\tilde{\mathbf{A}}$ and a different source signals vector $\tilde{\mathbf{s}}(n)$, whose observed signals vector $\tilde{\mathbf{x}}(n) = \tilde{\mathbf{A}}\tilde{\mathbf{s}}(n)$ possesses the same second-order statistical property as $\mathbf{x}(n) = \mathbf{A}\mathbf{s}(n)$.

1) Let us assume that, for a given τ, $\mathbf{p}_{s_1}^{(\tau)}$ and $\mathbf{p}_{s_2}^{(\tau)}$ are linearly dependent such that $\mathbf{p}_{s_1}^{(\tau)} = \mathbf{p}_{s_2}^{(\tau)}$. Then, consider two unknown mixing matrices $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3 \ldots, \mathbf{a}_N]$ and $\tilde{\mathbf{A}} = [\tilde{\mathbf{a}}_1, \tilde{\mathbf{a}}_2, \mathbf{a}_3 \ldots, \mathbf{a}_N]$ with $\tilde{\mathbf{a}}_1$ and $\tilde{\mathbf{a}}_2$ defined by

$$[\tilde{\mathbf{a}}_1, \tilde{\mathbf{a}}_2] = [\mathbf{a}_1, \mathbf{a}_2] \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}$$

where $\mathbf{a}_i$, $i = 1, 2, \ldots, N$, $\tilde{\mathbf{a}}_1$ and $\tilde{\mathbf{a}}_2$ are $N \times 1$ vectors. Also, consider two $N \times 1$ source signals vectors $\mathbf{s}(n) = [s_1(n), s_2(n), s_3(n), \ldots, s_N(n)]^T$ and $\tilde{\mathbf{s}}(n) = [\tilde{s}_1(n), \tilde{s}_2(n), s_3(n), \ldots, s_N(n)]^T$ with $\tilde{s}_1(n)$ and $\tilde{s}_2(n)$ given by

$$\begin{bmatrix} \tilde{s}_1(n) \\ \tilde{s}_2(n) \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} s_1(n) \\ s_2(n) \end{bmatrix}.$$

Due to the use of Jacobi rotation, we obtain $\tilde{\mathbf{x}}(n) = \tilde{\mathbf{A}}\tilde{\mathbf{s}}(n) = \mathbf{A}\mathbf{s}(n) = \mathbf{x}(n)$. Accordingly, it follows that $\mathbf{R}_{\mathbf{x}}^{(k,\tau)} = \mathbf{R}_{\tilde{\mathbf{x}}}^{(k,\tau)}, \forall k$. Next, we consider the second-order statistics of $\mathbf{s}(n)$ and $\tilde{\mathbf{s}}(n)$. Since the source signals are uncorrelated to each other and $\rho_{s_1}^{(1,\tau)} = \rho_{s_2}^{(1,\tau)}$, we thus have $\rho_{\tilde{s}_1}^{(1,\tau)} = \rho_{s_1}^{(1,\tau)} \cos^2\theta + \rho_{s_2}^{(1,\tau)} \sin^2\theta = \rho_{s_1}^{(1,\tau)}$. Also, it can be shown in a similar fashion that $\rho_{\tilde{s}_1}^{(k,\tau)} = \rho_{s_1}^{(k,\tau)}$ and $\rho_{\tilde{s}_2}^{(k,\tau)} = \rho_{s_2}^{(k,\tau)}, \forall k$. Accordingly, it follows that $\mathbf{\Lambda}_{\mathbf{s}}^{(k,\tau)} = \mathbf{\Lambda}_{\tilde{\mathbf{s}}}^{(k,\tau)}, \forall k$.

Now, we can see that the second-order statistical relation between $\mathbf{s}(n)$ and $\mathbf{x}(n)$ using a set of local time-average correlation matrices is similar to that of $\tilde{\mathbf{s}}(n)$ and $\tilde{\mathbf{x}}(n)$, even though these two pairs of relations originate from different source signals and different mixing matrices. Hence, it can be concluded that it is not possible to perform BSS using a set of local time-average correlation matrices in the case that $\mathbf{p}_{s_i}^{(\tau)}$ and $\mathbf{p}_{s_j}^{(\tau)}, \forall 1 \leq i \neq j \leq N$, are linearly dependent. In addition, this prerequisite generalizes the separability condition found in Proposition 1 of [32], where two correlation matrices of nonstationary signals at zero time lag obtained from two different time intervals are considered.

2) Let us similarly assume that, for a given k, $\tilde{\mathbf{p}}_{s_1}^{(k)} = \tilde{\mathbf{p}}_{s_2}^{(k)}$, which is the case of linear dependence between $\tilde{\mathbf{p}}_{s_1}^{(k)}$ and $\tilde{\mathbf{p}}_{s_2}^{(k)}$. Consider again the above-defined mixing processes $\mathbf{x}(n) = \mathbf{A}\mathbf{s}(n)$ and $\tilde{\mathbf{x}}(n) = \tilde{\mathbf{A}}\tilde{\mathbf{s}}(n)$. It can be readily seen that $\mathbf{R}_{\mathbf{x}}^{(k,\tau)} = \mathbf{R}_{\tilde{\mathbf{x}}}^{(k,\tau)}, \forall \tau$. For the second-order statistics of $\mathbf{s}(n)$ and $\tilde{\mathbf{s}}(n)$, since $\rho_{s_1}^{(k,0)} = \rho_{s_2}^{(k,0)}$, we obtain $\rho_{\tilde{s}_1}^{(k,0)} = \rho_{s_1}^{(k,0)} \cos^2\theta + \rho_{s_2}^{(k,0)} \sin^2\theta = \rho_{s_1}^{(k,0)}$. Likewise, it can be shown that $\rho_{\tilde{s}_1}^{(k,\tau)} = \rho_{s_1}^{(k,\tau)}$ and $\rho_{\tilde{s}_2}^{(k,\tau)} = \rho_{s_2}^{(k,\tau)}, \forall \tau$. Accordingly, it follows that $\mathbf{\Lambda}_{\mathbf{s}}^{(k,\tau)} = \mathbf{\Lambda}_{\tilde{\mathbf{s}}}^{(k,\tau)}, \forall \tau$.

We readily see that the second-order statistical relation between $\mathbf{s}(n)$ and $\mathbf{x}(n)$ is similar to the relation obtained from $\tilde{\mathbf{s}}(n)$ and $\tilde{\mathbf{x}}(n)$ no matter what time lag τ is chosen. Therefore, it can be concluded in a similar fashion as 1) that it is not possible to perform blind signal separation using a set of local time-average correlation matrices whenever $\tilde{\mathbf{p}}_{s_i}^{(\tau)}$ and $\tilde{\mathbf{p}}_{s_j}^{(\tau)}, \forall 1 \leq i \neq j \leq N$, are linearly dependent.

This prerequisite, when the local stationarity property is extended to the whole set of samples, can be viewed as a generalization of the separability condition on the normalized spectra of stationary but nonwhite source signals in [1], which utilizes variance normalization by considering $E[s_i^2(n)] = E[s_j^2(n)] = 1$ and as a special case of the condition found in [34] when a cycle frequency of the ith and jth cyclostationary source signals is zero. In addition, the nonwhiteness separability condition with proof for stationary but nonwhite complex signals is given in [35].

B. *Proof of Proposition 1*

1) is immediate. For 2), every matrix $\mathbf{W} \in \mathcal{W}_N$ is full rank by Definition 1 and thus has N independent rows and columns with at least one nonzero entry in each row and column. By using 1), an appropriate permutation and scaling of the rows of a matrix $\mathbf{W}$ gives a matrix whose properties satisfy $\mathcal{W}_N^*$. Accordingly, $\mathcal{W}_N^*$ is a case of 1) and hence $\mathcal{W}_N^*$ is a subset of $\mathcal{W}_N$.

C. *Proof of Lemma 1*

Using J_{JD} defined in (9) and making use of (12), we can express the natural gradient as

$$\tilde{\nabla}_{\mathbf{W}} J_{\mathrm{JD}} = 4 \sum_{k=1}^{K} \sum_{\tau=0}^{\Gamma} \beta^{(k,\tau)} \mathrm{off}\left(\underline{\mathbf{R}}_{\mathbf{y}}^{(k,\tau)}\right) \underline{\mathbf{R}}_{\mathbf{y}}^{(k,\tau)} \mathbf{W}. \quad (52)$$

First, if $\mathbf{W} = \mathbf{0}$, then it is straightforward that $\tilde{\nabla}_{\mathbf{W}} J_{\mathrm{JD}}$ is a zero matrix and, second, if $\mathbf{W} \in \mathcal{W}_l$, $l = 1, 2, \ldots, N$, then the term off $(\underline{\mathbf{R}}_{\mathbf{y}}^{(k,\tau)})$ is equal to a zero matrix for all (k, τ) and, as a result, $\tilde{\nabla}_{\mathbf{W}} J_{\mathrm{JD}}$ is also equal to a zero matrix.

Conversely, if $\tilde{\nabla}_{\mathbf{W}} J_{\mathrm{JD}} = \mathbf{0}$, we can rewrite (52) as

$$\sum_{k=1}^{K} \sum_{\tau=0}^{\Gamma} \beta^{(k,\tau)} \mathrm{off}\left(\underline{\mathbf{R}}_{\mathbf{y}}^{(k,\tau)}\right) \underline{\mathbf{R}}_{\mathbf{y}}^{(k,\tau)} \mathbf{W} = \mathbf{0}. \quad (53)$$

Expanding the term off$(\underline{\mathbf{R}}_{\mathbf{y}}^{(k,\tau)})$ using the relation off $(\underline{\mathbf{R}}_{\mathbf{y}}^{(k,\tau)}) = \underline{\mathbf{R}}_{\mathbf{y}}^{(k,\tau)} -$ diag $(\underline{\mathbf{R}}_{\mathbf{y}}^{(k,\tau)})$ and rearranging (53), we obtain

$$\sum_{k=1}^{K} \sum_{\tau=0}^{\Gamma} \beta^{(k,\tau)} \underline{\mathbf{R}}_{\mathbf{y}}^{(k,\tau)} \underline{\mathbf{R}}_{\mathbf{y}}^{(k,\tau)} \mathbf{W} = \sum_{k=1}^{K} \sum_{\tau=0}^{\Gamma} \beta^{(k,\tau)} \mathrm{diag}\left(\underline{\mathbf{R}}_{\mathbf{y}}^{(k,\tau)}\right) \underline{\mathbf{R}}_{\mathbf{y}}^{(k,\tau)} \mathbf{W}. \quad (54)$$

Using (8), we see that (54) holds, when either $\mathbf{W} = \mathbf{0}$, which also results in $\underline{\mathbf{R}}_{\mathbf{y}}^{(k,\tau)} = \mathbf{0}$, or $\underline{\mathbf{R}}_{\mathbf{y}}^{(k,\tau)}$ is a diagonal matrix. The latter case is the case where $\mathbf{W} \in \mathcal{W}_l$, $l = 1, 2, \ldots, N$. We therefore conclude that if $\tilde{\nabla}_{\mathbf{W}} J_{\mathrm{JD}} = \mathbf{0}$, then $\mathbf{W} = \mathbf{0}$ or $\mathbf{W} \in \mathcal{W}_l$, $l = 1, 2, \ldots, N$.

Authorized licensed use limited to: Imperial College London. Downloaded on January 4, 2010 at 08:09 from IEEE Xplore. Restrictions apply.

D. Proof of Theorem 1

By using the first-order conditions from (12) and (14), we obtain the following stationary points of J_{BP}: 1) $\mathbf{W} = \mathbf{0}$; 2) $\tilde{\nabla}_{\mathbf{W}} J_{\mathrm{JD}} = \tilde{\nabla}_{\mathbf{W}} J_C = \mathbf{0}$, and 3) $\tilde{\nabla}_{\mathbf{W}} J_{\mathrm{JD}} + \lambda_C \tilde{\nabla}_{\mathbf{W}} J_C = \mathbf{0}$ but $\tilde{\nabla}_{\mathbf{W}} J_{\mathrm{JD}}, \tilde{\nabla}_{\mathbf{W}} J_C \neq \mathbf{0}$.

To examine whether these points are a maximum, minimum, or saddle point, we follow [36] to investigate the effect of a small perturbation $\delta\mathbf{W} = [\delta w_{ij}]$. In practice, the small perturbation $\delta\mathbf{W}$ can be achieved by setting μ and λ_C to be sufficiently small. At the perturbed point $\mathbf{W} + \delta\mathbf{W}$, we obtain

$$\begin{aligned} J_{\mathrm{JD}}(\mathbf{W} + \delta\mathbf{W}) = \sum_{k=1}^{K} \sum_{\tau=0}^{\Gamma} \beta^{(k,\tau)} \Big\| \mathrm{off} \Big(& \mathbf{W} \underline{\mathbf{R}}_{\mathbf{x}}^{(k,\tau)} \mathbf{W}^T \\ & + \delta\mathbf{W} \underline{\mathbf{R}}_{\mathbf{x}}^{(k,\tau)} \delta\mathbf{W}^T + \delta\mathbf{W} \underline{\mathbf{R}}_{\mathbf{x}}^{(k,\tau)} \mathbf{W}^T \\ & + \mathbf{W} \underline{\mathbf{R}}_{\mathbf{x}}^{(k,\tau)} \delta\mathbf{W}^T \Big) \Big\|_F^2 \end{aligned} \tag{55}$$

$$J_C(\mathbf{W} + \delta\mathbf{W}) = \|\mathrm{diag}(\mathbf{W} + \delta\mathbf{W} - \mathbf{I})\|_F^2. \tag{56}$$

For sufficiently small $\delta\mathbf{W}$, the quadratic term $\delta\mathbf{W} \underline{\mathbf{R}}_{\mathbf{x}}^{(k,\tau)} \delta\mathbf{W}^T$ in (55) can be neglected. The difference $\Delta J_{\mathrm{BP}}(\delta\mathbf{W}) = J_{\mathrm{BP}}(\mathbf{W} + \delta\mathbf{W}) - J_{\mathrm{BP}}(\mathbf{W})$ at the stationary point 1) thus becomes

$$\Delta J_{\mathrm{BP}}(\delta\mathbf{W}) = \lambda_C \sum_{i=j}^{N} \delta w_{ij}^2 - 2\delta w_{ij} \approx -2\lambda_C \sum_{i=j}^{N} \delta w_{ij}. \tag{57}$$

The approximation in (57) results from dropping the small quadratic term δw_{ij}^2. Similarly, the difference $\Delta J_{\mathrm{BP}}(-\delta\mathbf{W}) = J_{\mathrm{BP}}(\mathbf{W} - \delta\mathbf{W}) - J_{\mathrm{BP}}(\mathbf{W})$ is of the form

$$\Delta J_{\mathrm{BP}}(-\delta\mathbf{W}) = \lambda_C \sum_{i=j}^{N} \delta w_{ij}^2 + 2\delta w_{ij} \approx 2\lambda_C \sum_{i=j}^{N} \delta w_{ij}. \tag{58}$$

It is clear that (57) is equal to (58) except for the sign. Therefore, we conclude that the stationary point 1) is a saddle point of J_{BP}.

By using Proposition 1 and Lemma 1, the stationary point 2) is nothing but $\mathbf{W} \in \mathcal{W}_N^*$. We now drop the quadratic terms $\delta\mathbf{W} \underline{\mathbf{R}}_{\mathbf{x}}^{(k,\tau)} \delta\mathbf{W}^T$ and $\delta\mathbf{W}_{ij}^2$. As a result, the differences $\Delta J_{\mathrm{BP}}(\delta\mathbf{W})$ and $\Delta J_{\mathrm{BP}}(-\delta\mathbf{W})$ at the stationary point 2), respectively, become

$$\begin{aligned} \Delta J_{\mathrm{BP}}(\delta\mathbf{W}) = \sum_{k=1}^{K}\sum_{\tau=0}^{\Gamma} \beta^{(k,\tau)} \Big\| \mathrm{off} \Big(& \delta\mathbf{W} \underline{\mathbf{R}}_{\mathbf{x}}^{(k,\tau)} \mathbf{W}^T \\ & + \mathbf{W} \underline{\mathbf{R}}_{\mathbf{x}}^{(k,\tau)} \delta\mathbf{W}^T \Big) \Big\|_F^2 \end{aligned} \tag{59}$$

$$\begin{aligned} \Delta J_{\mathrm{BP}}(-\delta\mathbf{W}) = \sum_{k=1}^{K}\sum_{\tau=0}^{\Gamma} \beta^{(k,\tau)} \Big\| \mathrm{off} \Big(& -\delta\mathbf{W} \underline{\mathbf{R}}_{\mathbf{x}}^{(k,\tau)} \mathbf{W}^T \\ & - \mathbf{W} \underline{\mathbf{R}}_{\mathbf{x}}^{(k,\tau)} \delta\mathbf{W}^T \Big) \Big\|_F^2 \\ = \Delta J_{\mathrm{BP}}(\delta\mathbf{W}). & \end{aligned} \tag{60}$$

From (60), we conclude that the stationary point 2) is the minimum of J_{BP}.

The stationary point 3), by Lemma 1, is *not* a desired demixing matrix but is the point that causes an undesired solution to the problem. By using (53), (14), and $\tilde{\nabla}_{\mathbf{W}} J_{\mathrm{JD}}, \tilde{\nabla}_{\mathbf{W}} J_C \neq \mathbf{0}$, the stationary point 3) is the point that satisfies

$$\begin{aligned} 2 \sum_{k=1}^{K} \sum_{\tau=0}^{\Gamma} \beta^{(k,\tau)} \mathrm{off} \left(\underline{\mathbf{R}}_{\mathbf{y}}^{(k,\tau)} \right) \underline{\mathbf{R}}_{\mathbf{y}}^{(k,\tau)} & \\ + \lambda_C \, \mathrm{diag} \left(\mathbf{W} - \mathbf{I} \right) \mathbf{W}^T & = \mathbf{0}. \end{aligned} \tag{61}$$

Next, we consider the neighborhood of $J_C = 0$, where the diagonal elements of $\mathbf{W}$ are defined as $1 \pm \delta w_{ii}$ with $\delta w_{ii} < 1$ being a small perturbation and $i = 1, 2, \ldots, N$. Since both $\beta^{(k,\tau)}$ and λ_C are always positive by setting, we see that all the diagonal elements of the first and the second terms on the left-hand side of (61) are always positive. Accordingly, (61) is not equal to a zero matrix when $1 \pm \delta w_{ii}$ is the diagonal elements of $\mathbf{W}$ and $\delta w_{ii} < 1$, which means that $\mathbf{W}$ is in the neighborhood of $J_C = 0$. Therefore, it can be said that the stationary point 3) does *not* exist in the vicinity of $\mathrm{diag}(\mathbf{W}) = \mathbf{I}$.

Since J_{BP} is a nonnegative function, and $J_{\mathrm{BP}} = 0$ at the stationary point 2), we therefore conclude that $\mathbf{W} \in \mathcal{W}_N^*$ is, by Definition 1, the desired demixing matrix as well as the global minimum of J_{BP} in the neighborhood of a matrix whose diagonal entries are all equal to unity.

ACKNOWLEDGMENT

The authors would like to thank the associate editor and anonymous reviewers whose suggestions have helped us to improve this paper significantly.

REFERENCES

[1] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines, “A blind source separation technique using second-order statistics,” *IEEE Trans. Signal Process.*, vol. 45, no. 2, pp. 434–444, Feb. 1997.

[2] H. Buchner, R. Aichner, and W. Kellermann, “A generalization of blind source separation algorithms for convolutive mixtures based on second- order statistics,” *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 120–134, Jan. 2005.

[3] L. Tong, R. Liu, V. C. Soon, and Y.-F. Huang, “Indeterminacy and identifiability of blind identification,” *IEEE Trans. Circuits Syst.*, vol. 38, no. 5, pp. 499–509, May 1991.

[4] L. Parra and C. Spence, “Convolutive blind separation of non-stationary sources,” *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 320–327, May 2000.

[5] S. Choi and A. Cichocki, “Blind separation of nonstationary sources in noisy mixtures,” *Electron. Lett.*, vol. 36, pp. 848–849, Apr. 2000.

[6] J.-F. Cardoso and A. Souloumiac, “Jacobi angles for simultaneous diagonalization,” *SIAM J. Matrix Anal. Applicat.*, vol. 17, no. 1, pp. 161–164, 1996.

[7] A. Ziehe, K.-R. Müller, G. Nolle, B.-M. Mackert, and G. Curio, “Artifact reduction in magnetoneurography based on lime-delayed second-order correlations,” *IEEE Trans. Biomed. Eng.*, vol. 47, no. 1, pp. 75–87, Jan. 2000.

[8] N. Murata, S. Ikeda, and A. Ziehe, “An approach to blind source separation based on temporal structure of speech signals,” *Neurcomputing*, vol. 41, pp. 1–24, Oct. 2001.

[9] J.-F. Cardoso, “On the performance of orthogonal source separation algorithms,” in *Proc. EUSIPCO’94*, Edinburgh, U.K., Sep. 1994, pp. 776–779.

[10] D. T. Pham, “Joint approximate diagonalization of positive definite Hermitian matrices,” *SIAM J. Matrix Anal. Applicat.*, vol. 22, no. 4, pp. 1136–1152, 2000.

[11] A. Yeredor, “Non-orthogonal joint diagonalization in the least-squares sense with application in blind source separation,” *IEEE Trans. Signal Process.*, vol. 50, no. 7, pp. 1545–1553, Jul. 2002.

Authorized licensed use limited to: Imperial College London. Downloaded on January 4, 2010 at 08:09 from IEEE Xplore. Restrictions apply.

[12] K. Matsuoka, M. Ohya, and M. Kawamoto, "A neural net for blind separation of nonstationary signals," *Neural Netw.*, vol. 8, no. 3, pp. 411–419, 1995.

[13] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1985.

[14] R. Aichner, H. Buchner, F. Yan, and W. Kellermann, "A real-time blind source separation scheme and its application to reverberant and noisy acoustic environment," *Signal Process.*, vol. 86, pp. 1260–1277, Jun. 2006.

[15] U. Manmontri and P. A. Naylor, "Blind identification using second-order statistics: a nonstationarity and nonwhiteness approach," in *Proc. IEEE ICASSP'05*, Mar. 2005, vol. 5, pp. 305–308.

[16] R. Aichner, H. Buchner, F. Yan, and W. Kellermann, "Blind signal separation using a criterion based on principle of minimal disturbance," in *Proc. IEEE ICASSP'06*, May 2006, vol. 5, pp. 829–832.

[17] S. Amari, "Natural gradient works efficiently in learning," *Neural Computat.*, vol. 10, pp. 251–276, Feb. 1998.

[18] J.-F. Cardoso and B. H. Laheld, "Equivariant adaptive source separation," *IEEE Trans. Signal Process.*, vol. 44, no. 12, pp. 3017–3030, Dec. 1996.

[19] D.-T. Pham and J.-F. Cardoso, "Blind separation of instantaneous mixtures of nonstationary sources," *IEEE Trans. Signal Process.*, vol. 49, no. 9, pp. 1837–1848, Sep. 2001.

[20] C. D. Meyer, *Matrix Analysis and Applied Linear Algebra*. Philadelphia, PA: SIAM, 2000.

[21] M. Joho, R. H. Lambert, and H. Mathis, "Elementary cost functions for blind separation of non-stationary source signals," in *Proc. IEEE ICASSP'01*, May 2001, pp. 2793–2796.

[22] M. Joho and H. Mathis, "Joint diagonalization of correlation matrices by using gradient methods with application to blind signal separation," in *Proc. IEEE SAM'02*, Aug. 2002, pp. 273–277.

[23] R. Thawonmas and A. Cichocki, "Blind signal extraction of arbitrarily distributed, but temporally correlated signals—A neural network approach," *IEICE Trans. Fundamentals Electron., Commun., Comput. Sci.*, vol. E82-A, pp. 1834–1844, Sep. 1999.

[24] W. Wang, S. Sanei, and J. A. Chambers, "Penalty function-based joint diagonalization approach for convolutive blind separation of nonstationary sources," *IEEE Trans. Signal Process.*, vol. 53, no. 5, pp. 1654–1669, May 2005.

[25] D. G. Luenberger, *Linear and Nonlinear Programming*. Reading, MA: Addison-Wesley, 1984, ch. 12.

[26] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear Programming Theory and Algorithms*. New York: Wiley, 1993, ch. 9.

[27] R. D. DeGroat, E. M. Dowling, and D. A. Linebarger, , V. K. Madisetti and D. B. Williams, Eds., "Subspace tracking," in *Digital Signal Processing Handbook*. Boca Raton, FL: CRC, 1998.

[28] B. Widrow and M. Lehr, "30 years of adaptive neural networks: perceptron, madaline, and backpropagation," *Proc. IEEE*, vol. 78, no. 9, pp. 1415–1442, Sep. 1990.

[29] J. Dattorro, *Convex Optimization and Euclidean Distance Geometry*. Palo Alto, CA: Meboo, 2005.

[30] E. Moreau, "A generalization of joint-diagonalization criteria for source separation," *IEEE Trans. Signal Process.*, vol. 49, no. 3, pp. 530–541, Mar. 2001.

[31] A. Souloumiac, "Blind source detection and separation using second order non-stationarity," in *Proc. IEEE ICASSP'95*, May 1995, pp. 1912–1915.

[32] M. K. Tsatsanis and C. Kweon, "Blind source separation os non-stationary sources using second-order statistics," in *Proc. 32nd Asilomar Conf. Signals, Syst., Comput.*, Nov. 1998, pp. 1574–1578.

[33] A. H. Karp, "Exponential and logarithm by sequential squaring," *IEEE Trans. Comput.*, vol. C-33, no. 5, pp. 462–464, May 1984.

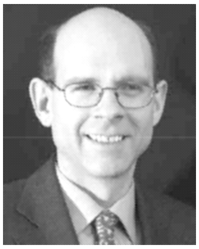
[34] K. Abed-Meraim, Y. Xiang, J. H. Manton, and Y. Hua, "Blind source-separation using second-order cyclostationary statistics," *IEEE Trans. Signal Process.*, vol. 49, no. 4, pp. 694–701, Apr. 2001.

[35] K. Abed-Meraim, Y. Xiang, and Y. Hua, "Generalized second order identifiability condition and relevant testing technique," in *Proc. IEEE ICASSP'00*, Jun. 2000, pp. 2989–2992.

[36] J. D. Erdogmus, Y. N. Rao, K. E. Hild, II, and J. C. Principe, "Simultaneous principal-component extraction with application to adaptive blind multiuser detection," *EURASIP J. Appl. Signal Process.*, vol. 2002, no. 12, pp. 1473–1484, 2002.

Uttachai Manmontri received the B.Eng. and M.Eng. degrees from Chulalongkorn University, Bangkok, Thailand, in 1996 and 1998, respectively, and the Ph.D. degree from Imperial College London, London, U.K., in 2006, all in electrical engineering.

From 1998 to 2001, he was with the Post and Telegraph Department, Ministry of Transport and Communications, Bangkok, Thailand, where he served as a Communications Engineer, and since 2001, he has been with the Thailand Institute of Scientific and Technological Research (TISTR), Pathumthani, Thailand. He was a recipient of the Royal Thai Government scholarship (2002–2006) to pursue the Ph.D. degree at Imperial College London. His research interests include blind signal separation, adaptive filter theory, and optimization theory.



Patrick A. Naylor (M'89) received the B.Eng. degree in electronics and electrical engineering from the University of Sheffield, Sheffield, U.K., in 1986 and the Ph.D. degree from Imperial College London, London, U.K., in 1990.

Since 1989, he has been a Member of Academic Staff in the Communications and Signal Processing Group, Imperial College London, where he is also the Director of Postgraduate Studies. His research interests are in the areas of speech and audio signal processing, and he has worked in particular on adaptive signal processing for acoustic echo control, speaker identification, multichannel speech enhancement, and speech production modeling. In addition to his academic research, he enjoys several fruitful links with industry in the U.K., USA, and in mainland Europe.

Dr. Naylor is an Associate Editor of IEEE SIGNAL PROCESSING LETTERS and a member of the IEEE Signal Processing Society Technical Committee on Audio and Electroacoustics.

Authorized licensed use limited to: Imperial College London. Downloaded on January 4, 2010 at 08:09 from IEEE Xplore. Restrictions apply.