GLOBALLY CONVERGENT NEWTON ALGORITHMS FOR BLIND DECORRELATION

Sergio Cruces

Signal Processing Group
Camino Descubrimientos,
41092-Seville, Spain.
http://viento.us.es/~sergio
sergio@us.es

Andrzej Cichocki

Brain Science Institute
RIKEN, 2-1 Hirosawa, Wako-shi,
Saitama, 351-0198 Japan.
http://www.bsp.brain.riken.go.jp/
cia@brain.riken.go.jp

ABSTRACT

This paper presents novel Newton algorithms for the blind adaptive decorrelation of real and complex processes. They are globally convergent and exhibit an interesting relationship with the natural gradient algorithm for blind decorrelation and the Goodall learning rule. Indeed, we show that these two later algorithms can be obtained from their Newton decorrelation versions when an exact matrix inversion is replaced by an iterative approximation to it.

1. INTRODUCTION

The problem of the decorrelation (or sphering) of the observations consist in finding algorithms that eliminate the second order redundancy among the different measured components that share a common temporal reference. Decorrelation of the data has several desired features: it increases the convergence speed of many algorithms and can be employed as a preprocessing step in other algorithms that restrict the search in the matrix parameters space. This indeed is the case in the Independent Component Analysis where one tries to obtain a linear transformation of the observations that makes the outputs the most independent as possible [1, 2]. In absence of noise, a preprocessing step that decorrelates or spheres the data, allows later to restrict one's attention to the unitary transformations of the data that drive the outputs towards their mutual independence.

In general, one wish to enforce the spatial decorrelation of the observations by modifying them the least possible in a certain sense. Thus, it is quite natural to impose that the decorrelating system be designed to preserve the general orientation of the observations, i.e., to allow them to be scaled in principal unitary directions but not to be rotated. The direct solution to this problem is the Mahalanobis or canonical transformation, i.e., the optimal decorrelation system is the principal square root of the sample covariance matrix.

Much of the research in the blind spatial decorrelation of the observations has been done in the field of the Principal Component Analysis (PCA) [1, 3]. However, when one is not mainly interested in obtaining the principal directions of the data, but in their decorrelation, other alternative algorithms can be designed to perform these task. This is the case of the Goodall algorithm [4, 5], of the Silva and Almeida symmetric decorrelation algorithm [6] (later identified with the natural gradient algorithm for blind decorrelation [7, 8]) and of the Least Squares prewhitening algorithms [9].

In this paper we propose a new Newton approach for the blind decorrelation problem. We show how the natural gradient for blind decorrelation and the Goodall decorrelation algorithm can be seen as low complexity approximations of Newton learning rules. These approximations hold true as long as certain parameters such as the learning step size and the initialization for the decorrelation system are properly chosen.

Throughout this work the following notation is used: $(\cdot)'$ and $(\cdot)^*$ denote the transpose and the Hermitian transpose operators, respectively; $\otimes$ denotes the Kronecker product between two given matrices; the $\mathrm{vec}(\cdot)$ operator is an arrangement that stacks the all the columns a matrix in a single vector; the decorrelation system is a matrix represented by $\mathbf{B}$ while the inverse decorrelation system $\mathbf{B}^{-1}$ is represented by $\mathbf{A}$.

The structure of the paper is the following. Section 2 presents the signal model and the direct solution of the decorrelation problem. Section 3 presents the proposed Newton approach to the blind decorrelation of real and complex processes. Section 4 analyzes the global stability of the Newton decorrelation algorithm. Section 5 establishes the relationship between the algorithm and other classical methods. Section 6 is devoted to the simulations and, finally, section 7 presents the conclusions.

This work has been supported by the CICYT project of the Government of Spain (Grant TIC2001-0751-C04-04).

2. SIGNAL MODEL

Let us define $\mathbf{x}[k] = [x_1[k], \ldots, x_N[k]]'$, $k = 0, \cdots, M-1$ as the M sample vectors of observations, each of dimension $N \times 1$, drawn from a wide sense stationary vector process associated with the recorded signals by N sensors. Without loss of generality we assume that the observation process is zero mean and has spatially correlated components.

The objective of the spatial decorrelation problem is to find a linear transformation $\mathbf{B}$ of the observations $\mathbf{x}[k]$ for which the components of the transformed vector

$$\mathbf{z}[k] = \mathbf{B}\mathbf{x}[k] \tag{1}$$

are mutually uncorrelated, i.e., being $\mathbf{R}_{xx} = E[\mathbf{x}[k](\mathbf{x}[k])^*]$ the correlation matrix of the original observations, then, for the desired decorrelation system $\mathbf{B}_0$ the correlation matrix of the transformed observations is equal to the identity matrix

$$E[\mathbf{z}[k](\mathbf{z}[k])^*] = \mathbf{B}_0\mathbf{R}_{xx}\mathbf{B}_0^* = \mathbf{I}\,. \tag{2}$$

Since the correlation matrix $\mathbf{R}_{xx}$ is Hermitian and thus normal, it verifies the Schur decomposition

$$\mathbf{R}_{xx} \quad = \quad \mathbf{U}\mathbf{\Lambda}\mathbf{U}^* \tag{3}$$

where $\mathbf{U}$ is the unitary matrix ($\mathbf{U}^*\mathbf{U} = \mathbf{I}$) whose columns are the N linearly independent eigenvectors of $\mathbf{R}_{xx}$ and $\mathbf{\Lambda}$ is the diagonal matrix with the associated non-negative eigenvalues. Assuming that $\mathbf{R}_{xx}$ is non-singular, the only analytic solution of decorrelation problem that preserves the orientation of the principal directions of the observations (see figure 1) is given by

$$\mathbf{B}_0 \quad = \quad \mathbf{R}_{xx}^{-1/2} = \mathbf{U}\mathbf{\Lambda}^{-1/2}\mathbf{U}^* \tag{4}$$

The inverse of the decorrelating system $\mathbf{A}_0 = \mathbf{B}_0^{-1} = \mathbf{R}_{xx}^{1/2}$ is the principal square root of the correlation matrix. Other possible decorrelating solutions like

$$\mathbf{B}_1 \quad = \quad \mathbf{Q}\mathbf{R}_{xx}^{-1/2}\,, \tag{5}$$

where $\mathbf{Q}$ is a unitary matrix different from the identity, will sphere, but also rotate the observations modifying their general orientation.

3. NEWTON DECORRELATION ALGORITHMS

For simplicity, we will consider first the problem of the decorrelation of real processes. In this section we will propose two different Newton algorithms for this task. The first one is obtained from a Newton approach in the space of the decorrelating systems while, the later one, originates from a similar approach in the space of the inverse decorrelating systems.

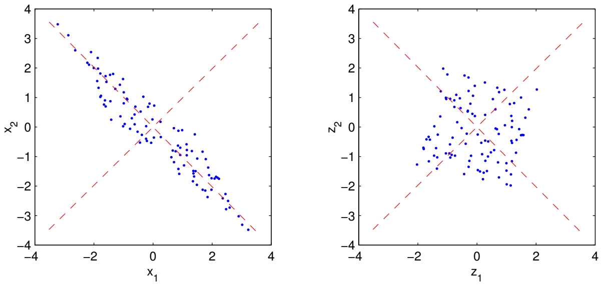

Fig. 1. Scatterplot of a sample set of observations before (left figure) and after (right figure) prewhitening by the decorrelation system of equation (4). Note that the transformation only scales the observations in the principal directions of the data (shown in dash lines) but does not rotate them.

3.1. A Newton algorithm in the space of the decorrelating systems

The sufficient statistic for decorrelation is the correlation matrix of the observations, thus, we needn't estimate any other parameters from the available data. A good criteria for modeling the observations is to assume the less biased model based on the given information, in our case, the correlation matrix. From the maximum entropy principle [10], we achieve this objective considering the process of the observations as Gaussian.

With the previous assumption, the likelihood of one sample vector of observations is easily obtained

$$p(\mathbf{x}[k]|\mathbf{B}) = \frac{1}{(2\pi)^{N/2}|\mathbf{R}_{xx}|^{1/2}}\, e^{-\frac{1}{2}(\mathbf{x}[k])'\mathbf{R}_{xx}^{-1}\mathbf{x}[k]} \tag{6}$$

being the dependence with the matrix $\mathbf{B}$ given through the tentative factorizations of the correlation matrix $\mathbf{R}_{xx} = \mathbf{A}\mathbf{A}'$ and its inverse $\mathbf{R}_{xx}^{-1} = \mathbf{B}\mathbf{B}'$. Note that this representation preserves the symmetry and non-negative definition of the correlation matrix. If, for simplicity, we further model the temporal sequence of observations as white, the normalized log-likelihood of the whole set of data $\{\mathbf{x}[0], \mathbf{x}[1], \ldots, \mathbf{x}[M-1]\}$ is given by

$$l(\mathbf{B}) \quad = \quad -\frac{1}{2}\mathrm{tr}\left\{\mathbf{B}\hat{\mathbf{R}}_{xx}\mathbf{B}'\right\} + \frac{1}{2}\ln|\mathbf{B}'\mathbf{B}| - \frac{N}{2}\ln(2\pi)$$

where

$$\hat{\mathbf{R}}_{xx} = \frac{1}{M}\sum_{k=0}^{M-1}\mathbf{x}[k](\mathbf{x}[k])' \tag{7}$$

denotes the sample estimate of the correlation matrix of the observations.

An iterative algorithm for the estimation of the decorrelating system, has the typical form

$$\mathbf{B}^{(n+1)} \quad = \quad \mathbf{B}^{(n)} + \eta\Delta_{\mathbf{B}}^{(n)} \tag{8}$$

The Newton update $\Delta_{\mathbf{B}}^{(n)}$ at the iteration n is obtained solving the following system of linear equations

$$-(\text{Hess}\ l|_{\mathbf{B}^{(n)}})\ \text{vec}(\Delta_{\mathbf{B}}^{(n)}) = \text{vec}\left(\left.\frac{\partial l(\mathbf{B})}{\partial \mathbf{B}}\right|_{\mathbf{B}^{(n)}}\right) \tag{9}$$

Defining $\mathbf{K}_N$ as the commutation matrix (which satisfies $\mathbf{K}_N \cdot \text{vec}\ \mathbf{A} = \text{vec}\ \mathbf{A}'$), after some matrix differential calculus the desired gradient and Hessian matrices are obtained

$$\begin{aligned} \frac{\partial l(\mathbf{B})}{\partial \mathbf{B}} &= -\mathbf{B}\,\hat{\mathbf{R}}_{xx} + \mathbf{A}' \qquad (10) \\ \text{Hess}\ l(\mathbf{B}) &= -(\hat{\mathbf{R}}_{xx} \otimes \mathbf{I}) - \mathbf{K}_N(\mathbf{A}' \otimes \mathbf{A}) \qquad (11) \end{aligned}$$

Substituting these results in equation (9) and solving for $\Delta_{\mathbf{B}}^{(n)}$ we obtain the 1st decorrelation algorithm[1]

$$\mathbf{B}^{(n+1)} = (1-\eta)\mathbf{B}^{(n)} + 2\eta\left(\mathbf{I} + \mathbf{B}^{(n)}\hat{\mathbf{R}}_{xx}(\mathbf{B}^{(n)})'\right)^{-1}\mathbf{B}^{(n)} \tag{12}$$

From a theoretical standpoint, other illustrative implementations of this algorithm can be obtained. Using a form of matrix inversion lemma we find the equivalent form

$$\begin{aligned} \mathbf{B}^{(n+1)} &= \mathbf{B}^{(n)} - \frac{\eta}{2}\left(\mathbf{I} + \frac{1}{2}\left(\mathbf{B}^{(n)}\hat{\mathbf{R}}_{xx}(\mathbf{B}^{(n)})' - \mathbf{I}\right)\right)^{-1} \\ &\quad \left(\mathbf{B}^{(n)}\hat{\mathbf{R}}_{xx}(\mathbf{B}^{(n)})' - \mathbf{I}\right)\mathbf{B}^{(n)} \qquad (13) \end{aligned}$$

Another implementation that will have some transcendence when establishing the connection with other algorithms, is obtained directly rewriting the 1st decorrelation algorithm of equation (12) using two nested iterations

$$\begin{aligned} \hat{\mathbf{A}}^{(n+1)} &= \frac{1}{2}\left(\mathbf{A}^{(n)} + \hat{\mathbf{R}}_{xx}(\mathbf{B}^{(n)})'\right) \qquad (14) \\ \mathbf{B}^{(n+1)} &= \mathbf{B}^{(n)} + \eta\left((\hat{\mathbf{A}}^{(n+1)})^{-1} - \mathbf{B}^{(n)}\right) \qquad (15) \end{aligned}$$

3.2. A Newton algorithm in the space of the inverse decorrelating systems

Let us consider now the Newton method in the space of the mixing system $\mathbf{A}$. We propose to obtain the square root factorization of $\hat{\mathbf{R}}_{xx}$ finding the zeros of the function

$$\mathbf{F}(\mathbf{A}) = (\mathbf{A})'\hat{\mathbf{R}}_{xx}^{-1}\mathbf{A} - \mathbf{I} = \mathbf{0}\ , \tag{16}$$

The Newton-Raphson update

$$\text{vec}\Delta_{\mathbf{A}} = -\left(\frac{\partial \text{vec}\mathbf{F}(\mathbf{A})}{\partial(\text{vec}\mathbf{A})'}\right)^{-1}\text{vec}\mathbf{F}(\mathbf{A}) \tag{17}$$

is expressed in terms of the Jacobian of $\mathbf{F}(\mathbf{A})$, which is given by

$$\frac{\partial \text{vec}\mathbf{F}(\mathbf{A})}{\partial(\text{vec}\mathbf{A})'} = (\mathbf{K}_N + \mathbf{I}_{N^2})(\mathbf{I} \otimes (\mathbf{A})'\hat{\mathbf{R}}_{xx}^{-1})\ . \tag{18}$$

Thus, we obtain the 2nd decorrelation algorithm as

$$\begin{aligned} \mathbf{A}^{(n+1)} &= \mathbf{A}^{(n)} + \Delta_{\mathbf{A}}^{(n)} \\ &= \mathbf{A}^{(n)} + \frac{\eta}{2}\,(\hat{\mathbf{R}}_{xx}(\mathbf{B}^{(n)})' - \mathbf{A}^{(n)}) \qquad (19) \\ \mathbf{B}^{(n+1)} &= (\mathbf{A}^{(n+1)})^{-1}\ , \qquad (20) \end{aligned}$$

Other alternative forms of expressing this algorithm are

$$\mathbf{B}^{(n+1)} = \left(\mathbf{I} + \frac{\eta}{2}\left(\mathbf{B}^{(n)}\hat{\mathbf{R}}_{xx}(\mathbf{B}^{(n)})' - \mathbf{I}\right)\right)^{-1}\mathbf{B}^{(n)} \tag{21}$$

and

$$\begin{aligned} \mathbf{B}^{(n+1)} &= \mathbf{B}^{(n)} - \frac{\eta}{2}\left(\mathbf{I} + \frac{\eta}{2}\left(\mathbf{B}^{(n)}\hat{\mathbf{R}}_{xx}(\mathbf{B}^{(n)})' - \mathbf{I}\right)\right)^{-1} \\ &\quad \left(\mathbf{B}^{(n)}\hat{\mathbf{R}}_{xx}(\mathbf{B}^{(n)})' - \mathbf{I}\right)\mathbf{B}^{(n)} \qquad (22) \end{aligned}$$

It is interesting to observe that, regardless of the fact that the 1st and 2nd decorrelation algorithms has been obtained in different spaces, when the optimal learning step-size $\eta = 1$ is chosen, both algorithms share the same adaptation

$$\mathbf{B}^{(n+1)} = 2(\mathbf{I} + \mathbf{B}^{(n)}\hat{\mathbf{R}}_{xx}(\mathbf{B}^{(n)})')^{-1}\mathbf{B}^{(n)}\ . \tag{23}$$

3.3. Algorithms for complex processes

In the previous section several forms of a Newton decorrelation algorithm has been obtained for real processes, however, complex processes are ubiquitous in *Signal Processing*, for instance, when working in the frequency domain.

The derivation of the Newton algorithms for the complex case needs a cumbersome notation and thus is omitted here, but the net result is that the complex algorithms only differ from their real versions in that one have to replace all the transpose operators $(\cdot)'$ by the conjugate transpose operator $(\cdot)^*$, also extending this change to the definition (7) of the sample correlation matrix.

3.4. Dual forms of the algorithms

Dual forms of the algorithms can be obtained by using inverse relations where the role of the decorrelation system $\mathbf{B}$ and the inverse system $\mathbf{A}$ is interchanged. A summary of these algorithms is presented in Table 1.

4. GLOBAL CONVERGENCE

This section we studies the local and global convergence behaviour of the proposed Newton approach when the optimal step size ($\eta = 1$) is chosen. A global convergence result is challenging since, in general, Newton algorithms do not possess this property.

For the purpose of the analysis, we define the following transformation $\mathbf{G}^{(n)} = \mathbf{B}^{(n)}\hat{\mathbf{R}}_{xx}^{1/2}$ that maps the sample

[1] Note that the optimal Newton step size is $\eta = 1$.

estimate of the decorrelation solution $\mathbf{B}_1 = \mathbf{Q}\hat{\mathbf{R}}_{xx}^{-1/2}$ of equation (5) to a global orthogonal matrix $\mathbf{Q}$. Note that for a number of observations M sufficient long, by the law of large numbers, the sample estimate of the correlation matrix $\hat{\mathbf{R}}_{xx}^{(n)}$ will converge in probability to their expectation $\mathbf{R}_{xx}$.

The Newton algorithm of equation (23) can be rewritten in terms of the matrix parameter $\mathbf{G}^{(n)}$ as

$$\mathbf{G}^{(n+1)} \quad = \quad 2(\mathbf{I} + \mathbf{G}^{(n)}(\mathbf{G}^{(n)})^*)^{-1}\mathbf{G}^{(n)} \qquad (24)$$

Using the singular value decomposition one can factor the global matrix as $\mathbf{G}^{(n)} = \mathbf{U}\mathbf{\Lambda}^{(n)}(\mathbf{V})^*$, where $\mathbf{U}$ and $\mathbf{V}$ are unitary matrices and $\mathbf{\Lambda}^{(n)}$ is the diagonal matrix of singular values. This simplifies iteration (24) as follows

$$\begin{aligned} \mathbf{G}^{(n+1)} &= \mathbf{U}^{(n)}\mathbf{\Lambda}^{(n+1)}(\mathbf{V}^{(n)})^* && (25) \\ \mathbf{\Lambda}^{(n+1)} &= 2\left(\mathbf{I} + \mathbf{\Lambda}^{(n)}(\mathbf{\Lambda}^{(n)})^*\right)^{-1}\mathbf{\Lambda}^{(n)} && (26) \end{aligned}$$

From these equations one observes that the left and right eigenvectors of the global matrix are preserved through iterations while the modulo of the singular values $\lambda_i^{(n)} = [\mathbf{\Lambda}^{(n)}]_{ii}$ evolve independently according to the recursion

$$|\lambda_i^{(n+1)}| \quad = \quad \frac{2|\lambda_i^{(n)}|}{1 + |\lambda_i^{(n)}|^2} \qquad i = 1, \ldots, N. \quad (27)$$

The previous expression can be identified with the Newton-Raphson method to find the only positive root $\lambda = 1$ of the function $F(\lambda) = \lambda - \lambda^{-1}$. This Newton algorithm has global quadratic convergence and a good local convergence rate of $1/2$. In order to prove the global convergence result we introduce the following theorem whose proof can be found in [11].

Theorem 1 *Given a twice differentiable continuous function $F(\lambda)$ in an interval $[a, b]$, if the following conditions are satisfied*

$$\begin{array}{rl} (i) & F(a)F(b) < 0 \\ (ii) & \frac{\partial F(\lambda)}{\partial \lambda} \neq 0\ ,\ \forall\lambda \in [a, b] \\ (iii) & \frac{\partial^2 F(\lambda)}{\partial \lambda^2} \geq 0 \; \mathit{or} \; \leq 0\ ,\ \forall\lambda \in [a, b] \\ (iv) & \left|\frac{F(c)}{\frac{\partial F(c)}{\partial c}}\right| \leq b - a \\ & \mathit{where}\ \ c = \mathit{arg}\left\{\min_{\lambda=\{a,b\}}\left|\frac{\partial F(\lambda)}{\partial \lambda}\right|\right\} \end{array} \qquad (28)$$

then the Newton-Raphson algorithm converges to the unique solution of $F(\lambda) = 0$ for any arbitrary initialization in the interval $[a, b]$.

In our case, by definition, $|\lambda| > 0$ and the theorem can be applied to the twice continuous differentiable function $F(|\lambda|) = |\lambda| - |\lambda|^{-1}$ in an interval $[a, a^{-1}]$ for any sufficient small a that verifies $min\{1, |\lambda_1|, \ldots, |\lambda_N|\} > a > 0$. It can be checked that all the four conditions hold true and, therefore, the proposed Newton-Rapshon algorithm is globally convergent. Then, at a particular given iteration n_0, all the iterations in (27) will have converged to $|\lambda| = 1$, value which, substituted in (26), implies that one obtains the decorrelation solution

$$\mathbf{G}^{(n_0)}(\mathbf{G}^{(n_0)})^* = (\mathbf{B}^{(n_0)})\hat{\mathbf{R}}_{xx}(\mathbf{B}^{(n_0)})^* = \mathbf{I} \qquad (29)$$

5. RELATIONS WITH EXISTING APPROACHES

5.1. Natural gradient

Silva and Almeida were the first to propose in [6] a distributed technique for the symmetric orthogonalization of the observations. The same technique was independently justified in [7], in the context of blind source separation, and later identified with the natural gradient algorithm for blind decorrelation (see Chapter 4 of [1] and references therein)

$$\begin{aligned} \mathbf{B}^{(n+1)} &= \mathbf{B}^{(n)} + \frac{\partial l(\mathbf{B})}{\partial \mathbf{B}}\mathbf{B}^T\mathbf{B} && (30) \\ &= \mathbf{B}^{(n)} - \mu\,(\mathbf{B}^{(n)}\hat{\mathbf{R}}_{xx}^{(n)}(\mathbf{B}^{(n)})^* - \mathbf{I})\mathbf{B}^{(n)}\ . \end{aligned}$$

Let us rewrite here, for convenience, the 1st decorrelation algorithm

$$\begin{aligned} \hat{\mathbf{A}}^{(n+1)} &= \frac{1}{2}\left(\mathbf{A}^{(n)} + \hat{\mathbf{R}}_{xx}^{(n)}(\mathbf{B}^{(n)})'\right) && (31) \\ \mathbf{B}^{(n+1)} &= \mathbf{B}^{(n)} + \eta\left((\hat{\mathbf{A}}^{(n+1)})^{-1} - \mathbf{B}^{(n)}\right) && (32) \end{aligned}$$

and the 2nd decorrelation algorithm

$$\begin{aligned} \mathbf{A}^{(n+1)} &= \mathbf{A}^{(n)} + \frac{\eta}{2}\,(\hat{\mathbf{R}}_{xx}^{(n)}(\mathbf{B}^{(n)})' - \mathbf{A}^{(n)}) && (33) \\ \mathbf{B}^{(n+1)} &= (\mathbf{A}^{(n+1)})^{-1}\ , && (34) \end{aligned}$$

One can observe that both algorithms need to compute the inversion of a matrix $\mathbf{M}^{(n+1)}$, being $\mathbf{M}^{(n+1)} = \mathbf{A}^{(n+1)}$ in the first case and $\mathbf{M}^{(n+1)} = \hat{\mathbf{A}}^{(n+1)}$ for the second one. This matrix inversion can be avoided replacing it by an estimate resulting from an iterative improvement algorithm [12] and, thus, obtaining an approximate algorithm. When $\mathbf{B}^{(n)}$ is not so far away from the desired inverse so that the following condition is satisfied

$$\|\mathbf{B}^{(n)}\mathbf{M}^{(n+1)} - \mathbf{I}\|_2 < 1\ , \qquad (35)$$

one can replace the exact inversion of $\mathbf{M}^{(n+1)}$ by the result of the following Newton algorithm for iterative matrix inversion

$$(\mathbf{M}^{(n+1)})^{-1} \quad \approx \quad 2\mathbf{B}^{(n)} - \mathbf{B}^{(n)}\mathbf{M}^{(n+1)}\mathbf{B}^{(n)} \qquad (36)$$

Substituting (36) in (32) and in (34), we obtain, in both cases, the natural gradient algorithm for blind decorrelation

$$\mathbf{B}^{(n+1)} \quad = \quad \mathbf{B}^{(n)} - \frac{\eta^{(n)}}{2}\left(\mathbf{B}^{(n)}\hat{\mathbf{R}}_{xx}^{(n)}(\mathbf{B}^{(n)})^* - \mathbf{I}\right)\mathbf{B}^{(n)}$$

Defining the update $\Delta_{NG} = \frac{1}{2}\left(\mathbf{B}^{(n)}\hat{\mathbf{R}}_{xx}^{(n)}(\mathbf{B}^{(n)})^* - \mathbf{I}\right)\mathbf{B}^{(n)}$, the optimal value for $\eta^{(n)}$ in the sense of maximizing the log-likelihood $l(\mathbf{B})$ up to the second order approximation is

$$\begin{aligned}\eta_{opt}^{(n)} &= \frac{tr\{\Delta_{NG}^* \frac{\partial l(\mathbf{B})}{\partial \mathbf{B}}\}}{(\text{vec}\Delta_{NG})^* \text{ Hess } l(\mathbf{B}) \text{ vec}\Delta_{NG}} \\ &= \frac{2\|\hat{\mathbf{R}}_{zz}^{(n)} - \mathbf{I}\|_F^2}{\|\hat{\mathbf{R}}_{zz}^{(n)} - \mathbf{I}\|_F^2 + \|\hat{\mathbf{R}}_{zz}^{1/2(n)}(\hat{\mathbf{R}}_{zz}^{(n)} - \mathbf{I})\|_F^2} \quad (37) \\ &\geq \frac{2}{1 + \|\hat{\mathbf{R}}_{zz}^{(n)}\|_p} \quad \forall p \in \mathbb{N} \quad (38)\end{aligned}$$

where the last lower bound holds true for any arbitrary p-norm, although, one can use those simpler to evaluate such as the 1-norm or ∞-norm. It is easy to check that (38) is a good and simple proposal for the step-size because it satisfies condition (35) and also approaches to the optimal Newton step-size as the algorithm converges.

5.1.1. Stochastic-Newton decorrelation algorithm

Using the instantaneous estimate in the sample correlation matrix $\hat{\mathbf{R}}_{xx}^{(n)} = \mathbf{x}[n]\mathbf{x}^*[n]$ and setting an small step-size $\eta \ll 1$, one obtains the stochastic form of the 2^{nd} decorrelation algorithm

$$\begin{aligned}\mathbf{A}^{(n+1)} &= (1 - \frac{\eta}{2})\mathbf{A}^{(n)} + \frac{\eta}{2}\mathbf{x}[n]\mathbf{z}^*[n] \quad (39) \\ \mathbf{B}^{(n+1)} &= (\mathbf{A}^{(n+1)})^{-1} \quad (40)\end{aligned}$$

Defining $\alpha = \eta/(2 - \eta)$ and using the Sherman-Morrison formula [12] we can avoid the explicit inversion

$$\begin{aligned}\mathbf{B}^{(n+1)} &= \mathbf{B}^{(n)} - \frac{\alpha}{1 + \alpha\|\mathbf{y}[n]\|_2^2}\left(\mathbf{y}[n]\mathbf{y}^*[n] - \mathbf{I}\right)\mathbf{B}^{(n)} \\ &\quad - \frac{\alpha^2}{1 + \alpha\|\mathbf{y}[n]\|_2^2}\left(\mathbf{y}^*[n]\mathbf{y}[n] - \|\mathbf{y}[n]\|_2^2\mathbf{I}\right)\mathbf{B}^{(n)}\end{aligned}$$

By assumption $\eta \ll 1$, then $\alpha^2 \ll \alpha$ and the algorithm is dominated by the linear part. This part of the stochastic-Newton algorithm coincides with the stochastic implementation of the natural gradient algorithm for decorrelation but, what is more interesting, it also provides the form of the learning step-size one should use

$$\mu^{(n)} = \frac{\eta^{(n)}}{2} = \frac{\alpha}{1 + \alpha\|\mathbf{y}[n]\|_2^2} \quad (41)$$

This normalized step-size, originally suggested by Cardoso as a good empirical choice for a fast convergence, has been now justified from the theoretical standpoint.

5.2. Goodall algorithm for decorrelation

The Goodall algorithm has been first proposed in [4] and later justified in [5] as gradient-like method to maximize $l(\mathbf{B})$. The algorithm was shown to be globally convergent for sufficient small learning step-sizes. It consists in updating the inverse decorrelation system according to

$$\hat{\mathbf{A}}^{(n+1)} = (1 - \mu_1)\hat{\mathbf{A}}^{(n)} + \mu_1\mathbf{x}[n]\mathbf{z}^*[n] \quad (42)$$

and the outputs according to

$$\mathbf{z}^{(n+1)} = \mu_2\mathbf{x}[n] + (\mathbf{I} - \mu_2\hat{\mathbf{A}}^{(n+1)})\mathbf{z}^{(n)} \quad (43)$$

After looking at (42) it is easy to show that the Goodall algorithm is an approximation to the stochastic form of the Newton decorrelation algorithms, where the inversions have been avoided by means of the following trick. Rewriting (43) to obtain the explicit form of the iteration in terms of the decorrelation system

$$\mathbf{B}^{(n+1)} = \mathbf{B}^{(n)} + \mu_2\hat{\mathbf{A}}^{(n+1)}\left((\hat{\mathbf{A}}^{(n+1)})^{-1} - \mathbf{B}^{(n)}\right) \quad (44)$$

one can see that it converges to the inverse of $\hat{\mathbf{A}}$ whenever condition $\|\mu_2\hat{\mathbf{A}} - \mathbf{I}\| < 1$ be true. This implies that $\hat{\mathbf{A}}$ should be positive definite and, therefore, one can regard (44) as a preconditioned form of the update in (15) or as an approximation to (20).

6. SIMULATIONS

Let us define the following experiment. Given five sensors that measure one thousand spatially correlated data we compute the sample correlation matrix $\hat{\mathbf{R}}_{xx}$ and use the proposed algorithms to find the inverse of the principal square root $\hat{\mathbf{R}}_{xx}^{1/2}$. Figure 2 shows, for 50 random experiments, the median Frobenious norm of the difference between the separation system $\mathbf{B}^{(n)}$, at iteration n, and the desired solution. From the figure, we can observe that the convergence of the proposed Newton algorithms is quadratic as well as the local convergence of the natural gradient algorithm with the proposed step-sizes in equations (37) and (38). The figure shows that the best convergence is for the Newton algorithms with step-size $\eta = 1$, however, the natural gradient algorithm with a properly chosen step-size has a convergence that approaches closely to that of the Newton algorithm. This experiment confirms the previously explained connection between both algorithms.

7. CONCLUSIONS

We have proposed novel Newton algorithms for blind decorrelation whose global convergence is guaranteed. Adaptive and batch implementations of the algorithm have been provided. Two classical blind decorrelation algorithms apparently unrelated (the natural gradient for decorrelation and the Goodall algorithms) have been shown to approximate these Newton algorithms through different simplification of

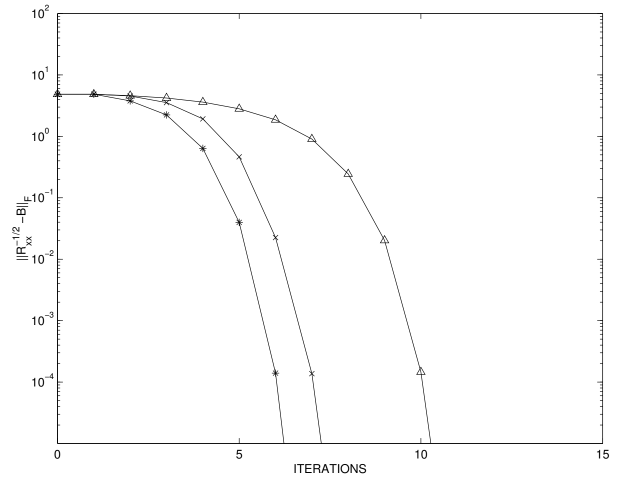

Fig. 2. Median of the performance index $\|\hat{\mathbf{R}}_{xx}^{-1/2} - \mathbf{B}\|_F$ versus iterations: symbol '*' denotes the 1st and 2nd Newton algorithms, '$^\times$' the natural gradient algorithm with step-size η_{opt} of equation (37) and '$^\triangle$' the same algorithm with the step-size of equation (38) is used.

a matrix inversion. This framework provides theoretical justification for the normalized learning step-sizes that guarantee and speed up the convergence of these algorithms.

8. REFERENCES

[1] A. Cichocki, S.-i. Amari, *Adaptive Blind Signal and Image Processing,* John Wiley & Sons, 2002.

[2] S. Amari and A. Cichocki, "Adaptive blind signal processing – neural network approaches," *Proceedings of the IEEE*, vol. 86, no. 10, pp. 2026–2048, 1998.

[3] K.L. Diamantaras, S.Y. Kung, *Principal Component Neural Networks,* John Wiley & Sons, 1996.

[4] M.C. Goodall, "Performance of stochastic net," *Nature (London)*, vol. 185, pp. 557–558, 1960.

[5] J.J. Attick, A.N. Redlich, "Convergent algorithm for sensory receptive field development," *Neural Computation*, vol. 5, pp. 45–60, 1993.

[6] L.B. Almeida, F.M. Silva, "Adaptive Decorrelation," *Artificial Neural Networks*, vol. 2, pp. 149–156, 1992.

[7] A. Cichocki, R. Unbehauen "Robust neural networks with on-line learning for blind identification and blind separation of sources," *IEEE Trans. on Circuits and Systems-I*, vol. 43(11), pp. 894–906, 1996.

[8] S.C. Douglas, A. Cichocki, "Neural Networks for blind decorrelation of signals," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2829–2842, 1997.

Table 1. Summary of the blind decorrelation algorithms.

Batch implementation: $\hat{\mathbf{R}}_{xx}^{(n)} = \hat{\mathbf{R}}_{xx}$
Adaptive " : $\gamma^{(n)} = \frac{n-1}{n}$ or $\gamma^{(n)} = \gamma$, $0 < 1-\gamma \ll 1$

$$\hat{\mathbf{R}}_{xx}^{(n)} = \gamma^{(n)}\hat{\mathbf{R}}_{xx}^{(n-1)} + (1-\gamma^{(n)})\mathbf{x}[n](\mathbf{x}[n])^*$$

$$(\hat{\mathbf{R}}_{xx}^{(n)})^{-1} = \frac{1}{\gamma^{(n)}}\left((\hat{\mathbf{R}}_{xx}^{(n-1)})^{-1} - \frac{(\hat{\mathbf{R}}_{xx}^{(n-1)})^{-1}\mathbf{x}[n](\mathbf{x}[n])^*(\hat{\mathbf{R}}_{xx}^{(n-1)})^{-1}}{\frac{\gamma^{(n)}}{1-\gamma^{(n)}}+(\mathbf{x}[n])^*(\hat{\mathbf{R}}_{xx}^{(n-1)})^{-1}\mathbf{x}[n]}\right)$$

1st Newton algorithm:

$$\mathbf{B}^{(n+1)} = (1-\eta)\mathbf{B}^{(n)} + 2\eta\left(\mathbf{I} + \mathbf{B}^{(n)}\hat{\mathbf{R}}_{xx}^{(n)}(\mathbf{B}^{(n)})^*\right)^{-1}\mathbf{B}^{(n)}$$

2nd Newton algorithm:

$$\mathbf{B}^{(n+1)} = (\mathbf{A}^{(n+1)})^{-1}\ ,\quad \mathbf{A}^{(n+1)} = (1-\tfrac{\eta}{2})\mathbf{A}^{(n)} + \tfrac{\eta}{2}\hat{\mathbf{R}}_{xx}^{(n)}(\mathbf{B}^{(n)})^*$$

Natural Gradient algorithm [6, 1]:

$$\mathbf{B}^{(n+1)} = \mathbf{B}^{(n)} - \mu\left(\mathbf{B}^{(n)}\hat{\mathbf{R}}_{xx}^{(n)}(\mathbf{B}^{(n)})^* - \mathbf{I}\right)\mathbf{B}^{(n)}$$

Dual of the 1st Newton algorithm:

$$\mathbf{A}^{(n+1)} = (1-\eta)\mathbf{A}^{(n)} + 2\eta\mathbf{A}^{(n)}\left(\mathbf{I} + (\mathbf{A}^{(n)})^*(\hat{\mathbf{R}}_{xx}^{(n)})^{-1}\mathbf{A}^{(n)}\right)^{-1}$$

Dual of the 2nd Newton algorithm:

$$\mathbf{A}^{(n+1)} = (\mathbf{B}^{(n+1)})^{-1}\ ,\quad \mathbf{B}^{(n+1)} = (1-\tfrac{\eta}{2})\mathbf{B}^{(n)} + \tfrac{\eta}{2}(\mathbf{A}^{(n)})^*(\hat{\mathbf{R}}_{xx}^{(n)})^{-1}$$

Dual of Natural Gradient algorithm:

$$\mathbf{A}^{(n+1)} = \mathbf{A}^{(n)} - \mu\mathbf{A}^{(n)}\left((\mathbf{A}^{(n)})^*(\hat{\mathbf{R}}_{xx}^{(n)})^{-1}\mathbf{A}^{(n)} - \mathbf{I}\right)$$

Goodall algorithm [4, 5]:

$$\mathbf{A}^{(n+1)} = (1-\mu_1)\mathbf{A}^{(n)} + \mu_1\ \ \mathbf{x}[n]\mathbf{z}^*[n]$$
$$\mathbf{z}[n+1] = \mu_2\mathbf{x}[n] + (\mathbf{I} - \mu_2\mathbf{A}^{(n+1)})\mathbf{z}[n]$$

[9] S.C. Douglas, "Numerically-robust $\mathcal{O}(N^2)$ RLS algorithms using Least-Squares prewhitening," *in Proceedings of ICASSP*, Istanbul, Turkey, pp. 412–415, 2000.

[10] E.T. Jaynes, "Information Theory and Statistical Mechanics", *The Physical Review*, vol. 106(4), pp. 620-630, 1957.

[11] P. Henrici, Elements of numerical analysis. Wiley, New York, 1964.

[12] W.H. Press, S.A. Teukosly, W.T. Vetterling, B.P. Flanney, *Numerical Recipes*, Second ed., Cambridge Press, pp. 55–58, 1992.

[13] S. Cruces, A. Cichocki, L. Castedo, "An iterative inversion approach to blind source separation," *IEEE Trans. on Neural Networks*, vol. 11(6), pp. 1423–1437, Nov. 2000.

[14] S. Cruces, A. Cichocki, and L. Castedo, "Blind source extraction in Gaussian noise," *proc. of the 2nd ICA and BSS workshop, Helsinki, Finland*, pp. 63–68, 2000.