LETTER

An Extension to the Natural Gradient Algorithm for Robust Independent Component Analysis in the Presence of Outliers*

Muhammad TUFAIL[a], Nonmember, Masahide ABE[†], and Masayuki KAWAMATA[‡], Members

SUMMARY In this paper, we propose to employ an extension to the natural gradient algorithm for robust Independent Component Analysis against outliers. The standard natural gradient algorithm does not exhibit this property since it employs nonrobust sample estimates for computing higher order moments. In order to overcome this drawback, we propose to use robust alternatives to higher order moments, which are comparatively less sensitive to outliers in the observed data. Some computer simulations are presented to show that the proposed method, as compared to the standard natural gradient algorithm, gives better performance in the presence of outlying data.

key words: Independent Component Analysis, robust higher order statistics, outliers

1. Introduction

In the simplest Blind Source Separation (BSS) problem, we observe \( n \) signals \( x(k) = [x_1(k), x_2(k), \ldots, x_n(k)]^T \) that are generated by the following multi-input/multi-output system

\[
    x(k) = As(k),
\]

where \( A \) is an \( n \times n \) nonsingular mixing matrix, and \( s(k) = [s_1(k), s_2(k), \ldots, s_n(k)]^T \) is the vector of source signals. The objective of BSS is to determine a separating matrix \( W = A^{-\dagger} \) such that the \( n \) dimensional transformed vector given by

\[
    y(k) = Wx(k) = WAs(k) = Hs(k),
\]

estimates the original input vector, given only the observed data, and certain assumptions about the statistics of sources. Here, \( A^{-\dagger} \) denotes an estimate of \( A^{-1} \), and \( H = WA \) represents the global transformation matrix from \( s(k) \) to \( y(k) \).

It may be noted that because of certain inherent ambiguities associated with BSS problem, source signals and mixing matrix can only be recovered up to scaling and permutation. This implies that, in general, an optimal separating matrix \( W_{opt} \) satisfies the following relation

\[
    H_{opt} = W_{opt}A = PD,
\]

where \( P \) is a permutation matrix, and \( D \) is an arbitrary diagonal matrix.

2. Independent Component Analysis

Independent component analysis (ICA) [1] is a statistical method for transforming an arbitrary random vector into another vector with (maximally) statistically independent components. In case of the observation model given by Eq. (1), it solves the BSS problem by only assuming statistical independence between the components of \( s(k) \). Since independence is much stronger condition than decorrelation, higher order statistics are employed in order to achieve BSS using ICA. Furthermore, because a Gaussian signal is completely characterized by its first and second order statistics, ICA is only valid for non-Gaussian signals. More specifically, at most one source is allowed to be Gaussian.

Various approaches for ICA, such as entropy maximization [2], maximum likelihood estimation [3], and negentropy maximization [4], obtain the separating matrix as a solution to an estimating equation

\[
    \frac{1}{N} \sum_{k=1}^{N} F(y(k), W) = 0, \tag{4}
\]

where \( F \) is an appropriate (matrix-valued) estimating function, and \( N \) is the total number of available realizations of \( x(k) \). In particular, the natural gradient learning rule [3], given by

\[
    \Delta W(l) = \mu(I_n - \langle \varphi(y)y^T \rangle)W(l), \tag{5}
\]

is associated to a class of estimating functions of the form

\[
    F(y, W) = (I_n - \varphi(y)y^T)W, \tag{6}
\]

where \( \varphi(y) = [\varphi_1(y_1), \ldots, \varphi_n(y_n)] \) is a vector of nonlinear activation functions, \( \mu \) is a step size parameter, \( \langle \cdot \rangle \) denotes time average, and \( l \) is the iteration number. The best choice of \( \varphi_i(y_i) \) is given by

\[
    \varphi_i(y_i) = \frac{q_i(y_i)}{q_i}, \tag{7}
\]

where \( q_i \) is the probability density function (pdf) of the source \( s_i \). With these activation functions, we obtain the maximum likelihood estimate of \( W_{opt} \). However, in many cases we do not have exact knowledge about the pdfs of sources. In these situations, we can still obtain a consistent estimate of \( W_{opt} \) by employing nonlinearities other than those given by Eq. (7). However, these nonlinear activation
functions have to be chosen to ensure that the true solution is a locally stable equilibrium point of the natural gradient algorithm [5].

Based on a local stability analysis of the natural gradient algorithm, we can use the following monomial functions [5]

$$\varphi_i(y_i) = |y_i|^p \text{sign}(y_i),$$  

(8)

in order to separate signals with a wide class of symmetric pdfs. In the above equation, $p$ is an adequately chosen parameter depending on the pdfs of sources. In particular, it can be easily shown that a cubic function ($p = 3$) for sub-Gaussian signals, and a sign function ($p = 0$) for super-Gaussian signals satisfy the local stability conditions of the natural gradient algorithm.

3. Robustness

In practical situations, it is important to employ a robust estimator/method which is not greatly influenced by potential outliers. Robustness of an M-estimator, like the one defined by Eq. (4), can be analyzed using the concept of influence function [6]. The influence function describes the effect of a single observation on the statistical estimation method. For a robust estimator, we require the corresponding influence function to be bounded. This ensures that even an extreme outlier has a bounded effect on the estimate. Such an estimator is called B-robust [7].

For the natural gradient algorithm (or any M-estimator), the influence function is proportional to the associated estimation function i.e.

$$\text{IF}(y, W) \propto (I_n - \varphi(y)y^T)W. \quad (9)$$

This implies that B-robustness (of an M-estimator) is equivalent to boundedness of the estimating function or its individual terms $\varphi_i(y_i)$ for $i = 1, \ldots, n$. It can easily be seen that the off diagonal terms $\varphi_i(y_i)y_j$ for $i \neq j$ cannot be made bounded (for arbitrary $y_i$ and $y_j$) irrespective of which activation functions are chosen [7]. Such a nonrobustness can also be observed by considering the batch natural gradient learning rule given by Eq. (5). In this equation, the term $\langle \varphi(y)y^T \rangle$ corresponds to sample estimates of some nonlinear moments. These sample estimates, however, are shown to be unstable in the presence of outliers.

The purpose of this paper is to employ an extension to the natural gradient algorithm for a robust ICA of an observed random vector. In the proposed method, a decaying factor is introduced so as to down-weight the effect of outliers. In the batch mode, the proposed approach is equivalent to using robust alternatives to higher order moments [8], which are relatively insensitive to outlying data.

4. The Proposed Approach

In order to perform a robust ICA, we consider the following modified natural gradient learning rule for $W$

$$\Delta W(l) = \mu((D_1(y)D_2(y) - (D_1(y)\varphi(y))) (D_2(y)y^T))W(l), \quad (10)$$

where $D_1(y)$ and $D_2(y)$ are two positive definite diagonal matrices. The above update rule can be derived from a more general natural gradient algorithm given by [9]

$$\Delta W(l) = (\bar{\mu}(y)(W(l))^T - \varphi(y)y^T)G(W(l), y)^T G(W(l), y), \quad (11)$$

where $G$ is an arbitrary (nonsingular) matrix–value function, and $\bar{\mu}$ is a symmetric positive definite matrix. The learning rule described by Eq. (10) can now be obtained by making the following substitutions

$$\bar{\mu}(y) = \mu D_1(y), \quad G(W, y) = D_2(y)^{1/2}W. \quad (12)$$

From Eq. (10), we note that the matrices $D_1(y)$ and $D_2(y)$ act as scaling matrices for vectors $\varphi(y)$ and $y$, respectively. Consequently, a (B-) robust algorithm can be obtained if we choose these scaling matrices such that the (components of) modified vectors $D_1(y)\varphi(y)$ and $D_2(y)y$ always remain bounded. Moreover, we also prefer to have an algorithm that almost completely rejects extreme outliers. In order to do so, we employ the following scaling matrices

$$D_1(y) = \text{diag}(d_1(y_1), \ldots, d_n(y_n)), \quad (14)$$

$$D_2(y) = \text{diag}(d_2(y_1), \ldots, d_2(y_n)), \quad (15)$$

where $d_1(y_i)$ and $d_2(y_i)$ are given by

$$d_2(y_i) = \frac{\alpha p(y_i) \beta_0}{p(y_i) \beta_0} = \alpha \exp \left( - (a^2 - 1) \frac{|y_i|}{\bar{\sigma}_y \beta(\theta)} \right), \quad (16)$$

$$d_1(y_i) = d_2(y_i) \theta \beta(\theta) = \sqrt{1/\theta} \Gamma(1/\theta) \Gamma(3/\theta). \quad (17)$$

In the above equations, $p(y_i) \beta(\theta) \beta(\theta) = (2\sigma_y \beta(\theta) \Gamma(1 + 1/\theta)^{-1} \exp \left( - \frac{|y_i|}{\bar{\sigma}_y \beta(\theta)} \right). \quad (18)$$

is the pdf of Generalized Gaussian Distribution (GGD) [10], $\alpha$ and $\theta$ are two positive parameters, and $\bar{\sigma}_y$ is an estimate of the standard deviation of $y_i$. With these scaling matrices, the learning rule given by Eq. (10) can be written as

$$\Delta W(l) = (\Lambda(y) - \langle \varphi(y)y^T \rangle)W(l), \quad (19)$$

where $\bar{y} = D_2(y)y$ and $\Lambda(y) = D_2(y)^{p+1}$. It can be easily seen that the above update rule reduces to standard natural gradient learning if we choose $\alpha = 1$. However, for $\alpha > 1$, the exponential decaying factor minimizes the effect of outliers (on the above algorithm) by giving less weights to observations with greater deviations from the mean. In particular, if $\theta = 2$, the term $\langle \varphi(y)y^T \rangle$ in the above algorithm corresponds to samples estimates of robust higher order statistics.
proposed in [8]. In this paper, however, we also tune the parameter $\theta$ so as to acquire a better separation performance in the presence of potential outliers.

Finally, it may be noted that by choosing the scaling matrices as given by Eq. (16) and Eq. (17), we inherently assume that the pdfs of sources are clustered around their means. This is due to the fact that the decaying term given by Eq. (16) is a monotonically decreasing function of $|y_i|$, i.e., the signal amplitudes with greater distances from the mean are (comparatively) given less weights.

4.1 Choice of $\theta$ and $\alpha$

Figure 1(a) plots the decaying function given by Eq. (16) for some values of $\theta$; $\alpha$ is fixed to 1.1. From this figure, we see that as $\theta$ is increased, the decaying factor becomes flatter around zero (mean), while it decreases sharply for relatively larger values of signal amplitudes. This implies that for larger values of $\theta$, the decaying function more effectively attenuates the observations lying far away from the rest of data. As pointed out earlier, a cubic function is generally employed as an activation function in order to separate signals with flat (sub-Gaussian) distributions. However, such a function is highly non-robust against outliers. Consequently, only a few far lying observations can make the algorithm unstable. It is therefore necessary to select a relatively larger value of $\theta$ for these signals (as compared to super-Gaussian signals that employ a more robust sign function as an activation function) so as to better diminish the effect of outliers. In particular, we propose to employ $\theta = 4$ for sub-Gaussian signals, and $\theta = 2$ for super-Gaussian signals.

With $\theta$ fixed as described above, the parameter $\alpha$ controls the spread of the decaying function around zero (mean) (see Fig. 1(b)). In general, we require a relatively larger value of $\alpha$ as the number of outliers in the observed vector increase. An optimal value of $\alpha$ depends on the (unknown) level of erroneous outlying data present in the observed vector, and must be tuned empirically. However, we cannot make $\alpha$ too large since this will result in loss of statistical efficiency when there are no outliers in the data. Nevertheless, we have found through computer simulations that a value of $\alpha$ between 1 and 1.1 can be chosen so as to obtain a good separation performance in the presence of outliers.

4.2 Estimation of Mean and Standard Deviation

In the above formulation, we assume that the observed vector has zero mean. Therefore, as a first step, we replace each observed signal $x_i$ by $x_i - v_k$ so that it is centered around zero. Here, $v_k$ is some robust estimate (like the median) of mean. Similarly, we estimate the standard deviation $\sigma$ by employing a highly robust MAD (median absolute deviation) estimator. In case of online implementation, we can obtain these parameters by using the corresponding moving average estimators.

5. Simulation Results

In this section, some computer simulations are presented in order to demonstrate the effectiveness of the proposed approach. The separation performance is evaluated by the following performance index [11]

$$PI = \left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{h_{ij}^2}{\max_{j}h_{ij}^2} - 1\right)\right),$$  

where $h_{ij}$ is the $ij$th element of the global transformation matrix $H$. Similarly, we define $PI$ (dB) as

$$PI (dB) = 10 \log(PI),$$

to measure $PI$ in decibels. The above performance index measures the average inter-channel interference. Therefore, a small value of $PI$ (dB) implies a good separation performance.

In the following computer simulations, outliers are modeled by a normal distribution with zero mean and standard deviation $\sigma_o$, where $\sigma_o$ is chosen to be much greater than the variances of the sources. The occurrence rate of outliers is denoted by $\epsilon$.

Experiment 1: In the first computer simulation, the sources consist of 5 sub-Gaussian signals including 3 uniformly distributed signals, and 2 signals each having a kurtosis value of −0.8. These signals are mixed together using a randomly chosen mixing matrix. The algorithm given by Eq. (19) is implemented in a batch mode with $N = 1500$, $\varphi(y_i^j) = y_i^j$, and $\theta = 4$. The results of this experiment are displayed in Fig. 2. Figure 2(a) plots the evolution of $PI$ (dB) (averaged over 100 realizations) for different values of $\alpha$ when $\epsilon = 0$. 

---

**Fig. 1** Plots of decaying function for (a) some values of $\theta$ ($\alpha = 1.1$), (b) some values of $\alpha$ ($\theta = 2$).

**Fig. 2** The evolution of performance index (for different values of $\alpha$) in Experiment 1 when (a) $\epsilon = 0$, and (b) $\epsilon = 0.01$; $\theta$ is fixed to 4.
The step size for each value of $\alpha$ is tuned so as to obtain the best achievable performance. From this figure, we see that in the absence of outliers, the algorithm converges to a $PI$ (dB) level of around $-26$ dB for values of $\alpha$ close to 1. On the other hand, Fig. 2(b) shows the evolution of averaged $PI$ (dB) for $\epsilon = 0.01$. From this figure, we can see that when there are outliers in the data, the proposed approach with $\alpha > 1$ results in much better separation performance as compared to the standard natural gradient algorithm ($\alpha = 1$). In particular, with $\alpha = 1.05$, the algorithm given by Eq. (19) attains a $PI$ (dB) level of $-25$ dB, which is very close to the case when there are no outliers in the data. In contrast to this, the standard natural gradient algorithm fails to achieve an acceptable value of $PI$ (dB) in the presence of outliers.

**Experiment 2**: In the next simulation, we perform a similar computer experiment with 5 super-Gaussian signals including 3 Laplace distributed signals and 2 speech signals. In this experiment, we select $\phi(y_i) = \text{sign}(y_i)$, and $\theta = 2$. The algorithm given by Eq. (19) is again implemented in a batch mode with data consisting of 1500 sample points. The simulation results are shown in Fig. 3(a) and Fig. 3(b). From these figures, we again observe that by choosing $\alpha$ greater than 1, an improved separation performance can be achieved in the presence of outliers. For instance, when $\epsilon = 0.01$ the algorithm with $\alpha = 1.02$ converges to a $PI$ (dB) value of $-27$ dB as compared to $-17$ dB for the standard natural gradient algorithm. Furthermore, when there are no outliers in the observed data i.e. $\epsilon = 0$, the asymptotic performance remains close to that of the standard natural gradient algorithm, provided that we do not make $\alpha$ too large as compared to 1.

**6. Conclusion**

In this paper, we propose to use a modified natural gradient update rule for robust independent component analysis against outliers. Such a robustness is achieved by introducing a decaying function that (comparatively) gives less weights to observations with greater deviations from the sample mean. Some computer simulations are presented to show that in the presence of potential outliers, the proposed approach gives better performance as compared to the standard natural gradient algorithm. The proposed approach works particularly well when the probability distributions of sources are dominant around their means.

**References**


