World Academy of Science, Engineering and Technology
International Journal of Mathematical and Computational Sciences
Vol:3, No:9, 2009

A completed adaptive de-mixing algorithm on Stiefel manifold for ICA

Jianwei Wu

Abstract—Based on the one-bit-matching principle and by turning the de-mixing matrix into an orthogonal matrix via certain normalization, Ma et al proposed a one-bit-matching learning algorithm on the Stiefel manifold for independent component analysis [8]. But this algorithm is not adaptive. In this paper, an algorithm which can extract kurtosis and its sign of each independent source component directly from observation data is firstly introduced. With the algorithm , the one-bit-matching learning algorithm is revised, so that it can make the blind separation on the Stiefel manifold implemented completely in the adaptive mode in the framework of natural gradient.

Keywords—Independent component analysis; kurtosis; Stiefel manifold; super-Gaussians or sub-Gaussians

I. Introduction

ONE of the main problems in independent component analysis (ICA) is to recover independent sources which have been mixed by an unknown channel. Generally, the noiseless linear model of ICA problem is assumed as

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \tag{1}$$

where $\mathbf{A}$ is an $n \times n$ mixing matrix, $\mathbf{s}$ is an original source vector which has n independent components and $\mathbf{x}$ is an $n \times 1$ observation vector.

ICA problem has been studied in the literature for many years with a number of results. For the detailed review, readers can refer to monographes [1, 2]. As well known, any decent learning algorithm, such as the natural or relative gradient one [3, 4], can work only in the cases that the components of $\mathbf{s}$ are all either super-Gaussians or sub-Gaussians. In order to separate observation data mixed with both super- and sub-Gaussian components, Lee et al. proposed the extended infomax algorithm [5]. This approach is to switch between the super- and sub-Gaussian model depending on the sign of a switching moment, which tests the stability of the solution [4], and it only requires the estimation of one binary parameter per source for each step of the algorithm. Moreover, for the general ICA problem, based on experiments, Xu et al. summarized the one-bit-matching principle [6] which states that "all the sources can be separated as long as there is a one-to-one same-sign-correspondence between the kurtosis signs of all source pdf's and the kurtosis signs of all model pdf's." Although the proof of the conjecture has not been obtained on some general assumptions, a large number of experiments show its correctness.

On the other hand, if the observed $\mathbf{x}$ and the output $\mathbf{y}$ are pre-whiten or normalized during each phase of the learning process, the de-mixing matrix $\mathbf{W}$ should be an orthogonal matrix. So on the Stiefel manifold, by introducing the general Gaussian distribution, Choi et al. proposed the flexible independent component analysis algorithm [7]. The algorithm can recover sources from their linear mixtures in the way of adaptive model matching without any prior knowledge of source distributions. Moreover, based on the one-bit-matching conjecture, Ma et al. gave the one-bit-matching learning algorithm on the Stiefel manifold in the framework of natural gradient [8]. This algorithm is simple in the form. But the number of super-Gaussian sources must be known before the algorithm is carried out. Obviously, the algorithm is not adaptive for the case which super- and sub-Gaussian components coexist in an unknown mode. In this paper, for noiseless observation data, we propose an algorithm which can directly estimate kurtosis' signs of components in original sources from whitening observation data before any de-mixing operation is carried out. By the estimation algorithm, we can recognize the number of the super- or sub-Gaussian components from noiseless observation data, and adaptively separate mixed sources with both the super- and sub-Gaussian components on condition that the number of super-Gaussion sources is unknown in the framework of natural gradient on the Stiefel manifold.

The author is with the Department of Information and Calculation Science, School of Sciences, Central University for Nationalities, Beijing 100081, P. R. of China (e-mail: wjw957@163.com).

II. The identification of the kurtosis or its sign for source components

The kurtosis of a random variable $\mathbf{Y}$(assuming $E(\mathbf{Y}) = 0$) is defined as

$$\kappa_4 = E(\mathbf{Y})^4 - 3(E(\mathbf{Y}^2))^2. \tag{2}$$

In this section, based on the fourth order blind identification [9], a theorem which expresses the kurtosis of independent source components with eigenvalues of some relative matrices is given as follows. By the theorem, an algorithm of directly exacting kurtosis signs from the noiseless observed data is then presented.

Theorem 1. *Suppose that the observation data are*

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s}, \tag{3}$$

where $\boldsymbol{A}$ is an $n \times n$ nonsingular mixing matrix, $\boldsymbol{s}$ is an $n \times 1$ source vector. The whitening matrix is $\boldsymbol{B}$, and let $\boldsymbol{U} = \boldsymbol{B}\boldsymbol{A}$, $\boldsymbol{v} = \boldsymbol{B}\boldsymbol{A}\boldsymbol{s} = \boldsymbol{U}\boldsymbol{s}$, $\bar{\boldsymbol{v}} = \boldsymbol{v} - E(\boldsymbol{v})$, then

$$E(\bar{\boldsymbol{v}}\bar{\boldsymbol{v}}^T) = \boldsymbol{U}diag(\lambda_1, \lambda_2, \ldots, \lambda_n)\boldsymbol{U}^T, \tag{4}$$

$$E(|\bar{\boldsymbol{s}}|^2\bar{\boldsymbol{v}}\bar{\boldsymbol{v}}^T) = \boldsymbol{U}diag(\mu_1, \mu_2, \ldots, \mu_n)\boldsymbol{U}^T, \tag{5}$$

TABLE I
MISTAKE IDENTIFICATION TIMES IN 20 OPERATIONS FOR EACH DATA GROUP.

data group	100	200	300	400	500	≥ 600
3 sources	3	3	0	0	0	0
5 sources	3	2	1	1	0	0
7 sources	7	3	3	0	1	0

where $\bar{\boldsymbol{s}} = \boldsymbol{s} - E(\boldsymbol{s}) = (\bar{s_1}, \bar{s_2}, \ldots, \bar{s_n})^T$, $\lambda_i = E(\bar{s_i}^2)$, $\mu_i = E(\bar{s_i}^4) + \sum_{l=1,l\neq i}^{n} E(\bar{s_l}^2)E(\bar{s_i}^2), i = 1, 2, ..., n$. *Then the kurtosis of all components in source are*

$$\mu_i - \sum_{j=1,j\neq i}^{n} \lambda_j\lambda_i - 3\lambda_i^2 \quad , \quad 1 \leq i \leq n, \tag{6}$$

where the eigenvalue order of two matrices is determined by the orthogonal decomposition of $E(\bar{\boldsymbol{v}}\bar{\boldsymbol{v}}^T)$ *and* $E(|\bar{\boldsymbol{s}}|^2\bar{\boldsymbol{v}}\bar{\boldsymbol{v}}^T)$ *with the same orthogonal matrix.*

The proof of the theorem can be found in author's another paper [10].

From the theorem, $\mathbf{v}$ and $|\mathbf{s}|^2$ can be obtained from whitened observertion data, and $E(\bar{\mathbf{v}}\bar{\mathbf{v}}^T)$ and $E(|\bar{\mathbf{s}}|^2\bar{\mathbf{v}}\bar{\mathbf{v}}^T)$ must be expressed in Eq.(4) and Eq.(5). Actually, columns in $\mathbf{U}$ just are common eigenvectors of $E(\bar{\mathbf{v}}\bar{\mathbf{v}}^T)$ and $E(|\bar{\mathbf{s}}|^2\bar{\mathbf{v}}\bar{\mathbf{v}}^T)$.

Based on the theorem, for s_i in $\mathbf{s}$, to extract its the kurtosis sign from their noiseless mixture, the following algorithm from the definition of the kurtosis can be obtained:

(i) whiten $\mathbf{x} = \mathbf{As}$ such that $\mathbf{v}$=$\mathbf{Us}$;
(ii) $\bar{\mathbf{v}} = \mathbf{v} - E(\mathbf{v})$;
(iii) computer eigenvalues of $E(\bar{\mathbf{v}}\bar{\mathbf{v}}^T) : \lambda_i$;
(iv) computer eigenvalues of $E(|\bar{\mathbf{s}}|^2\bar{\mathbf{v}}\bar{\mathbf{v}}^T) : \mu_i$;
(v) extract the sign of the kurtosis for s_i

$$k_i = Sign(\mu_i - \sum_{j=1,j\neq i}^{n} \lambda_j\lambda_i - 3\lambda_i^2). \tag{7}$$

From the algorithm, it is clear that the total number of super-Gaussian components in mixed data equals the number that the value of k_i is equal to 1, where $k_i (1 \leq i \leq n)$ which are defined in Eq.(7) can be obtained by the calculation of eigenvalues of $E(\bar{\mathbf{v}}\bar{\mathbf{v}}^T)$ and $E(|\bar{\mathbf{s}}|^2\bar{\mathbf{v}}\bar{\mathbf{v}}^T)$ with Eq.(7). In order to test the validity and reliability of the algorithm, some experiments on three sets of mixed data were conducted respectively, where the first is three independent sources from two super-Gaussian and one sub-Gaussian distributions, the second is five independent sources from three super-Gaussian and two sub-Gaussian distributions, and the third is seven independent sources from four super-Gaussian and three sub-Gaussian distributions.

For each set of mixed data, experiments were performed with 15 groups of iid data, and the i-th group consists of $i \times 100$ data (i=1,2,...,15). The algorithm in section 2 was carried out 20 times on each group of data, and in each time the mixing matrix was randomly generated . The times that the identification is not correct in 20 operations for each group are listed in Table 1.

From the table, one can see that if data are enough, the algorithm can correctly identify the numbers of super-Gaussian component in noiseless mixed data.

III. THE ADAPTIVE DE-MIXING ALGORITHM ON STIEFEL MANIFOLD

The Stiefel manifold $V_{n,p}$ consists of n-by-p($n \geq p$) orthogonal matrices. For each $\mathbf{X} \in V_{n,p}$, its p column vectors are pair-wised orthogonal in $\mathbb{R}^n$. In the case of $n = p$, the Stiefel manifold $V_{n,n}$ consists of $n \times n$ orthogonal matrices. For a smooth function $F(\mathbf{X})$ defined on $V_{n,n}$, its gradient is

$$\nabla F = F_X - XF_X^T X, \tag{8}$$

where F_X is the conventional gradient of $F(\mathbf{X})$ with respect to $\mathbf{X}$ [11].

With the $\mathbf{x}$ pre-whitened, that is $E(\mathbf{x}) = 0$, $E(\mathbf{xx}^T) = I_n$, for $\mathbf{y} = \mathbf{Wx}$ on the Stiefel manifold, then

$$E(\mathbf{y}) = 0 \quad , \quad E(\mathbf{yy}^T) = I_n. \tag{9}$$

So

$$I_n = E(\mathbf{yy}^T) = \mathbf{W}E(\mathbf{xx}^T)\mathbf{W}^T = \mathbf{WW}^T. \tag{10}$$

Thus,$\mathbf{W} \in V_{n,n}$. Therefore, if the observed $\mathbf{x}$ and the output $\mathbf{y}$ are pre-whitened during learning process, then for the orthogonal $\mathbf{W}$, one can solve it on the Stifel manifold.

On the other hand, the objective function $J(\mathbf{W})$ based on information theory is given as in Section 1, and its decent learning rule is (refer to [1] for details)

$$\triangle\mathbf{W} \propto [\mathbf{I}_n - \varphi(\mathbf{u})\mathbf{u}^T]\mathbf{W}. \tag{11}$$

where

$$\varphi(\mathbf{u}) = -\frac{\frac{\partial p(\mathbf{u})}{\partial \mathbf{u}}}{p(\mathbf{u})} = (-\frac{p_1'(u_1)}{p_1(u_1)}, ..., -\frac{p_n'(u_n)}{p_n(u_n)})^T, \tag{12}$$

and $p(\mathbf{u})$ is the model probability density function.

For the ICA problem on the manifold, Ma et al. proposed the one-bit-matching learning algorithm [8]. Under the condition of the one-bit-matching [6], The first p model pdfs are selected as super-Gaussians, and the left $n - p$ model pdfs are sub-Gaussians. Thus, after selecting $p(\mathbf{u})$, $\varphi(\mathbf{u})$ in Eq.(11) is

$$\varphi(\mathbf{u}) = \mathbf{K}_1 tanh(\mathbf{u}) - \mathbf{K}_2\mathbf{u}, \tag{13}$$

where $\mathbf{K}_1 = \text{diag}\ [-I_p, I_{n-p}]$, $\mathbf{K}_2 = \text{diag}\ [0_p, I_{n-p}]$, $tanh(\mathbf{u}) = (tanh(u_1), tanh(u_2), \ldots, tanh(u_n))^T$. With Eq.(11), the one-bit-matching learning algorithm of $\mathbf{W}$ on the Stiefel manifold as follows

$$\begin{aligned} \mathbf{J}_W &\propto [\mathbf{I}_n - (\mathbf{K}_1 tanh(\mathbf{u}) - \mathbf{K}_2\mathbf{u})\mathbf{u}^T]\mathbf{W}, \\ \Delta\mathbf{W} &\propto \mathbf{J}_W - \mathbf{W}\mathbf{J}_W^T\mathbf{W}. \end{aligned} \tag{14}$$

Note that the p in $\mathbf{K}_1$ or $\mathbf{K}_2$ is unknown now, and it is just the number of super-Gaussian sources in model pdfs. Before the algorithm is carried out, the p must be known.

This algorithm is different from the flexible ICA algorithm proposed by Choi et al [7], it is easy and understood. But the obvious drawback of the algorithm is that the number of super- or sub-Gaussian sources must be known before the operation of the algorithm. So on the manifold the algorithm is not used in the case that super- and sub-Gaussian sources coexist in an unknown mode. With the identifying algorithm of the independent component sign in section 2, the algorithm can be revised as follows.

With the identifying algorithm of the kurtosis sign in the above section, one can estimate the p at first before the operation of the above algorithm. When the p is obtained, $\mathbf{K}_1$ and $\mathbf{K}_2$ are immediately determined. Then, the gradient learning algorithm provided with Eq.(14) is carried out. Obviously, in this adaptive mode, the adaptive blind separation of sources on the manifold can completely be implemented.

To test the validity of the adaptive algorithm, an experiment on the noiseless ICA problem of five independent sources was conducted, in which there are three super-Gaussian sources generated from the exponential distribution E(0.5), the Chi-square distribution $\chi^2(6)$, and the the Gamma distribution $\gamma(1,4)$, and two sub-Gaussian sources generated from the Beta distribution $\beta(2,2)$ and the Uniform distribution U([0,1]),respectively. 100000 iid samples were generated to form a source from each distribution. The linearly mixed signals were generated via the following orthogonal matrix A:

$$A = \begin{pmatrix} -0.4466 & 0.3602 & -0.7987 & -0.0265 & -0.1844 \\ -0.3491 & 0.1407 & 0.0455 & 0.2944 & 0.8772 \\ -0.5950 & -0.1917 & 0.3217 & 0.5766 & -0.4162 \\ -0.4815 & 0.3369 & 0.4524 & -0.6694 & -0.0445 \\ -0.3076 & -0.8368 & -0.2277 & -0.3635 & 0.1456 \end{pmatrix}. \quad (15)$$

The learning rate was chosen as $\eta = 0.001$ and the algorithm operated in the adaptive mode and was stopped when all the 100000 data points of the mixed signals had been passed only once through our learning algorithm.

As a feasible solution of the ICA problem, the obtained **W** will make **WA** $=\Lambda$**P**, where $\Lambda =$diag$[\lambda_1, \lambda_2, \ldots, \lambda_n]$ with $\lambda_i \neq 0$, and P is a permutation matrix. The result of the adaptive learning algorithm on the Stiefel manifold is as follows:

$$WA = \begin{pmatrix} 0.0151 & 0.9998 & -0.0016 & -0.0088 & -0.0023 \\ -0.0006 & 0.0016 & 1.0000 & -0.0017 & -0.0037 \\ 0.9999 & -0.0151 & 0.0006 & -0.0041 & 0.0033 \\ 0.0042 & 0.0087 & 0.0017 & 0.9998 & 0.0172 \\ -0.0034 & 0.0022 & 0.0037 & -0.0172 & 0.9998 \end{pmatrix}. \quad (16)$$

Meanwhile, with the extended infomax algorithm processing the mixed signals, the result is

$$WA = \begin{pmatrix} -1.6024 & 0.0068 & 0.0024 & 0.0547 & -0.0574 \\ 0.0474 & -0.2291 & -0.0044 & -0.0842 & 0.0027 \\ 0.0065 & -0.0033 & -0.1918 & -0.0432 & 0.1068 \\ -0.0297 & -0.0035 & 0.0032 & 6.4640 & -0.1931 \\ -0.0084 & 0.0002 & 0.0017 & 0.1171 & -5.0005 \end{pmatrix}. \quad (17)$$

As a performance measure, the performance index defined by [2]

$$\sum_{i=1}^{n}(\sum_{j=1}^{n}\frac{|g_{ij}|}{max_k|g_{ik}|}-1)+\sum_{j=1}^{n}(\sum_{i=1}^{n}\frac{|g_{ij}|}{max_k|g_{kj}|}-1) \quad (18)$$

is often used in literature, where g_{ij} is the (i,j)th element of $n \times n$ matrix **G=WA**. For a perfect separation, this index is zero. The index using the adaptive learning algorithm on the Stiefel manifold and that using the extended infomax algorithm are 0.7117 and 1.8722, respectively. One can see that the performance of the adaptive learning algorithm on the Stiefel manifold is much better than that of the extended infomax algorithm.

IV. Conclusion

With the identifying algorithm of the kurtosis sign, a complement adaptive de-mixing algorithm on Stiefel manifold has been obtained when the super- and sub-Gaussian components coexist in an unknown mode in observation data, but this algorithm is still complicated. Actually, if we know the mixing matrix is orthogonal beforehand, we can estimated source signal with the simpler approach [10].

References

[1] Cichocki, A.,Amari, S.I., 2002. Adaptive Blind Signal and Image Processing. John Wiley & Sons, Ltd.
[2] Hyvärinen, A., Karhunen, J., Oja, E., 2001. Independent component analysis. John Wiley and Sons. Inc.
[3] Amari, S.I., Cichocki, A., Yang, H. 1996. A new learning algorithm for blind separation of sources. Advances in neural information processing, **8** (pp.757-763).Combridge, MA:MIT Press..
[4] Cardoso, J. F.and Laheld, B. 1996. Equivalent adaptive source separation. IEEE Trans.Signal Processing, 44(12), 3017-3030.
[5] T. W. Lee, M.Girolami, T. J. Sejnouski. 1999. Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. Neural Computation, 11(2),417-441.
[6] Xu, L., Cheung, C. C. and Amari, S. I. 1998. Learned parametric mixture based on ica algorithm. Neurocomputing, 22(1-3), 69-80.
[7] Choi, S., Cichocki, A. and Amari, S.I. 2000. Flexible Independent Component Analysis. Journal of VLSI Signal Processing, Vol.26, No.1,25-38.
[8] Jinwen Ma, Dengpan Gao, Fei Ge and Amari, S.I. 2006. A one-bit-matching learning algorithm for independent component analysis. Independent Component Analysis and Blind Signal Separation: 6th Interenational Conference, ICA 2006, Charleston, sc, USA, March 5-8, 2006, 173-180.
[9] Cardoso, J. F. 1989. Source separation using higher order moments. Proc.IEEE ICASSP, vol. 4, 2109-2112.
[10] Jianwei Wu 2009. Estimating Source Kurtosis Directly from Observation Data for ICA. to be submmitted to Signal Processing.
[11] Jianwei Wu 2008. A de-mixing algorithm based on the second order sample moment for independent component analysis. International Conference on Signal Processing Proceedings. 56-59.