[illegible] ★

Jinwen Ma[1,2], Dengpan Gao[1], Fei Ge[1], and Shun-ichi Amari[2]

[1] [illegible] 100871, [illegible]

[2] [illegible] 351-0198, [illegible]

jwma@math.pku.edu.cn

Abstract. [illegible]

[illegible]

Independent component analysis (ICA) [1]-[2] aims to blindly separate some independent sources $\mathbf{s}$ from their linear mixture $\mathbf{x} = \mathbf{A}\mathbf{s}$ via

$$\mathbf{y} = \mathbf{W}\mathbf{x}, \quad \mathbf{x} \in \mathbb{R}^m, \quad \mathbf{y} \in \mathbb{R}^n, \quad \mathbf{W} \in \mathbb{R}^{m\times n}, \tag{1}$$

where $\mathbf{A}$ is a mixing matrix, and $\mathbf{W}$ is the de-mixing matrix to be estimated. For simplicity of analysis, the number of mixed signals is required to be equal to the number of source signals, i.e., $m = n$, and $\mathbf{A}$ is an $n \times n$ nonsingular matrix. Although the ICA problem has been studied from different perspectives [3]-[5], it can be typically solved by minimizing the following objective function:

$$D = -H(\mathbf{y}) - \sum_{i=1}^{n} \int p \quad (y_i; \mathbf{W}) \log p_i(y_i) dy_i, \tag{2}$$

where $H(\mathbf{y}) = -\int p(\mathbf{y}) \log p(\mathbf{y}) d\mathbf{y}$ denotes the entropy of $\mathbf{y}$, $p_i(y_i)$ denotes the pre-determined model probability density function (pdf), and $p \quad (y_i; \mathbf{W})$ denotes the probability distribution on $\mathbf{y} = \mathbf{W}\mathbf{x}$.

★ [illegible] 60471054.

J. Rosca et al. (Eds.): ICA 2006, LNCS 3889, pp. 173–180, 2006.

In the literature, how to choose the model pdf's $p_i(y_i)$ remains a key issue for the ICA algorithms using the objective function Eq.(2). In general, any gradient descent learning algorithm, such as the relative or natural gradient algorithms [3]-[4], can work only in the cases that the components of $\mathbf{s}$ are either all super-Gaussians or all sub-Gaussians. Recently, many new algorithms (e.g., the extended Infomax algorithm [6] and the Fast-ICA algorithm [7]) have been proposed to solve the general ICA problem, but their theoretical foundations are yet unclear. In order to solve the general ICA problem, Xu et al. [8] proposed the one-bit-matching conjecture 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". Recently, Liu et al. [9] proved this conjecture by globally minimizing the objective function under certain assumptions on the model pdf's. Ma et al. [10] further proved the conjecture by locally minimizing the same objective function on the two-source ICA problems. It is generally believed that the one-bit-matching condition can serve as a reasonable principle for the design of the model pdf's. On the other hand, if the observed $\mathbf{x}$ and the output $\mathbf{y}$ are both normalized with zero mean and unit covariance matrix, the de-mixing matrix becomes orthogonal, which can be learned on the Stiefel manifold.

In this paper, under the condition that the model pdf's are designed according to the one-bit-matching principle, with the observed $\mathbf{x}$ and the output $\mathbf{y}$ being properly normalized, we propose a gradient-type ICA learning algorithm on the Stiefel manifold, which we call as one-bit-matching learning algorithm. It is shown by the simulated and audio experiments that the proposed algorithm works e ciently on the general blind source separation problems and outperforms the typical existing ICA algorithms.

a a

We start to introduce the Stiefel manifold. Roughly, the Stiefel manifold $V_{n,p}$ consists of n-by-p "tall skinny" orthogonal matrices. That is, the p column vectors of each matrix in $V_{n,p}$ are pair-wised orthogonal in $\mathbb{R}^n$. Here, we need only to consider the special Stiefel manifold $V_{n,n}$, i.e., the orthogonal group O_n consisting of n-by-n orthogonal matrices. For a smooth function $F(\mathbf{Z})$ on the Stiefel manifold O_n, i.e., $\mathbf{Z} \in O_n$, with the canonical Euclidean metric, its gradient on the manifold is computed by

$$\nabla F = F \quad - \mathbf{Z}F^T\mathbf{Z}, \tag{3}$$

where F is the conventional gradient of $F(\mathbf{Z})$ with respect to the matrix $\mathbf{Z}$. This gradient is consistent with the natural Riemannian gradient on the Stiefel manifold from information geometry.

We further pre-whiten the observed $\mathbf{x}$ and the output $\mathbf{y}$ so that the de-mixing matrix $\mathbf{W}$ can only be orthogonal, i.e, on the Stiefel manifold O_n. Clearly, we can easily pre-whiten the observed $\mathbf{x}$ such that

$$E(\mathbf{x}) = 0, \qquad E\mathbf{x}\mathbf{x}$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix. With each matrix $\mathbf{W}$, we can also pre-whiten the output $\mathbf{y} = \mathbf{W}\mathbf{x}$ such that

$$E(\mathbf{y}) = 0, \qquad\qquad E\mathbf{y}\mathbf{y}^T = \mathbf{I}_n. \tag{5}$$

In this way, we have

$$\mathbf{I}_n = E(\mathbf{y}\mathbf{y}^T) = \mathbf{W}E(\mathbf{x}\mathbf{x}^T)\mathbf{W}^T = \mathbf{W}\mathbf{W}^T. \tag{6}$$

Thus, $\mathbf{W}\mathbf{W}^T = \mathbf{I}_n$. That is, $\mathbf{W}$ must be an orthogonal matrix. Therefore, if we can pre-whiten or normalize the observed $\mathbf{x}$ and the output $\mathbf{y}$ during each phase of the learning process, the resulted $\mathbf{W}$ should be an orthogonal matrix. Therefore, we can solve it on the Stiefel manifold.

We now revisit the objective function defined in Eq.(2). Suppose that the observed $\mathbf{x}$ and the output $\mathbf{y}$ are both pre-whitened. Then, the required de-mixing $\mathbf{W}$ should be orthogonal. So, we can search the feasible solution $\mathbf{W}$ of the ICA problem via minimizing the objective function of $\mathbf{W}$ on the Stiefel manifold O_n. For the design of model pdf's, we use the one-bit-matching condition. Suppose that p is the number of super-Gaussian sources in the ICA problem. The model pdfs of sub- and super-Gaussians are selected as

$$p_{super}(u) = \frac{1}{\pi}\text{sech}(u), \qquad p_{sub}(u) = \frac{1}{2}[p_{N(1,1)}(u) + p_{N(-1,1)}(u)],$$

respectively, where $p_{N(\mu,\sigma^2)}(u)$ is the Gaussian probability density with mean μ and variance σ^2, and sech$(\cdot)$ is the hyperbolic secant function. That is, the first p model pdf's are selected as $p_{super}(u)$, while the rest $n-p$ model pdf's are selected as $p_{sub}(u)$. In this way, the conventional gradient of the objective function can be computed as follows.

We let $\mathbf{V} = (v_1, v_2, \cdots, v_n)^T$ be an n-dim vector. For each observed $\mathbf{x}$ and the corresponding output $\mathbf{y}$ via the relation $\mathbf{y} = \mathbf{W}\mathbf{x}$, we define

$$\begin{aligned} v_i &= -\tanh(y_i), \quad \text{for } \; i = 1, \cdots, p; \\ v_i &= \tanh(y_i) - y_i, \quad \text{for } \; i = p+1, \cdots, n. \end{aligned}$$

Then, the adaptive gradient J of the objective function Eq. (2) with respect to $\mathbf{W}$ is simplified as

$$J \quad = -\mathbf{W} - \mathbf{V}\mathbf{x}^T. \tag{7}$$

Given Eq.(7), we can construct the one-bit-matching learning algorithm as a local gradient-descent learning algorithm of $\mathbf{W}$ on the Stiefel manifold O_n as follows.

$$\triangle\mathbf{W} = -\eta(J \quad - \mathbf{W}J^T\ \mathbf{W}) = \eta(\mathbf{V}\mathbf{x}^T - \mathbf{W}\mathbf{x}\mathbf{V}^T\mathbf{W}), \tag{8}$$

where $\eta > 0$ is the learning rate parameter which is generally selected as a small positive constant.

Since J is just the adaptive gradient of the objective function, this algorithm is adaptive. As $\mathbf{W}$ keeps an orthogonal matrix during the learning process, the output $\mathbf{y}$ will be always normalized or whitened. Therefore, we need only to pre-whiten the observed $\mathbf{x}$ at the beginning of the algorithm. Certainly, we can establish the batch gradient learning algorithm on the Stiefel manifold with the batch gradient of the objective function.

a

In order to test our one-bit-matching learning algorithm, we conducted several simulated and audio experiments on three source separation problems: (i). mixed super-Gaussian and sub-Gaussian ones; (ii). uniform noises which are all sub-Gaussian; (iii). audio samples which are all super-Gaussian. We also compared it with the extended Informax and Fast-ICA algorithms.

3.1 On Separating Mixed Super-Gaussian and Sub-Gaussian Sources

We began to consider the ICA problem of seven independent sources in which there are four super-Gaussian sources generated from one Exponential distribution $E(0.5)$, one Chi-square distribution $\chi^2(6)$, one Gamma distribution $\gamma(1,4)$ and one F-distribution $F(10,50)$, respectively, and three sub-Gaussian sources generated from two β distributions $\beta(2,2)$, $\beta(0.5,0.5)$, and one Uniform distribution $U([0,1])$, respectively. From each source or distribution, 100000 i.i.d. samples were generated to form a source. Accordingly, these samples were further pre-whitened.

The first set of linearly mixed signals was generated from these seven source signals via a random orthogonal mixing matrix $\mathbf{A}_1$. We implemented the one-bit-matching learning algorithm ($p = 4, n = 7$) on the first set of linearly mixed signals with the learning rate being selected as $\eta = 0.001$ and the initial $\mathbf{W}$ being set as a randomly generated orthogonal matrix. The one-bit-matching learning algorithm operated adaptively and was stopped after 100000 iterations to ensure

For comparison, we also ran the extended Infomax algorithm [5] (a kind of natural or relative gradient learning with a switch criterion) on this set of linearly mixed signals and obtained the separation result given by Eq. (10).

$$
\mathbf{WA_1} = \begin{bmatrix}
0.0148 & \mathbf{-0.7588} & 0.0085 & -0.0005 & -0.0241 & -0.0189 & 0.0088 \\
0.0222 & 0.0167 & -0.0109 & 0.0135 & \mathbf{-1.4220} & -0.0111 & 0.0093 \\
0.0088 & -0.0042 & \mathbf{-0.7532} & -0.0197 & -0.0133 & 0.0336 & 0.0103 \\
-0.0144 & -0.0141 & 0.0037 & -0.0333 & -0.0280 & -0.0141 & \mathbf{1.4943} \\
\mathbf{-0.8065} & 0.0161 & -0.0018 & -0.0146 & -0.0581 & -0.0465 & 0.0777 \\
0.0176 & -0.0197 & -0.0057 & 0.0288 & -0.0210 & \mathbf{-1.4393} & 0.0343 \\
0.0001 & -0.0353 & 0.0284 & \mathbf{0.7675} & 0.0004 & -0.0537 & -0.0017
\end{bmatrix} \quad (10)
$$

From the above two tables, it can be found that the one-bit-matching learning algorithm is much better than that of the extended Infomax algorithm. Precisely, we calculated the performance index (introduced in [3]) defined by

$$
PI = \sum_{i=1}^{n}\Big(\sum_{j=1}^{n} \frac{|r_{ij}|}{{}_{k}\,|r_{ik}|} - 1\Big) + \sum_{j=1}^{n}\Big(\sum_{i=1}^{n} \frac{|r_{ij}|}{{}_{k}\,|r_{kj}|} - 1\Big),
$$

where $\mathbf{R} = (r_{ij})_{n\times n} = \mathbf{WA}$. For a perfect separation, this index should be zero. Actually, the performance indexes of the one-bit-matching learning and extended Infomax algorithms are 0.3411 and 1.6399, respectively, which quantitatively shows that the one-bit-matching learning algorithm is much better than the extended Infomax. Moreover, we implemented the Fast-ICA algorithm on this linearly mixed signal set and obtained the separation result with the performance index being 0.3028, which is slightly better than that of the one-bit-matching learning algorithm.

3.2 On Separating Uniform Noises

Next, we considered the ICA problem of separating eight independent uniform noises. That is, each source was sampled from a uniform distribution and contains 100000 samples. These sources are all sub-Gaussian, being recognized as the uniform noises. Our second set of linearly mixed signals was generated from these eight uniform noises via another random orthogonal mixing matrix $\mathbf{A}_2$. The signals were further pre-whitened. On this set of linearly mixed signals, we implemented the one-bit-matching learning and extended Infomax algorithms and their results are given by Eq. (11) and Eq. (12), respectively. It was found that their performance indices are 0.1713 and 2.2776, respectively, which also shows that the one-bit-matching learning algorithm also outperforms the extended Infomax. Moreover, it was also found that the the performance index of the separation results via the Fast-ICA algorithm on this set is 0.2342, which is considerably larger than 0.1713. That is, the one-bit-matching learning algorithm also outperforms the Fast-ICA algorithm in this case.

$$\mathbf{WA} = \begin{bmatrix} -0.003 & 0.000 & 0.001 & -0.001 & \mathbf{1.000} & 0.000 & -0.002 & -0.003 \\ -0.002 & 0.000 & -0.002 & 0.001 & -0.002 & 0.000 & \mathbf{-1.000} & -0.001 \\ 0.000 & 0.002 & 0.000 & \mathbf{-1.000} & -0.001 & -0.004 & -0.001 & -0.002 \\ 0.001 & -0.003 & \mathbf{-1.000} & 0.000 & 0.001 & 0.000 & 0.002 & 0.002 \\ -0.003 & \mathbf{-1.000} & 0.003 & -0.002 & 0.000 & 0.001 & 0.000 & 0.001 \\ -0.002 & -0.001 & -0.002 & 0.002 & -0.003 & 0.002 & 0.001 & \mathbf{-1.000} \\ -0.001 & 0.001 & 0.000 & -0.004 & 0.000 & \mathbf{1.000} & 0.000 & 0.002 \\ \mathbf{1.000} & -0.003 & 0.001 & 0.000 & 0.003 & 0.001 & -0.002 & -0.003 \end{bmatrix} \quad (11)$$

$$\mathbf{WA} = \begin{bmatrix} 0.021 & -0.002 & \mathbf{-1.449} & 0.029 & -0.014 & 0.038 & -0.017 & -0.053 \\ 0.009 & -0.057 & -0.039 & -0.033 & 0.053 & \mathbf{1.430} & -0.008 & -0.022 \\ \mathbf{1.450} & -0.040 & -0.014 & -0.010 & 0.033 & 0.052 & 0.009 & -0.076 \\ 0.050 & -0.035 & -0.047 & 0.045 & -0.014 & 0.025 & -0.008 & \mathbf{-1.499} \\ -0.037 & \mathbf{-1.452} & 0.031 & -0.040 & -0.048 & 0.037 & -0.038 & -0.023 \\ -0.031 & -0.046 & -0.031 & 0.007 & \mathbf{-1.444} & -0.047 & -0.005 & -0.023 \\ -0.014 & 0.015 & 0.037 & 0.022 & 0.037 & -0.016 & \mathbf{1.443} & 0.032 \\ -0.013 & 0.056 & -0.006 & \mathbf{1.404} & 0.014 & -0.018 & 0.057 & -0.013 \end{bmatrix} \quad (12)$$

3.3 On Separating Audio Sources

Finally, we considered the ICA problem of separating 8 independent real-life audio recordings (downloaded from Barak Pearlmutter's homepage: `http://www- bcl.cs.may.ie/~bap/demos.html`). Each audio source consists of 100000 data sampled at 22050 Hz. We pre-whitened these audio sources and then linearly mixed them via an 8×8 random orthogonal mixing matrix $\mathbf{A}_3$ to form the third set of linearly mixed signals. On such a data set, we implemented the one-bit-matching algorithm, obtaining a successful separation result shown in Fig. 1.

For comparison, we also implemented the extended Infomax and Fast-ICA algorithms on the data set. It was found by the experiments that the performance indices of the one-bit-matching learning, extended Infomax and Fast-ICA algorithms are 1.2979, 2.2746, and 1.3288, respectively, which again shows that the one-bit-matching learning algorithm outperforms the extended Infomax and Fast-ICA algorithms in this case.

For further comparison, we calculated the signal-to-noise ratios (SNRs) of the output signals of the one-bit-matching, extended Infomax, and Fast-ICA learning algorithms on the data set. Their results are listed in Table 1, which again shows our proposed one-bit-matching learning algorithm outperforms the other two popular ICA learning algorithms.

In addition, extensive experiments on the di erent ICA problems with mixed super- and sub-Gaussian sources also showed that the one-bit-matching learning algorithm always reaches an accurate feasible solution. It was even found that as the number of sources increases, the one-bit-matching learning algorithm can still maintain a similarly good performance on the source separation problems. By comparison, we have found that the one-bit-matching learning algorithm considerably outperforms the extended Infomax algorithm in the general case. As

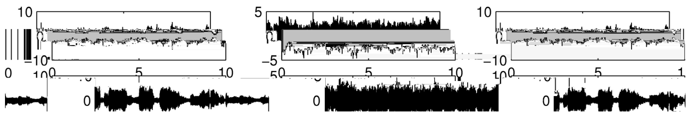

Fig. 1. s s s s g s ( ), s g s ( ), s g s (g) - - g g g

Table 1. s s s g s

	g - - s ( )									
A	1	2	3	4	5	6	7	8	_ .	A g.
- - g	23.06	37.25	21.84	25.69	28.71	31.19	36.64	30.15	29.43	29.32
	18.82	23.96	22.56	26.97	22.39	26.97	25.71	23.67	23.81	23.88
s - A	22.21	34.37	22.25	25.74	29.96	31.97	36.33	28.07	29.02	28.86

compared with the Fast-ICA algorithm, the one-bi-matching learning algorithm leads to a similar result in the case of mixed sub- and Super-Gaussian sources, but a better result in the case of all the sub- or super-Gaussian sources.

In practice, the number of super-Gaussian sources, p, may not be available in ceratin cases. In this situation, the one-bit-matching learning algorithm cannot work directly. However, we can implement the one-bit-matching learning algorithm on the pre-whitened observed $\mathbf{x}$ with p varying from zero to n, then there must be a feasible solution $\mathbf{W}$ with which the components of the output $\mathbf{y} = \mathbf{W}\mathbf{x}$ are independent, which can be checked by certain statistical independence test method. That is, for each p, we can check whether the n components of the output $\mathbf{y}$ by the resulted $\mathbf{W}$ are mutually independent. If they are, this $\mathbf{W}$ is just a feasible solution for the ICA problem. Otherwise, it is not a feasible solution for the ICA problem. Since the independence between the components of the output $\mathbf{y}$ is su cient for the feasible solution of the ICA problem, we can find out the feasible solution of the ICA problem by this test and checking procedure with the one-bit-matching learning algorithm. In fact, with a certain independence test criterion, we can use the one-bit-matching learning algorithm to obtain the feasible solution for all the above three cases without knowing the number of super-Gaussian sources.

In this paper, we have investigated the ICA problem from the point of view of the one-bit-matching principle, and established an e cient one-bit-matching ICA learning algorithm based on the Stiefel manifold gradient under the condition that the number of super-Gaussian sources is known and the observed signals are pre-whitened. It is demonstrated by the simulated and audio experiments that the proposed one-bit-matching learning algorithm can solve the source separation problem of mixed super- and sub-Gaussian sources e ciently and even outperforms the existing extended Infomax and Fast-ICA learning algorithms. Moreover, with certain independence test criterion, the one-bit-matching learning algorithm can be used to solve the source separation problem without knowing the number of super-Gaussian sources.

1. g ., ., .: - s g s s s. s. g ss g, 41(7)(1993) 2461-2470
2. .: s ?. g ss g, 36(3)(1994) 287-314
3. A. s .: A - s - . , 7(6)(1995) 1129-1159
4. A . ., A., g .: A g g s - s s. A s ss g, 8(1996) 757-763
5. s . . .: s s . s. g ss g, 44(12)(1996) 3017-3030
6. . ., ., s .: s s s g g s g ss s g ss s s. , 11(2)(1999) 417-441
7. A.: s s - g s s. s. s, 10(3)(1999) 626-634
8. ., g . ., A . .: s g . g, 22(1998) 69-80
9. . ., . ., .: - - g s. , 16(2)(2004) 383-399
10. ., ., .: A s A - - g . , 17(2)(2005) 331-334