Spurious Solution of the Maximum Likelihood Approach to ICA

Fei Ge and Jinwen Ma

***Abstract*—For the separation of linear instantaneous mixtures of independent sources, many Independent Component Analysis (ICA) algorithms can learn the separating matrix by optimizing some objective functions derived from various criteria. The Maximum Likelihood (ML) principle, with hypothesized model pdf's, provides an objective function which is commonly used. It is generally considered that the ML approach leads to a separating solution as long as the kurtosis signs of the model pdf's can correspond**

tiodinof

TCTI

FOR the blind separation of instantaneously mixed independent non-Gaussian signals, the Independent Component Analysis (ICA) [1] is a commonly utilized statistical technique, which exploits only the amplitude statistics of signals [2]. Under this model, the observed signals can be represented by an n-dimensional random vector $\mathbf{x} = \mathbf{A}\mathbf{s}$, which is simply a linear transformation of vector $\mathbf{s}$ of the latent source signals that are mutually independent. For simplicity we assume that $\mathbf{A}$ is square and invertible, then the sources can be reconstructed as $\mathbf{y} = \mathbf{W}\mathbf{x}$, if $\mathbf{W}$ is a separating matrix, i.e., $\mathbf{WA}$ has only one nonzero entry in each row and in each column.

Various approaches can lead to blind source separation, for example, minimizing mutual information (MMI) [1], [3], information maximization (Infomax) [4]. If the joint pdf of the sources is known as $r(\mathbf{s}) = \prod_{i=1}^{n} r_i(s_i)$, the Maximum Likelihood (ML) approach provides a consistent estimator of $\mathbf{A}$, by maximizing the normalized log-likelihood function [2]:

$$\mathcal{L}_N(\mathbf{A}) = \frac{1}{N}\sum_{t=1}^{N} \log r(\mathbf{A}^{-1}\mathbf{x}_t)\frac{1}{|\det(\mathbf{A})|} \tag{1}$$

under a given set of i.i.d. samples $\mathbf{x}_1, \ldots, \mathbf{x}_N$. However, in the case of ICA, $r(\cdot)$ in (1) is not known in practical situations.

Manuscript received January 09, 2010; revised March 13, 2010; accepted March 13, 2010. Date of publication May 03, 2010; date of current version May 26, 2010. This work was supported by the Ph.D. Programs Foundation of Ministry of Education of China for Grant 20070001042. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Alfred Mertins.

The authors are with the Department of Information Science, School of Mathematical Sciences and LMAM, Peking University, Beijing 100871, China (e-mail: felixge@math.pku.edu.cn; jwma@math.pku.edu.cn).

Digital Object Identifier 10.1109/LSP.2010.2049516

In the literature, when hypothesized model pdf's $q_i(\cdot)$ were used instead of $r_i(\cdot)$, maximizing the likelihood of the data could yield source separation as long as the true and hypothesized pdf's did not differ too much. By letting $\mathbf{W} = \mathbf{A}^{-1}$, $N \to \infty$, such ML-type objective function for source separation turns out to be

$$J(\mathbf{W}) = \sum_{i=1}^{n} E\left[\log q_i\left(\mathbf{w}_i^T\mathbf{x}\right)\right] + \log|\det(\mathbf{W})| \tag{2}$$

sepationI. I

where $\mathbf{w}_i^T$ is the i-th row of $\mathbf{W}$. Essentially, the Infomax algorithm [4] has the same optimizing criteria [2].

An important but unsettled question is: to what extent can each model pdf differ from true source pdf, while ensuring any maximum point of (2) to be a valid separating matrix? Based on experimental experiences, Xu [5] suggested the one-bit-matching (OBM) condition that "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" and conjectured that it is sufficient for any local maximum point of ML-type objective function to be a separating matrix.

The OBM conjecture asserts spurious-free solution for the ML approach. However, in [6] it has been illustrated for two trimodal sources that even when the source model is exact, spurious solutions may be encountered. This may be regarded as a counter example. Recently, there were some theoretical analyses [7], [8] on the OBM conjecture, but the objective function was simplified and quite different from (2).

In this letter, we give some examples showing that the OBM condition is not sufficient, even if the global maximum of the objective function is found. First, we revisit the stability conditions for adaptive algorithms and their link to the OBM condition is analyzed. Then, some specific examples evidencing the spurious maxima of ML-type objective function, with or without whiteness constraint, are presented with detailed analysis and numerical simulations.

II. Local Convergence Issues of Adaptive ML Algorithm

Since the OBM conjecture does not specify a particular optimization procedure, it applies to general adaptive gradient methods, for which the stability analysis has been established [9], [10]. An equilibrium of an adaptive learning rule for (2) is characterized by

$$E[\varphi(\mathbf{y})\mathbf{y}^T] - \mathbf{I} = \mathbf{0}, \tag{3}$$

where $\varphi(\mathbf{y}) = (\varphi_1(y_1), \ldots, \varphi_n(y_n))^T$, $y_i = \mathbf{w}_i^T\mathbf{x}$, and

$$\varphi_i(u) = -q_i'(u)/q_i(u) \tag{4}$$

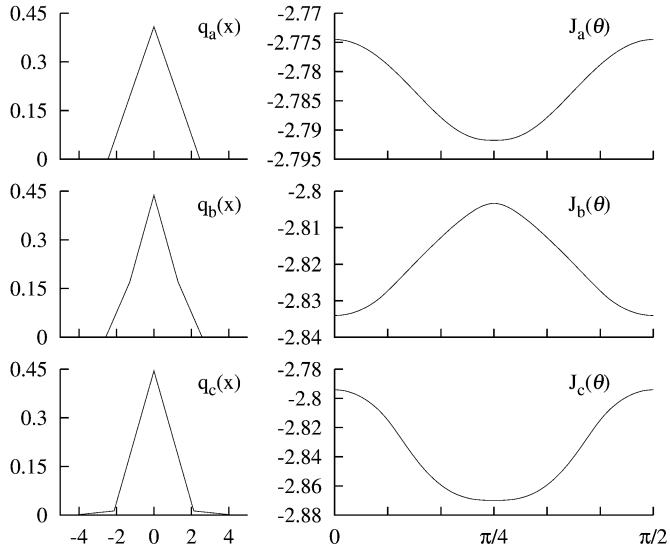

Fig. 1. Three different model pdf's and the corresponding objective functions, for the mixture of two uniformly distributed sources.

or characterized by

$$E[\mathbf{y}\mathbf{y}^T] = \mathbf{I}, \tag{5}$$

$$E[\varphi(y)\mathbf{y}^T - \mathbf{y}\varphi(\mathbf{y})^T] = \mathbf{0} \tag{6}$$

if whiteness constraint is imposed.

Let

$$\kappa_i \triangleq E[\varphi_i'(y_i)]E\left[y_i^2\right] - E[\varphi_i(y_i)y_i] \tag{7}$$

and then if $\mathbf{W}$ is a local maximum point of (2), it is necessary and sufficient that

$$1 + \kappa_i > 0, \quad \text{for} \quad 1 \le i \le n \tag{8}$$

$$(1 + \kappa_i)(1 + \kappa_j) > 1, \quad \text{for} \quad 1 \le i < j \le n \tag{9}$$

hold, or under whiteness constraint

$$\kappa_i + \kappa_j > 0, \quad \text{for} \quad 1 \le i < j \le 0 \tag{10}$$

hold [10].

Given the source pdf's and $\varphi_i(\cdot)$, it is not difficult to check the stability of separating matrices, but still, there is no guarantee that other solutions (matrices) are unstable—the global behavior of an adaptive algorithm is untouched. The OBM conjecture, if it was true, ensures global convergence or spurious-free solutions. It can be interpreted as: if the kurtoses of $q_i(\cdot)$ and those of s_i are matched in certain order, any equilibrium $\mathbf{W}$ satisfying the stability conditions is a separating matrix. But this is in doubt, considering the particular example in [6].

Here we want to stress that the OBM condition does not ensure local convergence either. Its requirement for the model pdf's is simply kurtosis sign matching to the sources. Along with scaling constraints (3) or (5), it does not generally imply the stability of a separating matrix. As pointed out by Cardoso [2], the stability depends on the kurtosis signs of the sources only when cubic nonlinearities $\varphi_i(\cdot)$ are used. Thus, we may suspect that there might be cases (combinations of source and model pdf's) when the OBM condition is satisfied but none of the separating matrices that meet (3) (or (6) under whiteness constraint) satisfy the stability conditions. Then an adaptive algorithm cannot converge to a separating matrix and must result in a spurious solution. Such cases have been identified and will be shown in the following sections.

III. Spurious Maxima of ML-Type Objective Function Under Whiteness Constraint

For ease of analysis we consider the simplest case involving only two sources with whiteness constraint on $\mathbf{s}$, $\mathbf{x}$ and $\mathbf{y}$. Then the transfer matrix $\mathbf{WA}$ is a rotation or reflection, which can be parameterized as

$$\mathbf{WA} = \begin{pmatrix} \cos\theta & \sin\theta \\ \pm\sin\theta & \mp\cos\theta \end{pmatrix}. \tag{11}$$

The pdf of $\mathbf{y} = \mathbf{Wx} = \mathbf{WAs}$ is now determined by θ. For a particular θ, the pdf of y_1 or y_2 can be represented in an integration form, because the pdf of linear combination $\alpha U + \beta V$ of two independent variables U and V is the convolution

$$p_{\alpha U+\beta V}(x) = \frac{1}{|\alpha\beta|}\int p_U\left(\frac{z}{\alpha}\right) p_V\left(\frac{x-z}{\beta}\right) dz. \tag{12}$$

In each of the following two examples, we utilize the same model pdf $q(\cdot)$ for both sources because they belong to the same class of either sub-Gaussian or super-Gaussian. Then, the ML-type objective function can be simplified as

$$J(\theta) = E[\log q(y_1(\theta))] + E[\log q(y_2(\theta))]. \tag{13}$$

If it has local maxima only at $k\pi/2$, $(k \in \mathbb{Z})$, the two sources can be separated by any local optimizing algorithm.

Here s_1 and s_2 are chosen to follow the uniform distribution on $[-\sqrt{3}, \sqrt{3}]$, whose kurtosis is $-6/5$. The pdf's of $y_1(\theta)$ and $y_2(\theta)$, which are piecewise linear functions, can be derived easily using (12). We consider the piecewise linear function

$$q(x) = \begin{cases} C(A - |x|)/A + B, & |x| \le A \\ B(2A - |x|)/A, & A < |x| \le 2A \\ 0, & |x| > 2A \end{cases} \tag{14}$$

as the model pdf, where A, B, C are parameters. To ensure that $q(x)$ is a pdf with unit variance, it must hold that $B = 1/(2A^3) - 1/(12A)$ and $C = 5/(4A) - 3/(2A^3)$, so there is just one free parameter.

We have tested three model pdf's (subscripted by a,b and c) with A equal to $\sqrt{6}, 9/7$ and $17/8$, repectively. The model pdf's and corresponding objective functions

$$J(\theta) = \int \left(p_{y_1(\theta)}(x) + p_{y_2(\theta)}(x)\right) \log q(x) dx \tag{15}$$

are plotted in Fig. 1. As $J(\theta)$ is $\pi/2$ periodic, we just show the curves in $[0, \pi/2]$.

These model pdf's are quite different in shape from that of the sources, but we can see that the maxima of $J_{\mathrm{a}}(\theta)$ and $J_{\mathrm{c}}(\theta)$ are all corresponding to separating solutions. Actually, $q_{\mathrm{a}}(x)$ is precisely $p_{y_1(\pi/4)}(x)$, with kurtosis $-3/5$, so it "matches" the source distribution, considering the OBM condtion. But the kurtosis of $q_{\mathrm{c}}(x)$ is 0.594, and thus the OBM condition is not satisfied. On the contrary, the kurtosis of $q_{\mathrm{b}}(x)$ is -0.423, which

TABLE I
SIMULATED SEPARATION QUALITY (DEVIATION OF $\hat{\theta}$) ON THE TWO UNIFORM SOURCES

model	N=500	N=1000	N=2000	N=5000	N=10000
a	3.54	2.45	1.70	1.05	0.76
b	42.94	43.62	44.10	44.46	44.63
c	1.00	0.72	0.51	0.31	0.22

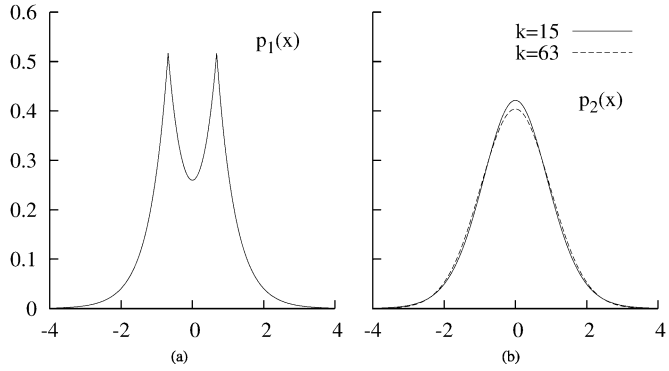

Fig. 2. Pdf's of the two super-Gaussian sources in our test example.

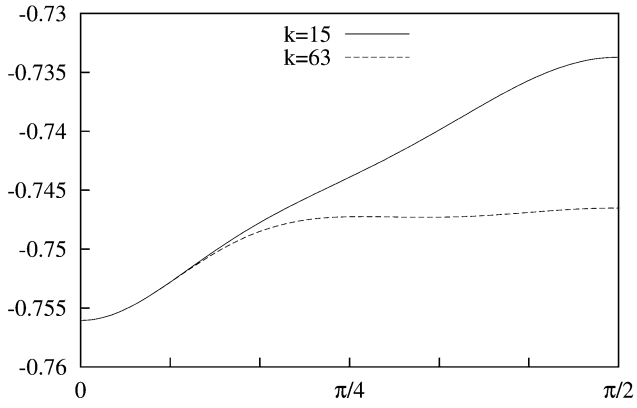

Fig. 3. Evolution of $-E[\log\cosh((\pi/\ )y_1(\theta))]$ versus θ, for the mixture of two specially chosen super-Gaussian sources.

"matches" that of the sources, but only the minima of $J_{\rm b}(\theta)$ are corresponding to separating solutions and in this case the ML approach must fail.

Numerical simulations with randomly generated data of varying sample size have been performed, using these three model pdf's, respectively. For each simulation, a local maximum of the likelihood function was sought and the corresponding $\hat{\theta}$ has been recorded. Table I lists the deviations (in degrees, averaged value over 1000 simulations) of $\hat{\theta}$ from $k\pi/2$ ($k \in \mathbb{Z}$). It can be seen that with the increase of sample size, the separation quality improves for $q_{\rm a}(x)$ and $q_{\rm c}(x)$, but spurious maxima maintain the same level of separation failure for $q_{\rm b}(x)$. Therefore, the simulation results confirmed the analysis from Fig. 1.

In this example the pdf for source s_1 is designed to be

$$p_1(x) = \frac{b}{4}\left(\exp(-b|x-a|) + \exp(-b|x+a|)\right). \qquad (16)$$

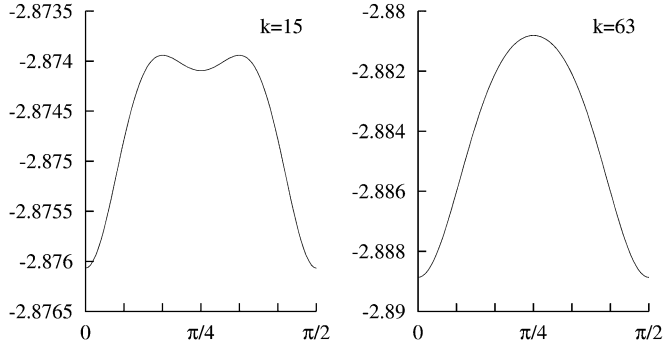

Fig. 4. Objective function $J(\theta)$ with hyperbolic secant model pdf for separating the two specially chosen super-Gaussian sources.

It has tails decaying like a Laplacian pdf, but has two peaks at $x = -a$ and $x = a$. The kurtosis sign of $p_1(x)$ depends on a and b. We set $a = 0.68$ here (b is set such that the variance is 1), and its kurtosis is 0.439. Fig. 2(left) sketches this pdf.

The distribution for s_2 is chosen as the normalized t-distribution:

$$p_2(x) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{(k-2)\pi}\,\Gamma\left(\frac{k}{2}\right)}\left(1+\frac{x^2}{k}\right)^{-\left(\frac{k+1}{2}\right)}. \qquad (17)$$

When the degree of freedom k goes to $+\infty$, t-distribution converges to Gaussian. If $k > 4$, it has kurtosis $6/(k-4)$. The pdf's for $k = 15$ and $k = 63$ are sketched in Fig. 2(right). Actually when $k = 63$ the pdf is very close to Gaussian, though it still has positive kurtosis.

We try to separate the two sources from their mixtures using hyperbolic secant model pdf of unit variance:

$$q(x) = \frac{1}{2\cosh\left(\frac{\pi}{2}x\right)}, \qquad (18)$$

which is a common choice for separating super-Gaussian sources (with $\varphi_i(x) = (\pi/2)\tanh((\pi/2)x)$). Its kurtosis is positive. The objective function in this example becomes

$$J(\theta) = -E\left[\log\cosh\left(\frac{\pi}{2}y_1(\theta)\right)\right] - E\left[\log\cosh\left(\frac{\pi}{2}y_2(\theta)\right)\right] - 2\log 2. \qquad (19)$$

Using numerical integration method, we can sketch out the curve of $-E[\log\cosh((\pi/2)y_1(\theta))]$, as shown in Fig. 3. For $p_2(x)$, the results with $k = 15$ and $k = 63$ are both shown. Since $y_1(\theta)$ and $y_2(\theta)$ have different pdf's, $J(\theta)$ does not have the same shape of $-E[\log\cosh((\pi/2)y_1(\theta))]$. Accordingly, Fig. 4 sketches $J(\theta)$. In both cases, spurious maxima exist, and the separating solutions are in fact corresponding to the minima of $J(\theta)$.

Numerical simulations with randomly generated data have also been performed using the source and model pdf's above. The deviations (in degrees, averaged value over 1000 simulations) of $\hat{\theta}$ from desired $k\pi/2$ ($k \in \mathbb{Z}$) are listed in Table II. As expected, the separation quality does not improve with increase of sample size.

IV. FURTHER DISCUSSION

In the previous section we have shown by example that in the ML approach under whiteness constraint, inappropriate model pdf can result in spurious solution, even if the kurtosis signs

TABLE II
SIMULATED SEPARATION QUALITY (DEVIATION OF $\hat{\theta}$) ON THE TWO SPECIALLY CHOSEN SUPER-GAUSSIAN SOURCES

	N=500	N=1000	N=2000	N=5000	N=10000
k=15	24.20	25.09	25.00	25.89	26.98
k=62	28.47	29.86	32.80	35.08	37.97

of real and assumed source distributions are consistent. That is, the OBM condition is not sufficient to ensure the global maxima of ML-type objective function are corresponding to separating solutions.

For regular ML approach without whiteness constraint, the surface of the objective function is more difficult to investigate, even for two-source mixing. But focusing on the stability conditions, we are able to construct some counter example for the OBM conjecture. Inspired by Douglas [11], we consider the symmetric discrete distribution $X \in \{\pm A_1, \pm A_2\}$, with $\Pr(X = A_1) = \Pr(X = -A_1) = p_1$ and $\Pr(X = A_2) = \Pr(X = -A_2) = p_2$. We want two sources s_1 and s_2 with identical distributions being super-Gaussian (having positive kurtosis) but not separable using the hyperbolic secant model pdf $q(u) = \mathrm{sech}(u)/\pi$, or equivalently the nonlinearity $\varphi_i(u) = \tanh(u)$.

For any separating matrix, it must yield $y_1 = c_1 s_1$ and $y_2 = c_2 s_2$ (or $y_1 = c_1 s_2$ and $y_2 = c_2 s_1$) where c_1 and c_2 are scaling factors. Taking into count the constraint (3) for an equilibrium,

$$p_1 c_i \tanh(c_i A_1) A_1 + p_2 c_i \tanh(c_i A_2) A_2 = 1/2 \qquad (20)$$

must hold for $i = 1, 2$. If p_1, p_2, A_1, A_2 are all fixed, this equation has only two opposite solutions for c_i. We want $c_i = \pm 1$ so that if the sources are separated, there is no amplitude change. The kurtosis constraint is that

$$2\left(p_1 A_1^4 + p_2 A_2^4\right) - 12\left(p_1 A_1^2 + p_2 A_2^2\right)^2 > 0 \qquad (21)$$

to satisfy the OBM condition. Together with $p_1 + p_2 = 1/2$, we have constructed a particular distribution with $A_1 = 1, A_2 = 3.745457, p_1 = 0.46$ and $p_2 = 0.04$. Its variance and normalized kurtosis are respectively 2.04 and 0.995.

However, according to (7),

$$\kappa_i = (1 - (2p_1 \tanh(A_1)^2 + 2p_2 \tanh(A_2)^2)) \cdot 2\left(p_1 A_1^2 + p_2 A_2^2\right) - 1 \approx -0.211$$

is negative, so (9) cannot be satisfied. Hence, no separating matrix can be stable. Furthermore, we have performed numerical simulation experiments with random mixtures of two sources generated from this distribution, with $\mathbf{W}$ initialized to be $\mathbf{A}^{-1}$. The natural gradient adaptive algorithm always converged to a spurious solution, which left the entries of $\mathbf{WA}$ almost equal in absolute value.

So far we have discussed ML-type algorithms with fixed model pdf's, or equivalently fixed nonlinearities for an adaptive implementation. We have not touched the circumstances when the model pdf is parameterized, i.e., ICA algorithms with flexible density model, such as [5], [12]. Whether such flexibility in source pdf modeling can avoid spurious solutions caused by inappropriate modeling needs further investigation.

Furthermore, for the information theoretic or entropy contrast function which does not assume any source model, the existence of spurious solutions has been identified [13]. Besides, there are other ICA approaches that have been proved to be free from spurious solutions (see, e.g., [14] and [15]).

V. CONCLUSION

We have presented some analyses on the ML-type objective function for the ICA with predetermined source model pdf's. It is shown by experimental evidences that spurious solutions can exist for both regular ML and ML under whiteness constraint, even if the kurtosis signs of the model pdfs' and those of the true sources are consistent. This makes clear that the one-bit-matching condition is not sufficient for an ML algorithm to get a separating solution and thus the one-bit-matching conjecture is untrue.

REFERENCES

[1] P. Comon, "Independent component analysis—A new concept?,"