LETTER Communicated by Max Welling

One-Bit-Matching Conjecture for Independent Component Analysis

Zhi-Yong Liu
zyliu@cse.cuhk.edu.hk
Kai-Chun Chiu
kcchiu@cse.cuhk.edu.hk
Lei Xu
lxu@cse.cuhk.edu.hk
Department of Computer Science and Engineering, Chinese University of Hong Kong, Shatin, New Territories, Hong Kong

The one-bit-matching conjecture for independent component analysis (ICA) could be understood from different perspectives but is basically stated as "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 probability density functions (pdf's) and the kurtosis signs of all model pdf's" (Xu, Cheung, & Amari, 1998a). This conjecture has been widely believed in the ICA community and implicitly supported by many ICA studies, such as the Extended Infomax (Lee, Girolami, & Sejnowski, 1999) and the soft switching algorithm (Welling & Weber, 2001). However, there is no mathematical proof to confirm the conjecture theoretically. In this article, only skewness and kurtosis are considered, and such a mathematical proof is given under the assumption that the skewness of the model densities vanishes. Moreover, empirical experiments are demonstrated on the robustness of the conjecture as the vanishing skewness assumption breaks. As a by-product, we also show that the kurtosis maximization criterion (Moreau & Macchi, 1996) is actually a special case of the minimum mutual information criterion for ICA.

1 Introduction

Independent component analysis (ICA) aims at blindly separating the independent sources $\mathbf{s}$ from their linear mixture $\mathbf{x} = \mathbf{A}\mathbf{s}$ via

$$\mathbf{y} = \mathbf{W}\mathbf{x},\ \mathbf{x} \in \mathbb{R}^m,\ \mathbf{y} \in \mathbb{R}^n,\ \mathbf{W} \in \mathbb{R}^{m \times n}. \tag{1.1}$$

The recovered $\mathbf{y}$ is required to be as component-wise independent as possible where independence is defined as

$$q(\mathbf{y}) = \prod_{j=1}^{n} q(y^{(j)}). \tag{1.2}$$

This effort is supported by Tong, Inouye, and Liu (1993). They showed that $\mathbf{y}$ recovers $\mathbf{s}$ up to constant scales and a permutation of components when the components of $\mathbf{y}$ become component-wise independent and at most one of them is gaussian. The problem is further formalized by Comon (1994) under the name ICA.

Although ICA has been studied from different perspectives, such as the minimum mutual information (MMI) (Bell & Sejnowski, 1995; Amari, Cichocki, & Yang, 1996) and maximum likelihood (ML) (Cardoso, 1997), in the case that $\mathbf{W}$ is invertible, all such approaches are equivalent to minimizing the following cost function,

$$D = -H(\mathbf{y}) - \sum_{i=1}^{n} \int p_{\mathbf{W}}(y_i; \mathbf{W}) \log p_i(y_i)\, dy_i \tag{1.3}$$

where

$$H(\mathbf{y}) = -\int p(\mathbf{y}) \log p(\mathbf{y})\, d\mathbf{y}$$

is the entropy of $\mathbf{y}$, $p_i(y_i)$ is the predetermined model probability density function (pdf), and $p_{\mathbf{W}}(y_i; \mathbf{W})$ is the distribution on $\mathbf{y} = \mathbf{W}\mathbf{x}$. However, with each model pdf $p_i(y_i)$ prefixed, this approach works only for the cases that the components of $\mathbf{y}$ are either all subgaussians (Amari et al., 1996) or all supergaussians (Bell & Sejnowski, 1995).

In the cases that sources of supergaussian and subgaussian coexist in a unknown manner, each model pdf $p_i(y_i)$ is suggested to be a flexibly adjustable density that is learned together with $\mathbf{W}$, with the help of either a mixture of sigmoid functions that learns the cumulative distribution function (cdf) of each source (Xu, Yang, & Amari, 1996; Xu, Cheung, Yang, & Amari, 1997) or a mixture of parametric pdf's (Xu, 1997; Xu, Cheung, & Amari, 1998b). A so-called learned parametric mixture–based ICA (LPM-ICA) algorithm is derived, with successful results on the sources that can be either subgaussian or supergaussian, as well as any combination of both types (Xu et al., 1997, 1998b). The mixture model was also adopted for the ICA algorithms by Pearlmutter and Parra (1996), although it did not explicitly target separating the mixed sub- and supergaussian sources. Later, Attias (1999) also studied the mixture model-based ICA, which is regarded as a noise-free degeneration of the independent factor analysis (IFA) model.

Interestingly it has also been found that a rough estimate of each source pdf or cdf may be enough for source separation. For instance, a simple sigmoid function such as $\tanh(x)$ seems to work well on the supergaussian sources (Bell & Sejnowski, 1995), and a mixture of only two or three gaussians may be enough already (Xu, Cheung, & Amari, 1998a; Xu et al., 1998b) for the mixed sub- and supergaussian sources. This leads to the so-called one-bit-matching conjecture (Xu et al., 1998a), 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." In recent years, this conjecture has also been implicitly supported by several other ICA studies (Girolami, 1998; Everson & Roberts, 1999; Lee, Girolami, & Sejnowski, 1999; Welling & Weber, 2001).

Although the one-bit-conjecture was widely accepted in the ICA community, there is no theoretical proof for it. In literature, a mathematical proof (Cheung & Xu, 2000) was given for the case involving only two subgaussian sources, but the result cannot be extended to a model with more than two sources or with mixed sub- and supergaussian sources. Moreover, to guarantee a general adaptive ICA algorithm that is stable at the correct separation points, the constraints for the nonlinear function $\varphi_i(y_i) = -\frac{d}{dy_i}\log p_i(y_i)$ were studied by Amari and Chen (1997), but it did not touch the circumstance under which the sources can be separated.

When only skewness and kurtosis are under consideration, this letter provides a mathematical proof on the one-bit-matching conjecture under the assumption of zero skewness for the model pdf's. The entire proof proceeds in three stages. First, the observed mixture and recovered source signals become prewhitened with zero mean and identity covariance matrix. Next, an equivalence is established between minimization of the cost function (1.3) and maximization of a weighted sum of matching scores between kurtoses of source pdf's and model pdf's. Finally, we show that maximizing the weighted sum will recover the sources up to a permutation and sign indeterminacy. Meantime, as a by-product, we also show that the kurtosis maximization (Moreau & Macchi, 1996)–based ICA can be taken as a special case of the ICA based on equation 1.3.

The rest of the letter is organized in the following way. Section 2 is devoted to a detailed proof of the conjecture. Section 3 empirically demonstrates the robustness of the conjecture against the vanishing skewness assumption for the model pdf's, and section 4 concludes the letter.

2 A Theorem on the One-Bit-Matching Conjecture

In this section, we prove the theorem on the one-bit-matching conjecture according to the three stages described above.

Lemma 1. *Assume the independent sources* $\mathbf{s}$, *observed samples* $\mathbf{x}$, *and the output* $\mathbf{y}$ *are all prewhitened with zero mean and identity covariance matrix. We have*

$$\gamma_i^m = \sum_{j=1}^{n} r_{ij}^3 \gamma_j^s \tag{2.1}$$

$$v_i^m = \sum_{j=1}^{n} r_{ij}^4 v_j^s, \tag{2.2}$$

where γ_i^m, v_i^y, γ_j^s, and v_j^s denote the skewness and kurtosis of y_i, and the skewness and kurtosis of the source s_j, respectively, and $(r_{ij})_{n\times n}$ is an orthonormal matrix.

Proof. Based on the prewhitened assumption, we have $\mathbf{y} = \mathbf{Wx} = \mathbf{WAs} = \mathbf{Rs}$ with

$$E(\mathbf{y}\mathbf{y}^T) = \mathbf{R}E(\mathbf{s}\mathbf{s}^T)\mathbf{R}^T \Rightarrow \mathbf{R}\mathbf{R}^T = \mathbf{I}.$$

That is, $\mathbf{R} = (r_{ij})_{n\times n}$ is an orthonormal matrix. Provided that $\mathbf{s} = [s_1, s_2, \ldots, s_n]^T$ are component-wise independent with $E(s_i) = 0$ and $E(s_i^2) = 1$, we can obtain

$$\gamma_i^y = E\left[\left(\sum_{j=1}^{n} r_{ij}s_j\right)^3\right] = \sum_{j=1}^{n} r_{ij}^3 E(s_j^3) = \sum_{j=1}^{n} r_{ij}^3 \gamma_j^s \tag{2.3}$$

$$\begin{aligned} v_i^y &= E(y_i^4) - 3 = E\left[\left(\sum_{j=1}^{n} r_{ij}s_j\right)^4\right] - 3 \\ &= \sum_{j=1}^{n} r_{ij}^4 E(s_j^4) + 6\sum_{j=1}^{n-1}\sum_{r=j+1}^{n} r_{ij}^2 r_{ir}^2 - 3\left(\sum_{j=1}^{n} r_{ij}^2\right)^2 \\ &= \sum_{j=1}^{n} r_{ij}^4 E(s_j^4) - 3\sum_{j=1}^{n} r_{ij}^4 = \sum_{j=1}^{n} r_{ij}^4 (E(s_j^4) - 3) = \sum_{j=1}^{n} r_{ij}^4 v_j^s, \end{aligned} \tag{2.4}$$

where $\sum_{j=1}^{n} r_{ij}^2 = 1$ due to the orthonormality of $\mathbf{R}$, γ_i^y and v_i^y, respectively, denote the skewness and kurtosis of y_i, which in practice are computed based on the samples; and γ_j^s and v_j^s denote the skewness and kurtosis of the source s_j, respectively.

Meanwhile, the orthonormality of $\mathbf{R}$ further results in the entropy $H(\mathbf{y}) = H(\mathbf{s})$ in equation 1.3 being a constant. Thus, minimizing 1.3 is equivalent to maximizing

$$\hat{D} = \sum_{i=1}^{n} \int p_{\mathbf{W}}(y_i; \mathbf{W}) \log p_i(y_i)\, dy_i, \tag{2.5}$$

where $p_{\mathbf{W}}(y_i; \mathbf{W})$ is obtained via $\mathbf{y} = \mathbf{Rs}$. Based on the truncated Gram-Charlier expansion (Stuart & Ord, 1994) up to kurtosis, we can then get the

following approximation for $p_{\mathbf{W}}(y_i; \mathbf{W})$,

$$p_{\mathbf{W}}(y_i; \mathbf{W}) \approx g(y_i)\left(1 + \frac{\gamma_i^y}{6}H_3(y_i) + \frac{\nu_i^y}{24}H_4(y_i)\right), \tag{2.6}$$

where $g(y_i)$ denotes the standard gaussian distribution density as $g(y_i) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{y_i^2}{2})$, γ_i^m and ν_i^y are given by equations 2.3 and 2.4, respectively, and the Chebyshev-Hermite polynomials $H_3(y_i)$ and $H_4(y_i)$ are defined as follows:

$$H_3(y_i) = y_i^3 - 3y_i \tag{2.7}$$

$$H_4(y_i) = y_i^4 - 6y_i^2 + 3. \tag{2.8}$$

We choose the Gram-Charlier expansion because it clearly shows how the higher-order statistics affect the pdf and also the polynomials involved, that is, $H_3(y_i)$ and $H_4(y_i)$, have an orthogonal property (Stuart & Ord, 1994).

Theorem 1. *When only the skewness and kurtosis are under consideration and when the skewness of the model pdf's is zero,*

1. *Maximizing equation 2.5 is equivalent to maximizing*

$$\sum_{i=1}^{n}\sum_{j=1}^{n} r_{ij}^4 \nu_j^s k_i^m \tag{2.9}$$

where k_i^m is defined as

$$k_i^m = \int g(y_i)\frac{H_4(y_i)}{24}\log\left(1 + \frac{\nu_i^m}{24}H_4(y_i)\right)dy_i.$$

2. *k_i^m is a constant that possesses the same sign as ν_i^m.*

Proof. Taking a zero skewness, the prefixed model pdf $p_i(y_i)$ in equation 2.5 can be approximated by the following truncated Gram-Charlier expansion,

$$p_i(y_i) \approx g(y_i)\left(1 + \frac{\nu_i^m}{24}H_4(y_i)\right) \tag{2.10}$$

where ν_i^m denotes the kurtosis of $p_i(y_i)$.

Putting equations 2.10 and 2.6 into 2.5, maximizes the following cost function,

$$\begin{aligned}
J(\mathbf{R}) &= \sum_{i=1}^{n} \int g(y_i) \left(1 + \frac{\gamma_j^y}{6} H_3(y_i) + \frac{\nu_i^y}{24} H_4(y_i)\right) \\
&\quad \times \log\left(g(y_i)\left(1 + \frac{\nu_i^m}{24} H_4(y_i)\right)\right) dy_i \\
&= \sum_{i=1}^{n} \int g(y_i) \left(1 + \frac{\gamma_j^y}{6} H_3(y_i) + \frac{\nu_i^y}{24} H_4(y_i)\right) \log\left(1 + \frac{\nu_i^m}{24} H_4(y_i)\right) dy_i \\
&\quad - \frac{n}{2}(1 + \log 2\pi) \\
&= \sum_{i=1}^{n} \int g(y_i) \frac{\nu_i^y}{24} H_4(y_i) \log\left(1 + \frac{\nu_i^m}{24} H_4(y_i)\right) dy_i \\
&\quad + \sum_{i=1}^{n} \int g(y_i) \frac{\gamma_i^y}{6} H_3(y_i) \log\left(1 + \frac{\nu_i^m}{24} H_4(y_i)\right) dy_i \\
&\quad + \sum_{i=1}^{n} \int g(y_i) \log\left(1 + \frac{\nu_i^m}{24} H_4(y_i)\right) dy_i - \frac{n}{2}\left(1 + \log 2\pi\right) \\
&= \sum_{i=1}^{n} \nu_i^y \int g(y_i) \frac{H_4(y_i)}{24} \log\left(1 + \frac{\nu_i^m}{24} H_4(y_i)\right) dy_i \\
&\quad + C - \frac{n}{2}(1 + \log 2\pi), \qquad (2.11)
\end{aligned}$$

where, because the term

$$\sum_{i=1}^{n} \int g(y_i) \frac{\gamma_i^y}{6} H_3(y_i) \log\left(1 + \frac{\nu_i^m}{24} H_4(y_i)\right) dy_i = 0,$$

and since the parameter under consideration is $\nu_i^y = \sum_{j=1}^{n} r_{ij}^4 \nu_j^s$, the term

$$\sum_{i=1}^{n} \int g(y_i) \log\left(1 + \frac{\nu_i^m}{24} H_4(y_i)\right) dy_i$$

can be treated as a constant C with respect to $\mathbf{R}$. Thus, the problem is further simplified as maximizing the following cost function $\hat{J}(\mathbf{R})$,

$$\hat{J}(\mathbf{R}) = \sum_{i=1}^{n} \nu_i^y \int g(y_i) \frac{H_4(y_i)}{24} \log\left(1 + \frac{\nu_i^m}{24} H_4(y_i)\right) dy_i$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} r_{ij}^4 v_j^s \int g(y_i)\frac{H_4(y_i)}{24}\log\left(1+\frac{v_i^m}{24}H_4(y_i)\right)dy_i$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} r_{ij}^4 k_i^m v_j^s, \tag{2.12}$$

where

$$k_i^m \triangleq \int g(y_i)\frac{H_4(y_i)}{24}\log\left(1+\frac{v_i^m}{24}(H_4(y_i))\right)dy_i. \tag{2.13}$$

Note the three terms $g(y_i)$, $\frac{H_4(y_i)}{24}$ and $\log(1+\frac{v_i^m}{24}H_4(y_i))$ involved in the integration in equation 2.13. The first standard gaussian $g(y_i) > 0$. For the last two terms, their product has the same sign as that of v_i^m whenever $H_4(y_i) > 0$ or < 0. Thus, the sign of the product of the three terms is always the same as v_i^m (except for the four isolated points of y_i that cause the product zero), and this then makes the constant k_i^m have the same sign as v_i^m.

As a by-product, theorem 1 also reveals that the conventional kurtosis maximization criterion can be taken as a special case of the MMI criterion given by equation 1.3. As shown in equation 2.12, by setting $v_1^m = v_2^m = \cdots = v_n^m = v^m$, which implies that $k_1^m = k_2^m = \cdots = k_n^m = k^m$, we have

$$\hat{J}(\mathbf{R}) = \sum_{i=1}^{n}\sum_{j=1}^{n} r_{ij}^4 k_i^m v_j^s = k^m \sum_{i=1}^{n}\sum_{j=1}^{n} r_{ij}^4 v_j^s = k^m \sum_{i=1}^{n} v_i^y. \tag{2.14}$$

This is exactly the kurtosis maximization criterion (Moreau & Macchi, 1996) by setting either k^m as 1 for supergaussian sources or -1 for subgaussian sources. Such a linkage was also observed by Cardoso (1999) under the following much stricter condition:

$$\log\left(1+\frac{v_i^m}{24}H_4(y_i)\right) \approx \frac{v_i^m}{24}H_4(y_i).$$

So far we have shown that the objective function 2.9 is equivalent to equation 2.5 when only the skewness and kurtosis are under consideration and the model skewness vanishes. Then we proceed to prove that maximizing equation 2.9 recovers the sources up to permutation and sign indeterminacy. The result is summarized by the following one-bit-matching theorem.

Theorem 2. *When only the skewness and kurtosis are under consideration and when the model skewness vanishes, 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. That is, maximization of equation 2.9 can be reachable only by an identity matrix up to permutation and sign indeterminacy.*

Proof. Let $\mathbf{R}_{n\times n} = [\mathbf{r}_1, \ldots, \mathbf{r}_n]^T$ with $\mathbf{r}_i = [r_{i1}, \ldots, r_{in}]$ be an orthonormal matrix, the objective (2.9) can be rewritten as follow,

$$\begin{aligned} &\max \quad \hat{J}(\mathbf{R}) \;\; \text{s.t.} \quad \mathbf{R}^T\mathbf{R} = \mathbf{I} \\ &\hat{J}(\mathbf{R}) = \sum_{i=1}^{n} \hat{J}(\mathbf{r}_i), \quad \hat{J}(\mathbf{r}_i) = k_i^m \sum_{j=1}^{n} r_{i1}^4 v_j^s \end{aligned} \tag{2.15}$$

Consider an ordering $(i_1, \ldots, i_n)$ that is any permutation of $(1, \ldots, n)$, and consider the constraint $\mathbf{c} = \{c_{i_j}\}_{j=1}^n$, where $c_{i_j} = \{\mathbf{r}_{i_j}$ is orthogonal to $\mathbf{W}_{i_j}, \mathbf{W}_{i_j} = [\mathbf{r}_{i_1}, \ldots, \mathbf{r}_{i_{j-1}}]^T, \|\mathbf{r}_{i_j}\|^2 = 1\}$ $(j = 1, \ldots, n)$ with c_{i_1} degenerated to $\|\mathbf{r}_{i_1}\|^2 = 1$. There are totally $n!$ such constraints that are jointly equivalent to the orthonormality constraint $\mathbf{R}^T\mathbf{R} = \mathbf{I}$. Thus, the maximization of $\hat{J}(\mathbf{R})$ under the orthonormality constraint is equivalent to the maximization of $\hat{J}(\mathbf{R})$ under these $n!$ constraints all together. Since the subscript of k_i^m is the same as $\mathbf{r}_i$ in (2.15), the problem is equivalent to consider the $n!$ orderings $\mathbf{k} \in \mathcal{K}$ of $\{k_i^m\}_{i=1}^n$ under the fixed ordering of $\{c_i\}_{i=1}^n$. Thus, we have,

$$\max_{\text{s.t. } \mathbf{R}^T\mathbf{R}=\mathbf{I}} \{\hat{J}(\mathbf{R})\} = \max_{\mathbf{k}\in\mathcal{K}} \left\{ \sum_{i=1}^{n} \max_{\text{s.t. } c_i} \left\{ \hat{J}(\mathbf{r}_i) \right\} \right\} \tag{2.16}$$

where, for each $\mathbf{k} \in \mathcal{K}$, the maximization is implemented sequentially from the first term to the last term in $\{\sum_{i=1}^n \max_{\text{s.t. } c_i}\{\hat{J}(\mathbf{r}_i)\}\}$.

For every $\mathbf{k}$, under the one-bit-matching condition, there is at least one v_j^s $(1 \le j \le n)$ that possesses the same sign as k_1^m. Since $\sum_{j=1}^n r_{1j}^2 = 1$, $\max\{\hat{J}(\mathbf{r}_1)\} = k_1^m v_{l^{(1)}}^s$ is reached at $\mathbf{r}_i = [0, \ldots, r_{il^{(1)}}, \ldots, 0]$ with $r_{il^{(1)}} = \pm 1$, where $l^{(1)} = \max_i\{v_i^s\}_{i=1}^n$ if $k_1^m > 0$; otherwise, $l^{(1)} = \min_i\{v_i^s\}_{i=1}^n$ if $k_1^m < 0$. Then, under the constraints of $\sum_{j=1}^n r_{2j}^2 = 1$ and $\mathbf{r}_2 \perp \mathbf{r}_1$ that implies $r_{2l^{(1)}} = 0$, $\max\{\hat{J}(\mathbf{r}_2)\} = k_2^m v_{l^{(2)}}^s$, where $l^{(2)} = \max_{i\neq l^{(1)}}\{v_i^s\}_{i=1}^n$ if $k_2^m > 0$; otherwise, $l^{(2)} = \min_{i\neq l^{(1)}}\{v_i^s\}_{i=1}^n$ if $k_1^m < 0$. The one-bit-matching condition can guarantee that the above process can be sequentially proceeded until we get $\max\{\hat{J}(\mathbf{r}_n)\} = k_n^m v_{l^{(n)}}^s$. As a result, $\sum_{i=1}^n \max_{\text{s.t. } c_i}\{\hat{J}(\mathbf{r}_i)\} = \hat{J}(\Pi_{\mathbf{k}}) = \sum_{i=1}^n k_i^m v_{l^{(i)}}^s$, where $\Pi_{\mathbf{k}}$ is a permutation matrix. Moreover, considering all the orderings in $\mathcal{K}$, we can finally reach a set $\{\Pi_{\mathbf{k}}\}_{\mathbf{k}\in\mathcal{K}}$ of at most $n!$ permutation matrices.

We further find one Π such that $\hat{J}(\Pi) = \max_{k\in\mathcal{K}}\{\hat{J}(\Pi_{\mathbf{k}})\}$, for which we show $\Pi = \mathbf{I}$ that corresponds to the particular ordering of $k_1^m > \cdots > k_n^m$ and $v_1^s > \cdots > v_n^s$. Such an ordering has the following nature,

$$k_i^m v_i^s + k_j^m v_q^s - k_i^m v_q^s - k_j^s v_i^s = (k_i^m - k_j^m)(v_i^s - v_q^s) > 0 \;\; \text{for} \;\; j, q > i \tag{2.17}$$

Now consider any $\Pi = (\pi_{ij})_{n\times n} \in \{\Pi_{\mathbf{k}}\}_{\mathbf{k}\in\mathcal{K}}$, $\Pi \neq \mathbf{I}$. If $\pi_{11}^4 = 1$, we directly go to consider π_{22}; otherwise, there must exist $\pi_{1j}^4 = 1$ and $\pi_{i1}^4 = 1$ for $i, j > 1$.

We modify Π to $\Pi^{(1)}$ by letting $\pi_{11}^{(1)} = 1, \pi_{ij}^{(1)} = 1, \pi_{i1}^{(1)} = 0, \pi_{1j}^{(1)} = 0$ such that $\hat{J}(\Pi) < \hat{J}(\Pi^{(1)})$ due to (2.17). Continuing the same process, we get that $\hat{J}(\Pi) < \hat{J}(\Pi^{(1)}) < \cdots < \hat{J}(\Pi^{(n)})$ with $\Pi^{(n)} = \mathbf{I}$.

In a summary, for any orthonormal matrix $\mathbf{R} \neq \mathbf{I}$ we can conclude that $\hat{J}(\mathbf{R}) < \hat{J}(\mathbf{I})$. In a special case as, for instance, $k_i^m = k_{i+1}^m$, the maximization of $\hat{J}(\mathbf{R})$ can be also reached by the permutation matrix $\hat{\Pi}$ that is obtained by switching the ith and $(i+1)$th row of $\mathbf{I}$. Similarly, when there are more two k_i^m equal to each other, the maximization of $\hat{J}(\mathbf{R})$ can be also reached by the permutation matrix obtained by switching the corresponding rows. In other words, the maximization of $\hat{J}(\mathbf{R})$ is also reachable by such permutation matrices in these special cases.

Remark 1. In the proof above it can be observed that the one-bit-matching condition takes a key role. Without it, we cannot always ensure that there is a $v_{q_i}^s$ that possesses the same sign as k_i^m, and thus maximization of $\hat{J}(\mathbf{r}_i)$ cannot cause the $\mathbf{r}_i$ take the form $[0, \ldots, \pm 1, \ldots, 0]$.

For example, as $k_1^m = 3, k_2^m = 2, k_3^m = -1$ and $v_1^s = 1, v_2^s = -4, v_3^s = -5$, the maximum of (2.9) is $\frac{5}{3}$ obtained by setting $(r_{ij}^4)_{n\times n} = \begin{pmatrix} 4/9 & 1/9 & 0 \\ 1/9 & 4/9 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ instead of a permutation matrix.

Remark 2. The proof above consists of a continuous optimization that leads to one permutation matrix and a combinatorial optimization that reaches $\hat{J}(\mathbf{I})$. Actually, any permutation matrix already corresponds to a separable solution for ICA. This also means that a local maximization of $\hat{J}(\mathbf{R})$ that leads to a permutation can provide a successful solution for ICA already. In fact, it explains the success of the gradient-based algorithms for a continuous optimization such as the natural gradient algorithm (Amari et al. 1996).

3 An Empirical Study on the Conjecture with Nonvanishing Model Skewness

The one-bit-matching theorem is proved under the assumption that the model skewness vanishes. Although the assumption is quite weak in reality (for instance, any symmetric pdf satisfies the assumption of zero skewness), it is still interesting to study the case when the assumption of zero model skewness is not satisfied. We give some empirical evidence for it based on two experiments.

3.1 On Demonstrating the Case with Small Model Skewness. This experiment aims to demonstrate the case as the model densities with small skewness are chosen. The experiment is based on a two-source model, with the 50,000 source samples s_i generated by the following pdf approximation,

$$p(s) = g(s)\left(1 + \frac{\gamma^s}{6}H_3(s) + \frac{\nu^s}{24}H_4(s)\right) \tag{3.1}$$

where γ^s and ν^s denote the skewness and kurtosis, respectively. The observed samples $\mathbf{x}$ are mixed from $\mathbf{s}$ by the rotation matrix

$$\mathbf{A} = \begin{pmatrix} \cos(\xi) & -\sin(\xi) \\ \sin(\xi) & \cos(\xi) \end{pmatrix}$$

with $\xi = \pi/6$ as follows,

$$\mathbf{x} = \mathbf{A}\mathbf{s}. \tag{3.2}$$

Based on the output $\mathbf{y}$ obtained by a rotation of $\mathbf{x}$ via

$$\mathbf{y} = \mathbf{W}\mathbf{x} = \begin{pmatrix} \cos(\phi) & -\sin(\phi) \\ \sin(\phi) & \cos(\phi) \end{pmatrix}\mathbf{x}, \tag{3.3}$$

objective 2.5 can be empirically obtained as follows,

$$\phi = \arg\max_{\phi} \hat{D}(\phi),\ \hat{D}(\phi) = \frac{1}{50000}\sum_{i=1}^{2}\sum_{t=1}^{50000}\log p_i(y_i(t)), \tag{3.4}$$

where the model pdf $p_i(y_i(t))$ is constrained in the approximate form

$$p(y) = g(y)\left(1 + \frac{\gamma^m}{6}H_3(y) + \frac{\nu^m}{24}H_4(y)\right), \tag{3.5}$$

with γ^y and ν^y denoting the skewness and kurtosis, respectively. Typically, in the experiment, we demonstrate the following three cases:

Case 1: $\gamma_1^s = \gamma_2^s = 0$, and $\gamma_1^m = 0.2$, $\gamma_2^m = -0.2$

Case 2: $\gamma_1^s = 0.2$, $\gamma_2^s = -0.2$, and $\gamma_1^m = \gamma_2^m = 0$

Case 3: $\gamma_1^s = 0.2$, $\gamma_2^s = -0.2$, and $\gamma_1^m = 0.2$, $\gamma_2^m = -0.2$

Actually, the proved theorem includes case 2, which we use here for comparison. Meanwhile, in all cases we fix $\nu_1^s = \nu_1^m = 0.5$ and $\nu_2^s = \nu_2^m = 1$.

The $\hat{D}(\phi)$ versus the $\phi \in [0, 2\pi]$ obtained for the three cases are shown in Figures 1, 2, and 3, respectively.

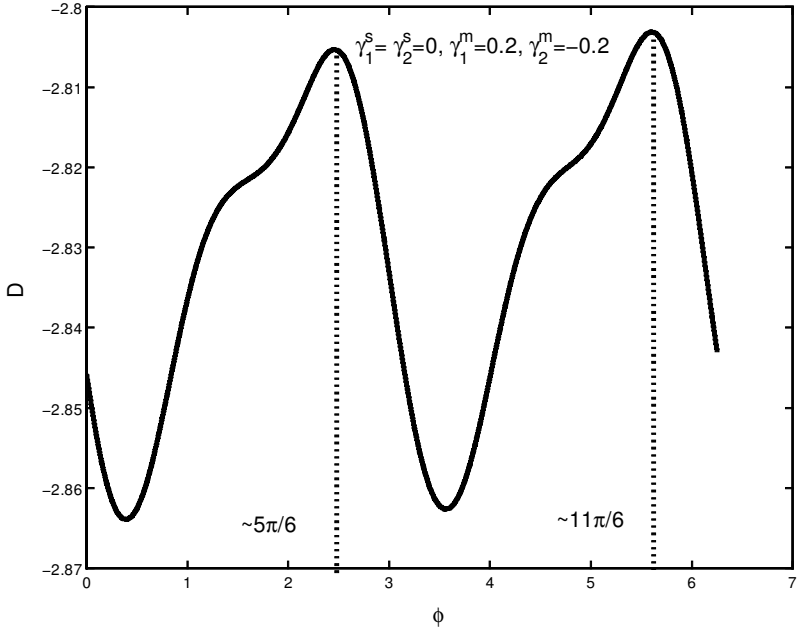

Figure 1: $\hat{D}$ vs. θ for case 1.

From the experiment results, we see that the obtained ϕ's corresponding to the maximized $\hat{D}$ in all three cases are around $5\pi/6$ or $11\pi/6$, which then makes the rotation matrix $\mathbf{R} = \mathbf{WA}$ approximately equal to $-\mathrm{I}_2$ or I_2, respectively. That is, all of the three settings recovered the original two sources up to sign indeterminacy, even when the model pdf's are with a small nonzero skewness, as in cases 1 and 3.

3.2 On Demonstrating the Breakpoint of the Conjecture. We proceed to demonstrate the conjecture by gradually increasing the skewness of the model densities to get empirical evidence regarding to what extent the conjecture holds or breaks down in practice. In this experiment, the 50,000 source samples of s_1 and s_2 are generated by the following mixture of two gaussians,

$$p(y \mid \theta) = 0.5G(y \mid \theta, (2-\theta)^2) + 0.5G(y \mid -\theta, 1), \qquad 0 \le \theta \le 1, \quad (3.6)$$

with $\theta_1 = 0.1$ and $\theta_2 = 0.9$, respectively, where $G(y \mid m, \sigma^2)$ denotes a gaussian pdf with mean m and variance σ^2. By a standard technique (Stuart & Ord, 1994), the skewness γ and kurtosis ν of the mixture of k gaussians

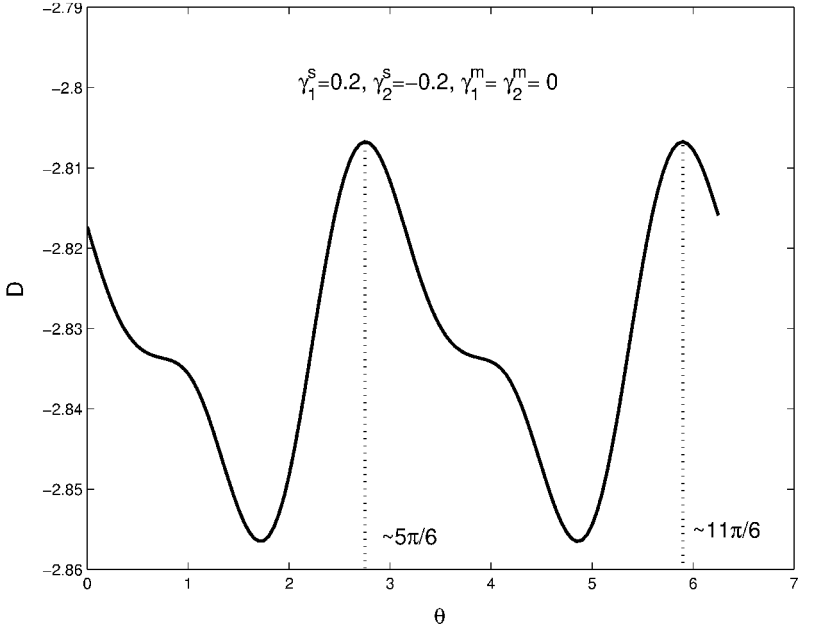

Figure 2: $\hat{D}$ vs. θ for case 2.

can be obtained as follows,

$$\begin{aligned}\gamma &= \mu_{30} + 3\mu_{12} + 2\mu_{10}^3 - 3\mu_{10}\mu_{02} - 3\mu_{10}\mu_{20}\\ \nu &= \mu_{40} + 6\mu_{22} + 3\mu_{04} + 12\mu_{10}^2\mu_{02} + 12\mu_{10}^2\mu_{20}\\ &\quad - 12\mu_{10}\mu_{12} - 4\mu_{10}\mu_{30} - 3\mu_{02}^2 - 3\mu_{20}^2 - 6\mu_{02}\mu_{20} - 6\mu_{10}^4, \end{aligned} \tag{3.7}$$

where $\mu_{pq} \triangleq \sum_{j=1}^{k} \alpha_j m_j^p \sigma_j^q$. In our case, they become, respectively,

$$\gamma = 1.5\theta^3 - 6\theta^2 + 4.5\theta, \tag{3.8}$$

$$\nu = -1.25\theta^4 - 6\theta^3 + 16.5\theta^2 - 18\theta + 6.75, \tag{3.9}$$

with the change of γ and k as θ varies as shown in Figure 4. As a result, the skewness γ and kurtosis ν of the two sources are as follows,

$$\gamma_1^s = 0.39,\ \nu_1^s = 5.11,\ \gamma_2^s = 0.28,\ \nu_2^s = -1.28.$$

That is, s_1 is supergaussian and s_2 is subgaussian. The observations $\mathbf{x}$ and

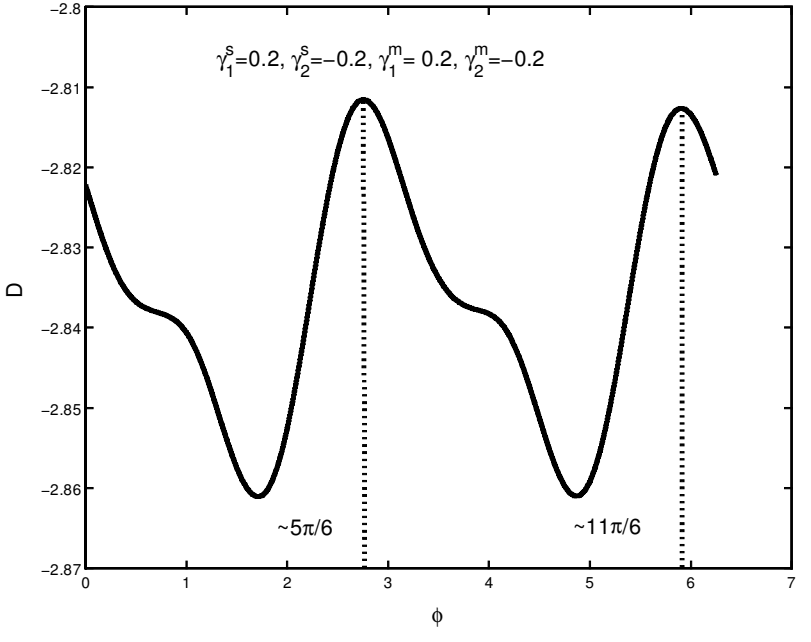

Figure 3: $\hat{D}$ vs. θ for case 3.

outputs **y** are obtained in the same way as in the previous experiment using equations 3.2 and 3.3, respectively, but now with $\xi = \frac{\pi}{4}$.

Meanwhile, the ϕ is also obtained using equation 3.4, and the two model pdf's $p_1(y_1 \mid \theta_1)$ and $p_2(y_2 \mid \theta_2)$ are still with the form given by equation 3.6 but with θ_1 increased from 0 to 0.6 using

$$\theta_1 = 0.6\tau,\ 0 \le \tau \le 1, \tag{3.10}$$

and θ_2 decreased from 1 to 0.7 using

$$\theta_2 = 1 - 0.3\tau,\ 0 \le \tau \le 1. \tag{3.11}$$

In this way, $p_1(y_1 \mid \theta_1)$ is constrained as a supergaussian while $p_2(y_2 \mid \theta_2)$ remains always subgaussian, and meanwhile, the increase of τ will result in the decrease of the kurtosis (absolute value) coupled with the increase of the skewness for both p_1 and p_2, as illustrated in Figure 6.

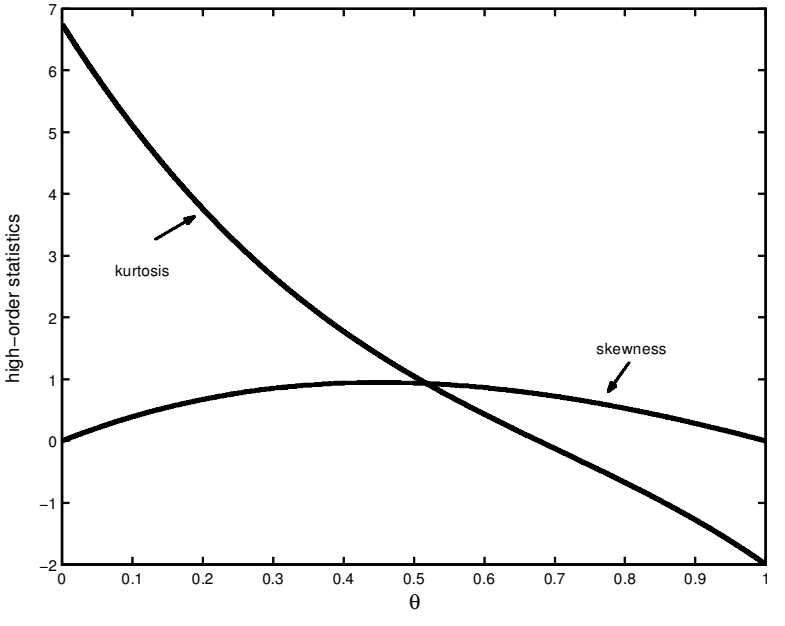

Figure 4: Skewness and kurtosis vs. θ

The performance of the recovery is measured by the following error metric (Amari et al., 1996),

$$E = \sum_{i=1}^{d} \left(\sum_{j=1}^{d} \frac{|p_{ij}|}{\max_k |p_{ik}|} - 1 \right) + \sum_{j=1}^{d} \left(\sum_{i=1}^{d} \frac{|p_{ij}|}{\max_k |p_{kj}|} - 1 \right), \tag{3.12}$$

where $\mathbf{P} \triangleq \mathbf{WA}$. The obtained error E versus τ is shown in Figure 5.

From Figure 5, we notice that the performance worsens as τ increases. Typically, if we set the threshold as $E = 0.40$, which corresponds to the matrix

$$\mathbf{P} = \mathbf{WA} = \begin{pmatrix} 0.9934 & -0.1144 \\ 0.1144 & 0.9934 \end{pmatrix},$$

source separation by ICA via maximizing equation 1.3 would fail when $\tau > 0.89$. This corresponds to

$$0.6 \geq \theta_1 \geq 0.53 \quad \text{and} \quad 0.7 \leq \theta_2 \leq 0.73$$

or

$$0.93 \geq \gamma_1^m \geq 0.86, \ 0.43 \leq \nu_1^m \leq 0.85$$

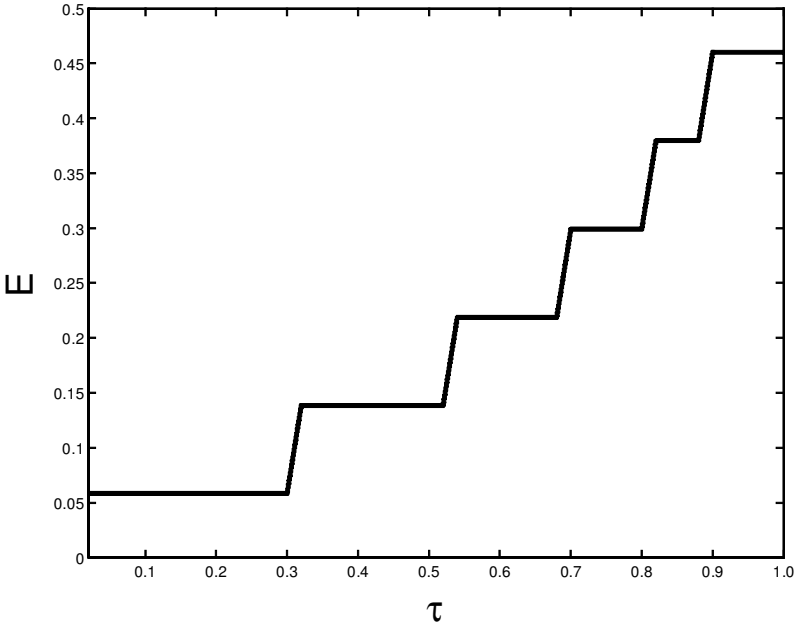

Figure 5: The obtained error E vs. τ.

and

$$0.73 \geq \gamma_2^m \geq 0.67, \ -0.12 \geq \nu_2^m \geq -0.29,$$

as illustrated by Figure 6.

Apparently, small kurtosis coupled with large skewness poses a threat to the one-bit-matching conjecture. An intuitive explanation for the breakdown of the conjecture in the experiment might be as follows. As the model skewness becomes relatively big enough, it can no longer, even roughly, guarantee that k_i^m in equation 2.9 possesses the same sign as ν_i^m, and thus make the one-bit-matching conjecture break down in practice.

4 Conclusions

When only skewness and kurtosis are under consideration, we have theoretically proved the so-called one-bit-matching conjecture for ICA under the assumption of varnishing skewness for the model pdf's. We also empirically demonstrated the robustness of conjecture against the vanishing skewness assumption for the model pdf's and, as a by-product, showed that the kurtosis maximization criterion is actually a special case of the MMI criterion.

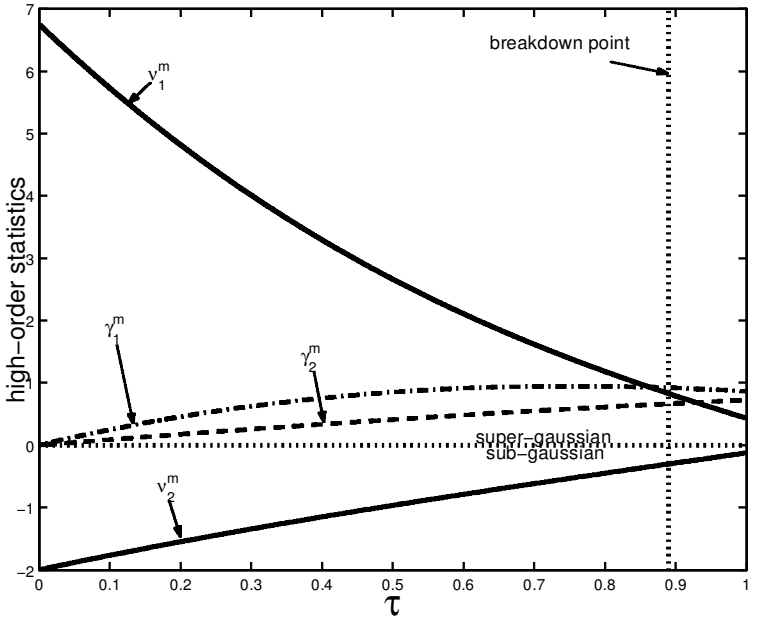

Figure 6: Illustration of the ICA performance as the model pdf varies.

Acknowledgments

We thank the anonymous reviewers for their valuable suggestions, which improved the original manuscript. Z.-Y. Liu thanks Jinwen Ma for some helpful discussions. The work described in this article was fully supported by a grant from the Research Grant Council of the Hong Kong SAR (Project No: CUHK 4336/02E).

References

Amari, S. I., & Chen, T. P. (1997). Stability analysis of adaptive blind source separation. *Neural Networks Letter, 10*, 1345–1351.

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). Cambridge, MA: MIT Press.

Attias, H. (1999). Independent factor analysis. *Neural Computation, 11*, 803–851.

Bell, A., & Sejnowski, T. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation, 7*, 1129–1159.

Cardoso, J. F. (1997). Infomax and maximum likelihood for source separation. *IEEE Letters on Signal Processing, 4*, 112–114.

Cardoso, J. F. (1999). High order constrasts for independent component analysis. *Neural Computation, 11*, 157–192.

Cheung, C. C., & Xu, L. (2000). Some global and local convergence analysis on the information-theoretic independent component analysis approach. *Neurocomputing, 30*, 79–102.

Comon, P. (1994). Independent component analysis—a new concept? *Signal Processing*, *36*, 287–314.

Everson, R., & Roberts, S. (1999). Independent component analysis: A flexible nonlinearity and decorrelating manifold approach. *Neural Computation*, *11*, 1957–1983.

Girolami, M. (1998). An alternative perspective on adaptive independent component analysis algorithms. *Neural Computation*, *10*, 2103–2114.

Lee, T. W., Girolami, M., & Sejnowski, T. J. (1999). Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. *Neural Computation*, *11*, 417–441.

Moreau, E., & Macchi, O. (1996). High order constrasts for self-adaptive source separation. *International Journal of Adaptive Control and Signal Processing*, *10*, 19–46.

Pearlmutter, B. A., & Parra, L. C. (1996). A context-sensitive genaralization of ICA. In *Proc. of Int. Conf. on Neural Information Processing*. Hong Kong: Springer-Verlag.

Stuart, A., & Ord, J. (1994). *Kendall's advanced theory of statistics, Vol. 1: Distribution theory*. London: Edward Arnold.

Tong, L., Inouye, Y., & Liu, R. (1993). Waveform-preserving blind estimation of multiple independent sources. *Signal Processing*, *41*, 2461–2470.

Welling, M., & Weber, M. (2001). A constrained EM algorithm for independent component analysis. *Neural Computation, 13*, 677–689.

Xu, L. (1997). Bayesian ying-yang learning based ICA models. *Proc. 1997 IEEE Signal Processing Society Workshop on Neural Networks for Signal Processing VII* (pp. 476–485). Florida.

Xu, L., Cheung, C. C., & Amari, S. I. (1998a). Further results on nonlinearity and separtion capability of a liner mixture ICA method and learned lpm. In C. Fyfe (Ed.), *Proceedings of the I&ANN'98* (pp. 39–45).

Xu, L., Cheung, C. C., & Amari, S. I. (1998b). Learned parametric mixture based ICA algorithm. *Neurocomputing, 22*, 69–80.

Xu, L., Cheung, C. C., Yang, H. H., & Amari, S. I. (1997). Independent component analysis by the information-theoretic approach with mixture of density. *Proc. of 1997 IEEE Intl. Conf on Neural Networks (IEEE-INNS IJCNN97)* (Vol. 3, pp. 1821–1826). Houston, TX.

Xu, L., Yang, H. H., & Amari, S. I. (1996). *Signal source separation by mixtures: accumulative distribution functions or mixture of bell-shape density distribution functions*. Research proposal presented at FRONTIER FORUM. Japan: Institute of Physical and Chemical Research.

Received January 2, 2003; accepted July 10, 2003.