Lecture Notes in Computer Science, Vol 3496: Advances in Neural Networks - ISNN 2005: A Plenary talk on Second International Symposium on Neural Networks, May 30 - June 1, 2005,

One-Bit-Matching ICA Theorem, Convex-Concave Programming, and Combinatorial Optimization*

Lei Xu

Department of Computer Science and Engineering,
The Chinese University of Hong Kong, Shatin, NT, Hong Kong
lxu@cse.cuhk.edu.hk

Abstract. Recently, a mathematical proof is obtained in (Liu, Chiu, Xu, 2004) on the so called one-bit-matching conjecture that all the sources can be separated as long as there is an 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), which is widely believed and implicitly supported by many empirical studies. However, this proof is made only in a weak sense that the conjecture is true when the global optimal solution of an ICA criterion is reached. Thus, it can not support the successes of many existing iterative algorithms that usually converge at one of local optimal solutions. In this paper, a new mathematical proof is obtained in a strong sense that the conjecture is also true when anyone of local optimal solutions is reached, in help of investigating convex-concave programming on a polyhedral-set. Theorems have also been proved not only on partial separation of sources when there is a partial matching between the kurtosis signs, but also on an interesting duality of maximization and minimization on source separation. Moreover, corollaries are obtained from the theorems to state that seeking a one-to-one same-sign-correspondence can be replaced by a use of the duality, i.e., super-gaussian sources can be separated via maximization and sub-gaussian sources can be separated via minimization. Also, a corollary is obtained to confirm the symmetric orthogonalization implementation of the kurtosis extreme approach for separating multiple sources in parallel, which works empirically but in a lack of mathematical proof. Furthermore, a linkage has been set up to combinatorial optimization from a Stiefel manifold perspective, with algorithms that guarantee convergence and satisfaction of constraints.

1 Introduction

Independent component analysis (ICA) aims at blindly separating the independent sources s from a unknown linear mixture $x = As$ via $y = Wx$. It has been shown in [18] that y recovers s up to constant scales and a permutation of components when the components of y become component-wise independent and at

* The work described in this paper was fully supported by a grant from the Research Grant Council of the Hong Kong SAR (Project No: CUHK4225/04E).

J. Wang, X. Liao, and Z. Yi (Eds.): ISNN 2005, LNCS 3496, pp. 5–20, 2005.

most one of them is gaussian. The problem is further formalized by Comon [7] under the name ICA. Although ICA has been studied from different perspectives, such as the minimum mutual information (MMI) [1, 4] and maximum likelihood (ML) [5], in the case that W is invertible, all such approaches are equivalent to minimizing the following cost function

$$D(W) = \int p(y;W) \ln \frac{p(y,W)}{\prod_{i=1}^{n} q(y_i)} dy, \tag{1}$$

where $q(y_i)$ is the pre-determined model probability density function (pdf), and $p(y,W)$ is the distribution on $y = Wx$. With each model pdf $q(y_i)$ prefixed, however, this approach works only for the cases that the components of y are either all sub-Gaussians [1] or all super-Gaussians [4].

To solve this problem, it is suggested that each model pdf $q(y_i)$ is a flexibly adjustable density that is learned together with W, with the help of either a mixture of sigmoid functions that learns the cumulative distribution function (cdf) of each source [24, 26] or a mixture of parametric pdfs [23, 25], and a so-called learned parametric mixture based ICA (LPMICA) algorithm is derived, with successful results on sources that can be either sub-Gaussian or super-Gaussian, as well as any combination of both types. The mixture model was also adopted in a so called context-sensitive ICA algorithm [17], although it did not explicitly target at separating the mixed sub- and super-Gaussian sources.

On the other hand, 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 super-Gaussian sources [4], and a mixture of only two or three Gaussians may be enough already [23] for the mixed sub- and super-Gaussian sources. This leads to the so-called one-bit-matching conjecture [22], which states that "all the sources can be separated as long as there is an 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 past years, this conjecture has also been implicitly supported by several other ICA studies [10, 11, 14, 19]. In [6], a mathematical analysis was given for the case involving only two sub-Gaussian sources. In [2], stability of an ICA algorithm at the correct separation points was also studied via its relation to the nonlinearity $\phi(y_i) = d \ln q_i(y_i)/dy_i$, but without touching the circumstance under which the sources can be separated.

Recently, the conjecture on multiple sources has been proved mathematically in a weak sense [15]. When only sources' skewness and kurtosis are considered with $Es = 0$ and $Ess^T = I$, and the model pdf's skewness is designed as zero, the problem $\min_W D(W)$ by eq.(1) is simplified via pre-whitening into the following problem

$$\max_{RR^T=I} J(R), \;\; J(R) = \sum_{i=1}^{n} \sum_{j=1}^{n} r_{ij}^4 \nu_j^s k_i^m, \;\; n \geq 2, \tag{2}$$

where $R = (r_{ij})_{n\times n} = WA$ is an orthonormal matrix, and ν_j^s is the kurtosis of the source s_j, and k_i^m is a constant with the same sign as the kurtosis ν_i^m of

the model $q(y_i)$. Then, it is further proved that the global maximization of eq. (2) can only be reachable by setting R a permutation matrix up to certain sign indeterminacy. That is, the one-bit-matching conjecture is true when the global minimum of $D(W)$ in eq.(1) with respect to W is reached. However, this proof still can not support the successes of many existing iterative ICA algorithms that typically implement gradient based local search and thus usually converge to one of local optimal solutions.

In the next section of this paper, all the local maxima of eq.(2) are investigated via a special convex-concave programming on a polyhedral set, from which we prove the one-bit-matching conjecture in a strong sense that it is true when anyone of local maxima by eq.(2) or equivalently local minima by eq.(1) is reached in help of investigating convex-concave programming on on a polyhedral-set. Theorems have also been provided on separation of a part of sources when there is a partial matching between the kurtosis signs, and on an interesting duality of maximization and minimization. Moreover, corollaries are obtained from theorems to state that the duality makes it possible to get super-gaussian sources via maximization and sub-gaussian sources via minimization. Another corollary is also to confirm the symmetric orthogonalization implementation of the kurtosis extreme approach for separating multiple sources in parallel, which works empirically but in a lack of mathematical proof [13].

In section 3, we further discuss that eq. (2) with R being a permutation matrix up to certain sign indeterminacy becomes equivalent to a special example of the following combinatorial optimization:

$$\begin{aligned} &\min_{V} E_o(V),\ V = \{v_{ij}, i = 1, \cdots, N, j = 1, \cdots, M\}, \quad \textit{subject to} \\ &C^c: \quad \sum_{i=1}^{N} v_{ij} = 1, j = 1, \cdots, M, \quad C^r: \quad \sum_{j=1}^{M} v_{ij} = 1, i = 1, \cdots, N; \\ &C^b: \quad v_{ij} \textit{ takes either } 0 \textit{ or } 1. \end{aligned} \tag{3}$$

This connection suggests to investigate combinatorial optimization from a perspective of gradient flow searching within the Stiefel manifold , with algorithms that guarantee convergence and constraint satisfaction.

2 One-Bit-Matching Theorem and Extension

2.1 An Introduction on Convex Programming

To facilitate mathematical analysis, we briefly introduce some knowledge about convex programming. A set in R^n is said to be *convex*, if $x_1 \in S, x_2 \in S$, we have $\lambda x_1 + (1 - \lambda)x_2 \in S$ for any $0 \leq \lambda \leq 1$. Shown in Fig.1 are examples of convex sets. As an important special case of convex sets, a set in R^n is called a *polyhedral set* if it is the intersection of a finite number of closed half-spaces, that is, $S = \{x : a_i^t x \leq \alpha_i,\ for\ i = 1, \cdots, m\}$, where a_i is a nonzero vector and α_i is a scalar for $i =, \cdots, m$. The second and third ones in Fig.1 are two examples. Let S be a nonempty convex set, a vector $x \in S$ is called an extreme

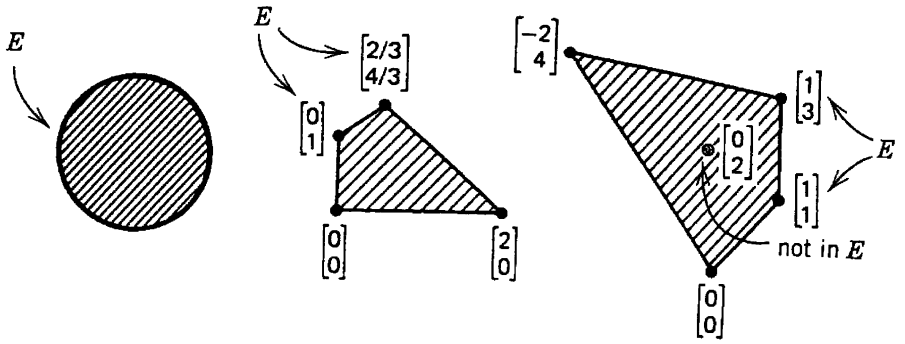

Fig. 1. Convex set and polyhedral set.

point of S if $x = \lambda x_1 + (1-\lambda)x_2$ with $x_1 \in S, x_2 \in S$, and $0 < \lambda < 1$ implies that $x = x_1 = x_2$. We denote the set of extreme point by E and illustrate them in Fig.1 by dark points or dark lines as indicated.

Let $f : S \to R$, where S is a nonempty convex set in R^n. As shown in Fig.1, the function f is said to be convex on S if

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2) \tag{4}$$

for $x_1 \in S, x_2 \in S$ and for $0 < \lambda < 1$. The function f is called *strictly convex* on S if the above inequality is true as a strict inequality for each distinct $x_1 \in S, x_2 \in S$ and for $0 < \lambda < 1$. The function f is called concave (strictly concave) on S if $-f$ is convex (strict convex) on S.

Considering an optimization problem $\min_{x \in S} f(x)$, if $\bar{x} \in S$ and $f(x) \geq f(\bar{x})$ for each $x \in S$, then $\bar{x}$ is called a global optimal solution. If $\bar{x} \in S$ and if there exists an ε-neighborhood $N_\varepsilon(\bar{x})$ around $\bar{x}$ such that $f(x) \geq f(\bar{x})$ for each $x \in S \cap N_\varepsilon(\bar{x})$, then $\bar{x}$ is called a local optimal solution. Similarly, if $\bar{x} \in S$ and if $f(x) > f(\bar{x})$ for all $x \in S \cap N_\varepsilon(\bar{x}), x \neq \bar{x}$, for some ε, then $\bar{x}$ is called a strict local optimal solution. Particularly, an optimization problem $\min_{x \in S} f(x)$ is called a *convex programming problem* if f is a convex function and S is a convex set.

Lemma 1

(a) Let S be a nonempty open convex set in R^n, and let $f : S \to R$ be twice differentiable on S. If its Hessian matrix is positive definite at each point in S, the f is strictly convex.

(b) Let S be a nonempty convex set in R^n, and let $f : S \to R$ be convex on S. Consider the problem of $\min_{x \in S} f(x)$. Suppose that $\bar{x}$ is a local optimal solution to the problem. Then (i) $\bar{x}$ is a global optimal solution. (ii) If either $\bar{x}$ is a strict local minimum or if f is strictly convex, then $\bar{x}$ is the unique global optimal solution.

(c) Let S be a nonempty compact polyhedral set in R^n, and let $f : S \to R$ be a strict convex function on S. Consider the problem of $\max_{x \in S} f(x)$. All the local maxima are reached at extreme points of S.

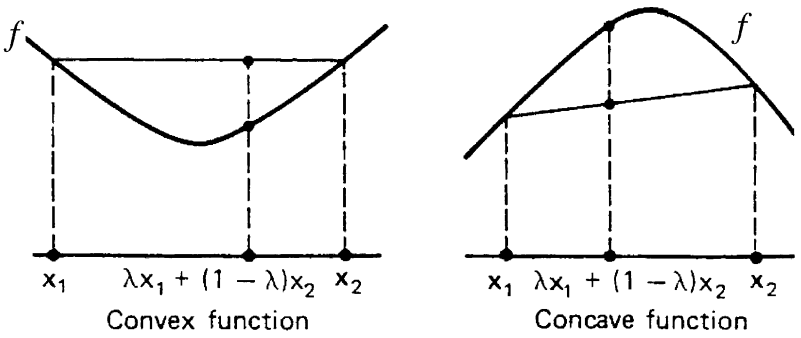

Fig. 2. Convex and concave function.

The above (a)(b) are basically known from a foundation course on mathematics during a undergraduate study. Though the statement (c) may not be included, it is not difficult to understand. Assume $\bar{x}$ is a local maximum but not an extreme point, we may find $x_1 \in N_\varepsilon(\bar{x}), x_2 \in N_\varepsilon(\bar{x})$ such that $\bar{x} = \lambda x_1 + (1-\lambda)x_2$ for $0 < \lambda < 1$. It follows from eq.(4) that $f(\bar{x}) < \lambda f(x_1) + (1-\lambda)f(x_2) \leq \max[f(x_1), f(x_2)]$, which contradicts to that $\bar{x}$ is a local maximum, while at an extreme point x of S, $x = \lambda x_1 + (1-\lambda)x_2$ with $x_1 \in S, x_2 \in S$ and $0 < \lambda < 1$ implies that $x = x_1 = x_2$, which does not contradict the definition of a strict convex function made after eq.(4). That is, a local maximum can only be reached at one of the extreme points of S.

Details of the above knowledge about convex programming are referred to one of textbooks on nonlinear programming, e.g., [3].

2.2 One-Bit-Matching Theorem

For the problem by eq. (2), neither the set $RR^T = I$ is convex nor $J(R)$ is always convex. To use the knowledge given in the previous section as a tool, we let $p_{ij} = r_{ij}^2$ and considering $RR^T = I$ via keeping the part of normalization conditions but ignoring the part of orthogonal conditions, then we can relax the problem by eq. (2) as follows:

$$\max_{P \in S} J(P), \;\; J(P) = \sum_{i=1}^{n}\sum_{j=1}^{n} p_{ij}^2 \nu_j^s k_i^m, \quad P = (p_{ij})_{n\times n}, \;\; n \geq 2$$

$$S = \{p_{ij}, i, j = 1, \cdots, n : \sum_{j=1}^{n} p_{ij} = 1, \;\; for \; i = 1, \cdots, n, \;\; and \; p_{ij} \geq 0\}, \quad (5)$$

where ν_j^s and k_i^m are same as in eq. (2), and S become a convex set or precisely a polyhedral set. Moreover, we stack P into a vector $vec[P]$ of n^2 elements and compute the Hessian H_P with respect to $vec[P]$, resulting in that

H_P is a $n^2 \times n^2$ diagonal matrix with each diagonal element being $\nu_j^s k_i^m$. (6)

Thus, whether $J(P)$ is convex can be checked simply via all the signs of $\nu_j^s k_i^m$.

We use $\mathcal{E}_{n\times k}$ to denote a family of matrices, with each $E_{n\times k}\in\mathcal{E}_{n\times k}$ being a $n\times k$ matrix with every row consisting of zero elements except that one and only one element is 1.

Lemma 2

(a) When either $\nu_i^s>0, k_i^m>0, \forall i$ or $\nu_i^s<0, k_i^m<0, \forall i$, every local maximum of $J(P)$ is reached at a $P\in\mathcal{E}_{n\times n}$.

(b) For a unknown $0<k<n$ with $\nu_i^s>0, k_i^m>0, i=1,\cdots,k$ and $\nu_i^s<0, k_i^m<0, i=k+1,\cdots,n$, every local maximum of $J(P)$ is reached at $P=\begin{bmatrix} P_1^+ & \mathbf{0} \\ \mathbf{0} & P_2^- \end{bmatrix}$, $P_1^+\in\mathcal{E}_{k\times k}$, $P_2^-\in\mathcal{E}_{(n-k)\times(n-k)}$.

Proof. (a) In this case, we have every $\nu_j^s k_i^m>0$ and thus it follows from eq. (6) and Lemma 1(a) that $J(P)$ is strictly convex on the polyhedral set S. It further follows from Lemma 1 (c) that all the local maxima of $J(P)$ are reached at the polyhedral set's extreme points that satisfy $\sum_{j=1}^n p_{ij}=1,\ for\ i=1,\cdots,n$, i.e., each local maximum $P\in\mathcal{E}_{n\times n}$.

(b) Notice that the constraint $\sum_{j=1}^n p_{ij}=1$ effects only on the i-th row, and $J(P)$ is additive, we see that the task by eq. (5) is solved by separately considering the following two tasks:

$$\begin{aligned} T_1: \max_{P_1} J(P_1),\ J(P_1)&=\sum_{i=1}^{k}\sum_{j=1}^{n} p_{ij}^2\nu_j^s k_i^m,\\ P_1&=(p_{ij})_{i=1,\cdots,k,j=1,\cdots,n}\ with\ every\ p_{ij}\geq 0,\\ Subject\ to&\sum_{j=1}^{n} p_{ij}=1,\ for\ i=1,\cdots,k. \end{aligned} \tag{7}$$

$$\begin{aligned} T_2: \max_{P_2} J(P_2),\ J(P_2)&=\sum_{i=k+1}^{N}\sum_{j=1}^{n} p_{ij}^2\nu_j^s k_i^m,\\ P_2&=(p_{ij})_{i=k+1,\cdots,n,j=1,\cdots,n}\ with\ every\ p_{ij}\geq 0,\\ Subject\ to&\sum_{j=1}^{n} p_{ij}=1,\ for\ i=k+1,\cdots,N. \end{aligned} \tag{8}$$

First, we consider T_1. Further let $J(P_1)=J_+^+(P_1^+)+J_+^-(P_1^-)$ with

$$\begin{aligned} J_+^+(P_1^+)&=\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}^2\nu_j^s k_i^m,\quad P_1^+=(p_{ij})_{i=1,\cdots,k,j=1,\cdots,k},\\ J_+^-(P_1^-)&=\sum_{i=1}^{k}\sum_{j=k+1}^{n} p_{ij}^2\nu_j^s k_i^m,\quad P_1^-=(p_{ij})_{i=1,\cdots,k,j=k+1,\cdots,n}, \end{aligned} \tag{9}$$

we see that J_+^+ and J_+^- are decoupled if ignoring the constraints $\sum_{j=1}^n p_{ij}=1,\ for\ i=1,\cdots,k$. So, the key point is considering the roles of the constraints.

Without the constraints $\sum_{j=1}^{n} p_{ij} = 1,\ for\ i = 1, \cdots, k,\ J_+^-(P_1^-) \le 0$ is strictly concave from Lemma 1(a) by observing $\nu_j^s k_i^m < 0$ for every term, and thus has only one maximum at $P_1^- = \mathbf{0}$. Then, the constraints can be re-taken in consideration via written as $\sum_{j=k+1}^{n} p_{ij} = c_i,\ for\ i = 1, \cdots, k$ with a unknown $c_i = 1 - \sum_{j=1}^{k} p_{ij}$. For $c_i > 0$, the boundary $\sum_{j=k+1}^{n} p_{ij} = c_i$ is inactive and will not affect that $J_+^- \le 0$ reaches its only maximum at $P_1^- = \mathbf{0}$. For $c_i = 0$, $J_+^- \le 0$ reaches its only maximum at the boundary $\sum_{j=k+1}^{n} p_{ij} = 0$ which is still $P_1^- = \mathbf{0}$. Thus, all the local maxima of $J(P_1)$ are reached at $P_1 = [P_1^+, P_1^-] = [P_1^+, \mathbf{0}]$ and thus determined by all the local maxima of $J_+^+(P_1^+)$ on the polyhedral set of $p_{ij} \ge 0, i = 1, \cdots, k, j = 1, \cdots, k$ and $\sum_{j=1}^{k} p_{ij} = 1,\ for\ i = 1, \cdots, k$ (because $p_{ij} = 0,\ for\ i = 1, \cdots, k, j = k+1, \cdots, n$). It follows from Lemma 1(b) that $J_+^+(P_1^+)$ is strictly convex on this polyhedral set since $\nu_j^s k_i^m > 0$ for every term. Similar to the above (a), each of the local maxima of $J_+^+(P_1^+)$ is $P_1^+ \in \mathcal{E}_{k\times k}$.

Second, we can consider T_2 in a same way and have $J(P_2) = J_-^+(P_2^+) + J_-^-(P_2^-)$ with

$$J_-^+(P_2^+) = \sum_{i=k+1}^{n} \sum_{j=1}^{k} p_{ij}^2 \nu_j^s k_i^m, \quad P_2^+ = (p_{ij})_{i=k+1,\cdots,n, j=1,\cdots,k},$$
$$J_-^-(P_2^-) = \sum_{i=k+1}^{n} \sum_{j=k+1}^{n} p_{ij}^2 \nu_j^s k_i^m, \quad P_2^- = (p_{ij})_{i=k+1,\cdots,n, j=k+1,\cdots,n}. \quad (10)$$

Now, J_-^+ is strictly concave and J_-^- is strictly convex. As a result, all the local maxima of $J(P_2)$ are reached at $P_2 = [P_2^+, P_2^-] = [\mathbf{0}, P_2^-]$ with $P_2^- \in \mathcal{E}_{(n-k)\times(n-k)}$. **Q.E.D.**

Further considering $p_{ij} = r_{ij}^2$ and the part of orthogonal conditions in $RR^T = I$, we get

Theorem 1. *Every local maximum of $J(R)$ on $RR^T = I$ by eq. (2) is reached at R that is an permutation matrix up to sign indeterminacy at its nonzero elements, as long as there is a one-to-one same-sign-correspondence between the kurtosis of all source pdf's and the kurtosis of all model pdf's.*

Proof. From $p_{ij} = r_{ij}^2$ and Lemma 2, we have $r_{ij} = 0$ for $p_{ij} = 0$ and either $r_{ij} = 1$ or $r_{ij} = -1$ for $p_{ij} = 1$. All the other choices of P in Lemma 2(a) or of P_1^+ and P_2^- in Lemma 2(b) can not satisfy the part of orthogonal conditions in $RR^T = I$ and thus should be discarded, except that P is a $n \times n$ permutation matrix for Lemma 2(a) or P_1^+ is a $k \times k$ permutation matrix and P_2^- is a $(n-k) \times (n-k)$ permutation matrix for Lemma 2(b). That is, R should be an permutation matrix up to sign indeterminacy at its nonzero elements. On the other hand, any other R on $RR^T = I$ with the corresponding P being not a local maximum of $J(P)$ is also not a local maxima of $J(R)$ on $RR^T = I$. Thus, we get the theorem proved by noticing that k_i^m has the same sign as the kurtosis ν_i^m of the model density $q_i(y_i)$. **Q.E.D.**

The above theorem is obtained from eq. (2) that is obtained from eq. (1) by approximately only considering the skewness and kurtosis and with the model

pdfs without skewness. Thus, in such an approximative sense, all the sources can also be separated by a local searching ICA algorithm (e.g., a gradient based algorithm) obtained from eq.(1) as long as there is a one-to-one same-sign-correspondence between the kurtosis of all source pdf's and the kurtosis of all model pdf's.

Though how seriously such an approximation will affect the separation performance by an ICA algorithm obtained from eq.(1) is unclear yet, this approximation can be removed by an ICA algorithm obtained directly from eq. (2). Under the one-to-one kurtosis sign matching assumption, we can derive a local search algorithm that is equivalent to maximize the problem by eq.(2) directly. A pre-whitening is made on observed samples such that we can consider the samples of x with $Ex = 0, Exx^T = I$. As a results, it follows from $I = Exx^T = AEss^TA^T$ and $Ess^T = I$ that $AA^T = I$, i.e., A is orthonormal. Thus, an orthonormal W is considered to let $y = Wx$ become independent among its components via

$$\max_{WW^T=I} J(W),\ \ J(W) = \sum_{i=1}^{n} k_i^m \nu_i^y, \tag{11}$$

where $\nu_i^y = Ey_i^4 - 3, i = 1, \cdots, n$, $\nu_j^x = Ex_j^4 - 3, j = 1, \cdots, n$, and $k_i^m, i = 1, \cdots, n$ are pre-specified constants with the same signs as the kurtosis ν_i^m. We can derive its gradient $\nabla_W J(W)$ and then project it onto $WW^T = I$, which results in an iterative updating algorithm for updating W in a way similar to eq.(19) and eq.(20) at the end of the next section. Such an ICA algorithm actually maximizes the problem by eq.(2) directly by considering $y = Wx = WAs = Rs, R = WA, RR^T = I$, and thus

$$\nu_i^y = \sum_{j=1}^{n} r_{ij}^4 \nu_j^s, i = 1, \cdots, n. \tag{12}$$

That is, the problem by eq.(11) is equivalent to the problem by eq.(2). In other words, under the one-to-one kurtosis sign matching assumption, it follows from Theorem 1 that all the sources can be separated by an ICA algorithm not in an approximate sense, as long as eq.(12) holds.

However, Theorem 1 does not tell us how such a kurtosis sign matching is built, which is attempted via eq.(1) through learning each model pdf $q_i(y_i)$ together with learning W [23, 24, 26] as well as further advances either given in [14, 19] or given by eqn. (103) in [20]. Still, it remains an open problem whether these efforts or the possibility of developing other new techniques can guarantee such an one-to-one kurtosis sign matching surely or in certain probabilistic sense, which deserves future investigations.

2.3 Cases of No Matching and Partial Matching

Next, we consider what happens when one-to-one kurtosis-sign-correspondence does not hold. We start at the extreme situation via the following Lemma.

Lemma 3. (no matching case)

When either $\nu_i^s > 0, k_i^m < 0, \forall i$ *or* $\nu_i^s < 0, k_i^m > 0, \forall i$, $J(P)$ *has only one maximum that is reached usually not in* $\mathcal{E}_{n\times n}$.

Proof. From eq.(6) and Lemma 1(a) that $J(P)$ is strictly concave since $\nu_j^s k_i^m < 0$ for every term. Thus, it follows from Lemma 1(b) that it has only one maximum usually at an interior point in S (thus not in $\mathcal{E}_{n\times n}$) instead of at the extreme points of S. **Q.E.D.**

Lemma 4. (partial matching case)

Given two unknown integers k, m *with* $0 < k < m < n$, *and provided that* $\nu_i^s > 0, k_i^m > 0, i = 1, \cdots, k$, $\nu_i^s k_i^m < 0, i = k+1, \cdots, m$, *and* $\nu_i^s < 0, k_i^m < 0, i = m+1, \cdots, n$, *every local maximum of* $J(P)$ *is reached either at* $P = \begin{bmatrix} P_1^+ & \mathbf{0} \\ \mathbf{0} & P_2^- \end{bmatrix}$, *where either* $P_1^+ \in \mathcal{E}_{k\times m}, P_2^- \in \mathcal{E}_{(n-k)\times(n-m)}$ *when* $\nu_i^s > 0, k_i^m < 0, i = k+1, \cdots, m$ *or* $P_1^+ \in \mathcal{E}_{m\times k}, P_2^- \in \mathcal{E}_{(n-m)\times(n-k)}$ *when* $\nu_i^s < 0, k_i^m > 0, i = k+1, \cdots, m$.

Proof. The proof is made similar to proving Lemma 2. The difference is that both P_1^+ and P_2^- are not square matrices. **Q.E.D.**

Theorem 2. *Given two unknown integers* k, m *with* $0 < k < m < n$, *and provided that* $\nu_i^s > 0, k_i^m > 0, i = 1, \cdots, k$, $\nu_i^s k_i^m < 0, i = k+1, \cdots, m$, *and* $\nu_i^s < 0, k_i^m < 0, i = m+1, \cdots, n$, *every local maximum of* $J(R)$ *on* $RR^T = I$ *by eq.(2) is reached at* $R = \begin{bmatrix} \Pi & \mathbf{0} \\ \mathbf{0} & \bar{R} \end{bmatrix}$ *subject to a* 2×2 *permutation, where* Π *is a* $(k+n-m) \times (k+n-m)$ *permutation matrix up to sign indeterminacy at its nonzero elements; while* $\bar{R}$ *is a* $(m-k) \times (m-k)$ *orthonormal matrix with* $\bar{R}\bar{R}^T = I$, *but usually not a permutation matrix up to sign indeterminacy.*

Proof. By Lemma 2, putting $p_{ij} = r_{ij}^2$ in P we can directly select a $(k+n-m)\times(k+n-m)$ sub-matrix Π that is of full rank in both row and column, also automatically with $\Pi\Pi^T = I$ satisfied. The remaining part in P must be linear dependent of Π with $RR^T = I$ still satisfied. Thus, the entire R should be the above form with $\bar{R}\bar{R}^T = I$. As a result, $\max_{RR^T=I} J(R)$ in eq. (2) is decoupled with $\bar{R}$ maximized via $\max_{\bar{R}\bar{R}^T=I} J(\bar{R})$, $J(\bar{R}) = \sum_{i=k+n-m+1}^{n} \sum_{j=k+n-m+1}^{n} \bar{r}_{ij}^4 \nu_j^s k_i^m$ with every $\nu_j^s k_i^m < 0$, which is a situation similar to Lemma 3. That is, $\bar{R}$ is usually not a permutation matrix up to sign indeterminacy. On the other hand, if the second row of R is not $[\mathbf{0}, \bar{R}]$ but in a form $[A, B]$ with both A, B being nonzero and $[A, B][A, B]^T = I$, the first row of R will non longer be $[\Pi, \mathbf{0}]$ and the resulting P deviates from a local maximum of $J(P)$. Thus, the corresponding R is not a local maxima of $J(R)$ on $RR^T = I$. **Q.E.D.**

In other words, there will be $k+n-m$ sources that can be successfully separated in help of a local searching ICA algorithm when there are $k+n-m$ pairs of matching between the kurtosis signs of source pdf's and of model pdf's. However, the remaining $m-k$ sources are not separable. Suppose that the kurtosis sign of each model is described by a binary random variable ξ_i with 1 for + and 0 for −, i.e., $p(\xi_i) = 0.5^{\xi_i} 0.5^{1-\xi_i}$. When there are k sources

with their kurtosis signs in positive, there is still a probability $p(\sum_{i=1}^{n}\xi_i = k)$ to have an one-to-one kurtosis-sign-correspondence even when model pdf's are prefixed without knowing the kurtosis signs of sources. Moreover, even when an one-to-one kurtosis-sign-correspondence does not hold for all the sources, there will still be $n - |\ell - k|$ sources recoverable with a probability $p(\sum_{i=1}^{n}\xi_i = \ell)$. This explains not only why those early ICA studies [1, 4], work in some case while fail in other cases due to the pre-determined model pdf's, but also why some existing heuristic ICA algorithms can work in this or that way.

2.4 Maximum Kurtosis vs Minimum Kurtosis

Interestingly, it can be observed that changing the maximization in eq. (2), eq. (5) and eq. (11) into the minimization will lead to similar results, which are summarized into the following Lemma 5 and Theorem 3.

Lemma 5

(a) When either $\nu_i^s > 0, k_i^m > 0, \forall i$ or $\nu_i^s < 0, k_i^m < 0, \forall i$, $J(P)$ has only one minimum that is reached usually not in $\mathcal{E}_{n\times n}$.

(b) When either $\nu_i^s > 0, k_i^m < 0, \forall i$ or $\nu_i^s > 0, k_i^m < 0, \forall i$, every local minimum of $J(P)$ is reached at a $P \in \mathcal{E}_{n\times n}$.

(c) For a unknown $0 < k < n$ with $\nu_i^s > 0, k_i^m > 0, i = 1, \cdots, k$ and $\nu_i^s < 0, k_i^m < 0, i = k+1, \cdots, n$, every local minimum of $J(P)$ is reached at $P = \begin{bmatrix} \mathbf{0} & P_1^- \\ P_2^+ & \mathbf{0} \end{bmatrix}$, $P_1^- \in \mathcal{E}_{k\times(n-k)}$, $P_2^+ \in \mathcal{E}_{(n-k)\times k}$.

(d) For two unknown integers k, m with $0 < k < m < n$ with $\nu_i^s > 0, k_i^m > 0, i = 1, \cdots, k$, $\nu_i^s k_i^m < 0, i = k+1, \cdots, m$, and $\nu_i^s < 0, k_i^m < 0, i = m+1, \cdots, n$, every local minimum of $J(P)$ is reached either at $P = \begin{bmatrix} \mathbf{0} & P_1^- \\ P_2^+ & \mathbf{0} \end{bmatrix}$, where either $P_1^- \in \mathcal{E}_{k\times(n-m)}, P_2^+ \in \mathcal{E}_{(n-k)\times m}$ when $\nu_i^s > 0, k_i^m < 0, i = k+1, \cdots, m$ or $P_1^+ \in \mathcal{E}_{m\times(n-k)}, P_2^- \in \mathcal{E}_{(n-m)\times k}$ when $\nu_i^s < 0, k_i^m > 0, i = k+1, \cdots, m$.

Proof. The proof can be made similar to those in proving Lemma 2, Lemma 3, and Lemma 4. The key difference is shifting our focus from the maximization of a convex function on a polyhedral set to the minimization of a concave function on a polyhedral set, with switches between 'minimum' and 'maximum', 'maxima' and 'minima', 'convex' and 'concave', and 'positive' and 'negative', respectively. The key point is that Lemma 1 still remains to be true after these switches. **Q.E.D.**

Similar to Theorem 2, from the above lemma we can get

Theorem 3

(a) When either $\nu_i^s k_i^m < 0, i = 1, \cdots, n$ or $\nu_i^s > 0, k_i^m > 0, i = 1, \cdots, k$ and $\nu_i^s < 0, k_i^m < 0, i = k+1, \cdots, n$ for a unknown $0 < k < n$, every local minimum of $J(R)$ on $RR^T = I$ by eq. (2) is reached at R that is an permutation matrix up to sign indeterminacy at its nonzero elements.

(b) For two unknown integers k, m with $0 < k < m < n$ with $\nu_i^s > 0, k_i^m > 0, i = 1, \cdots, k$, $\nu_i^s k_i^m < 0, i = k+1, \cdots, m$, and $\nu_i^s < 0, k_i^m < 0, i = m+1, \cdots, n$,

every local minimum of $J(R)$ *on* $RR^T = I$ *by eq. (2) is reached at* $R = \begin{bmatrix} \Pi & \mathbf{0} \\ \mathbf{0} & \bar{R} \end{bmatrix}$ *subject to a* 2×2 *permutation. When* $m + k \geq n$, Π *is a* $(n - m + n - k) \times (n - m + n - k)$ *permutation matrix up to sign indeterminacy at its nonzero elements, while* $\bar{R}$ *is a* $(m + n - k) \times (m + n - k)$ *orthonormal matrix with* $\bar{R}\bar{R}^T = I$, *but usually not a permutation matrix up to sign indeterminacy. When* $m + k < n$, Π *is a* $(k + m) \times (k + m)$ *permutation matrix up to sign indeterminacy at its nonzero elements, while* $\bar{R}$ *is a* $(n - k - m) \times (n - k - m)$ *orthonormal matrix with* $\bar{R}\bar{R}^T = I$, *but usually not a permutation matrix up to sign indeterminacy.*

In a comparison of Theorem 2 and Theorem 3, when $m + k \geq n$, comparing $n - m + n - k$ with $k + n - m$, we see that more source can be separated by minimization than maximization if $k < 0.5n$ while maximization is better than minimization if $k > 0.5n$. When $m + k < n$, comparing $k + m$ with $k + n - m$, we see that more source can be separated by minimization than maximization if $m > 0.5n$ while maximization is better than minimization if $m < 0.5n$.

We further consider a special case that $k_i^m = 1, \forall i$. In this case, eq. (2) is simplified into

$$J(R) = \sum_{i=1}^{n} \sum_{j=1}^{n} r_{ij}^4 \nu_j^s, \ n \geq 2, \tag{13}$$

From Theorem 2 at $n = m$, we can easily obtain

Corollary 1. *For a unknown integer* $0 < k < n$ *with* $\nu_i^s > 0, i = 1, \cdots, k$ *and* $\nu_i^s < 0, i = k + 1, \cdots, n$, *every local maximum of* $J(R)$ *on* $RR^T = I$ *by eq. (13) is reached at* $R = \begin{bmatrix} \Pi & \mathbf{0} \\ \mathbf{0} & \bar{R} \end{bmatrix}$ *subject to a* 2×2 *permutation, where* Π *is a* $k \times k$ *permutation matrix up to sign indeterminacy at its nonzero elements, while* $\bar{R}$ *is a* $(n - k) \times (n - k)$ *orthonormal matrix with* $\bar{R}\bar{R}^T = I$, *but usually not a permutation matrix up to sign indeterminacy.*

Similarly, from Theorem 3 we also get

Corollary 2. *For a unknown integer* k *with* $0 < k < n$ *with* $\nu_i^s > 0, i = 1, \cdots, k$ *and* $\nu_i^s < 0, i = k + 1, \cdots, n$, *every local minimum of* $J(R)$ *on* $RR^T = I$ *by eq.(2) is reached at* $R = \begin{bmatrix} \bar{R} & \mathbf{0} \\ \mathbf{0} & \Pi \end{bmatrix}$ *subject to a* 2×2 *permutation, where* Π *is a* $(n - k) \times (n - k)$ *permutation matrix up to sign indeterminacy at its nonzero elements, while* $\bar{R}$ *is a* $k \times k$ *orthonormal matrix with* $\bar{R}\bar{R}^T = I$, *but usually not a permutation matrix up to sign indeterminacy.*

It follows from Corollary 1 that k super-gaussian sources can be separated by $\max_{RR^T=I} J(R)$, while it follows from Corollary 2 that $n - k$ sub-gaussian sources can be separated by $\min_{RR^T=I} J(R)$. In implementation, from eq. (11) we get

$$J(W) = \sum_{i=1}^{n} \nu_i^y, \tag{14}$$

and then make $\max_{WW^T=I} J(W)$ to get k super-gaussian source and make $\min_{WW^T=I} J(W)$ to get $n-k$ sub-gaussian source. Thus, instead of learning an one-to-one kurtosis sign matching, the problem can also be equivalently turned into a problem of selecting super-gaussian components from $y = Wx$ with W obtained via $\max_{WW^T=I} J(W)$ and of selecting sub-gaussian components from $y = Wx$ with W obtained via $\min_{WW^T=I} J(W)$. Though we know neither k nor which of components of y should be selected, we can pick those with positive signs as super-gaussian ones after $\max_{WW^T=I} J(W)$ and pick those with negative signs as sub-gaussian ones after $\min_{WW^T=I} J(W)$. The reason comes from $\nu_i^y = \sum_{j=1}^n r_{ij}^4 \nu_j^s$ and the above corollaries. By Corollary 1, the kurtosis of each super-gaussian component of y is simply one of $\nu_j^s > 0, j = 1, \cdots, k$. Though the kurtosis of each of the rest components in y is a weighted combination of $\nu_j^s < 0, j = k+1, \cdots, n$, the kurtosis signs of these rest components will all remain negative. Similarly, we can find out those sub-gaussian components according to Corollary 2.

Anther corollary can be obtained from eq.(11) by considering a special case that $k_i^m = sign[\nu_i^y], \forall i$. That is, eq.(11) becomes

$$\max_{WW^T=I} J(W), \;\; J(W) = \sum_{i=1}^{n} |\nu_i^y|. \tag{15}$$

Actually, this leads to what is called kurtosis extreme approach and extensions [8, 13, 16], where studies were started at extracting one source by a vector w and then extended to extracting multiple sources by either sequentially implementing the one vector algorithm such that the newly extracted vector is orthogonal to previous ones or in parallel implementing the one vector algorithm on all the vectors of W separately together with a symmetric orthogonalization made at each iterative step. In the literature, the success of using one vector vector w to extract one source has been proved mathematically and the proof can be carried easily to sequentially extracting a new source with its corresponding vector w being orthogonal to the subspace spanned by previous. However, this mathematical proof is not applicable to implementing the one vector algorithm in parallel on all the vectors of W separately together with a symmetric orthogonalization, as suggested in Sec.8.4.2 of [13] but with no proof. Actually, what was suggested there can only ensure a convergence of such a symmetric orthogonalization based algorithm but is not able to guarantee that this local searching featured iterative algorithm will surely converge to a solution that can separate all the sources, though experiments usually turned out with successes.

When $\nu_i^y = \sum_{j=1}^n r_{ij}^4 \nu_j^s$ holds, from eq.(15) we have $\min_{RR^T=I} J(R), \; J(R) = \sum_{i=1}^n \sum_{j=1}^n r_{ij}^4 |\nu_j^s|$, which is covered by Lemma 2(a) and Theorem 2. Thus, we can directly prove the following corollary:

Corollary 3. *As long as $\nu_i^y = \sum_{j=1}^n r_{ij}^4 \nu_j^s$ holds, every local minimum of the above $J(R)$ on $RR^T = I$ is reached at a permutation matrix up to sign indeterminacy.*

Actually, it provides a mathematical proof on the success of the above symmetric orthogonalization based algorithm on separating all the sources.

The last but not least, it should be noticed that the above corollaries are true only when the relation $\nu_i^y = \sum_{j=1}^n r_{ij}^4 \nu_j^s, i = 1, \cdots, n$ holds, which is true only when there is a large size of samples such that the pre-whitening can be made perfectly.

3 Combinatorial Optimization in Stiefel Manifold

The combinatorial optimization problem by eq.(3) has been encountered in various real applications and still remains a hard task to solve. Many efforts have also been made on in the literature of neural networks since Hopfield and Tank [12]. As summarized in [21], these efforts can be roughly classified according to the features on dealing with C_e^{col}, C_e^{row} and C_b. Though having a favorable feature of being parallel implementable, almost all the neural network motivated approaches share one unfavorable feature that these intuitive approaches have no theoretical guarantees on convergence to even a feasible solution. Being different from several existing algorithms in the literature, a general LAGRANGE-enforcing iterative procedure is proposed firstly in [27] and further developed in the past decade, and its convergence to even a feasible solution is guaranteed. Details are referred to [21].

Interestingly, focusing at local maxima only, both eq.(2) and eq.(5) can be regarded as special examples of the combinatorial optimization problem by eq.(3) simply via regarding p_{ij} or r_{ij} as v_{ij}. Though such a linkage is not useful for ICA since we need not to seek a global optimization for making ICA, linking from eq.(3) reversely to eq.(2) and even eq.(1) leads to one motivation. That is, simply let $v_{ij} = r_{ij}^2$ and then use $RR^T = I$ to guarantee the constraints C_e^{col}, C_e^{row} as well as a relaxed version of C_b (i.e., $0 \leq v_{ij} \leq 1$). That is, the problem eq.(3) is relaxed into

$$\min_{RR^T=I \ for \ N \leq M} E_o(\{r_{ij}^2\}_{i=1,j=1}^{i=N,j=M}), \ R = \{r_{ij}\}_{i=1,j=1}^{i=N,j=M}. \tag{16}$$

We consider the problems with

$$\frac{\partial^2 E_o(V)}{\partial vec[V] \partial vec[V]^T} = \mathbf{0}, \ \frac{\partial^2 E_o(V)}{\partial vec[V] \partial vec[V]^T} \text{ is negative definite,} \tag{17}$$

or $E_o(V)$ in a form similar to $J(P)$ in eq.(5), i.e.,

$$E_o(V) = -\sum_{i=1}^n \sum_{j=1}^n v_{ij}^2 a_j b_i, \tag{18}$$

with $a_i > 0, b_i > 0, i = 1, \cdots, k$ and $a_i < 0, b_i < 0, i = 1, \cdots, k$ after an appropriate permutation on $[a_1, \cdots, a_n]$ and on $[b_1, \cdots, b_n]$. Similar to the study of eq.(5), maximizing $E_o(V)$ under the constraints C_e^{col}, C_e^{row} and $v_{ij} \geq$ will imply the satisfaction of C_b. In other words, the solutions of eq.(16) and of eq.(3)

are same. Thus, we can solve the hard problem of combinatorial optimization by eq.(3) via a gradient flow on the Stiefel manifold $RR^T = I$ to maximize the problem by eq.(16). At least a local optimal solution of eq.(3) can be reached, with all the constraints C_e^{col}, C_e^{row}, and C_b guaranteed automatically.

To get an appropriate updating flow on the Stiefel manifold $RR^T = I$, we first compute the gradient $\nabla_V E_o(V)$ and then get $G_R = \nabla_V E_o(V) \circ R$, where the notation $\circ$ means that

$$\begin{bmatrix} a_{11}\ a_{12} \\ a_{21}\ a_{22} \end{bmatrix} \circ \begin{bmatrix} b_{11}\ b_{12} \\ b_{21}\ b_{22} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11}\ a_{12}b_{12} \\ a_{21}b_{21}\ a_{22}b_{22} \end{bmatrix}.$$

Given a small disturbance δ on $RR^T = I$, it follows from $RR^T = I$ that the solution of $\delta R R^T + R\delta R^T = 0$ must satisfy

$$\delta R = ZR + U(I - R^T R), \tag{19}$$

where U is any $m \times d$ matrix and $Z = -Z$ is an asymmetric matrix.

From $Tr[G_R^T \delta R] = Tr[G_R^T(ZR + UI - R^T R)] = Tr[(G_R R^T)^T Z] + Tr[(G_R(I - R^T R))^T U]$, we get

$$Z = G_R R^T - R G_R^T,\ U = G_R(I - R^T R),\ \delta R = \begin{cases} U(I - R^T R) = U, & \text{(a)}, \\ ZR, & \text{(b)}, \\ ZR + U, & \text{(c)}. \end{cases}$$
$$R^{new} = R^{old} + \gamma_t \delta R. \tag{20}$$

That is, we can use anyone of the above three choices of δR as the updating direction of R. A general technique for optimization on the Stiefel manifold was elaborately discussed in [9], which can also be adopted for implementing our problem by eq.(16).

3.1 Concluding Remarks

The one-to-one kurtosis sign matching conjecture has been proved in a strong sense that every local maximum of $\max_{RR^T=I} J(R)$ by eq.(2) is reached at a permutation matrix up to certain sign indeterminacy if there is an 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, all the sources can be separated by a local search ICA algorithm. Theorems have also been proved not only on partial separation of sources when there is a partial matching between the kurtosis signs, but also on an interesting duality of maximization and minimization on source separation. Moreover, corollaries are obtained from the theorems to state that seeking a one-to-one same-sign-correspondence can be replaced by a use of the duality, i.e., super-gaussian sources can be separated via maximization and sub-gaussian sources can be separated via minimization. Furthermore, a corollary is also obtained to provide a mathematical proof on the success of symmetric orthogonalization implementation of the kurtosis extreme approach.

Due to the results, the open problem of the one-to-one kurtosis sign matching conjecture [22] can be regarded as closed. However, there still remain problems

to be further studied. First, the success of those eq.(1) based efforts along this direction [14, 19, 20, 23, 24, 26] can be explained as their ability of building up an one-to-one kurtosis sign matching. However, we still need a mathematical analysis to prove that such a matching can be achieved surely or in certain probabilistic sense by these approaches. Second, as mentioned at the end of Sec. 2.4, a theoretical guarantee on either the kurtosis extreme approach or the approach of extracting super-gaussian sources via maximization and sub-gaussian sources via minimization is true only when there is a large size of samples such that the pre-whitening can be made perfectly. In practice, usually with only a finite size of samples, it remains to be further studied on comparison of the two approaches as well as of those eq.(1) based approaches. Also, comparison may deserve to made on convergence rates of different ICA algorithms.

The last but not the least, the linkage of the problem by eq. (3) to eq.(2) and eq.(5) leads us to a Stiefel manifold perspective of combinatorial optimization with algorithms that guarantee convergence and satisfaction of constraints, which also deserve further investigations.

References

1. Amari, S. I., Cichocki, A., Yang, H.: A New Learning Algorithm for Blind Separation of Sources. Advances in Neural Information Processing. Vol. 8. MIT Press, Cambridge, MA (1996) 757-763
2. Amari, S., Chen, T.-P., & Cichocki, A.: Stability analysis of adaptive blind source separation, Neural Networks, **10** (1997) 1345-1351
3. Bazaraa, M.S., Sherall, H.D., Shetty, C.M.: Nonlinear Programming: Theory and Algorithms, John Wileys & Sons, Inc., New York (1993)
4. Bell, A., Sejnowski, T.: An Information-maximization Approach to Blind Separation and Blind Deconvolution. Neural Computation, **7** (1995) 1129-1159
5. Cardoso, J.-F. Blind signal separation: Statistical Principles, Proc. of IEEE, **86** (1998) 2009-2025
6. Cheung, C. C., Xu, L.: Some Global and Local Convergence Analysis on the Information-theoretic Independent Component Analysis Approach. Neurocomputing, **30** (2000) 79-102
7. Comon, P.: Independent component analysis - a new concept ? Signal Processing **36** (1994) 287-314
8. Delfosse, N., Loubation, P.: Adaptive Blind Separation of Independent Sources: A Deflation Approach. Signal Processing, **45** (1995) 59-83
9. Edelman, A., Arias, T.A., Smith, S.T.: The Geometry of Algorithms with Orthogonality Constraints, SIAM J. Matrix Anal. APPL., **20** (1998) 303-353
10. Everson, R., Roberts, S.: Independent Component Analysis: A Flexible Nonlinearity and Decorrelating Manifold Approach. Neural Computation, **11** (1999) 1957-1983
11. Girolami, M.: An Alternative Perspective on Adaptive Independent Component Analysis Algorithms. Neural Computation, **10** (1998) 2103-2114
12. Hopfield, J. J. & Tank, D. W.: Neural computation of decisions in optimization problems, Biological Cybernetics 52, 141-152 (1985).
13. Hyvarinen, A., Karhunen, J., Oja, A.: Independent Component Analysis, John Wileys, Sons, Inc., New York (2001)

14. Lee, T. W., Girolami, M., Sejnowski, T. J.: Independent Component Analysis Using an Extended Infomax Algorithm for Mixed Subgaussian and Supergaussian Sources. Neural Computation, **11** (1999) 417-441
15. Liu, Z.Y., Chiu, K.C., Xu, L.: One-Bit-Matching Conjecture for Independent Component Analysis, Neural Computation, **16** (2004) 383-399
16. Moreau, E., Macchi, O.: High Order Constrasts for Self-adaptive Source Separation. International Journal of Adaptive Control and Signal Processing, **10** (1996) 19?6
17. Pearlmutter, B. A., Parra, L. C.: A Context-sensitive Genaralization of ICA. In Proc. of Int. Conf. on Neural Information Processing. Springer-Verlag, Hong Kong (1996)
18. Tong, L., Inouye, Y., Liu, R.: Waveform-preserving Blind Estimation of Multiple Independent Sources. Signal Processing, **41** (1993) 2461-2470
19. Welling, M., Weber, M.: A Constrained EM Algorithm for Independent Component Analysis. Neural Computation, **13** (2001) 677-689
20. Xu, L.: Independent Component Analysis and Extensions with Noise and Time: A Bayesian Ying-Yang Learning Perspective, Neural Information Processing Letters and Reviews, **1** (2003) 1-52
21. Xu, L.: Distribution Approximation, Combinatorial Optimization, and Lagrange-Barrier, Proc. of International Joint Conference on Neural Networks 2003 (IJCNN '03), July 20-24, Jantzen Beach, Portland, Oregon, (2003) 2354-2359
22. Xu, L., Cheung, C. C., Amari, S. I.: 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?**8** (pp39-45) (1998a)
23. Xu, L., Cheung, C. C., & Amari, S. I.: Learned Parametric Mixture Based ICA Algorithm. Neurocomputing, **22** 69-80 (1998b)
24. Xu, L., Cheung, C. C., Yang, H. H., Amari, S. I.: Independent component analysis by the information-theoretic approach with mixture of density. Proc. of 1997 IEEE Intl. Conf on Neural Networks, Houston, TX. **3** 1821-1826 (1997)
25. Xu, L. Bayesian Ying-Yang Learning Based ICA Models, Proc. 1997 IEEE Signal Processing Society Workshop on Neural Networks for Signal Processing **VI** , Florida, 476-485 (1997).
26. Xu, L., Yang, H. H., Amari, S. I.: Signal Source Separation by Mixtures: Accumulative Distribution Functions or Mixture of Bell-shape Density Distribution Functions. Rresentation at FRONTIER FORUM. Japan: Institute of Physical and Chemical Research, April (1996)
27. Xu, L.: Combinatorial optimization neural nets based on a hybrid of Lagrange and transformation approaches, Proc. of World Congress on Neural Networks, San Diego, 399-404 (1994).