LETTER Communicated by Te-Won Lee

One-Bit-Matching Theorem for ICA, Convex-Concave Programming on Polyhedral Set, and Distribution Approximation for Combinatorics

Lei Xu
lxu@cse.cuhk.edu.hk
Department of Computer Science and Engineering,
Chinese University of Hong Kong, Shatin, NT, Hong Kong, P. R. C.

According to the proof by Liu, Chiu, and Xu (2004) on the so-called one-bit-matching conjecture (Xu, Cheung, and Amari, 1998a), 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, 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 independent component analysis criterion is reached. Thus, it cannot support the successes of many existing iterative algorithms that usually converge at one of the local optimal solutions. This article presents a new mathematical proof that is obtained in a strong sense that the conjecture is also true when any one of local optimal solutions is reached in helping to investigating convex-concave programming on a polyhedral set. Theorems are also provided 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 on an interesting duality, with supergaussian sources separated by maximization and subgaussian sources separated by 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 lacks mathematical proof. Furthermore, a linkage has been set up to combinatorial optimization from a distribution approximation perspective and a Stiefel manifold perspective, with algorithms that guarantee convergence as well as satisfaction of constraints.

1 Introduction

Independent component analysis (ICA) aims at blindly separating the independent sources s from an unknown linear mixture $x = As$ via $y = Wx$. Tong, Inouye, and Liu (1993) showed that y recovers s up to constant scales and a permutation of components when the components of y become

component-wise independent and at most one of them is gaussian. The problem is formalized by Comon (1994) as 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, 1998), in the case that W is invertible, all such approaches are equivalent to minimizing the following cost function,

$$D(W) = \int p(y; W) \ln \frac{p(y, W)}{\prod_{i=1}^{n} q(y_i)} dy, \tag{1.1}$$

where $q(y_i)$ is the predetermined 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 when the components of y are either all subgaussians (Amari et al., 1996) or all supergaussians (Bell & Sejnowski, 1995).

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 (Xu, Yang, & Amari, 1996; Xu, Cheung, Yang, & Amari, 1997) or a mixture of parametric pdf's (Xu, 1997; Xu, Cheung, & Amari, 1998b). A learned parametric mixture–based ICA (LPMICA) algorithm is derived, with successful results on sources that can be either subgaussian or supergaussian, as well as any combination of both types. The mixture model was also adopted in the ICA algorithm by Pearlmutter and Parra (1996), although it did not explicitly target separating the mixed sub- and supergaussian sources.

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 (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 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." This conjecture has been implicitly supported by several other ICA studies (Girolami, 1998; Everson & Roberts, 1999; Lee, Girolami, & Sejnowski, 1999; Welling & Weber, 2001). Cheung and Xu (2000) gave a mathematical analysis for the case involving only two subgaussian sources. Amari, Chen, and Cichock: (1997) also studied the stability of an ICA algorithm at the correct separation points 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 was proved mathematically in a weak sense (Liu et al. 2004). When only skewness and kurtosis of

sources 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 equation 1.1 is as simplified as

$$\max_{RR^T=I} J(R), \quad J(R) = \sum_{i=1}^{n}\sum_{j=1}^{n} r_{ij}^4 \nu_j^s k_i^m, \quad n \geq 2, \tag{1.2}$$

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 $q(y_i)$.[1] Then it is further proved that the global maximization of equation 1.2 can be reachable only by setting R a permutation matrix up to a certain sign indeterminacy. However, this proof still cannot 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 article, all the local maxima of equation 1.2 are investigated using 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 any one of local maxima by equation 1.2 is reached. Theorems have also been provided on separation 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 to state that the duality makes it possible to get supergaussian sources by maximization and subgaussian sources by minimization. Another corollary also confirms the symmetric orthogonalization implementation of the extreme approach of the kurtosis for separating multiple sources in parallel, which works empirically but without mathematical proof (Hyvarinen, Karhunen, & Oja, 2001).

In section 3, we discuss that equation 1.2, with R being a permutation matrix up to certain sign indeterminacy, becomes equivalent to a special example of the following combinatorial optimization:

$$\begin{aligned}
&\min_V E_o(V), \ V = \{v_{ij}, i = 1, \ldots, N, \ j = 1, \ldots, M\}, \quad \text{subject to} \\
&C^c: \quad \sum_{i=1}^{N} v_{ij} = 1, \ j = 1, \ldots, M, \quad C^r: \quad \sum_{j=1}^{M} v_{ij} = 1, i = 1, \ldots, N; \\
&C^b: \quad v_{ij} \text{ takes either 0 or 1.}
\end{aligned} \tag{1.3}$$

[1] The details refer to theorem 1 of Liu et al. (2004) for the specific expression k_i^m and the proof that k_i^m is a constant with the same sign as the kurtosis ν_i^m of the model $q(y_i)$.

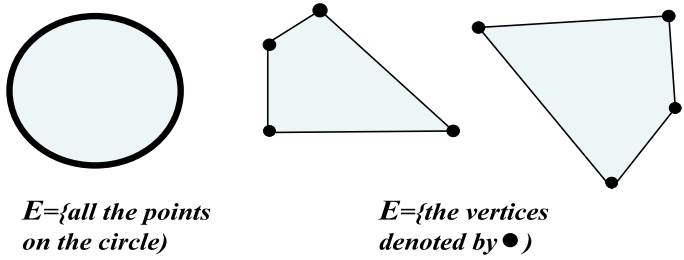

Figure 1: Convex set and polyhedral set.

This connection suggests investigating combinatorial optimization not only from a distribution approximation perspective of finding a simple distribution to approximate the Gibbs distribution induced from $E_o(V)$, but also a perspective of gradient flow searching within the Stiefel manifold, with algorithms that guarantee convergence as well as constraint satisfaction.

2 One-Bit-Matching Theorem and Extension

2.1 An Introduction to Convex Programming. To facilitate mathematical analysis, we briefly introduce convex programming. A set in R^n is said to be convex if $x_1 \in S, x_2 \in S$; we then have $\lambda x_1 + (1-\lambda)x_2 \in S$ for any $0 \le \lambda \le 1$. Shown in Figure 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, $S = \{x : a_i^t x \le \alpha_i, \ \ for\ i = 1, \ldots, m\}$, where a_i is a nonzero vector and α_i is a scalar for $i = 1, \ldots, m$. The second and third imagestion Figure 1 are two examples. Let S be a nonempty convex set. A vector $x \in S$ is called an extreme 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, illustrated in Figure 1.

Let $f : S \to R$, where S is a nonempty convex set in R^n. As shown in Figure 2, the function f is said to be convex on S if

$$f(\lambda x_1 + (1-\lambda)x_2) \le \lambda f(x_1) + (1-\lambda) f(x_2) \tag{2.1}$$

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 $f(x) \ge f(\bar{x})$ 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 $f(x) \ge f(\bar{x})$ for

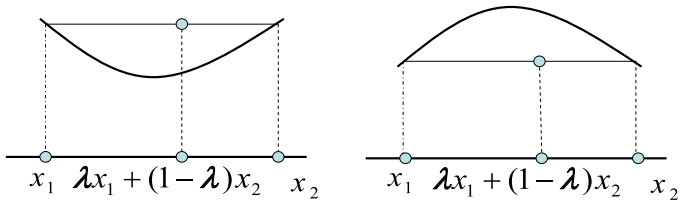

Figure 2: Convex and concave function.

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. 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.

Statements a and b are common knowledge, and statement c is not difficult to understand. As illustrated in Figure 2, 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 equation 2.1 that $f(\bar{x}) < \lambda f(x_1) + (1-\lambda) f(x_2) \leq \max[f(x_1), f(x_2)]$, which contradicts 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 equation 2.1. That is, a local maximum can be reached at only one of the extreme points of S.

(For details about convex programming, see, e.g., Bazaraa, Sherali, Shetty, 1993.)

2.2 One-Bit-Matching Theorem. This section aims at showing that every local maximum of $J(R)$ on $RR^T = I$ by equation 1.2 is reached at R,

which is a 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.

The proving line is sketched by two key steps. First, we divide the set C of constraint equations $RR^T = I$ into two nonoverlapped sets $C_n \cup C_o$. That is, for $R^T = [r_1, r_2, \ldots, r_n]$ with $r_j = [r_{1j}, r_{2j}, \ldots, r_{nj}]^T$, we have

$$\begin{aligned} C_n: &\quad \sum_{i=1}^{n} r_{ij}^2 = 1,\ j = 1, \ldots, n, && \text{for normalization,} \\ C_o: &\quad r_i^T r_j = 0, i = 1, \ldots, n-1,\ j = i, \ldots, n, && \text{for orthogonalization.} \end{aligned} \tag{2.2}$$

We find all the local maxima of $\max_{s.t.\ C_n} J(R)$ by proving lemma 2. Second, we consider how these local maxima will be affected by adding the constraint set C_o. Local maxima of $\max_{s.t.\ C_n} J(R)$ that do not satisfy C_o will be discarded. We find that adding C_o will not bring any extra local maximum. As a result, we conclude our aim by theorem 1.

We let $p_{ij} = r_{ij}^2$ and turn $\max_{s.t.\ C_n} J(R)$ into the following problem:

$$\begin{aligned} &\max_{P \in S} J(P),\ J(P) = \sum_{i=1}^{n}\sum_{j=1}^{n} p_{ij}^2 v_j^s k_i^m, \quad P = (p_{ij})_{n\times n},\ n \geq 2, \\ &S = \left\{ p_{ij}, i, j = 1, \ldots, n : \sum_{j=1}^{n} p_{ij} = 1,\ for\ i = 1, \ldots, n,\ \text{and every } p_{ij} \geq 0 \right\}, \end{aligned} \tag{2.3}$$

where v_j^s and k_i^m are same as in equation 1.2 and S becomes a convex set or precisely a polyhedral set. For every local maximum solution $P = [p_{ij}]$ in equation 2.3, we can get a subset of local maxima $R = [r_{ij}]$ of $\max_{s.t.\ C_n} J(R)$ with $r_{ij} = 0$ for $p_{ij} = 0$ and either $r_{ij} = \pm 1$ for $p_{ij} = 1$.

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,

$$H_P \text{ is a } n^2 \times n^2 \text{ diagonal matrix with each diagonal element being } v_j^s k_i^m. \tag{2.4}$$

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

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 an $n \times k$ matrix, with every row consisting of zero elements except that one and only one element is 1.

Lemma 2. *When either* $v_i^s > 0, k_i^m > 0, \forall i$ *or* $v_i^s < 0, k_i^m < 0, \forall i$, *every local maximum of* $J(P)$ *is reached at a* $P \in \mathcal{E}_{n\times n}$.

Proof. We have every $v_j^s k_i^m > 0$, and thus it follows from equation (2.4) and lemma 1a 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 $\sum_{j=1}^n p_{ij} = 1,\ for\ i = 1, \ldots, n$, that is, each local maximum $P \in \mathcal{E}_{n\times n}$.

Lemma 3.

a. Given $\boldsymbol{p} = [p_1, \ldots, p_n] \in \mathcal{P} = \{\boldsymbol{p} : p_j \geq 0,\ j = 1, \ldots, n, \sum_{j=1}^n p_j = 1\}$, *a local maximum of* $J(\boldsymbol{p}) = \sum_{j=1}^n p_j^2 \beta_j$ *is reached at:*

i. $\boldsymbol{p} = [e_k : \mathbf{0}]$ *for* $k \geq 1$ *with* $\beta_j > 0,\ j = 1, \ldots, k$ *and* $\beta_j < 0,\ j = k+1, \ldots, n$, *where the row vector* $e_k \in \mathcal{E}_{1\times k}$ *and* $\mathbf{0}$ *is a row vector consisting of* $n-k$ *zeros.*[2]

ii. One $\boldsymbol{p} \in \mathcal{P}$ *that is not in the form* $[e_k : \mathbf{0}]$, *when* $\beta_j < 0,\ j = 1, \ldots, n$.[3]

b. For an unknown $0 < k < n$ *with* $v_i^s > 0, k_i^m > 0, i = 1, \ldots, k$ *and* $v_i^s < 0, k_i^m < 0, i = k+1, \ldots, n$, *every local maximum of* $J(P)$ *in equation 2.3 is reached at*

$$P = \begin{bmatrix} P_1 & \mathbf{0} \\ \mathbf{0} & P_2 \end{bmatrix},\ P_1 \in \mathcal{E}_{k\times k},\ P_2 \in \mathcal{E}_{(n-k)\times(n-k)}. \tag{2.5}$$

Proof.

a. Let $J(\mathbf{p}) = J^+(\mathbf{p}^+) + J^-(\mathbf{p}^-)$ with $\mathbf{p} = [\mathbf{p}^+ : \mathbf{p}^-]$ and

$$J^+(\mathbf{p}^+) = \sum_{j=1}^{k} p_j^2 \beta_j, \quad J^-(\mathbf{p}^-) = \sum_{j=k+1}^{n} p_j^2 \beta_j. \tag{2.6}$$

$J^-(\mathbf{p}^-) \leq 0$ is strictly concave from lemma 1a because of $\beta_j < 0$, and thus it has only one maximum at $\mathbf{p}^- = \mathbf{0}$. Therefore, all the local maxima of $J(\mathbf{p})$ are reached at $[\mathbf{p}^+ : \mathbf{0}]$ and determined by all the local maxima of $J^+(\mathbf{p}^+)$ on the polyhedral set $\Gamma = \{\sum_{j=1}^k p_j = 1,\ p_j \geq 0,\ j = 1, \ldots, k\}$. It follows from lemma 1b that $J^+(\mathbf{p}^+)$ is strictly convex on Γ since $\beta_j > 0$. It further follows from lemma 1c that all its local maxima are reached at the extreme points of Γ, that is, each local maximum is reached at $\mathbf{p}^+ = e_k$.

[2] In the rest of this letter, we use $\mathbf{0}$ to denote a matrix of all its elements in zeros without explicitly indicating its dimension, including being either a row vector or a column vector.

[3] To illustrate, we observe $f(x, y) = -ax^2 - by^2, a > 0, b > 0$ subject to $x + y = 1, x \geq 0, y \geq 0$, A local maximum of $f(x, y)$ is reached at $x = \frac{b}{a+b}, y = \frac{a}{a+b}$.

Particularly for $\beta_j < 0$, $j = 1, \ldots, n$, it follows from lemma 1a that $J(\mathbf{p}) = \sum_{j=1}^{n} p_j^2 \beta_j$ is strictly concave. However, its only maximum at $\mathbf{p} = \mathbf{0}$ is not reachable because the constraint $\sum_{j=1}^{n} p_j = 1$. Instead, the maximum of $J(\mathbf{p})$ subject to this constraint is reached at one $\mathbf{p} \in \mathcal{P}$ that is not in the form $[e_k : \mathbf{0}]$.

b. Notice that the constraint $\sum_{j=1}^{n} p_{ij} = 1$ is imposed on only the ith row $\mathbf{p}^{(i)} = [p_{i1}, \ldots, p_{in}]$, and $J(P)$ in equation 2.3 is additive. The problem of finding the local maxima of $J(P)$ *s.t.* $P \in S$ is equivalent to separately finding the local maxima

$$J(\mathbf{p}^{(i)}) = \sum_{j=1}^{n} p_{ij}^2 v_j^s k_i^m, \text{ subject to } \sum_{j=1}^{n} p_{ij} = 1, \; p_{ij} \geq 0, \; j = 1, \ldots, n. \tag{2.7}$$

For each $i \leq k$, we have $\beta_{ij} = v_j^s k_i^m > 0$, *and* $j = 1, \ldots, k$ and $\beta_{ij} < 0$, $j = k+1, \ldots, n$. Therefore, it follows from lemma 3a that $\mathbf{p}^{(i)} = [e_k^{(i)} : \mathbf{0}]$ with $e_k^{(i)} \in \mathcal{E}_{1\times k}$, for $i = 1, \ldots, k$. For each $i > k$, we have $\beta_{ij} < 0$, $j = 1, \ldots, k$ and $\beta_{ij} > 0$, $j = k+1, \ldots, n$. Similarly, we have $\mathbf{p}^{(i)} = [\mathbf{0} : e_{n-k}^{(i)}]$ with $e_{n-k}^{(i)} \in \mathcal{E}_{1\times(n-k)}$, for $i = k+1, \ldots, n$. In summary, we get P given by equation 2.5.

From every local maximum solution $P = [p_{ij}]$ in lemmas 2 and 3b, all the matrices $R = [r_{ij}]$ with $r_{ij} = \pm\sqrt{p_{ij}}$ are local maxima $R = [r_{ij}]$ of $\max_{s.t.\ C_n} J(R)$. In other words, all the local maxima of $\max_{s.t.\ C_n} J(R)$ can be summarized as follows:

$$\mathcal{R} = \{R = [r_{ij}] : r_{ij} = \begin{cases} 0, & \text{if } p_{ij} = 0, \\ \pm 1 & \text{if } p_{ij} = 1. \end{cases}\}. \tag{2.8}$$

We will return to consider $J(R)$ on $RR^T = I$ by equation 1.2 by adding the orthogonal constraint C_o in equation 2.2 to $\max_{s.t.\ C_n} J(R)$. Local maxima in $\mathcal{R}$ that do not satisfy C_o will be discarded. Also, we show that adding in C_o will not bring any extra local maximum. Along this line, we can prove the following theorem:

Theorem 1. *Every local maximum of $J(R)$ on $RR^T = I$ by equation 1.2 is reached at R, which is a 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. We check every local maximum $J(R)$ *s.t.* C_n, that is, every solution in $\mathcal{R}$ by equation 2.8, whether it satisfies the orthogonal constraint C_o in equation 2.2. If not, it is discarded. If it does, it is put into the following solution set,

$$\mathcal{P} = \{R : \text{ every } R \in \mathcal{R} \text{ that satisfies } C_o\}, \tag{2.9}$$

which is not empty and each $P \in \mathcal{P}$ is either a permutation matrix or a variant with one or more nonzero elements switched to a negative sign. We check whether every $R^* \in \mathcal{P}$ is a local maximum of $J(R)$ on $RR^T = I$ by equation 1.2. Since R^* is a local maximum $J(R)$ *s.t.* C_n, there is a small enough neighbor area N_{R^*} with $R^* \in N_{R^*}$ such that every $R' \in N_{R^*}$, $R' \neq R^*$ satisfies C_n and $J(R') < J(R^*)$. If R^* also satisfies C_o, R^* must belong to the intersection $N_{R^*} \cap N_{C_o}$, where N_{C_o} denotes the set of all the points that satisfy C_o. It further follows from equation 2.2 that $N_{R^*} \cap N_{C_o}$ is also a neighbor area of R^*, within which we have $J(R) < J(R^*)$ for every $R \in N_{R^*} \cap N_{C_o}$, $R \neq R^*$. That is, each $R^* \in \mathcal{P}$ by equation 2.9 is a local maximum of $J(R)$ on $RR^T = I$ by equation 1.2.

We show that adding C_o to $J(R)$ *s.t.* C_n will not create any extra local maximum. Without C_o, it is linked by $p_{ij} = r_{ij}^2$ that the problem is equivalent to separately finding the local maxima of $J(\mathbf{p}^{(i)})$ by equation 2.7 for $i = 1, \ldots, n$, respectively. These individual optimizations may occur within the same n-dimensional space or across several different n-dimensional spaces that may have certain overlap. With C_o added in, these individual optimizations are forced to be implemented in the n different n-dimensional spaces that are orthogonal to each other. The effect of this process is equivalent to forcing the valid local maxima of $J(\mathbf{p}^{(i)})$ by equation 2.7 to be located on a more restricted polyhedral set. Without C_o, it follows from lemma 3a (i) that the valid local maxima of $J(\mathbf{p}^{(i)})$ by equation 2.7 should be located on a polyhedral set as follows:

$$\Gamma_i = \left\{ p_{ij} = 0 \text{ for } v_j^s k_i^m < 0, \;\; p_{ij} \geq 0 \text{ for } v_j^s k_i^m > 0 \text{ and } \sum_{p_{ij} \geq 0} p_{ij} = 1 \right\}. \tag{2.10}$$

With C_o added, Γ_i becomes more restrictive, with one or more constraints of type $p_{ij} \geq 0$ being further restricted into those of type $p_{ij} = 0$. Since there will be at least one P by equation 2.5 in an $n \times n$ nonsingular matrix, for every i there will be at least one j with its corresponding p_{ij} remaining to be type $p_{ij} \geq 0$; that is, no situation stated by lemma 3a (ii) occurs. Thus, the valid local maxima of $J(\mathbf{p}^{(i)})$ by equation 2.7 on a more restricted polyhedral set is merely a subset of those local maxima before imposing the extra restrictions. In other words, no extra local maximum will be created.

Summarizing the above two aspects and noticing that k_i^m has the same sign as the kurtosis v_i^m of the model density $q_i(y_i)$, the theorem is proved.

Equation 1.2 is obtained from equation 1.1 by considering only the skewness and kurtosis and with the model pdf's without skewness. 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 equation 1.1 as long as there is an one-to-one same-sign correspondence between the kurtosis of all source pdf's and the kurtosis of all model pdf's. Moreover, this approximation can be removed by an ICA algorithm obtained directly from equation 1.2.

Under the one-to-one kurtosis sign-matching assumption, we can derive a local search algorithm that is equivalent to maximizing the problem by equation 1.2 directly. A prewhitening is made on observed samples such that we can consider the samples of x with $Ex = 0$, $Exx^T = I$. As a result, it follows from $I = Exx^T = AEss^T A^T$ and $Ess^T = I$ that $AA^T = I$, that is, A is orthonormal. Thus, an orthonormal W is considered to let $y = Wx$ become independent among its components by

$$\max_{WW^T=I} J(W), \; J(W) = \sum_{i=1}^{n} k_i^m v_i^y, \tag{2.11}$$

where $v_i^y = Ey_i^4 - 3, i = 1, \ldots, n$, and $k_i^m, i = 1, \ldots, n$ are prespecified constants with the same sign as the kurtosis v_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 equations 3.17 and 3.18 at the end of section 3.3. Such an ICA algorithm actually maximizes the problem by equation 1.2 directly by noticing $y = Wx = WAs = Rs$, $R = WA$, $RR^T = I$, and thus

$$v_i^y = \sum_{j=1}^{n} r_{ij}^4 v_j^s, i = 1, \ldots, n. \tag{2.12}$$

That is, the problem by equation 2.11 is equivalent to the problem by equation 1.2. 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 in an exact sense as long as equation 2.12 holds.

However, theorem 1 does not tell us how such a kurtosis sign matching is built, which is attempted via equation 1.1 through learning each model pdf $q_i(y_i)$ together with learning W (Xu et al., 1996; Xu et al., 1997; Xu et al., 1998b) as well as further advances given in Lee et al. (1999) and Welling and Weber (2001) or by equation 103 in Xu (2003). Still, it remains an open problem whether these efforts or the possibility of

developing other techniques can guarantee such a one-to-one kurtosis sign-matching either surely or in some probabilistic sense, which deserves future investigations.

2.3 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 with the following lemma:

Lemma 4 (no matching). *When either* $v_i^s > 0, k_i^m < 0, \forall i$ *or* $v_i^s < 0, k_i^m > 0, \forall i$, $J(P)$ *has only one maximum that is not in* $\mathcal{E}_{n\times n}$.

Proof. From equation 2.4 and lemma 1a, $J(P)$ is strictly concave since $v_j^s k_i^m < 0$ for every term. Thus, it follows from lemma 1b that it has only one maximum that is not at the extreme points of S.

Lemma 5 (partial matching). *Given two unknown integers* k, m *with* $0 < k < m < n$ *and provided that* $v_i^s > 0, k_i^m > 0, i = 1, \ldots, k$, $v_i^s k_i^m < 0, i = k+1, \ldots, m$, *and* $v_i^s < 0, k_i^m < 0, i = m+1, \ldots, n$, *every local maximum of* $J(P)$ *is reached at* P *by equation 2.5 with*

$$P_1 \in \mathcal{E}_{m\times k}, P_2 \in \mathcal{E}_{(n-m)\times(n-k)}, \text{ when } v_i^s < 0, k_i^m > 0, i = k+1, \ldots, m;$$
$$P_1 \in \mathcal{E}_{k\times m}, P_2 \in \mathcal{E}_{(n-k)\times(n-m)}, \text{ when } v_i^s > 0, k_i^m < 0, i = k+1, \ldots, m. \quad (2.13)$$

Proof. When $v_i^s < 0, k_i^m > 0, i = k+1, \ldots, m$, for each $i \le m$, we have $\beta_{ij} = v_j^s k_i^m > 0$, $j = 1, \ldots, k$ and $\beta_{ij} < 0$, $j = k+1, \ldots, n$, and it follows from equation 2.7 and lemma 3a that $\mathbf{p}^{(i)} = [e_k^{(i)} : \mathbf{0}]$ with $e_k^{(i)} \in \mathcal{E}_{1\times k}$, for $i = 1, \ldots, m$. For each $i > m$, we have $\beta_{ij} < 0$, $j = 1, \ldots, k$ and $\beta_{ij} > 0$, $j = k+1, \ldots, n$, and we have $\mathbf{p}^{(i)} = [\mathbf{0} : e_{n-k}^{(i)}]$ with $e_{n-k}^{(i)} \in \mathcal{E}_{1\times(n-k)}$, for $i = m+1, \ldots, n$. Thus, we get P with $P_1 \in \mathcal{E}_{m\times k}$, $P_2 \in \mathcal{E}_{(n-m)\times(n-k)}$.

The situation of $v_i^s > 0, k_i^m < 0, i = k+1, \ldots, m$ is just a swap of the position (i, j). We can get P simply by matrix transposition.

Theorem 2. *Given two unknown integers* k, m *with* $0 < k < m < n$, *and provided that* $v_i^s > 0, k_i^m > 0, i = 1, \ldots, k$, $v_i^s k_i^m < 0, i = k+1, \ldots, m$, *and* $v_i^s < 0, k_i^m < 0, i = m+1, \ldots, n$, *every local maximum of* $J(R)$ *on* $RR^T = I$ *by equation 1.2 is reached at* $R = \begin{bmatrix} \Pi & 0 \\ 0 & \bar{R} \end{bmatrix}$ *subject to a* 2×2 *block 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 an* $(m-k) \times (m-k)$ *orthonormal matrix with* $\bar{R}\bar{R}^T = I$, *but usually not a permutation matrix up to sign indeterminacy.*

Proof. Similar to the proof of theorem 1, we consider the effect of imposing the orthogonal constraint C_o in equation 2.2 to $\mathcal{R}$ by equation 2.8 with P by equation 2.5 but P_1, P_2 by equation 2.13. We give more details for the cases of $v_i^s < 0, k_i^m > 0, i = 1, \ldots, k$, for each $i \leq m$, while the cases of $v_i^s > 0, k_i^m < 0, i = k+1, \ldots, m$ can be understood simply by matrix transposition.

Although the proof is similar to that for the proof of theorem 1, there is a key difference. Considering the row rank, $P_1 \in \mathcal{E}_{m\times k}$ is $k < m$, $P_2 \in \mathcal{E}_{(n-m)\times(n-m)}$ is $n-m$, and thus P by equation 2.5 is $k+n-m < n$. Since a permutation matrix remains a permutation matrix after any permutation, without losing generality, we can say that the first $k+n-m$ rows of $\mathcal{R}$ are at the rank $k+n-m$. C_o will not only force the rest of the $m-k$ columns of the $k+n-m$ rows to become $\mathbf{0}$ but also force every $p_{ij} \geq 0$ *for* $v_j^s k_i^m > 0$ in Γ_i by equation 2.10 to become type $p_{ij} = 0$ in the rest of the $m-k$ rows of $\mathcal{R}$ such that the first $k+n-m$ columns of the $m-k$ rows also become $\mathbf{0}$. To keep C_n in equation 2.2 satisfied, it has to be imposed on each row of the rest of the $m-k$ columns with all the $(m-k)\times(m-k)$ elements featured by $v_j^s k_i^m < 0$. As a result, we get $R = \begin{bmatrix} \Pi & 0 \\ 0 & \bar{R} \end{bmatrix}$ subject to a 2×2 block permutation. Being the same as that in theorem 2, Π is a $(k+n-m)\times(k+n-m)$ permutation matrix up to sign indeterminacy, while for $\bar{R}$, we have $\bar{R}\bar{R}^T = I$ from $RR^T = I$. Moreover, it follows from lemma 3a (ii) that the local maxima of $J(\mathbf{p}^{(i)})$ by equation 2.7 are reached at one $\mathbf{p}^{(i)}$ not in $\mathcal{E}_{1\times n}$. That is, $\bar{R}$ is usually not a permutation matrix up to sign indeterminacy.

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 and 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 $-$, that is, $p(\xi_i) = 0.5^{\xi_i} 0.5^{1-\xi_i}$. When there are k sources with their kurtosis signs positive, there is a probability $p(\sum_{i=1}^n \xi_i = k)$ of having a one-to-one kurtosis sign correspondence even when model pdf's are prefixed without knowing the kurtosis signs of sources. Moreover, even when a 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 early ICA studies (Amari et al., 1996; Bell & Sejnowski, 1995) work in some cases while failing in other cases due to the predetermined model pdf's, but also why some existing heuristic ICA algorithms always work to some extent.

2.4 Maximum Kurtosis versus Minimum Kurtosis. Interestingly, it can be observed that changing the maximization in equations 1.2, 2.3, and 2.11, into the minimization will lead to similar results, which are summarized in lemma and theorem.

Lemma 6.

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

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

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

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

Proof. The proof is similar to those in proving lemmas 2 to 5. The key difference is a shift in focus from the maximization of a convex function on a polyhedral set to the minimization of a concave function on a polyhedral set, with swaps between minimum and maximum, maxima and minima, convex and concave, and positive and negative, respectively. The key point is that lemma 1 remains correct after these swaps.

Similar to theorem 2, from the above lemma we get:

Theorem 3.

a. When either $v_i^s k_i^m < 0, i = 1, \ldots, n$ *or* $v_i^s < 0, k_i^m > 0, i = 1, \ldots, k$ *and* $v_i^s > 0, k_i^m < 0, i = k+1, \ldots, n$ *for an unknown* $0 < k < n$, *every local minimum of* $J(R)$ *on* $RR^T = I$ *by equation 1.2 is reached at a permutation matrix* R *up to sign indeterminacy at its nonzero elements.*

b. For two unknown integers k, m *with* $0 < k < m < n$ *with* $v_i^s > 0, k_i^m > 0, i = 1, \ldots, k$, $v_i^s k_i^m < 0, i = k+1, \ldots, m$, *and* $v_i^s < 0, k_i^m < 0, i = m+1, \ldots, n$, *every local minimum of* $J(R)$ *on* $RR^T = I$ *by equation 1.2 is reached at* $R = \begin{bmatrix} \Pi & 0 \\ 0 & \bar{R} \end{bmatrix}$ *subject to a* 2×2 *block permutation. When* $m + k \geq n$, Π *is an* $(n - m + n - k) \times (n - m + n - k)$ *permutation matrix up to sign indeterminacy at its nonzero elements, while* $\bar{R}$ *is an* $(m + k - n) \times (m + k - n)$ *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 an* $(n - k - m) \times (n - k - m)$ *orthonormal matrix with* $\bar{R}\bar{R}^T = I$, *but usually not a permutation matrix up to sign indeterminacy.*

In a comparison of theorems 2 and 3, when $m + k \geq n$, comparing $n - m + n - k$ with $k + n - m$, more sources can be separated by minimization

than maximization if $k < 0.5n$ and by maximization than minimization if $k > 0.5n$. When $m + k < n$, comparing $k + m$ with $k + n - m$, more sources can be separated by minimization than maximization if $m > 0.5n$, and by maximization than minimization if $m < 0.5n$.

We further consider a special case that $k_i^m = 1, \forall i$. In this case, equation 1.2 is simplified into

$$J(R) = \sum_{i=1}^{n} \sum_{j=1}^{n} r_{ij}^4 v_j^s, \ n \geq 2. \tag{2.14}$$

From theorem 2 at $n = m$, we can easily obtain the following corollary:

Corollary 1. *For an unknown integer $0 < k < n$ with $v_i^s > 0, i = 1, \ldots, k$ and $v_i^s < 0, i = k + 1, \ldots, n$, every local maximum of $J(R)$ on $RR^T = I$ by equation 2.14 is reached at $R = \begin{bmatrix} \Pi & \mathbf{0} \\ \mathbf{0} & \bar{R} \end{bmatrix}$ subject to a 2×2 block permutation, where Π is a $k \times k$ permutation matrix up to sign indeterminacy at its nonzero elements, while $\bar{R}$ is an $(n-k) \times (n-k)$ 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 an unknown integer k with $0 < k < n$ with $v_i^s > 0, i = 1, \ldots, k$ and $v_i^s < 0, i = k + 1, \ldots, n$, every local minimum of $J(R)$ on $RR^T = I$ by equation 1.2 is reached at $R = \begin{bmatrix} \bar{R} & \mathbf{0} \\ \mathbf{0} & \Pi \end{bmatrix}$ subject to a 2×2 block permutation, where Π is an $(n-k) \times (n-k)$ 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.*

According to corollary 1, k sources of supergaussian components are separated by $\max_{RR^T=I} J(R)$, while $n - k$ sources of subgaussian components are separated by $\min_{RR^T=I} J(R)$ according to corollary 2. In implementation, from equation 2.11, we get

$$J(W) = \sum_{i=1}^{n} v_i^y. \tag{2.15}$$

Then, from $\max_{WW^T=I} J(W)$ we get k sources of super gaussian components and from $\min_{WW^T=I} J(W)$ we get $n - k$ sources of subgaussian components. Thus, instead of learning one-to-one kurtosis sign matching, the problem can be turned into one of selecting supergaussian components from $y = Wx$ with W obtained via $\max_{WW^T=I} J(W)$ and of selecting subgaussian components from $y = Wx$ with W obtained via $\min_{WW^T=I} J(W)$. Though we know neither k nor which components of y should be selected, we can pick those

with positive signs as supergaussian ones after $\max_{WW^T=I} J(W)$ and pick those with negative signs as subgaussian ones after $\min_{WW^T=I} J(W)$. The reason comes from $v_i^y = \sum_{j=1}^n r_{ij}^4 v_j^s$ and the above corollaries. By corollary 1, the kurtosis of each supergaussian component of y is simply one of $v_j^s > 0$, $j = 1, \ldots, k$. Although the kurtosis of each of the other components in y is a weighted combination of $v_j^s < 0$, $j = k+1, \ldots, n$, the kurtosis signs of these will all remain negative. Similarly, we can find out those subgaussian components according to corollary 2.

Anther corollary can be obtained from equation 2.11 by considering a special case that $k_i^m = sign[v_i^y], \forall i$:

$$\max_{WW^T=I} J(W), \; J(W) = \sum_{i=1}^{n} \left| v_i^y \right|. \tag{2.16}$$

This leads to what is called a kurtosis extreme approach and extensions (Delfosse & Loubation, 1995; Moreau & Macchi, 1996; Hyvarinen et al., 2001), where studies began by extracting one source by a vector w and then extracting multiple sources by either sequentially implementing the one-vector algorithm such that the newly extracted vector is orthogonal to previous ones or implementing the one-vector algorithm on all the vectors of W in parallel together with a symmetric orthogonalization at each iterative step, as suggested by (Hyvarinen et al. 2001). In the literature, the success of using one 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 its previous ones. However, this mathematical proof is not applicable to the above symmetric orthogonalization-based implementation of the one-vector algorithm in parallel with all the vectors of W. Actually, what Hyvarinen et al., (2001) suggested can only ensure a convergence of a symmetric orthogonalization-based algorithm but cannot guarantee that this local searching-featured iterative algorithm will converge to a solution that can separate all the sources, though experiments have usually been successful.

When $v_i^y = \sum_{j=1}^n r_{ij}^4 v_j^s$ holds, which is true only when the prewhitening can be made perfectly, it follows from equation 2.16 that

$$\min_{RR^T=I} J(R), \; J(R) = \sum_{i=1}^{n} \sum_{j=1}^{n} r_{ij}^4 \left| v_j^s \right|, \tag{2.17}$$

which is covered by lemma 2 and theorem 2. Thus, we can directly prove the following corollary:

Corollary 3. *As long as* $v_i^y = \sum_{j=1}^n r_{ij}^4 v_j^s$ *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 symmetric orthogonalization-based implementation of the one-vector algorithm in parallel on separating all the sources.

3 Combinatorial Optimization, Distribution Approximation, and Stiefel Manifold

The combinatorial optimization problem by equation 1.3 has been encountered in various applications and still remains difficult to solve. Many efforts have also been made in the literature of neural networks since Hopfield and Tank (1985). As summarized in Xu (2003), these efforts can be roughly classified according to the features on dealing with C_e^{col}, C_e^{row}, and C_b. Although almost all the neural network motivated approaches are parallel implementable, they share one unfavorable feature: these intuitive approaches have no theoretical guarantee on convergence to even a feasible solution.

Interestingly, focusing on local maxima only, equations 1.2 and 2.3 can be regarded as special examples of the combinatorial optimization problem by equation 1.3 simply by regarding p_{ij} or r_{ij} as v_{ij}. Although such a linkage is not useful for ICA, where we do not need to seek a global optimization, a link from equation 1.3 to 1.2 and even 1.1 leads to two interesting questions. Could we consider the combinatorial optimization by equation 1.3 from the perspective of approximating one distribution by another simple model distribution, as in equation 1.1? Can we replace the constraints C_e^{col}, C_e^{row}, and C_b by the Stiefel manifold $RR^T = I$ for developing a more effective implementation?

3.1 A Distribution Approximation Perspective of Combinatorial Optimization.

Finding a global minimization solution of $E_o(V)$ under a set of constraints is equivalent to finding a global peak of the following Gibbs distribution,

$$p(V, \beta) = \frac{e^{-\frac{1}{\beta} E_o(V)}}{Z_\beta}, \quad Z_\beta = \sum_V e^{-\frac{1}{\beta} E_o(V)}, \tag{3.1}$$

subject to the constraints, since $\max_V p(V, \beta)$ is equivalent to $\max_V \ln p(V, \beta)$ or $\min_V E_o(V)$. Usually this $p(V, \beta)$ has many local maximums, so it is difficulty, to get the peak V_p. To avoid this difficulty, we use a simple distribution $q(V)$ to approximate $p(V, \beta)$ on a domain D_v such that the global peak of $q(V)$ is easy to find and that $p(V, \beta)$ and $q(V)$ share the

same peak $V_p \in D_v$, where D_v is considered using the following support of $p(V, \beta)$,

$$D_\varepsilon(\beta) = \{V : p(V, \beta) > \varepsilon,\ a\ small\ constant\ \varepsilon > 0\}, \tag{3.2}$$

under the control of a parameter β. For a sequence $\beta_0 > \beta_1, \ldots > \beta_t$, we have $D_\varepsilon(\beta_t) \subset \ldots D_\varepsilon(\beta_1) \subset D_\varepsilon(\beta_0)$, which includes the global minimization solution of $E_o(V)$, since the equivalence of $\max_V p(V, \beta)$ to $\min_V E_o(V)$ is irrelevant to β. Therefore, we can find a sequence $q_0(V), q_1(V), \ldots, q_t(V)$ that approximates $p(V, \beta)$ on the shrinking domain $D_\varepsilon(\beta)$. For a large β_t, $p(V, \beta)$ has large support, and thus $q(V)$ adapts the overall configuration of $p(V, \beta)$ in the large domain $D_\varepsilon(\beta)$. As β_t reduces, $q_t(V)$ concentrates more and more on adapting the detailed configuration of $p(V, \beta)$ around the global peak solution $V_p \in D_\varepsilon$. As long as β_0 is large enough and β reduces slowly enough toward zero, we can find the global minimization solution of $E_o(V)$. We adopt the following for implementing such a distribution approximation:

$$\min_p\ KL(q, p), \quad KL(q, p) = \sum_{V \in D_v} q(V) \ln \frac{q(V)}{p(V, \beta)}. \tag{3.3}$$

Alternatively, we can also consider $\min_p\ KL(q, p)$, which leads us to a class of Metroplois sampling-based mean-field approaches. (For details, see section II(B) in Xu, 2003.) Here, we consider only equation 3.3 with $q(V)$ in the following simple forms:

$$\begin{aligned} q_1(V) &= Z_1^{-1} \prod_{i,j} e^{v_{ij} \ln q_{ij}}, \quad 0 \le q_{ij} \le \infty, \quad Z_1 = \sum_{i,j} \prod_{i,j} e^{v_{ij} \ln q_{ij}}, \\ q_2(V) &= \prod_{i,j} q_{ij}^{v_{ij}} (1 - q_{ij})^{1 - v_{ij}}, \quad 0 \le q_{ij} \le 1, \end{aligned} \tag{3.4}$$

and from the constraints in equation 1.3, we have

$$C^c: \quad \sum_{i=1}^{N} \langle v_{ij} \rangle = 1,\ j = 1, \ldots, M,\ C^r: \quad \sum_{j=1}^{M} \langle v_{ij} \rangle = 1, i = 1, \ldots, N;$$

$$\langle v_{ij} \rangle = \begin{cases} q_{ij} \frac{Z_{ij}}{Z_1}, & \text{for } q_1(V), \\ q_{ij}, & \text{for } q_2(V), \end{cases} \quad Z_{ij} = \sum_{k \neq i, l \neq j} \prod_{k,l} e^{v_{kl} \ln q_{kl}}, \tag{3.5}$$

where $\langle x \rangle$ denotes the expectation of the random variable x. When N, M are large, we have $Z_{ij} \approx Z_1$, and thus $\langle v_{ij} \rangle \approx q_{ij}$ for the case of $q_1(V)$.

Putting equation 3.4 into equation 3.3 and after certain derivations as shown in Xu (2003), equation 3.3 becomes equivalent to

$$\min_{q_{ij}} E(\{q_{ij}\}), \text{ subject to equation 3.5,}$$

$$E(\{q_{ij}\}) = \frac{1}{\beta} E_o(\{q_{ij}\}) + \begin{cases} \sum_{ij} q_{ij} \ln q_{ij}, & \text{for } q_1(V), \\ \sum_{ij} [q_{ij} \ln q_{ij} + (1 - q_{ij}) \ln (1 - q_{ij})]. & \text{for } q_2(V). \end{cases}$$

The case for $q_1(V)$ interprets the Lagrange transform approach with the barrier $\sum_{i,j} v_{ij} \ln v_{ij}$ in Xu (1994) and justifies the intuitive treatment of simply regarding the discrete v_{ij} as an analog variable between the interval $[0, 1]$. From this perspective, these analog variables are the parameters of the simple distribution that we use to approximate the Gibbs distribution induced from the cost $E_o(V)$ of the discrete variables. Subsequently, a discrete solution will be recovered from these analog parameters of q_{ij}.

Similarly, the case for $q_2(V)$ interprets and justifies the Lagrange transform approach with the barrier $\sum_{i,j} [v_{ij} \ln v_{ij} + (1 - v_{ij}) \ln (1 - v_{ij})]$ in Xu (1995), where this barrier is intuitively argued to be better than the barrier $\sum_{i,j} v_{ij} \ln v_{ij}$ because it gives a U shape curve. Here, this intuitive preference can also be justified from equation 3.5 since there is an approximation $Z_{ij} \approx Z_1$ used for $q_1(V)$ while transforming v_{ij} into q_{ij}, but no approximation for $q_2(V)$. Moreover, both barriers are the special cases $S(v_{ij}) = v_{ij}$ and $S(v_{ij}) = v_{ij}/(1 - v_{ij})$ of a family of barrier functions equivalent to minimizing the leaking energy in the classical Hopfield network (Xu, 1995).

3.2 Lagrange-Enforcing Algorithms. A general iterative procedure was proposed in Xu (1994) and then refined in Xu (1995) for minimizing $E(\{q_{ij}\})$ by considering the following Lagrange barrier costs:

$$\begin{aligned} E(\{q_{ij}\}) = \frac{1}{\beta} E_o(\{q_{ij}\}) + \sum_{j=1}^{M} \lambda_j^{col} \left[\sum_{i=1}^{N} q_{ij} - 1 \right] \\ + B(q_{ij}) + \sum_{i=1}^{N} \lambda_i^{row} \left[\sum_{j=1}^{M} q_{ij} - 1 \right], \end{aligned}$$

$$B(q_{ij}) = \begin{cases} \sum_{ij} q_{ij} \ln q_{ij}, & q_1(V), \\ \sum_{ij} [q_{ij} \ln q_{ij} + (1 - q_{ij}) \ln (1 - q_{ij})], & q_2(V). \end{cases} \tag{3.6}$$

Following the derivation made in Xu (1994, 1995), it follows from $\frac{\partial E(\{q_{ij}\})}{\partial q_{ij}} = 0$ that

$$q_{ij}^e = \begin{cases} \frac{1}{e\ a_i b_j \exp\left(\frac{1}{\beta}\frac{\partial E_o(\{q_{ij}\})}{\partial q_{ij}}\right)}, & \text{for } q_1(V), \\ \frac{1}{1+a_i b_j \exp\left(\frac{1}{\beta}\frac{\partial E_o(\{q_{ij}\})}{\partial q_{ij}}\right)}, & \text{for } q_2(V); \end{cases} \quad a_i = \exp\left(\lambda_i^{row}\right), \quad b_j = \exp\left(\lambda_j^{col}\right) \tag{3.7}$$

when $\frac{\partial E_o(\{q_{ij}\})}{\partial q_{ij}}$ is irrelevant to q_{ij} and $\frac{\partial^2 E(\{q_{ij}\})}{\partial^2 q_{ij}} > 0$. If other variables are fixed at their old values, $E(\{q_{ij}\})$ is minimized at $q_{ij} = q_{ij}^e$, given by equation 3.7. Thus, we can update

$$\textit{either } q_{ij}^{new} = q_{ij}^e \textit{ or } q_{ij}^{new} = q_{ij}^{old} + \eta\left(q_{ij}^e - q_{ij}^{old}\right), \quad \text{for a } \eta > 0 \text{ small enough,} \tag{3.8}$$

which will reduce $E(\{q_{ij}\})$ monotonically, since each $q_{ij}^e - q_{ij}^{old}$ is the descending direction of $E(\{q_{ij}\})$ along the coordinate q_{ij}.

We can see that the constraints C^c, C^r are satisfied by q_{ij}^{new} as long as they are satisfied by both q_{ij}^{old} and q_{ij}^e. What needs to be done is to enforce

$$\sum_{i=1}^{N} q_{ij}^e = 1,\ j = 1, \ldots, M;\ \sum_{j=1}^{M} q_{ij}^e = 1, i = 1, \ldots, N, \tag{3.9}$$

which is achieved by an ENFORCING-LAGRANGE iteration loop, for example, by iteratively solving the above $N + M$ nonlinear equations with respect to $\{a_i\}_{i=1}^N$, $\{b_j\}_{j=1}^M$ or equivalently finding $\{a_i\}_{i=1}^N$, $\{b_j\}_{j=1}^M$ that reaches a global minimum (i.e., zero) of

$$P(\{q_{ij}\}) = \sum_{j=1}^{M}\left(\sum_{i=1}^{N} q_{ij}^e - 1\right)^2 + \sum_{i=1}^{N}\left(\sum_{j=1}^{M} q_{ij}^e - 1\right)^2. \tag{3.10}$$

As this ENFORCING-LAGRANGE loop converges, we have

$$\left|\sum_{j=1}^{M} \lambda_j^{col}\left[\sum_{i=1}^{N} q_{ij} - 1\right] + \sum_{i=1}^{N} \lambda_i^{row}\left[\sum_{j=1}^{M} q_{ij} - 1\right]\right| < \varepsilon \tag{3.11}$$

for an arbitrarily small ε. Thus, the fact that equation 3.8 can reduce $E(\{q_{ij}\})$ monotonically is equivalent to being able to reduce $\frac{1}{\beta}E_o(\{q_{ij}\}) + B(q_{ij})$

monotonically. When β becomes small enough, it is also equivalent to equation 3.8 reducing $E_o(\theta)$ monotonically under equation 3.11, because $B(q_{ij})$ is bounded for $q_{ij} \in [0, 1]$.

In summary, we get the following iterative procedure that guarantees convergence to a feasible solution that is a minimum of $E_o(\{q_{ij}\})$ and satisfies C^c and C^r:

Step 0: Initialize $\{q_{ij}\}$ such that they satisfy the constraints C_e^c, C_e^r by equation 3.5,

Step 1: Get $\{q_{ij}^e\}$ by equation 3.7 and then call an inner ENFORCING-LAGRANGE loop until it converges or terminates according to a given checking criterion on the satisfaction of equation 3.11. It results in a specific setting on a_i, b_j, $j = 1, \ldots, M, i = 1, \ldots, N$.

Step 2: Update q_{ij}^{old} to q_{ij}^{new} by using equation 3.8 either sequentially or in parallel.

Step 3: Check whether the procedure converges according to a pregiven criterion. If yes, stop; otherwise, go to step 1.

Based on the obtained $\{q_{ij}\}$, we can get a discrete solution simply by a threshold $T_h > 0$ as follows:

$$v_{ij} = \begin{cases} 1, & \text{if } q_{ij} > T_h, \\ 0, & \text{otherwise,} \end{cases} \quad \text{and resolve a tie heuristically.} \tag{3.12}$$

Moreover, we can also consider explicitly the constraints C_e^c and get

$$v_{ij} = \begin{cases} 1, & \text{if } j = arg\max_k q_{ik}, \\ 0, & \text{otherwise,} \end{cases} \quad \text{and resolve a tie heuristically.} \tag{3.13}$$

Similarly, we can also consider explicitly the constraints C_e^r, as well as both C_e^c and C_e^r.

3.3 Stiefel Gradient Flow Algorithms. Alternatively, the study on equation 1.2 via equation 2.3 motivates another way to handle the constraints C_e^{col}, C_e^{row}, and C_b: 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., $0 \le v_{ij} \le 1$). That is, problem equation 1.3 becomes

$$\min_{RR^T = I \ for \ N \le M} E_o\left(\left\{r_{ij}^2\right\}_{i=1,j=1}^{i=N,j=M}\right), \quad R = \{r_{ij}\}_{i=1,j=1}^{i=N,j=M}. \tag{3.14}$$

We consider the problems with

$$\frac{\partial^2 E_o(V)}{\partial vec[V] \partial vec[V]^T} \text{ is negative definite,} \tag{3.15}$$

or $E_o(V)$ in a form similar to $J(P)$ in equation 2.3,

$$E_o(V) = -\sum_{i=1}^{n}\sum_{j=1}^{n} v_{ij}^2 a_j b_i, \tag{3.16}$$

with $a_i > 0, b_i > 0, i = 1, \ldots, k$ and $a_i < 0, b_i < 0, i = k+1, \ldots, n$ after an appropriate permutation on either or both of $[a_1, \ldots, a_n]$ and $[b_1, \ldots, b_n]$. Similar to the study of equation 2.3, maximizing $E_o(V)$ under the constraints C_e^{col}, C_e^{row}, and $v_{ij} \geq 0$ will imply satisfying C_b. In other words, the solutions of equation 3.14 and equation 1.3 are the same. Thus, we can solve the hard problem of combinatorial optimization by equation 1.3 using a gradient flow on the Stiefel manifold $RR^T = I$ to maximize the problem by equation 3.14. At least a local optimal solution of equation 1.3 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

$$\delta R = ZR + U(I - R^T R), \; U \text{ is a } m \times d \text{ matrix and}$$
$$Z = -Z \text{ is an asymmetric matrix.} \tag{3.17}$$

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

$$Z = G_R R^T - R G_R^T, \; U = G_R(I - R^T R), \delta R = \begin{cases} U(I - R^T R) = U, & \text{(a)} \\ ZR, & \text{(b)} \\ ZR + U, & \text{(c)} \end{cases}$$

$$R^{new} = R^{old} + \gamma_t \delta R. \tag{3.18}$$

That is, we can use one of the above three choices of δR as the updating direction of R. A general technique for optimization on the Stiefel manifold

that was elaborately discussed by Edelman, Arias, and Smith (1998) can also be adopted for implementing the problem by equation 3.14.

4 Conclusion

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 equation 1.2 is reached at a permutation matrix up to a certain sign indeterminacy if there is a one-to-one same-sign correspondence between the kurtosis signs of all source pdfs 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 been provided 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 to state that seeking a one-to-one same-sign correspondence can be replaced by using the duality; supergaussian sources can be separated by maximization and subgaussian sources can be separated by minimization. Furthermore, a corollary is obtained to provide a mathematical proof of the success of symmetric orthogonalization implementation of the kurtosis extreme approach.

There still remain problems for study. First, the success of the efforts based on equation 1.1 (Xu et al., 1996; Xu et al., 1997; Xu et al., 1998b; Lee et al., 1999; Welling & Weber, 2001; Xu, 2003) can be explained as their ability to build up a one-to-one kurtosis sign matching. However, we still need a mathematical analysis to guarantee that these approaches can achieve this matching exactly or in some probabilistic sense. Second, a theoretical guarantee on either the kurtosis extreme approach or the approach of extracting supergaussian sources via maximization and subgaussian sources via minimization is true only when the prewhitening can be made perfectly. It remains to be studied on comparison of the two approaches as well as approaches based on equation 1.1. Also, comparison may deserve to be made on convergence rates of different ICA algorithms.

Last, but not least, the linkage of the problem by equation 1.3 to equation 1.2 and equation 2.3, as well as equation 1.1, leads us to both a distribution approximation and a Stiefel manifold perspective of combinatorial optimization with algorithms that guarantee both convergence and satisfaction of constraints, which also deserve further investigation.

Acknowledgments

The work was done in preparation for a possible RGC earmarked grant (CUHK 417707E) to be supported by the Research Grant Council of the Hong Kong SAR. A preliminary version of this work was presented as a plenary talk at the Second International Symposium of Neural Networks (Xu, 2005).

References

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

Amari, S. I., Cichocki, A., & Yang, H. (1996). A new learning algorithm for blind separation of sources. In D. Toureteky, M. Mozer, & M. Hasselmo (Eds.), *Advances in neural information processing, 8* (pp. 757–763). Cambridge, MA: MIT Press.

Bazaraa, M. S., Sherall, H. D., & Shetty, C. M. (1993). *Nonlinear programming: Theory and algorithms*. New York: Wiley.

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

Cardoso, J.-F. (1998). Blind signal separation: Statistical principles. *Proc. IEEE, 86*(10), 2009–2025.

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.

Delfosse, N., & Loubation, P. (1995). Adaptive blind separation of independent sources: A deflation approach. *Signal Processing, 45*, 59–83.

Edelman, A., Arias, T. A., & Smith, S. T. (1998). The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl., 20*, 303–353.

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.

Hopfield, J. J., & Tank, D. W. (1985). Neural computation of decisions in optimization problems. *Biological Cybernetics, 52*, 141–152.

Hyvarinen, A., Karhunen, J., & Oja, A. (2001), *Independent component analysis*. New York: Wiley.

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.

Liu, Z. Y., Chiu, K. C., & Xu, L. (2004). One-bit-matching conjecture for independent component analysis. *Neural Computation, 16*, 383–399.

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

Pearlmutter, B. A., & Parra, L. C. (1996). A context-sensitive genaralization of ICA. In *Proc. of Int. Conf. on Neural Information Processing*, 151–157. Hong Kong: Springer-Verlag.

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. (1994). Combinatorial optimization neural nets based on a hybrid of Lagrange and transformation approaches. In *Proc. of World Congress on Neural Networks* (pp. 399–404). San Diego, CA.

Xu, L., (1995). On the hybrid LT combinatorial optimization: New U-shape barrier, sigmoid activation, least leaking energy and maximum entropy. In *Proc. of Intl. Conf. on Neural Information Processing* (pp. 309–312). Beijing, China.

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

Xu, L. (2003). Distribution approximation, combinatorial optimization, and Lagrange-barrier. In *Proc. of International Joint Conference on Neural Networks 2003* (pp. 2354–2359). Piscataway, NJ: IEEE.

Xu, L. (2005). One-bit-matching ICA theorem, convex-Concave programming, and Combinatorial optimization. In *Advances in neural networks* (pp. 5–20). Berlin: Springer-Verlag.

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 International ICSC Workshop on Independence and Artificial Neural Networks* (pp. 39–44). Tenerife, Spain.

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. In *Proc. of 1997 IEEE-INNS International Joint Conference on Neural Networks* (Vol. 3, pp. 1821–1826). Piscataway, NJ: IEEE.

Xu, L., Yang, H. H., & Amari, S. I. (1996, April). *Signal source separation by mixtures: Accumulative distribution functions or mixture of bell-shape density distribution functions*. Paper presented at the Frontier Forum, Institute of Physical and Chemical Research, Japan.

Received January 18, 2005; accepted June 20, 2006.