

# Cross-Entropy Optimization for Independent Process Analysis

Zoltán Szabó, Barnabás Póczos, and András Lőrincz

Department of Information Systems,  
Eötvös Loránd University, Budapest, Hungary,  
Research Group on Intelligent Information Systems,  
Hungarian Academy of Sciences  
szzoli@cs.elte.hu, pbarn@cs.elte.hu, lorincz@inf.elte.hu  
<http://nipg.inf.elte.hu>

**Abstract.** We treat the problem of searching for hidden multi-dimensional independent auto-regressive processes. First, we transform the problem to Independent Subspace Analysis (ISA). Our main contribution concerns ISA. We show that under certain conditions, ISA is equivalent to a combinatorial optimization problem. For the solution of this optimization we apply the cross-entropy method. Numerical simulations indicate that the cross-entropy method can provide considerable improvements over other state-of-the-art methods.

## 1 Introduction

Search for independent components is in the focus of research interest. There are important applications in this field, such as blind source separation, blind source deconvolution, feature extraction and denoising. Thus, a variety of particular methods have been developed over the years. For a recent review on these approaches and for further applications, see [1] and the references therein.

Originally, Independent Component Analysis (ICA) is 1-dimensional in the sense that all sources are assumed to be independent real valued stochastic variables. The typical example of ICA is the so called *cocktail-party problem*, where there are  $n$  sound sources and  $n$  microphones and the task is to separate the original sources from the observed mixed signals. However, applications where not all, but only certain groups of the sources are independent may have high relevance in practice. In this case, independent sources can be multi-dimensional. For example, consider the following generalization of the cocktail-party problem. There could be independent groups of people talking about independent topics, or more than one group of musicians may be playing at a party. This is the Independent Subspace Analysis (ISA) extension of ICA, also called Multi-dimensional Independent Component Analysis [2]. An important application is, e.g., the processing of EEG-fMRI data [3]. However, the motivation of our work stems from the fact that most practical problems, alike to the analysis of EEG-fMRI signals, exhibit considerable temporal correlations. In such cases, one may take

advantage of Independent Process Analysis (IPA) [4], a generalization of ISA for auto-regressive (AR) processes, similar to the AR generalization of the original ICA problem [5].

Efforts have been made to develop ISA algorithms [2, 3, 6, 7, 8, 9, 10]. Theoretical problems are mostly connected to entropy and mutual information estimations. Entropy estimation by Edgeworth expansion [3] has been extended to more than 2 dimensions and has been used for clustering and mutual information testing [11]. Other recent approaches search for independent subspaces via kernel methods [7], joint block diagonalization [10],  $k$ -nearest neighbor [8], and geodesic spanning trees [9].

Here, we shall explore a particular approach that tries to solve the ISA problem by ICA transformation and then searches for an optimal permutation of the ICA components. We shall investigate sufficient conditions that justify this algorithm. Different methods for solving the IPA and the related ISA problems will be compared. The paper is constructed as follows: Section 2 formulates the problem domain and suggests a novel approach for solving the related ISA task. Section 3 contains computer studies. Conclusions are also drawn here.

## 2 The IPA Model

We shall treat the generative model of mixed independent AR processes. Assume that we have  $M$  hidden and independent AR processes and that only the mixture of these  $M$  components is available for observation:

$$\mathbf{s}^m(t+1) = \mathbf{F}^m \mathbf{s}^m(t) + \mathbf{e}^m(t), \quad m = 1, \dots, M \quad (1)$$

$$\mathbf{z}(t) = \mathbf{A} \mathbf{s}(t), \quad (2)$$

where  $\mathbf{s}(t) := [\mathbf{s}^1(t); \dots; \mathbf{s}^M(t)]$  is the vector concatenated form of the components  $\mathbf{s}^m$ ,  $\mathbf{s}^m(t), \mathbf{e}^m(t) \in \mathbb{R}^d$ ,  $\mathbf{e}^m(t)$  is i.i.d. in  $t$ ,  $\mathbf{e}^i(t)$  is independent from  $\mathbf{e}^j(t)$ , if  $i \neq j$ , and  $\mathbf{F}^m \in \mathbb{R}^{d \times d}$ . The total dimension of the components is  $D := d \cdot M$ ,  $\mathbf{s}(t), \mathbf{z}(t) \in \mathbb{R}^D$  and  $\mathbf{A} \in \mathbb{R}^{D \times D}$  is the so called *mixing matrix* that, according to our assumptions, is invertible. It is easy to see that the invertibility of  $\mathbf{A}$  and the reduction step using innovations (see later in Section 2.1) allow, without any loss of generality, to restrict (i) to *whitened* noise process  $\mathbf{e}(t) := [\mathbf{e}^1(t); \dots; \mathbf{e}^M(t)]$ , and (ii) to orthogonal matrix  $\mathbf{A}$ . That is,

$$E[\mathbf{e}(t)] = \mathbf{0}, E[\mathbf{e}(t)\mathbf{e}(t)^T] = \mathbf{I}_D, \quad \forall t \quad (3)$$

$$\mathbf{I}_D = \mathbf{A} \mathbf{A}^T, \quad (4)$$

where  $\mathbf{I}_D$  is the  $D$ -dimensional identity matrix, superscript  $T$  denotes transposition and  $E[\cdot]$  is the expectation value operator. The goal of the IPA problem is to estimate the original source  $\mathbf{s}(t)$  and the unknown mixing matrix  $\mathbf{A}$  (or its inverse  $\mathbf{W}$ , which is called the *separation matrix*) by using observations  $\mathbf{z}(t)$  only. If  $\forall \mathbf{F}^m = \mathbf{0}$  then the task reduces to the ISA task. The ICA task is recovered if both  $\forall \mathbf{F}^m = \mathbf{0}$  and  $d = 1$ .

### 2.1 Uncertainties of the IPA Model

The identification of the IPA model, alike to the identification of the ICA and ISA models, is ambiguous. First, we shall reduce the IPA task to the ISA task [5, 12, 4] by means of *innovations*. The innovation of a stochastic process  $\mathbf{u}(t)$  is the error of the optimal quadratic estimation of the process using its past, i.e.,

$$\tilde{\mathbf{u}}(t) := \mathbf{u}(t) - E[\mathbf{u}(t)|\mathbf{u}(t-1), \mathbf{u}(t-2), \dots]. \tag{5}$$

It is easy to see that for an AR process, the innovation is identical to the noise that drives the process. Therefore, constructing a block-diagonal matrix  $\mathbf{F}$  from matrices  $\mathbf{F}^m$ , the IPA model assumes the following form

$$\mathbf{s}(t+1) = \mathbf{F}\mathbf{s}(t) + \mathbf{e}(t), \tag{6}$$

$$\mathbf{z}(t) = \mathbf{A}\mathbf{F}\mathbf{A}^{-1}\mathbf{z}(t-1) + \mathbf{A}\mathbf{e}(t-1), \tag{7}$$

$$\tilde{\mathbf{z}}(t) = \mathbf{A}\mathbf{e}(t-1) = \mathbf{A}\tilde{\mathbf{s}}(t). \tag{8}$$

Thus, applying ISA to innovation  $\tilde{\mathbf{z}}(t)$  of the observation, mixing matrix  $\mathbf{A}$  and thus  $\mathbf{e}(t)$  as well as  $\mathbf{s}(t)$  can be determined.

Concerning the ISA task, if we assume that both the components and the observation are white, that is,  $E[\mathbf{s}] = \mathbf{0}$ ,  $E[\mathbf{s}\mathbf{s}^T] = \mathbf{I}_D$  and  $E[\mathbf{z}] = \mathbf{0}$ ,  $E[\mathbf{z}\mathbf{z}^T] = \mathbf{I}_D$ , the ambiguity of the problem is lessened: apart from permutations, the components are determined up to orthogonal transformations within the subspaces. It also follows from the whitening assumption that mixing matrix  $\mathbf{A}$  (and thus matrix  $\mathbf{W} = \mathbf{A}^{-1}$ ) are orthogonal, because:

$$\mathbf{I}_D = E[\mathbf{z}\mathbf{z}^T] = \mathbf{A}E[\mathbf{s}\mathbf{s}^T]\mathbf{A}^T = \mathbf{A}\mathbf{I}_D\mathbf{A}^T = \mathbf{A}\mathbf{A}^T. \tag{9}$$

Identification ambiguities of the ISA task are detailed in [13].

### 2.2 Reduction of ISA to ICA and Permutation Search

Here, we shall reduce the original IPA task further. The ISA task can be seen as the minimization of mutual information between the components. That is, we should minimize cost function  $J(\mathbf{W}) := \sum_{m=1}^M H(\mathbf{y}^m)$  in the space of  $D \times D$  orthogonal matrices, where  $\mathbf{y} = \mathbf{W}\mathbf{z}$ ,  $\mathbf{y} = [\mathbf{y}^1; \dots; \mathbf{y}^M]$ ,  $\mathbf{y}^m$  ( $m = 1, \dots, M$ ) are the estimated components and  $H$  is Shannon's (multi-dimensional) differential entropy (see, e.g., [4]). Now, we present our main result:

**Theorem (Separation theorem for ISA).** *Let us suppose, that all the  $\mathbf{u} = [u_1; \dots; u_d] = \mathbf{s}^m$  components of source  $\mathbf{s}$  in the ISA task satisfy*

$$H\left(\sum_{i=1}^d w_i u_i\right) \geq \sum_{i=1}^d w_i^2 H(u_i), \forall \mathbf{w} : \sum_{i=1}^d w_i^2 = 1. \tag{10}$$

Now, processing observation  $\mathbf{z}$  by ICA, and assuming that the ICA separation matrix  $\mathbf{W}_{ICA}$  is unique up to permutation and sign of the components, then

$\mathbf{W}_{ICA}$  is also the separation matrix of the ISA task up to permutation and sign of the components. In other words, the  $\mathbf{W}$  separation matrix of the ISA task assumes the following form  $\mathbf{W} = \mathbf{P}\mathbf{W}_{ICA}$ , where  $\mathbf{P} (\in \mathbb{R}^{D \times D})$  is a permutation matrix to be determined.

The proof of the theorem can be found in a technical report [14] because of lack of space. Sources that satisfy the conditions of the theorem are also provided in [14], where we show that elliptically symmetric sources, among others, satisfy the condition of the theorem.

In sum, the IPA model can be estimated by applying the following steps:

1. observe  $\mathbf{z}(t)$  and estimate the AR model,
2. whiten the innovation of the AR process and perform ICA on it,
3. solve the combinatorial problem: search for the permutation of the ICA sources that minimizes the cost function  $J$ .

Thus, after estimating the AR model and performing ICA on its estimated innovation process, IPA needs only two steps: (i) estimation of multi-dimensional entropies, and (ii) optimization of the cost function  $J$  in  $S_D$ , the permutations of length  $D$ .

A recent work [4] provides an algorithm to solve the IPA task. To our best knowledge, this is the only algorithm for this task at present. This algorithm applies Jacobi rotations for any pairs of the elements received after ICA preprocessing. We shall call it the *ICA-Jacobi* method and compare it with our novel algorithm that we refer to as the *ICA-TSP* method for reasons to be explained later. For entropy estimation, we shall apply the method suggested in [4], which is the following:

### 2.3 Multi-dimensional Entropy Estimation by the $k$ -Nearest Neighbor Method

Shannon's entropy can be estimated by taking the limit of Rényi's entropy, which has efficient estimations. Let  $f$  denote the probability density of  $d$ -dimensional stochastic variable  $\mathbf{u}$ . Rényi's  $\alpha$ -entropy of variable  $\mathbf{u}$  ( $1 \neq \alpha > 0$ ) is defined as:

$$H_\alpha(\mathbf{u}) := \frac{1}{1-\alpha} \log \int_{\mathbb{R}^d} f^\alpha(v) dv \xrightarrow{\alpha \rightarrow 1} H(\mathbf{u}). \quad (11)$$

Assume that we have i.i.d. samples of  $T$  elements from the distribution of  $\mathbf{u}$ :  $\mathbf{u}(1), \dots, \mathbf{u}(T)$ . For each sample  $\mathbf{u}(t)$  let us choose the  $k$  samples, which are the closest to  $\mathbf{u}(t)$  in Euclidean norm ( $\|\cdot\|$ ). Let this set be denoted by  $\mathcal{N}_{k,t}$ . Let us choose  $\alpha := \frac{d-\gamma}{d}$ , and thus  $\alpha \rightarrow 1$  corresponds to  $\gamma \rightarrow 0$ . Then, under mild conditions, the Beadword-Halton-Hammersley theorem holds [15, 16]:

$$\hat{H}(k, \gamma) := \frac{1}{1-\alpha} \log \left( \frac{1}{T^\alpha} \sum_{t=1}^T \sum_{\mathbf{v} \in \mathcal{N}_{k,t}} \|\mathbf{v} - \mathbf{u}(t)\|^\gamma \right) \xrightarrow{T \rightarrow \infty} H_\alpha(\mathbf{u}) + c, \quad (12)$$

where  $c$  is an irrelevant constant. This entropy estimation is asymptotically unbiased and strongly consistent [15]. In the numerical studies, we shall use  $\gamma = 0.01$  and  $k = 3$  alike to [4].

## 2.4 Cross-Entropy Method for Combinatorial Optimization

The CE method has been found efficient for combinatorial optimization problems [17]. The CE technique operates as a two step procedure: First, the problem is converted to a stochastic problem and then the following two-phases are iterated (for detailed description, see [17]):

1. Generate  $\mathbf{x}_1, \dots, \mathbf{x}_N \in \mathcal{X}$  samples from a distribution family parameterized by a  $\theta$  parameter and choose the *elite* of the samples. The elite is the best  $\rho\%$  of the samples according to the cost function  $J$ .
2. Modify the sample generation procedure ( $\theta$ ) according to the elite samples. In practice, smoothing, i.e.,  $\theta^{new} = \beta \cdot \theta^{proposed} + (1 - \beta) \cdot \theta^{old}$  is utilized in the update of  $\theta$ .

This technique will be applied in our search for permutation matrix  $\mathbf{P}$ . Our method is similar to the CE solution suggested for the Travelling Salesman Problem (TSP) (see [17]) and we call it *ICA-TSP* method. In the TSP problem, a permutation of cities is searched for. The objective is to minimize the cost of the travel. We are also searching for a permutation, but now travel cost is replaced by  $J(\mathbf{W})$ . Thus, in our case,  $\mathcal{X} = S_D$  and  $\mathbf{x}$  is an element of this permutation group. Further, CE cost  $J$  equals to  $J(\mathbf{P}_x \mathbf{W}_{ICA})$ , where  $\mathbf{P}_x$  denotes the permutation matrix associated to  $\mathbf{x}$ . Thus, optimization concerns permutations in  $\mathcal{X}$ . In the present work,  $\theta$  contains transition probabilities  $i \rightarrow j$  ( $1 \leq i, j \leq D$ ), called *node transition* parametrization in the literature [17].

The above iteration is stopped if there is no change in the cost (in the last  $L$  steps), or the change in parameter  $\theta$  is negligibly small (smaller than  $\epsilon$ ).

## 3 Numerical Studies

### 3.1 Databases

Computer simulations are presented here. We defined four different databases. They were whitened and were used to drive the AR processes of Eq. (1). Then the AR processes were mixed. Given the mixture, an AR process was identified and its innovation was computed. The innovation was analyzed by ISA. We note that this reduction step using innovations based on AR estimation ('AR-trick') can also work for non-AR processes, as it was demonstrated in [4].

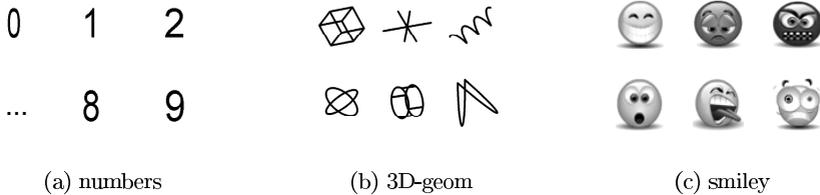
Three of the four computational tasks are shown in Fig. 1. In these test examples (i) dimensions  $D$  and  $d$  were varied ( $D = 12, 18, 20$ ,  $d = 2, 3, 4$ ), (ii) sample number  $T$  was incremented by 100 between 300 and 1500. For all tests, we averaged the results of 10 computer runs. In the fourth task  $M (= 5)$  pieces of  $d (= 4)$ -dimensional components were used and the innovation for each  $d$ -dimensional process was created as follows: coordinates  $u_i(t)$  ( $i = 1, \dots, k$ ), were uniform random variables on the set  $\{0, \dots, k-1\}$ , whereas  $u_{k+1}$  was set to  $\text{mod}(u_1 + \dots + u_k, k)$ . In this construction, every  $k$ -element subset of  $\{u_1, \dots, u_{k+1}\}$  is made of independent variables. This database is called the *all- $k$ -independent* problem [9]. In our simulations  $d = k + 1$  was set to 4.

Numerical values of the CE parameters were chosen as  $\rho = 0.05$   $\beta = 0.4$ ,  $L = 7$ ,  $\epsilon = 0.005$ . The quality of the algorithms was measured by the generalized Amari-distance.

**Generalized Amari-Distance.** The optimal estimation of the IPA model provides matrix  $\mathbf{B} := \mathbf{W}\mathbf{A}$ , a permutation matrix made of  $d \times d$  sized blocks. Let us decompose matrix  $\mathbf{B} \in \mathbb{R}^{D \times D}$  into  $d \times d$  blocks:  $\mathbf{B} = [\mathbf{B}^{ij}]_{i,j=1,\dots,M}$ . Let  $b^{i,j}$  denote the sum of the absolute values of the elements of matrix  $\mathbf{B}^{i,j} \in \mathbb{R}^{d \times d}$ . Then the normalized version of the generalized Amari-distance (see also [9, 10]) is defined as:

$$r(\mathbf{B}) := \frac{1}{2M(M-1)} \cdot \left( \sum_{i=1}^M \left( \frac{\sum_{j=1}^M b^{ij}}{\max_j b^{ij}} - 1 \right) + \sum_{j=1}^M \left( \frac{\sum_{i=1}^M b^{ij}}{\max_i b^{ij}} - 1 \right) \right) \quad (13)$$

For matrix  $\mathbf{B}$  we have that  $0 \leq r(\mathbf{B}) \leq 1$ , and  $r(\mathbf{B}) = 0$  if, and only if  $\mathbf{B}$  is a block-permutation matrix with  $d \times d$  sized blocks.



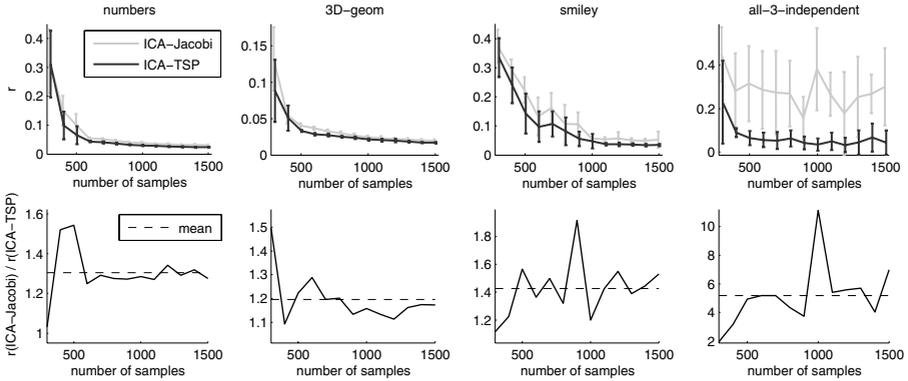
**Fig. 1.** 3 test databases: densities of  $e^m$ . Each object represents a probability density. Left: *numbers*:  $10 \times 2 = 20$ -dimensional problem, uniform distribution on the images of numbers. Middle: *3D-geom*:  $6 \times 3 = 18$ -dimensional problem, uniform distribution on 3-dimensional geometric objects. Right: *smiley*: 6 basic facial expressions [18], non-uniform distribution defined in 2 dimensions,  $6 \times 2 = 12$ -dimensional problem.

### 3.2 Results and Discussion

The precision of the procedures is shown in Fig. 2 as a function of the sample number. In the ICA-Jacobi method we applied exhaustive search for all Jacobi pairs with 50 angles between  $[0, \pi/2]$  several times until convergence. Still, the ICA-TSP is superior in all of the studied examples. Quantitative results are shown in Table 1. The innovations estimated by the ICA-TSP method on facial expressions are illustrated in Fig. 3.

We observed that the greedy ICA-Jacobi method seems to be similar or sometimes inferior to the global ICA-TSP, in spite of the much smaller search space available for the latter. We established rigorous conditions when the ICA-TSP is sufficient to find a global minimum, which justifies our finding. In the reduced search space of permutations, the global CE method was very efficient.

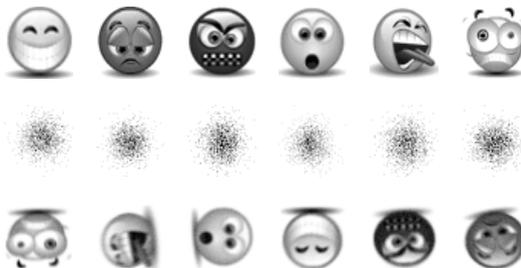
We make two notes: (1) Simulations indicate that conditions of the ‘Separation Theorem’ may be too restrictive. (2) For the IPA problem, the subspaces (the



**Fig. 2.** Mean±standard deviation of generalized Amari-distances as a function of sample number (upper row). Gray: ICA-Jacobi, black: ICA-TSP. In the lower row, black: precision of relative estimation, dashed: average over the different sample numbers. Columns from left to right correspond to databases ‘numbers’, ‘3D-geom’, ‘smiley’, ‘all-3-independent’, respectively.

**Table 1.** Column 1: test databases. Columns 2 and 3: average Amari-errors (in  $100 \cdot r\% \pm$  standard deviation) for 1500 samples on the different databases. Column 4: precision of the ICA-TSP relative to that of ICA-Jacobi in sample domain 300 – 1500.

Database	ICA-Jacobi	ICA-TSP	Improvement (min - mean - max)
numbers	3.06% ( $\pm 0.22$ )	2.40% ( $\pm 0.11$ )	1.03 - 1.30 - 1.54
3D-geom	1.99% ( $\pm 0.17$ )	1.69% ( $\pm 0.10$ )	1.09 - 1.20 - 1.50
smiley	5.26% ( $\pm 2.76$ )	3.44% ( $\pm 0.36$ )	1.16 - 1.43 - 1.92
all-3-independent	30.05% ( $\pm 17.90$ )	4.31% ( $\pm 5.61$ )	1.96 - 5.18 - 11.12



**Fig. 3.** Illustration of the ICA-TSP algorithm on the ‘smiley’ database. Upper row: density function of the sources (using  $10^6$  data points). Middle row: 1,500 samples of the observed mixed signals ( $z(t)$ ). The ICA-TSP algorithm works on these data. Lower row: Estimated separated sources (recovered up to permutation and orthogonal transformation).

optimal permutation of the ICA components) may be found by transforming the observations with the learned ICA matrix followed by an AR estimation that serves to identify the predictive matrices of Eq. (1), which – under certain conditions – allows one to *list* the components of the connected subspaces [19].

## References

1. Choi, S., Cichocki, A., Park, H.M., Lee, S.Y.: Blind Source Separation and Independent Component Analysis. *Neural Inf. Proc. Lett. and Reviews* (2005)
2. Cardoso, J.: Multidimensional Independent Component Analysis. In: ICASSP'98, Seattle, WA. (1998)
3. Akaho, S., Kiuchi, Y., Umeyama, S.: MICA: Multimodal Independent Component Analysis. In: IJCNN. (1999) 927–932
4. Póczos, B., Takács, B., Lőrincz, A.: Independent Subspace Analysis on Innovations. In: ECML, Porto, Portugal. (2005) 698–706
5. Hyvärinen, A.: Independent Component Analysis for Time-dependent Stochastic Processes. In: ICANN 1998. (1998) 541–546
6. Vollgraf, R., Obermayer, K.: Multi-Dimensional ICA to Separate Correlated Sources. In: NIPS. Volume 14. (2001) 993–1000
7. Bach, F.R., Jordan, M.I.: Finding Clusters in Independent Component Analysis. In: ICA2003. (2003) 891–896
8. Póczos, B., Lőrincz, A.: Independent Subspace Analysis Using k-Nearest Neighborhood Distances. ICANN 2005 (2005) 163–168
9. Póczos, B., Lőrincz, A.: Independent Subspace Analysis Using Geodesic Spanning Trees. In: ICML. (2005) 673–680
10. Theis, F.J.: Blind Signal Separation into Groups of Dependent Signals Using Joint Block Diagonalization. In: Proc. ISCAS 2005, Kobe, Japan (2005) 5878–5881
11. Van Hulle, M.M.: Edgeworth Approximation of Multivariate Differential Entropy. *Neural Comput.* **17** (2005) 1903–1910
12. Cheung, Y., Xu, L.: Dual Multivariate Auto-Regressive Modeling in State Space for Temporal Signal Separation. *IEEE Tr. on Syst. Man Cyb. B* **33** (2003) 386–398
13. Theis, F.J.: Uniqueness of Complex and Multidimensional Independent Component Analysis. *Signal Proc.* **84** (2004) 951–956
14. Szabó, Z., Póczos, B., Lőrincz, A.: Separation Theorem for Independent Subspace Analysis. Technical report, Eötvös Loránd University, Budapest (2005) <http://people.inf.elte.hu/lorincz/Files/TR-ELU-NIPG-31-10-2005.pdf>.
15. Yukich, J.E.: Probability Theory of Classical Euclidean Optimization Problems. Volume 1675 of *Lecture Notes in Math.* Springer-Verlag, Berlin (1998)
16. Costa, J.A., Hero, A.O.: Manifold Learning Using k-Nearest Neighbor Graphs. In: ICASSP, Montreal, Canada. (2004)
17. Rubinstein, R.Y., Kroese, D.P.: *The Cross-Entropy Method.* Springer (2004)
18. Ekman, P.: *Emotion in the Human Face.* Cambridge Univ. Press, New York (1982)
19. Póczos, B., Lőrincz, A.: Non-combinatorial Estimation of Independent Autoregressive Sources. (2005) submitted.