NONLINEAR INDEPENDENT COMPONENT ANALYSIS BY HOMOMORPHIC TRANSFORMATION OF THE MIXTURES

Deniz Erdogmus, Yadunandana N. Rao, Jose C. Principe
CNEL, Electrical & Computer Engineering Department, University of Florida, Gainesville, FL 32611, USA

Abstract – Independent component analysis is often approached from an information theoretic perspective employing specific sample estimates for the mutual information between the separated outputs. These approximations involve the nonparametric estimation of signal entropies. The common approach involves the estimation of these quantities and adaptation based on these criteria. In contrast, in this paper, we propose a Gaussianization-based approach, where the separation is performed in two stages: Gaussianization of the mixtures using a homomorphic nonlinearity and separation of the independent components using principal component analysis (both stages possibly adaptive). Due to the rotation uncertainty in nonlinear ICA, the original sources cannot be recovered solely by the independence assumption. The proposed ICA methodology is applicable to instantaneous linear and nonlinear mixtures. The idea also generalizes easily to complex-valued nonlinear ICA.

I. INTRODUCTION

Blind source separation (BSS) has recently become a mainstream research in signal processing due to the encouraging observation that numerous contemporary engineering problems involve the extraction of *unknown* source signals that are *mixed* by unknown physical processes. A wide class of array signal processing problems where multiple sensors are employed can be put in the BSS framework. Typically, BSS is used as a pre-processing stage in signal processing, estimation, and detection.

A common assumption in adaptive BSS algorithms is that the sources of interest are statistically independent [1-3], which gives rise to independent components analysis (ICA). Alternative assumptions that are exploited in various algorithms include the nonstationarity or coloredness of the sources [2,4,5]. The literature on BSS algorithms is vast and various algorithms exploit one or more of these criteria [6-9]. In this paper, we focus on the independence assumption for which the use of information theoretic optimization criteria becomes natural, since "mutual information is a canonical contrast for ICA" [9]. Due to the availability of many excellent sources in the literature where these basics of ICA and BSS can be found [1-3,10-13], we shall not dwell further on such introductory material and literature review.

More recently, research efforts have been concentrating more on the convolutive mixtures and nonlinear mixtures of sources. In this paper, we deal with nonlinear ICA, which is a necessary but not sufficient condition for the separation of nonlinearly mixed independent signals. For the latter problem, certain existence and uniqueness criteria have recently been demonstrated by Hyvarinen and Pajunen [14]. Several different techniques include minimum mutual information [15], variational Bayesian learning [16], symplectic transformations and nonparametric entropy estimates [17], higher order statistics [18], temporal decorrelation [19], and kernel-based methods [20]. A review of the current state-of-the-art in nonlinear ICA is provided recently by Jutten and Karhunen [21].

The method proposed in this paper is a novel technique based on utilizing a homomorphic nonlinear function on the available mixtures; whether they are obtained from linear or nonlinear mixtures is not of consequence. In essence, the homomorphic transformation converts the marginal distributions of the mixtures to Gaussian. This makes the mixtures jointly Gaussian, which are then made independent simply using principal components analysis (PCA).

II. HOMOMORPHIC ICA

The nonlinear ICA problem is described by a generative signal model that assumes the observed signals, denoted by $\mathbf{x}$, are a nonlinear instantaneous function of some unknown independent source signals, denoted by $\mathbf{s}$. In particular,

$$\mathbf{x}_k = \mathbf{h}(\mathbf{s}_k) \tag{1}$$

where k is the sample index. Let the observation vector be n-dimensional, $\mathbf{x}_k \in \Re^n$. Then, according to the existence results on nonlinear ICA, it is always possible to construct a function $\mathbf{g}: \Re^n \to \Re^n$, such that the outputs $\mathbf{y}=\mathbf{g}(\mathbf{x})$ are mutually independent [14]. Furthermore, this separation function is not unique. Clearly, there exist a number of operations that one might employ to change the distributions of these outputs individually without introducing mutual dependence; thus an uncertainty regarding the independent component densities exists. Furthermore, as will be shown later, in accordance with the rotation uncertainty reported in [14], the Homomorphic ICA solution will separate the observation into independent components, which are possibly a related to the original sources by an unknown rotation matrix. Also, by partitioning the variables in $\mathbf{y}$ to disjoint sets and taking various nonlinear combinations of the variables in these partitions, it is possible to generate a random vector $\mathbf{z} \in \Re^m$, where $m<n$ is the number of partitions. Thus, $\mathbf{z}=\mathbf{f}(\mathbf{y})=\mathbf{f}(\mathbf{g}(\mathbf{x}))$ also has independent components. Hence it is, in fact, possible to come up with infinitely many separating solutions that result in a smaller number of outputs than the inputs. A number of possible regularization conditions have been proposed before [14,16] to ensure uniqueness and the actual separation of the unknown sources.

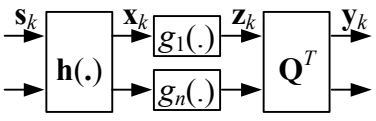

Figure 1. A schematic diagram of the proposed homomorphic independent components analysis topology.

Due to these uncertainties, we will consider the problem of determining n independent components from $\mathbf{x} \in \Re^n$, which is a necessary condition for source separation, but not sufficient. In particular, the essence of the proposed solution is to generate n independent Gaussian distributed outputs and this can be achieved quite easily. Consider the ideal case where an observation vector $\mathbf{x} \in \Re^n$ is available and the marginal cumulative distribution functions (cdf) of each observed signal is known. Let $\mathbf{x}=[x^1,\ldots,x^n]^T$ and let $F_d(.)$ denote the cdf of x^d. Also let $\phi_\sigma(.)$ denote the cdf of a zero-mean Gaussian random variable with variance σ^2. According to the fundamental theorem of probability (p. 93, [22]), z^d has a zero-mean, unit-variance Gaussian pdf:

$$z^d = \phi_1^{-1}(F_d(x^d)) \overset{def}{=} g_d(x^d) \tag{2}$$

Combining these random variables into a random vector $\mathbf{z}=[z^1,\ldots,z^n]^T$, we observe that the joint distribution of $\mathbf{z}$ is also zero-mean Gaussian with covariance $\mathbf{\Sigma}_\mathbf{z}$. Now consider the principal components of $\mathbf{z}$. Let $\mathbf{y}=\mathbf{Q}^T\mathbf{z}$, where $\mathbf{Q}$ is the orthonormal eigenvector of $\mathbf{\Sigma}_\mathbf{z}$, such that $\mathbf{\Sigma}_\mathbf{z}=\mathbf{Q}\mathbf{\Delta}\mathbf{Q}^T$ and $\mathbf{\Delta}$ is the diagonal eigenvalue matrix. Then the covariance of $\mathbf{y}$ is $\mathbf{\Sigma}_\mathbf{y}=\mathbf{\Delta}$. Hence, since $\mathbf{z}$ is zero-mean jointly Gaussian, $\mathbf{y}$ is zero-mean and jointly Gaussian with covariance $\mathbf{\Delta}$. It is well known that uncorrelated Gaussian random variables are also independent. Therefore, the components of $\mathbf{y}$ are mutually independent.[1] The overall scheme of the proposed nonlinear ICA topology is illustrated in Fig. 1.

Certain conditions must be met by the nonlinear mixing function for the separated outputs and the original sources to have maximal mutual information. In the most restrictive case, for the reconstruction of independent components that are related to the original sources by an invertible function, the mixing function must be invertible, i.e. its Jacobian must be non-singular when evaluated at any point in its input space.[2] The following theorem summarizes this fact.

[1] After developing this principle for nonlinear ICA, it came to the authors' attention that the importance of Gaussianization for breaking the curse of dimensionality was independently recognized by Chen *et al.* [23].

[2] Notice that for a broad class of nonlinear mixtures, the condition that at most one source can have a Gaussian distribution is not necessary, as the nonlinear mixture will not preserve the Gaussianity. The commonly considered post-nonlinear mixtures are easily excluded from this group. In fact, to the best knowledge of authors, there is no result available in the literature about the general conditions that the nonlinear mixture should satisfy for the non-Gaussianity condition to be lifted. Clearly, when applying the Homomorphic ICA principle to linear source separation using ICA, the non-Gaussianity conditions must still hold.

Theorem 1. If the mixing nonlinearity is invertible and the marginal probability distributions of the observed vector are always positive except possibly at a set of points whose measure is zero, then, with probability one, there is a one-to-one function between the source signals and the independent components when the outputs are constructed according to Homomorphic ICA rules.

Proof. By assumption $\mathbf{h}$ is invertible. By construction the PCA matrix $\mathbf{Q}^T$ is invertible and the Gaussianizing function $\mathbf{g}$ is monotonically increasing in all principal directions with probability one since the measure of the set on which its Jacobian has zero eigenvalues is zero. Similarly, due to the same reason, the probability of having source signals in this zero-measure set is zero. Therefore, with probability one, the Jacobian of the overall nonlinear function from $\mathbf{s}$ to $\mathbf{y}$ is invertible. Hence, there is a one-to-one relationship between these two vectors. □

Another possible scenario is that the mixing nonlinearity is only locally invertible (i.e., its Jacobian is invertible in a set $S \subset \Re^n$). In this case, if S is the support of the source distribution, one can achieve maximum mutual information between the separated outputs and the original sources.

It is well known that the nonlinear ICA problem is ill-posed and the original sources can be at most resolved up to a rotation uncertainty with the independence assumptions alone. That is, even if the mixing function is invertible, one can arrive at independent components that are not necessarily the separated versions of the original sources. This can easily be observed by examining the Homomorphic ICA output. Suppose a set of independent components are obtained from an observed vector $\mathbf{x}$ by $\mathbf{y}=\mathbf{Q}^T\mathbf{g}(\mathbf{x})$, where $\mathbf{g}(.)$ consists of individual Gaussianizing functions for each components of $\mathbf{x}$ and $\mathbf{Q}$ is the orthonormal eigenvector that is the solution to the PCA problem after Gaussianization. If the covariance of $\mathbf{y}$ is $\mathbf{\Lambda}$, by selecting an arbitrary orthonormal matrix $\mathbf{R}$, one can generate the output $\mathbf{z}=\mathbf{R}\mathbf{\Lambda}^{-1/2}$, which still has independent components (since it is jointly Gaussian with identity covariance matrix), however, different choices of $\mathbf{R}$ result in different independent components. In order to resolve this ambiguity, one needs additional information about the sources or the mixing process.

III. A Practical Homomorphic ICA Algorithm

In practice, the marginal cdf of the observed signals are not known analytically, therefore, the Gaussianizing functions must be estimated using the available samples. An asymptotically unbiased and consistent nonparametric probability density function (pdf) estimator can be employed, from which the marginal cdf can be deduced. In particular, we will consider the Parzen density estimator in this paper.

Given N independent and identically distributed (iid) samples of a random variable x with pdf $f(.)$, the Parzen window estimate for this distribution is simply [24]

$$\hat{f}(x) = \frac{1}{N}\sum_{k=1}^{N} \kappa_\sigma(x - x_k) \tag{3}$$

where $\kappa_\sigma(.)$ is the kernel function and σ denotes the kernel size (the window width). Generally, Gaussian kernels as in (4) are used in Parzen windowing and the standard deviation naturally becomes the kernel size parameter.

$$G_\sigma(\xi) = e^{-\xi^2/(2\sigma^2)} \Big/ (\sqrt{2\pi}\sigma) \tag{4}$$

Recall that Parzen windowing is an asymptotically unbiased and consistent estimator if the kernel size is reduced to zero as the number of samples approaches infinity [25].

In the case of Gaussian kernels, the estimated cdf of the random variable x using Parzen windowing is

$$\hat{F}(x) = \frac{1}{N}\sum_{k=1}^{N} \phi_\sigma(x - x_k) \tag{5}$$

Consequently, the estimated Gaussianizing function for x is

$$\hat{g}(x) = \phi_1^{-1}(\hat{F}(x)) = \phi_1^{-1}\left(\frac{1}{N}\sum_{k=1}^{N} \phi_\sigma(x - x_k)\right) \tag{6}$$

In practice, given the observed vector samples $\{\mathbf{x}_1,\ldots,\mathbf{x}_N\}$, the Gaussianizing functions for each component x^d must be estimated using the samples $\{x_1^d,\ldots,x_N^d\}$. Without loss of generality, the observed data may first be normalized to have zero mean and unit variance in all of its components using

$$\overline{\mathbf{x}} = \mathbf{\Lambda}_\mathbf{x}^{-1}(\mathbf{x} - \mathbf{m}_\mathbf{x}) \tag{7}$$

where $\mathbf{m}_\mathbf{x}$ is the sample mean of $\mathbf{x}$ and $\mathbf{\Lambda}_\mathbf{x}$ is the diagonal matrix consisting of sample variances of the components of $\mathbf{x}$.

A. Selecting the Kernel Size in Parzen Windowing

An important consideration in estimating the data pdf using Parzen windowing is the kernel size. A vast literature exists on the optimization of kernel density estimates [26]. Extreme approaches involve assigning an individual kernel size for every sample and then optimizing all these parameters. These procedures are computationally very expensive and will hinder the practicality of the proposed ICA algorithm. Instead, here we will focus on the single kernel size situation formulated in (3). The kernel size of the Parzen density estimate can be optimized to minimize the Kullback-Leibler divergence between the estimated density and the true underlying density using solely the samples available. Consider the following definition of Kullback-Leibler divergence between the true and estimated pdf [27]:

$$\begin{aligned} D_{KL}(f;\hat{f}) &= \int f(x)\log\left(f(x)/\hat{f}(x)\right)dx \\ &= -H_S(x) - E_x\left[\log\hat{f}(x)\right] \end{aligned} \tag{8}$$

where $H_S(x)$ is the true Shannon entropy of the underlying pdf, which is a constant in the optimization problem. Approximating the expectation with sample mean and using the Parzen density estimate of (3) in the second term of (8), the following optimality criterion is obtained [28]:

$$\hat{H}_S(x;\sigma) = -\frac{1}{N}\sum_{j=1}^{N}\log\left(\frac{1}{N-1}\sum_{i=1,i\neq j}^{N}\kappa_\sigma(x_j - x_i)\right) \tag{9}$$

In extensive case studies using generalized Gaussian random variables [28] and in many linear ICA simulations [29], the optimal value of the kernel size was observed to lie in the interval [0.1,0.3] when the data standard deviation was normalized to unity. In addition, the results on generalized Gaussian variables demonstrated that the cost function in (9) as well as the true Kullback-Leibler divergence in (8) are quite flat over this wide range of kernel size values [28], reducing the sensitivity of the ICA solutions to this parameter. Nevertheless, a standard minimization algorithm could be applied to determine the optimal kernel size.

B. Time-Varying Mixtures & Non-Stationary Sources

Time-variability in the mixture and the source statistics constitute an important class of source separation problems. The Homomorphic ICA algorithm can be extended to handle time-variations in the mixture statistics by employing the forgetting principle from adaptive filtering theory. The Parzen density estimator in (3) can be modified as [30]:

$$\hat{f}_k(x) = (1-\lambda)\hat{f}_{k-1}(x) + \lambda G_\sigma(x - x_k) \tag{10}$$

where $\hat{f}_k(x)$ is the current pdf estimate obtained from the incorporation of the new sample in the previous estimate. Consequently, the cdf estimate in (5) will become

$$\hat{F}_k(x) = (1-\lambda)\hat{F}_{k-1}(x) + \lambda\phi_\sigma(x - x_k) \tag{11}$$

This recursive cdf estimator can be incorporated in (6) to obtain the *forgetting Gaussianization function*.

C. Complex-Valued Homomorphic ICA

In some applications, such as fMRI analysis, the observed data is complex-valued. The extension of Homomorphic ICA to these situations is trivial. Note that an n-dimensional complex ICA problem becomes a $2n$-dimensional real valued one. In the Gaussianization step, the real and imaginary parts of the observed complex mixtures are regarded as individual real-valued observations. Once all the channels are transformed, the PCA step must consider the complex-valued nature of the problem. That is, the separated real and imaginary parts must be combined into complex Gaussian channels for PCA. The complex principal components are the Gaussian distributed solutions to the complex-valued nonlinear ICA problem at hand. The following theorem states the condition for the mixing nonlinearity to satisfy so that there is an invertible function between the separated complex outputs and the original sources (similar to the situation described in Theorem 1). In the following, consider the following nonlinear mixing function written in terms of real and imaginary parts:

$$\mathbf{x}_r + i\mathbf{x}_i = \mathbf{h}_r(\mathbf{s}) + i\mathbf{h}_i(\mathbf{s}) \tag{12}$$

where $\mathbf{s}=\mathbf{s}_r+i\mathbf{s}_i$. The Gaussianizing homomorphic transformations are denoted by $g_{dr}(.)$ and $g_{di}(.)$ for the real and imaginary parts of the d^{th} observed signal in $\mathbf{x}$. The result of Gaussianization is the complex Gaussian vector $\mathbf{z}=\mathbf{z}_r+i\mathbf{z}_i$, whose covariance is $\mathbf{\Sigma}_\mathbf{z}=\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^H$. The separated outputs are given by $\mathbf{y}=\mathbf{Q}^H\mathbf{z}$.

Theorem 2. If the marginal probability distributions of the observed vector are always positive except possibly at a set of points whose measure is zero, and the function $\overline{\mathbf{h}}(\mathbf{s}_r,\mathbf{s}_i) = [\mathbf{h}_r^T(\mathbf{s}_r,\mathbf{s}_i)\ \mathbf{h}_i^T(\mathbf{s}_r,\mathbf{s}_i)]^T$ is invertible then, with

probability one, the mutual information between the original source vector **s** and the separated output vector **y** is maximized.

Proof. Note that, the output is explicitly given by

$$\mathbf{y}_r + i\mathbf{y}_i = (\mathbf{Q}_r^T - i\mathbf{Q}_i^T)(\mathbf{g}_r(\mathbf{h}_r(\mathbf{s})) + i\mathbf{g}_i(\mathbf{h}_i(\mathbf{s}))) \quad (13)$$

We construct the vectors $\bar{\mathbf{y}} = [\mathbf{y}_r^T \ \mathbf{y}_i^T]^T$ and $\bar{\mathbf{s}} = [\mathbf{s}_r^T \ \mathbf{s}_i^T]^T$. From (13), the Jacobian of $\bar{\mathbf{y}}$ with respect to $\bar{\mathbf{s}}$ is:

$$\frac{\partial \bar{\mathbf{y}}}{\partial \bar{\mathbf{s}}} = \begin{bmatrix} \mathbf{Q}_r^T & \mathbf{Q}_i^T \\ -\mathbf{Q}_i^T & \mathbf{Q}_r^T \end{bmatrix} \cdot \begin{bmatrix} \nabla \mathbf{g}_r(\mathbf{h}_r(\mathbf{s})) & \mathbf{0} \\ \mathbf{0} & \nabla \mathbf{g}_i(\mathbf{h}_i(\mathbf{s})) \end{bmatrix} \cdot \begin{bmatrix} \partial \mathbf{h}_r(\mathbf{s}) / \partial \mathbf{s}_r & \partial \mathbf{h}_r(\mathbf{s}) / \partial \mathbf{s}_i \\ \partial \mathbf{h}_i(\mathbf{s}) / \partial \mathbf{s}_r & \partial \mathbf{h}_i(\mathbf{s}) / \partial \mathbf{s}_i \end{bmatrix} \quad (14)$$

This Jacobian is nonsingular at every possible value of **s** if and only if the third term on the right hand side of (14) is nonsingular for every value, since the other two terms are nonsingular (the second term is nonsingular with probability one as discussed in Theorem 1). Thus with Homomorphic ICA, the function from the original sources to the outputs is invertible with probability one, which equivalently means maximum mutual information between these vector signals. □

D. Homomorphic ICA for Linear Instantaneous Mixtures

If the mixture is known to be linear, there are two methods one can use to obtain the original sources. The first approach is, in principle, equivalent to the minimum mutual information algorithms in the literature. Suppose **x**=**Hs**, where **H** is the invertible mixing matrix. By sphering the observations, we reduce the mixture to a pure rotation, say **R**, so that the Gaussianized vector is **y**=**g**(**Rs**). Since **y** is (zero-mean) jointly Gaussian, this joint distribution is characterized by its covariance, which can be estimated from samples. Then, using the fundamental theorem of probability, the joint distribution of the observation vector is found to be $p_{\mathbf{x}}(\mathbf{x})=p_{\mathbf{y}}(\mathbf{g}(\mathbf{x}))/|\nabla \mathbf{g}^{-1}(\mathbf{g}(\mathbf{x}))|$, where $p_{\mathbf{y}}(.)$ is given by the determined joint Gaussian density. Then, one can minimize the divergence between $p_{\mathbf{z}}(\mathbf{z})$ and the product of its marginals, where **z**=**Qx**, **Q** being the adaptive rotation matrix.

The second approach involves training an MLP and it requires the knowledge of the source distributions. Suppose **y**=**g**(**Hs**) is the Gaussianized observation vector. In theory, the following neural network should be able to regenerate the original mixtures **x** from **y** provided that the mixture is linear: $\mathbf{x} = \mathbf{H}\mathbf{f}(\mathbf{P}\boldsymbol{\Lambda}^{-1/2}\mathbf{Q}^T\mathbf{y})$, where **f**(.) is a function that individually converts zero-mean unit-variance Gaussian distributed variables to the original source distributions. Clearly, with the correct choice of hidden layer nonlinearities, i.e., **f**(.) as described above, the network $\mathbf{z} = \mathbf{W}_2\mathbf{f}(\mathbf{W}_1\mathbf{y})$ could be trained using **x** as the desired output. Consequently, $\mathbf{W}_2$ would be an estimate of **H** (with the usual scaling and permutation uncertainties), and $\mathbf{W}_1$ will encompass $\mathbf{P}\boldsymbol{\Lambda}^{-1/2}\mathbf{Q}^T$. This approach is more practical than the first one and it requires training a single-hidden-layer MLP using a least squares criterion. Besides, in the case of noisy observations, assuming that the measurement noise is Gaussian, this least squares approach becomes maximum likelihood as well.

E. Two Interesting Special Situations

There are two interesting special situations that might occur when using the Homomorphic ICA approach. This section is devoted to a short discussion of these cases.

The first one is the case when, after Gaussianization, the covariance matrix of the Gaussianized mixtures is *I* (assuming that each observation is transformed to a zero-mean unit-variance Gaussian. In this case, there is no need for the PCA stage, and the eigenvector matrix can be automatically set to the identity matrix since the Gaussianized mixtures are already independent. This situation might arise both in linear and nonlinear mixtures.

The second interesting observation is for a 2-by-2 linear mixture. In this case, assuming that each mixture is Gaussianized to have zero-mean and unit-variance, the covariance matrix of the Gaussianized mixture vector becomes $\boldsymbol{\Sigma}_{\mathbf{z}}=[1\ \beta;\beta\ 1]$, where β is the cross-correlation of these mixtures. Provided that $\beta\neq 0$, the eigenvectors of this covariance matrix are [1;1] and [1;-1], regardless of β (if $\beta=0$, then the first degenerate case is obtained where the eigenvectors are the principal vectors). Thus, in these cases solving a PCA problem from the sample-estimated covariance matrix is not necessary. In simulations, it was observed that even with a small number of samples, the PCA solution also approached these eigenvectors closely.

IV. Simulations

In order to demonstrate the performance of the proposed Homomorphic ICA approach using the Parzen window Gaussianization functions described above, computer simulations of nonlinear ICA are performed using two speech waveforms. The signals are two different sentences uttered by two different male speakers and the means are subtracted and the powers are normalized. The correlation between the original normalized sources are about 0.01, indicating their approximate independence (although this assumption does not hold exactly). The sources are mixed using the following nonlinear invertible mixture:

$$\mathbf{x} = \mathbf{R}_2\mathbf{g}(\mathbf{R}_1\mathbf{s}) \quad (15)$$

where $\mathbf{R}_1$ and $\mathbf{R}_2$ are randomly selected 2x2 rotation (orthonormal) matrices and the nonlinear function **g**(.) is

$$\mathbf{g}(\xi_1, \xi_2) = [\arctan(\alpha_1\xi_1) \ \ \arctan(\alpha_2\xi_2)]^T \quad (16)$$

where α_i are uniformly random in [5,10] to guarantee some clipping/saturation effects.

Two experiments are carried out using 5000-sample and 20000-sample training sets both randomly selected from the available 35000-sample data set. The Gaussianizing functions and the PCA matrices are determined using the training data. One major problem with this nonstationary source signal scenario was observed to be the scarcity of the training data at the peak values of the source signals. This typically resulted in the Gaussianizing function estimates to be poor

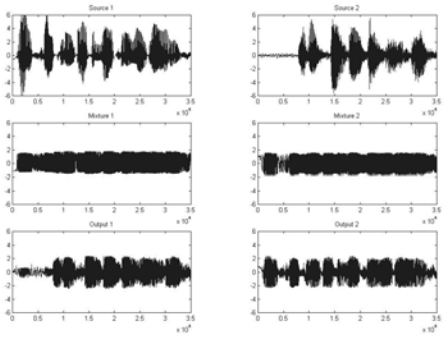

Figure 2. Original sources (top), mixtures (middle), and separated Gaussianized outputs (bottom), when the separation function is trained using only 5000 samples.

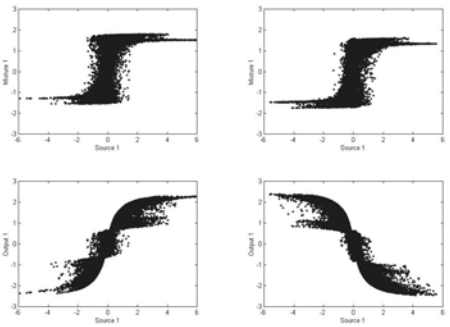

Figure 3. Scatter plots of the mixture samples versus source samples (top) and scatter plots of the Gaussianized separated outputs versus source samples (bottom), when the separation function is trained using only 5000 samples.

for the overall signal duration although the training set samples are perfectly Gaussianized. As an example demonstration, the two source signals, the two mixtures, and the two separated Gaussian-distributed outputs (using 5000 samples and Gaussian kernels with size 0.1) are shown in Fig. 2. The mixtures and the separated outputs are also shown in the scatter plots in Fig. 3 versus the source signals.[3] Ideally, a clearly visible, narrow monotonic function between the outputs and the sources indicate perfect separation. In this case (5000-sample training), obviously, the solution determined by the training set does not generalize well to new data from the speech segment. The same plots are repeated in Fig. 4&5 using the solution determined using 20000 samples this time (using a different mixing function and a kernel size of 0.01). Relative to the scatter plots between the mixtures and the sources, the separated outputs yield *slimmer* scatter plots around the origin indicating improvement of separability in this regime. This improvement is audible. Although compared to the 5000-sample case this result is better, the clipping effects are still visible in the separated outputs in Fig. 4. The separability in the saturation regime is not as successful as the linear regime. This means even at 20000 samples, the saturations in the mixture are not

[3] The signal-to-interference ratio cannot be used meaningfully in nonlinear mixtures so the scatter plots are preferred for the lack of a better measure.

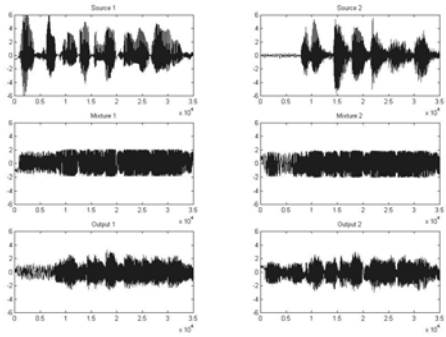

Figure 4. Original sources (top), mixtures (middle), and separated Gaussianized outputs (bottom), when the separation function is trained using only 20000 samples.

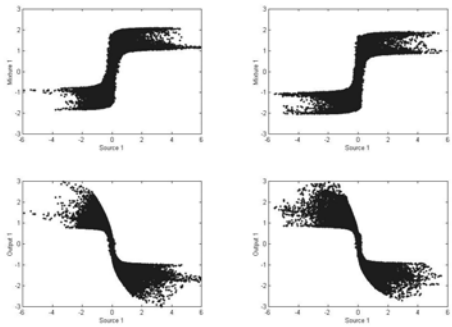

Figure 5. Scatter plots of the mixture samples versus source samples (top) and scatter plots of the Gaussianized separated outputs versus source samples (bottom), when the separation function is trained using only 20000 samples.

sufficiently represented by the training data; therefore, a reliable separating solution is not obtained for these regimes.

The phenomenon observed here is a generalization of the large-sample requirements imposed by higher order cumulants (e.g., kurtosis) in linear ICA. It is clear from the scatter plots of Fig. 3&5 that although the separation performance is relatively high in small-signal operating regimes, the saturation regimes of the mixing nonlinearity are not excited sufficiently, thus it becomes more difficult to statistically analyze the function behavior. Especially when the sources are highly kurtotic (as in speech), many samples from the tails are necessary to saturate the nonlinearity.

In simulations with linear mixtures, whose results are not shown here, much better separation was achieved (although still not comparable to existing methods). Consequently, the scatter plots of sources versus outputs were much *slimmer*. In fact, Homomorphic ICA should not be the first choice for solving linear ICA problems, since many solutions working only with adaptive linear matrices can solve this problem.

V. Conclusions

In this paper, we have presented a novel approach to nonlinear ICA, called Homomorphic ICA. The proposed approach is based on utilizing a homomorphic nonlinear transformation such that the marginal distributions of the

mixtures are converted to Gaussian, which effectively transforms the joint mixture density to Gaussian. Then, independent components analysis reduces to principal components analysis of the transformed mixtures. Due to the nature of this solution, it is possible to obtain batch, or on-line separating function solutions. Although in this paper, we have demonstrated one possible way of implementing the Homomorphic ICA approach under nonparametric density estimation principles, the proposed methodology could yield a family of solutions depending on the mixture density estimation technique: parametric density estimation could lead to analytical nonlinear ICA solutions, non-parametric density estimation leads to separation solutions found by batch training, while adaptive density estimation techniques (such as neural networks or support vector machines) could lead to on-line nonlinear ICA.

In the simulations, the proposed algorithm based on Parzen window density estimates was applied to speech separation from invertible nonlinear mixtures. The mixing nonlinearities involved saturating functions, which in turn lead to the observation that, in nonlinear ICA, there exists a generalization of the large-sample condition that typically accompanies kurtosis or other cumulant-based methods in linear ICA. This phenomenon needs to be studied further to understand the generalization capabilities of the nonlinear ICA solutions obtained from finite number of samples.

ACKNOWLEDGMENTS

DE thanks E. Todorov for asking the question at AS-SPCC 2003 that sparked the idea, and K.E. Hild and T. Bell for comments that helped understand the method better. This work was supported by NSF grant ECS-0300340.

REFERENCES

[1] A. Hyvarinen, J. Karhunen, E. Oja, *Independent Component Analysis*, Wiley, New York, 2001.

[2] A. Cichocki, S.I. Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*, Wiley, New York, 2002.

[3] T.W. Lee, *Independent Component Analysis: Theory and Applications*, Kluwer, New York, 1998.

[4] A. Ziehe, K.R. Müller, "TDSEP - An Efficient Algorithm for Blind Separation Using Time Structure," Proceedings of ICANN'98, pp. 675-680, Skovde, Sweden, 1998.

[5] S. Choi, A. Cichocki, Y. Deville, "Differential Decorrelation for Nonstationary Source Separation," Proceedings of ICA'01, pp. 319-322, San Diego, California, 2001.

[6] T. Bell, T. Sejnowski, "An Information-Maximization Approach to Blind Separation and Blind Deconvolution," Neural Computation, vol. 7, pp. 1129-1159, 1995.

[7] K.E. Hild II, D. Erdogmus, J.C. Principe, "Blind Source Separation Using Renyi's Mutual Information," IEEE Signal Processing Letters, vol. 8, pp. 174-176, 2001.

[8] A. Hyvarinen, "Fast and Robust Fixed-Point Algorithms for Independent Component Analysis," IEEE Transactions on Neural Networks, vol. 10, pp. 626-634, 1999.

[9] J.F. Cardoso, A. Souloumiac, "Blind Beamforming for Non-Gaussian Signals," IEE Proceedings F: Radar and Signal Processing, vol. 140, pp. 362-370, 1993.

[10] J.F. Cardoso, "Blind Signal Separation: Statistical Principles," Proceedings of IEEE, vol. 86, pp. 2009-2025, 1998.

[11] J.F. Cardoso, "High-Order Contrasts for Independent Component Analysis," Neural Computation, vol. 11, pp. 157-192, 1999.

[12] P. Comon, "Independent Component Analysis: A New Concept?" Signal Processing, vol. 36, pp. 287-314, 1994.

[13] A. Hyvarinen, "Survey on Independent Component Analysis," Neural Computing Surveys, vol. 2, pp. 94-128, 1999.

[14] A. Hyvarinen, P. Pajunen, "Nonlinear Independent Component Analysis: Existence and Uniqueness Results," Neural Networks, vol. 12, no. 3, pp. 429–439, 1999.

[15] L.B. Almeida, "MISEP – Linear and Nonlinear ICA Based on Mutual Information", Journal of Machine Learning Research, vol. 4, pp. 1297-1318, 2003.

[16] H. Valpola, E. Oja, A. Ilin, A. Honkela, J. Karhunen, "Nonlinear Blind Source Separation by Variational Bayesian Learning," IEICE Transactions on Fundamentals of Electronics, Communications, and Computer Sciences, vol. 86, no. 3, pp. 532-541, 2003.

[17] L. Parra, "Symplectic Nonlinear Independent Component Analysis," Proceedings of NIPS'96, pp. 437-443, 1996.

[18] Y. Tan, J. Wang, "Nonlinear Blind Source Separation Using Higher Order Statistics and a Genetic Algorithm," IEEE Transactions on Evolutionary Computation, vol. 5, no. 6, 2001.

[19] A. Ziehe, M. Kawanabe, S. Harmeling, K.R. Muller, "Blind Separation of Post-Nonlinear Mixtures Using Linearizing Transformations and Temporal Decorrelation," Journal of Machine Learning Research, vol. 4, pp. 1319-1338, 2003.

[20] S. Harmeling, A. Ziehe, M. Kawanabe, K.R. Müller, "Kernel-Based Nonlinear Blind Source Separation," Neural Computation, vol. 15, pp. 1089-1124, 2003.

[21] C. Jutten, J, Karhunen, "Advances in Nonlinear Blind Source Separation," Proceedings of ICA'03, pp. 245-256, Nara, Japan, 2003.

[22] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 3rd ed., McGraw-Hill, New York, 1991.

[23] S. Chen, R.A. Gopinath, "Gaussianization," Proceedings of NIPS'01, pp. 423-429, Denver, Colorado, 2001.

[24] E. Parzen, "On Estimation of a Probability Density Function and Mode", in *Time Series Analysis Papers*, Holden-Day, San Diego, California, 1967.

[25] D. Erdogmus, *Information Theoretic Learning: Renyi's Entropy and its Applications to Adaptive System Training*, PhD Dissertation, University of Florida, Gainesville, Florida, 2002.

[26] J. Weston, A. Gammerman, M.O. Stitson, V. Vapnik, V. Vovk, C. Watkins, "Support Vector Density Estimation," in *Advances in Kernel Methods: Support Vector Learning*, B. Scholkopf, C.J.C. Burges, A.J. Smola (eds.), pp. 293-305, MIT Press, Cambridge, Massachusetts, 1998.

[27] T. Cover, J. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.

[28] D. Erdogmus, K.E. Hild II, J.C. Principe, "Kernel Size Selection in Parzen Density Estimation," unpublished work, Nov 2003.

[29] K.E. Hild II, *Blind Separation of Convolutive Mixtures Using Renyi's Divergence*, Ph.D. Dissertation, University of Florida, Gainesville, Florida, 2003.

[30] D. Erdogmus, J.C. Principe, S.P. Kim, J.C. Sanchez, "A Recursive Renyi's Entropy Estimator," Proceedings of NNSP'02, pp. 209-217, Martigny, Switzerland, 2002.