Gaussianizing Transformations for ICA

Deniz Erdogmus, Yadunandana N. Rao, and José Carlos Príncipe

CNEL, Electrical and Computer Engineering Department,
University of Florida, Gainesville, Florida 32611, USA
{deniz,yadu,principe}@cnel.ufl.edu
http://www.cnel.ufl.edu

Abstract. Nonlinear principal components analysis is shown to generate some of the most common criteria for solving the linear independent components analysis problem. These include minimum kurtosis, maximum likelihood and the contrast score functions. In this paper, a topology that can separate the independent sources from a linear mixture by specifically utilizing a Gaussianizing nonlinearity is demonstrated. The link between the proposed topology and nonlinear principal components is established. Possible extensions to nonlinear mixtures and several implementation issues are also discussed.

1 Introduction

Independent components analysis (ICA) is now a mature field with numerous approaches and algorithms to solve the basic instantaneous linear mixture case as well as a variety of extensions of these basic principles to solve the more complicated problems involving convolutive or nonlinear mixtures [1-3]. Due to the existence of a wide literature and excellent survey papers [4,5], in addition to the books listed above, we shall not go into a detailed literature survey. Interested readers are referred to the references mentioned above and the references therein.

In this paper, we will focus on a special type of homomorphic transformation, called the Gaussianizing function. Several interesting observations about this transformation and its utility in ICA will be addressed in this paper. Especially, we will establish a link between a Gaussianizing function based topology for solving linear instantaneous mixture problems and the established technique of nonlinear principal components analysis (NPCA) [6], which has already been shown to encompass a number of linear ICA optimization criteria as special cases [1] corresponding to certain choices of the *nonlinear functions of projection*. Nevertheless, the selection of these nonlinear projection functions stemming from the principal of mutual independence has not been yet addressed. Determining such a function is intellectually appealing since "mutual information is a canonical contrast for ICA" [7]. Finally, we would like to stress that the goal of this paper is *not* to present yet another linear ICA algorithm, but to demonstrate an interesting selection of the nonlinearity in NPCA as this method is applied to solving the ICA problem.

C.G. Puntonet and A. Prieto (Eds.): ICA 2004, LNCS 3195, pp. 26–32, 2004.

2 Gaussianizing Transformations

Given an n-dimensional random vector $\mathbf{Y}$ with joint probability density function (pdf) $p_{\mathbf{Y}}(\mathbf{y})$, there exist many functions $\mathbf{g}:\Re^n \to \Re^n$ such that $\mathbf{Z}=\mathbf{g}(\mathbf{Y})$ is jointly Gaussian. In particular we are interested in the elementwise Gaussianization of $\mathbf{Y}$. Suppose Y_i has marginal pdf $p_i(y_i)$, whose corresponding cumulative distribution function (cdf) is $P_i(y_i)$. Let $\phi(.)$ denote the cdf of a zero-mean unit-variance single dimensional Gaussian variable, i.e.,

$$\phi(\xi) = \int_{-\infty}^{\xi} \frac{1}{\sqrt{2\pi}} e^{-\alpha^2/2} d\alpha \tag{1}$$

Then, according to the fundamental theorem of probability [8], $Z_i=\phi^{-1}(P_i(Y_i))$ is a zero-mean and unit-variance Gaussian random variable.

We define $g_i(\xi)=\phi^{-1}(P_i(\xi))$ and call this the Gaussianizing transformation for Y_i. Combining $g_i(.)$ into a vector valued function, we get the elementwise Gaussianizing transformation for $\mathbf{Y}$ as $\mathbf{Z}=\mathbf{g}(\mathbf{Y})$. Since this $\mathbf{g}:\Re^n \to \Re^n$ is acting on each argument separately, its Jacobian matrix is *diagonal* at every point in its domain. Furthermore, since every Z_i is zero mean and unit-variance Gaussian, the vector $\mathbf{Z}$ is jointly Gaussian denoted by $\mathbf{G}(\mathbf{z},\Sigma)$ with zero mean and covariance

$$\mathbf{\Sigma} = E[\mathbf{Z}\mathbf{Z}^T] = \begin{bmatrix} 1 & & \rho_{ij} \\ & \ddots & \\ \rho_{ji} & & 1 \end{bmatrix} \tag{2}$$

The utility of this Gaussianizing transformation was pointed out earlier for multidimensional pdf estimation [9]. Clearly, if one estimates the marginal pdfs of $\mathbf{Y}$ and the covariance of $\mathbf{Z}$ after Gaussianizing $\mathbf{Y}$ as described above, then an estimate of the joint pdf of $\mathbf{Y}$ can be obtained using the fundamental theorem of probability [8].

$$\begin{aligned} p_{\mathbf{Y}}(\mathbf{y}) &= \frac{\mathbf{G}(\mathbf{g}(\mathbf{y}),\mathbf{\Sigma})}{\left|\nabla \mathbf{g}^{-1}(\mathbf{g}(\mathbf{y}))\right|} = \mathbf{G}(\mathbf{g}(\mathbf{y}),\mathbf{\Sigma})\left|\nabla \mathbf{g}(\mathbf{y})\right| \\ &= \mathbf{G}(\mathbf{g}(\mathbf{y}),\mathbf{\Sigma}) \cdot \prod_{i=1}^{n} g_i'(y_i) = \mathbf{G}(\mathbf{g}(\mathbf{y}),\mathbf{\Sigma}) \cdot \prod_{i=1}^{n} \frac{p_i(y_i)}{G(g_i(y_i),1)} \end{aligned} \tag{3}$$

3 Homomorphic Linear ICA Topology

The linear ICA problem is described by a generative signal model that assumes the observed signals, denoted by $\mathbf{x}$, and the sources, denoted by $\mathbf{s}$, are obtained by a *square* linear system of equations. The sources are assumed to be statistically independent. In summary, assuming an unknown mixing matrix $\mathbf{H}$, we have

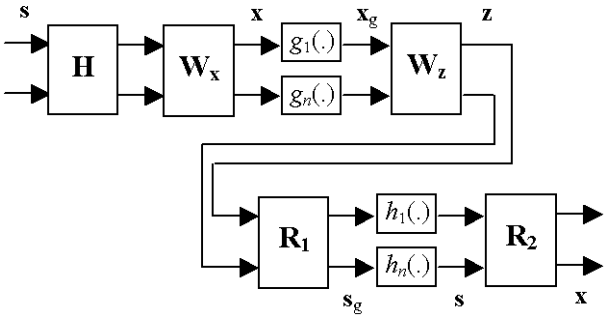

Fig. 1. A schematic diagram of the proposed homomorphic ICA topology.

$$\mathbf{x}_k = \mathbf{H}\mathbf{s}_k \tag{4}$$

where the subscript k is the sample/time index. The linear ICA problem exhibits the following uncertainties, which cannot be resolved by the independence assumption alone: permutation of separated source estimates and scaling factors (including sign changes).

The goal is to recover the sources from the observed mixtures. For the sake of simplicity in the following arguments, we will assume that the marginal pdfs of the sources and the mixtures are known and all are strictly positive valued (to guarantee the invertibility of Gaussinizing transformations). It is assumed without loss of generality that the sources are already zero-mean.

Consider the topology shown in Fig. 1 as a solution to linear ICA. The observed mixtures are first spatially whitened by $\mathbf{W}_\mathbf{x}$ to generate the whitened mixture vector $\mathbf{x}$. Since whitening reduces the mixing matrix to only a coordinate rotation, without loss of generality, we can always focus on mixing matrices that are orthonormal. In this case, we assume that the mixing matrix is $\mathbf{R}_2=\mathbf{W}_\mathbf{x}\mathbf{H}$. Since the marginal pdfs of the mixtures are known, one can construct the Gaussianizing functions $g_i(.)$ according to the previous section to obtain the Gaussianized mixtures $\mathbf{x}_\mathbf{g}$. Whitening the Gaussianized mixtures will yield zero-mean unit-variance and uncorrelated signals $\mathbf{z}$. Since $\mathbf{z}$ is jointly Gaussian, uncorrelatedness corresponds to mutual independence. However, considering the function from the sources ($\mathbf{s}$) to the Gaussianized mixtures ($\mathbf{x}_\mathbf{g}$) as a post-nonlinear mixture, we notice that although by obtaining $\mathbf{z}$ we have obtained independent components, due to the inherent rotation ambiguity of nonlinear mixtures in the ICA framework [10], we have not yet achieved source separation. Consequently, there is still an unknown orthonormal matrix $\mathbf{R}_\mathbf{1}$ that will transform $\mathbf{z}$ into Gaussianized versions of the original sources. If the marginal source pdfs are known, the inverse of the Gaussianizing transformations for the sources could be obtained in accordance with the previous section (denoted by $h_i(.)$ in the figure), which would transform $\mathbf{s}_\mathbf{g}$ to the original source distribution, thus yield the separated source signals (at least their estimates).

In summary, given the whitened mixtures, their marginal pdfs and the marginal pdfs of the sources (up to permutation and scaling ambiguities in accordance with the theory of linear ICA), it is possible to obtain an estimate of the orthonormal mixing matrix $\mathbf{R_2}$ and the sources $\mathbf{s}$ by training a constrained multilayer perceptron (MLP) topology with first layer weights given by $\mathbf{R_1}$ and second layer weights given by $\mathbf{R_2}$. The nonlinear functions of the hidden layer processing elements (PE) are determined by the inverse Gaussianizing transformations of the source signals. This MLP with square first and second layer weight matrices would be trained according to the following constrained optimization problem:

$$\min_{\mathbf{R_1},\mathbf{R_2}} E\left[\left\|\mathbf{x}-\mathbf{R_2}\mathbf{h}(\mathbf{R_1}\mathbf{z})\right\|^2\right] \quad \text{subject to} \quad \mathbf{R_1}\mathbf{R_1}^T=\mathbf{I}, \mathbf{R_2}\mathbf{R_2}^T=\mathbf{I}, \tag{5}$$

Constrained neural structures of this type have been considered previously by Fiori [11]. Interested readers are referred to his work and the references therein to gain a detailed understanding of this subject.

4 Relationship with Nonlinear PCA

NPCA is known to solve the linear (and nonlinear) ICA problem when the nonlinear projection functions are properly selected. Various choices of these functions correspond to different ICA criteria ranging from kurtosis to maximum likelihood (ML) [1]. In the most general sense, the NPCA problem is compactly defined by the following optimization problem:

$$\min_{\mathbf{W}} E\left[\left\|\mathbf{x}-\mathbf{W}\mathbf{f}(\mathbf{W}^T\mathbf{x})\right\|^2\right] \tag{6}$$

where $\mathbf{f}(.)$ is an elementwise function (i.e. with a diagonal Jacobian at every point) that is selected *a priori*. For the special case of $\mathbf{f}(\mathbf{z})=\mathbf{z}$, this optimization problem reduces to the linear bottleneck topology, which is utilized by Xu to obtain the LMSER algorithm for linear PCA [12].

Returning to the topology in Fig. 1, under the assumptions of invertibility (which is satisfied if and only if the source pdfs are strictly greater than zero[1]) we observe that $\mathbf{z}=\mathbf{W_z}\mathbf{g}(\mathbf{x})$ and $\mathbf{x}=\mathbf{R_2}\mathbf{s}$, therefore, the cost function in (5) is $E[\|\mathbf{R_2}\mathbf{s}-\mathbf{R_2}\mathbf{h}(\mathbf{R_1}\mathbf{W_z}\mathbf{g}(\mathbf{R_2}\mathbf{s}))\|^2]$. Being orthonormal, $\mathbf{R_2}$ does not affect the Euclidean norm, and the cost becomes $E[\|\mathbf{s}-\mathbf{h}(\mathbf{R_1}\mathbf{W_z}\mathbf{g}(\mathbf{R_2}\mathbf{s}))\|^2]$. In the ICA setting, $\mathbf{s}$ is approximated by its estimate, the separated outputs $\mathbf{y}$, which is the output of the $\mathbf{h}(.)$ stage of Fig. 1. In the same setting, assuming whitened mixtures, NPCA would optimize

[1] In the case of zero probability densities, the Gaussianizing functions will not be invertible in general, since locally at these points the Jacobian might become singular. However, since the probability of occurrence of such points is also zero for the same reason, for the given signal-mixture case global invertibility is not necessary. However, it is assumed for simplicity.

$$\min_{\mathbf{W}} E\left[\left\|\mathbf{y}-\mathbf{f}(\mathbf{y})\right\|^2\right] \tag{7}$$

where $\mathbf{y}=\mathbf{W}\mathbf{x}$, in accordance with (6) [1]. A direct comparison of (7) and the expression given above that is equivalent to (5) yields $\mathbf{f}(\mathbf{y})=\mathbf{h}(\mathbf{R_1}\mathbf{W_z}\mathbf{g}(\mathbf{R_2}\mathbf{y}))$.

In summary, the homomorphic ICA approach described in the previous section and formulated in (5) tries to determine a nonlinear subspace projection of the separated outputs such that the projections become independent. While an arbitrary selection of the nonlinear projection functions would not necessarily imply independence of the separated outputs, the proposed approach specifically exploits homomorphic Gaussianizing transformations of the signals such that orthogonality (uncorelatedness of zero-mean signals) is equivalent to mutual independence.

5 Alternative Approaches

The Gaussianizing transformations could be utilized in alternative linear ICA solution strategies. Here, we will briefly discuss a few. The obvious approach would be to utilize the Gaussianizing transformation to estimate the joint density of the mixtures or the separated outputs. This leads to two possible approaches.

Estimating the joint density of the mixtures: Suppose the whitened mixtures are related to the sources by $\mathbf{x}=\mathbf{R}\mathbf{s}$ and the marginal source distributions are known. Since the sources are independent, the joint source distribution, denoted by $p_{\mathbf{S}}(\mathbf{s})$, is simply the product of the marginals. Due to the fundamental theorem of probability, the joint pdf of the mixtures could be determined as $p_{\mathbf{X}}(\mathbf{x})=p_{\mathbf{S}}(\mathbf{R}^T\mathbf{x})$. At the same time, from (3), we have $p_{\mathbf{X}}(\mathbf{x})=\mathbf{G}(\mathbf{g}(\mathbf{x}),\Sigma)|\nabla\mathbf{g}(\mathbf{x})|$. These two joint distributions must be identical, therefore one can determine $\mathbf{R}$ by minimizing any suitable divergence measure between the two representations of the mixture pdf. If the appropriate definition of Kullback-Leibler (KL) divergence is utilized as the measure, then the estimate would also be asymptotically maximum likelihood, due to the well-known relationships between ML and KL divergence.

Estimating the joint density of the separated outputs: Suppose that $\mathbf{x}=\mathbf{H}\mathbf{s}$ and $\mathbf{y}=\mathbf{W}\mathbf{x}$. Suppose that an estimate of the marginal pdfs of $\mathbf{y}$ is available at every step of learning iterations (nonparametric density estimations could be utilized at this stage). Then, one could construct the elementwise Gaussianizing functions of $\mathbf{y}$ to estimate its joint density using (3). The separation matrix $\mathbf{W}$ can be optimized to minimize the mutual information in $\mathbf{y}$ estimating Shannon's definition using the nonparametric marginal and joint distribution estimates of $\mathbf{y}$.

6 Extension to Nonlinear Mixtures

With some modifications, the topology shown in Fig. 1 could also be utilized to obtain independent components from mixtures generated by invertible nonlinear functions of the sources. In fact, given any n dimensional random vector $\mathbf{x}$ (regardless of

it being generated from independent sources or not) one can determine *n* independent components. A proof of existence is provided in [10]. A much simpler proof of existence is as follows: Given $\mathbf{x}$, $\mathbf{z}=\mathbf{W}_z\mathbf{g}(\mathbf{x})$ are independent components, where $\mathbf{W}_z$ and $\mathbf{g}(.)$ are obtained as described above and in Fig. 1. In [10], the rotation ambiguity of nonlinear ICA is also addressed. This ambiguity is also readily observed in Fig. 1. Since $\mathbf{z}$ are independent components, $\mathbf{R}_1\mathbf{z}$ for any orthonormal matrix $\mathbf{R}_1$ also yields independent components for $\mathbf{x}$. Nevertheless, if one is not concerned about these ambiguities, nonlinear ICA is reduced to estimating the marginal pdfs of the mixture and applying whitening to the Gaussianized mixtures.

Actual separation of sources in the nonlinear mixture case requires additional constraints. For example if the mixture is post-nonlinear and the source distributions are known, the structure in Fig. 1 can be used as described in (5) with some modifications to solve the problem. Since the nonlinearities would be absorbed by the initial Gaussianizing transformation $\mathbf{g}(.)$, similar Gaussianizing functions must be employed at the output stage and the desired output should be $\mathbf{x}_g$. The latter Gaussianizing functions will be required to change at every learning iteration as they include the most current estimate of the nonlinearities of the post-nonlinear mixture and the following Gaussianizing function $\mathbf{g}(.)$. An approach along these lines was also proposed by Ziehe *et al.* [13].

7 Conclusions

In this paper, we have presented a topology based on using Gaussianizing homomorphic transformations that allows handling higher order statistics by considering only second order statistics in the ICA problem setup. The proposed topology is extremely interesting in that it lies at the intersection of nonlinear principal component analysis and learning in neural networks with orthonormality constraints on weight matrices.

Some alternative approaches that basically correspond to directly minimizing an estimate of the mutual information between the separated outputs are also sketched based on the density estimates obtained through the Gaussianizing transformations.

Extensions of the proposed topology to solve nonlinear ICA problems is discussed with special emphasis on post-nonlinear mixtures. The proposed topology also points out much simpler proofs for the existence of nonlinear ICA and its rotation ambiguity.

Acknowledgments

This work is supported by NSF grant ECS-0300340. The authors would like to thank K.E. Hild for useful discussions.

References

1. Hyvarinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. Wiley, New York (2001)

2. Cichocki, A., Amari, S.I.: Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications. Wiley, New York (2002)
3. Lee, T.W.: Independent Component Analysis: Theory and Applications. Kluwer, New York (1998)
4. Hyvarinen, A.: Survey on Independent Component Analysis. Neural Computing Surveys. 2 (1999) 94-128
5. Jutten, C., Karhunen, J.: Advances in Nonlinear Blind Source Separation. Proceedings of ICA'03, Nara, Japan. (2003) 245-256
6. Karhunen, J., Joutsensalo, J.: Representation ans Separation of Signals Using Nonlinear PCA Type Learning. Neural Networks. 7 (1994) 113-127
7. Cardoso, J.F., Souloumiac, A.: Blind Beamforming for Non-Gaussian Signals. IEE Proceedings F: Radar and Signal Processing. 140 (1993) 362-370
8. Papoulis, A.: Probability, Random Variables, and Stochastic Processes. 3rd edn. McGraw-Hill, New York (1991)
9. Chen, S., Gopinath, R.A.: Gaussianization. Proceedings of NIPS'01, Denver, Colorado. (2001) 423-429
10. Hyvarinen, A., Pajunen, P.: Nonlinear Independent Component Analysis: Existence and Uniqueness Results. Neural Networks. 12 (1999) 429–439
11. Fiori, S.: A Theory for Learning by Weight Flow on Stiefel-Grassman Manifold. Neural Computation. 13 (2001) 1625-1647
12. Xu, L.: Least Mean Square Error Reconstruction Principle for Self- Organizing Neural Nets. Neural Networks. 6 (1993) 627-648
13. Ziehe, A., Kawanabe, M., Harmeling, S., Muller, K.R.: Blind Separation of Post-nonlinear Mixtures Using Linearizing Transformations and Temporal Decorrelation. Journal of Machine Learning Research. 4 (2003) 1319-1338