ON THE UNIQUENESS OF THE MINIMUM OF THE INFORMATION-THEORETIC COST FUNCTION FOR THE SEPARATION OF MIXTURES OF NEARLY GAUSSIAN SIGNALS

Riccardo Boscolo and Vwani P. Roychowdhury

Electrical Engineering Department
University of California, Los Angeles
Los Angeles, CA 90095
{riccardo,vwani}@ee.ucla.edu

ABSTRACT

A large number of Independent Component Analysis (ICA) algorithms are based on the minimization of the statistical mutual information between the reconstructed signals, in order to achieve the source separation. While it has been demonstrated that a global minimum of such cost function will result in the separation of the statistically independent sources, it is an open problem to show that such cost function has a unique minimum (up to scaling and permutations of the signals). Without such result, there is no guarantee that the related ICA algorithms will not get stuck in local minima, and hence, return signals that are statistically dependent. We derive a novel result showing that for the special case of mixtures of two independent and identically distributed (i.i.d.) signals with symmetric, nearly gaussian probability density functions, such objective function has no local minima. This result is shown to yield a useful extension of the well-known entropy power inequality.

1. INTRODUCTION

In the classic independent component analysis (ICA) framework, a generative model is assumed where N independent stationary signals $\mathbf{s} = \{s_1, \ldots, s_N\}$ are mixed through a linear transformation $\mathbf{x} = A\mathbf{s}$. It has been shown (for example in [1]) that, in absence of noise, there always exist an inverse linear transformation of the type $\mathbf{y} = B\mathbf{x}$, through which the reconstruction of the original signals is possible, up to an arbitrary scaling and permutations of the signals themselves. In particular, if we consider the statistical mutual information [2] between the reconstructed signals as a function of the unmixing matrix B[1], such

---

[1] $I(y_1, \ldots, y_N) \triangleq \int p_y(\mathbf{y}) \log \frac{p_y(\mathbf{y})}{\prod_{i=1}^{N} p_{y_i}(y_i)} d\mathbf{y}$

a function has a global minimum, yielding the source separation [3][4].

Therefore, a vast number of independent component analysis frameworks are designed to solve the following optimization problem:

$$B_{opt} = \arg \min_B I(y_1, \ldots, y_N) \qquad (1)$$

or an approximate version thereof. When the mixture data $\mathbf{x}$ is sphered prior to the reconstruction ($Cov(\mathbf{x}\mathbf{x}^T) = I$), one can show that the unmixing matrix B must belong to the manifold of orthogonal matrices[5]. Using some basic information theory inequalities, the problem posed in (1) can be re-written as:

$$\min_B \sum_{i=1}^{N} h(y_i) \qquad (2)$$

$$\text{s.t. } BB^T = I, \qquad (3)$$

where $h(a) = -\int p_a(u) \log p_a(u) du$ is the differential entropy of the continuous random variable a. The equality constraints (3) define a sub-group of the Stiefel manifold for the case of square matrices. If we define $F(B) \triangleq \sum_{i=1}^{N} h(y_i)$, then the gradient of the cost function defined on such manifold is given by [6]:

$$\nabla_m F(B) \triangleq \nabla F(B) - B \nabla F(B)^T B. \qquad (4)$$

where $\nabla F(B)$ is the conventional gradient of $F(B)$ in the Euclidean space. The extrema of the optimization problem (2) are found in correspondence to all the matrices satisfying the condition:

$$\nabla_m F(B) = 0 \quad \Rightarrow \quad \nabla F(B)B^T = B\nabla F(B)^T. \quad (5)$$

Several ICA algorithms optimizing different approximated versions of the cost function (1) have been shown to possess good local convergence properties [7][8]. Although the global minimum of (1) is known to yield the desired source separation [1], no proof is available to show that such a function has no local minima. On the other hand, because of the uniqueness of the separation matrix (up to permutations and scaling), proved by Comon in [1], convergence to any solution other than the global would result in a failure to separate the source signals. The problem of convergence to suboptimal solutions was recently investigated for example in [9] and in [10].

In this paper, we address the fundamental problem of the uniqueness of the minimum (up to scaling and permutation of the solution) of the information-theoretic cost function in the case of linear mixtures. We show that in the case of mixtures of two symmetric i.i.d. nearly gaussian signals, such cost function is indeed free from spurious local minima. In addition, we derive an interesting connection between the problem defined by (2) and the well-known *entropy power inequality*, showing that, under the aforementioned hypotheses, not only this inequality does not hold for dependent random variables, but it is, in fact, always violated (converse entropy power inequality).

2. EXTREMA FOR MIXTURES OF TWO NEARLY GAUSSIAN SOURCES

We consider the traditional linear framework, where we assume that the mixing matrix A is the 2×2 identity matrix and the original signals are zero-mean, and unit variance. The reconstructed signals can be written as:

$$y_1 = b_{11}s_1 + b_{12}s_2 \qquad (6)$$
$$y_2 = b_{21}s_1 + b_{22}s_2. \qquad (7)$$

The general case where the mixing matrix is not the identity matrix can be mapped to this special case through an orthogonal transformation [5], as long as the mixture data is sphered, thus preserving the characteristics of the solution space of (2) (in particular, the number of extrema). We restrict our analysis to those cases where the probability density functions of s_1 and s_2 are symmetric and they can be approximated using a Gram-Charlier [11] expansion of the type:

$$f_{s_i}(u) = g(u)\left(1 + \frac{\kappa_{4,s_i}}{24}H_4(u)\right) \quad i = 1, 2. \qquad (8)$$

where $H_4(u)$ is the 4th order Chebyshev-Hermite polynomial and $g(u)$ is the zero-mean, unit-variance, normal probability density function. The probability density functions of y_1 and y_2, can be approximated as[2]:

$$f_{y_i}(u) \approx g(u)\left(1 + \frac{\kappa_{4,y_i}}{24}H_4(u)\right) \quad i = 1, 2. \qquad (9)$$

The cumulants κ_{4,y_i} can be computed as:

$$\kappa_{4,y_1} = E[y_1^4] - 3 = b_{11}^4\mu_{4,s_1} + 6b_{11}^2 b_{12}^2 + b_{12}^4\mu_{4,s_2} - 3 \quad (10)$$
$$\kappa_{4,y_2} = E[y_2^4] - 3 = b_{21}^4\mu_{4,s_1} + 6b_{21}^2 b_{22}^2 + b_{22}^4\mu_{4,s_2} - 3 \quad (11)$$

where μ_{4,s_i} is the 4th order central moment of s_i.

The extrema of the cost function (2) must satisfy (5). For mixtures of two sources these conditions can be written as:

$$\nabla h(\mathbf{b}_1)\mathbf{b}_2^T = \nabla h(\mathbf{b}_2)\mathbf{b}_1^T, \qquad (12)$$

where $\mathbf{b}_i$ is the ith row of B, and in order to make explicit the dependence of the entropy $h(y_i)$ on $\mathbf{b}_i$, we can define $h(\mathbf{b}_i) \triangleq h(y_i)$, $i = 1, 2$. Given that:

$$\frac{\partial h(b_i)}{\partial b_{ij}} = -\int_{-\infty}^{\infty}(1 + \log f_{y_i}(u))\frac{\partial f_{y_i}(u)}{\partial b_{ij}}du \qquad (13)$$

the identity (12) can be written as:

$$\int_{-\infty}^{\infty}\log f_{y_1}(u)\left[b_{21}\frac{\partial f_{y_1}(u)}{\partial b_{11}} + b_{22}\frac{\partial f_{y_1}(u)}{\partial b_{12}}\right]du = \qquad (14)$$
$$= \int_{-\infty}^{\infty}\log f_{y_2}(u)\left[b_{11}\frac{\partial f_{y_2}(u)}{\partial b_{21}} + b_{12}\frac{\partial f_{y_2}(u)}{\partial b_{22}}\right]du.$$

Using (9) we can compute explicitly ($i = 1, 2$):

$$\frac{\partial f_{y_i}(u)}{\partial b_{i1}} = g(u)\left(\frac{1}{6}b_{i1}^3\mu_{4,s_1} + \frac{1}{2}b_{i1}b_{i2}^2\right)H_4(u) \qquad (15)$$

$$\frac{\partial f_{y_i}(u)}{\partial b_{i2}} = g(u)\left(\frac{1}{6}b_{i2}^3\mu_{4,s_2} + \frac{1}{2}b_{i1}^2 b_{i2}\right)H_4(u) \qquad (16)$$

Now define:

$$D_1(u, B) \triangleq \frac{1}{g(u)}\left[b_{21}\frac{\partial f_{y_1}(u)}{\partial b_{11}} + b_{22}\frac{\partial f_{y_1}(u)}{\partial b_{12}}\right] \qquad (17)$$

---

[2]Only the 8th order term of this Gram-Charlier expansion is non-zero and it is neglected.

$$= c_{4,y_1} H_4(u),$$

where:

$$c_{4,y_1} = \frac{1}{6}\left(b_{11}^3 b_{21}\mu_{4,s_1} + b_{12}^3 b_{22}\mu_{4,s_2}\right) + \tag{18}$$

$$+\frac{1}{2}\left(b_{11}b_{12}^2 b_{21} + b_{11}^2 b_{12} b_{22}\right)$$

and:

$$D_2(u,B) \triangleq \frac{1}{g(u)}\left[b_{11}\frac{\partial f_{y_2}(u)}{\partial b_{21}} + b_{12}\frac{\partial f_{y_2}(u)}{\partial b_{22}}\right] \tag{19}$$

$$= c_{4,y_2} H_4(u).$$

where:

$$c_{4,y_2} = \frac{1}{6}\left(b_{11}b_{21}^3 \mu_{4,s_1} + b_{12}b_{22}^3 \mu_{4,s_2}\right) + \tag{20}$$

$$+\frac{1}{2}\left(b_{11}b_{21}b_{22}^2 + b_{12}b_{21}^2 b_{22}\right),$$

The following integrals need to be evaluated:

$$\int_{-\infty}^{\infty} g(u)\log f_{y_i}(u)D_i(u,B)du \quad i = 1,2. \tag{21}$$

where:

$$\log f_{y_i}(u) = -\frac{1}{2}\log(2\pi) - \frac{u^2}{2}\log(e) + \tag{22}$$

$$+\log\left(1 + \frac{\kappa_{4,y_i}}{24}H_4(u)\right) \quad i = 1,2.$$

Substituting this expression in (21), we obtain:

$$\int_{-\infty}^{\infty} g(u)\left[-\frac{1}{2}\log(2\pi) - \frac{u^2}{2}\log(e) + \tag{23}\right.$$

$$\left.+\log\left(1 + \frac{\kappa_{4,y_i}}{24}H_4(u)\right)\right]D_i(u,B)du$$

Now notice that:

$$\int_{-\infty}^{\infty} g(u)H_4(u)du = 0, \tag{24}$$

and:

$$\int_{-\infty}^{\infty} u^2 g(u)H_4(u)du = 0. \tag{25}$$

The integral (23) simplifies as:

$$\int_{-\infty}^{\infty} g(u)\log\left(1 + \frac{\kappa_{4,y_i}}{24}H_4(u)\right)D_i(u)du. \tag{26}$$

Using the following useful indefinite integral:

$$\int g(u)D_i(u)du = -c_{4,y_i}H_3(u), \tag{27}$$

we can integrate (26) per parts. If we define $X_i(u) \triangleq \kappa_{4,y_i}H_4(u)/24$, we obtain:

$$\int_{-\infty}^{\infty} g(u)\log\left(1 + X_i(u)\right)D_i(u)du = \tag{28}$$

$$= c_{4,y_i}\int_{-\infty}^{\infty} g(u)H_3(u)\frac{X_i'(u)}{1 + X_i(u)}du,$$

where $X_i'(u) = \kappa_{4,y_i}H_3(u)/6$. Using (28), we find that (14) reduces to:

$$c_{4,y_1}\kappa_{4,y_1}\int_{-\infty}^{\infty}\frac{H_3^2(u)}{1 + \kappa_{4,y_1}/24\,H_4(u)}g(u)du = \tag{29}$$

$$= c_{4,y_2}\kappa_{4,y_2}\int_{-\infty}^{\infty}\frac{H_3^2(u)}{1 + \kappa_{4,y_2}/24\,H_4(u)}g(u)du.$$

In particular, when the sources are i.i.d. ($\mu_{4,s_1} = \mu_{4,s_2} \triangleq \mu_4$), we have that $k_{4,y_1} = k_{4,y_2} \neq 0$, and the two integrals on the left-hand-side and on the right-hand-side of (29) are always equal. Moreover, because their integrands are non-negative, these integrals are also strictly positive. Thus, the conditions for the gradient to be zero become simply:

$$c_{4,y_1} = c_{4,y_2} \tag{30}$$

We can now study the solutions of (30) in the space of orthogonal matrices. This is achieved by operating the substitution:

$$\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}. \tag{31}$$

Substituting in the expressions for c_{4,y_1} and c_{4,y_2}, we obtain:

$$c_{4,y_1} = -\frac{1}{6}\sin\theta\cos\theta\left[(\mu_4 - 3)(\cos^2\theta - \sin^2\theta)\right] \tag{32}$$

$$c_{4,y_2} = \frac{1}{6}\sin\theta\cos\theta\left[(\mu_4 - 3)(\cos^2\theta - \sin^2\theta)\right] \tag{33}$$

Thus, (30) is satisfied if and only if:

$$(\mu_4 - 3) \sin\theta \cos\theta \cos 2\theta = 0. \qquad (34)$$

Because of the symmetry of the problem, it suffices to study the zeros of (34) in the interval $[0, \pi/2]$. The solutions found in $[\pi/2, 2\pi)$, correspond, in fact, to a permutation or sign change of the rows of B. In this interval, (34) has only two zeros, one for $\theta = 0$ corresponding to a minimum of (2), and one for $\theta = \pi/4$, corresponding to a maximum of the objective function, thus proving that (2) has no local minima.

3. AN EXTENSION OF THE ENTROPY POWER INEQUALITY

In this section we will illustrate the connection between the result we just proved and the well-known entropy power inequality [2].

The *entropy power* of a scalar random variable s is defined as:

$$N(s) = \frac{1}{2\pi e} \, e^{2h(s)} \qquad (35)$$

Given two independent random variables s_1 and s_2, the entropy power inequality states that:

$$N(s_1 + s_2) \geq N(s_1) + N(s_2), \qquad (36)$$

with equality holding if and only if s_1 and s_2 are both normal. The inequality (36) can be used to prove the convexity of the entropy under a covariance preserving transformation, i.e. given $0 \leq \lambda \leq 1$, it holds that [12]:

$$h(\lambda s_1 + \sqrt{1 - \lambda^2} s_2) \geq \lambda^2 h(s_1) + (1 - \lambda^2) h(s_2). \qquad (37)$$

Now simply define:

$$\lambda = \cos\theta \;\; \Rightarrow \;\; \sqrt{1 - \lambda^2} = \sin\theta \quad 0 \leq \theta \leq \pi/2 \qquad (38)$$

Thus one can write:

$$h(\cos\theta \, s_1 + \sin\theta \, s_2) \geq \cos^2\theta \, h(s_1) + \sin^2\theta \, h(s_2) \qquad (39)$$

and analogously:

$$h(-\sin\theta s_1 + \cos\theta s_2) \geq \sin^2\theta h(s_1) + \cos^2\theta h(s_2) \qquad (40)$$

(note that $h(as) = h(s) + \log|a|$, a being a scalar parameter). Simply by adding (39) and (40) we obtain:

$$h(y_1) + h(y_2) \geq h(s_1) + h(s_2). \qquad (41)$$

In particular (41) proves that the extremum corresponding to $\theta = 0$ is a global minimum of (2), regardless of the actual distributions of s_1 and s_2. The uniqueness of this minimum, proved in the previous section, extends the inequality theorem showing that there are no local minima of $h(y_1) + h(y_2)$, for $0 \leq \lambda < 1$.

This result can be used to show that a *converse entropy power inequality* holds, if certain hypotheses are satisfied. Define two random variables z_1 and z_2 as follows:

$$z_1 = \lambda y_1 + \sqrt{1 - \lambda^2} y_2 \qquad (42)$$

$$z_1 = \sqrt{1 - \lambda^2} y_1 + \lambda y_2, \qquad (43)$$

for $0 \leq \lambda < 1$. Because of the uniqueness of the minimum of $h(y_1) + h(y_2)$ in this interval, it follows that the following inequality never holds:

$$h(z_1) + h(z_2) \not\geq h(y_1) + h(y_2), \qquad (44)$$

unless y_1 and y_2 are obtained from s_1 and s_2, solely through scaling or permutation. In other words, the entropy power inequality is *always* violated by two dependent random variables obtained through an orthogonal projection of independent random variables.

4. CONCLUSIONS

We introduced a novel result proving the uniqueness of the minimum of the information-theoretic cost function, for the special case of linear mixtures of independent and identically distributed signals with symmetric probability density functions. Such a result, the first of its kind, can be used to show that a converse entropy power inequality holds for this particular class of distributions. In process of deriving a proof for our result, we introduced a useful framework that can potentially be extended in order to investigate the problem for more general classes of distributions. In particular, the method can be used to study whether a converse entropy power inequality, proved for this special case, holds in general. So far, in fact, examples of source distributions for which the uniqueness property is systematically violated have not been identified.

5. REFERENCES

[1] P. Comon, "Independent component analysis, a new concept?," *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

[2] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, John Wiley & Sons, 1991.

[3] Jean-François Cardoso, "Infomax and maximum likelihood for source separation," *IEEE Letters on Signal Processing*, vol. 4, no. 4, pp. 112–114, Apr. 1997.

[4] S. Amari S. Cruces, A. Cichocki, "The minimum entropy and cumulant based contrast functions for blind source extraction," in *Bio-Inspired Applications of Connectionism, Lecture Notes in Computer Science, Springer-Verlag. [6th International Work-Conference on Artificial and Natural Neural Networks (IWANN'2001)]*, J. Mira and A. Prieto editors, Eds., Granada, Spain, June 2001, vol. II, pp. 786–793.

[5] D. Obradovic and G. Deco, "Information maximization and independent component analysis: Is there a difference?," *Neural Computation*, vol. 10, pp. 2085–2101, 1998.

[6] S.T. Smith A. Edelman, T.A. Arias, "The geometry of algorithms with orthogonality constraints," *SIAM J. Matrix Anal. Appl.*, vol. 20, no. 2, pp. 303–353, 1999.

[7] Aapo Hyvärinen, "Survey on independent component analysis," *Neural Computing Surveys*, vol. 2, pp. 94–128, 1999.

[8] S. i. Amari, A. Chichoki, and H.H. Yang, "A new learning algorithm for blind source separation," in *Advances in Neural Information Processing Systems*. 1996, vol. 8, pp. 757–763, MIT Press, Cambridge, MA.

[9] M. Rattray and Gleb Basalyga, "Scaling laws and local minima in hebbian ica," in *Advances in Neural Information Processing Systems 14*, S. Becker T.G. Dietterich and Z. Ghahramani Editors, Eds., Vancouver, Canada, Dec.

[10] M. Rattray and Gleb Basalyga, "Stochastic trapping in a solvable model of on-line Independent Component Analysis," *Neural Computation*, vol. 14, pp. 421–435, 2002.

[11] D. L. Wallace, "Asymptotic approximations to distributions.," *Ann. Math. Stat.*, vol. 29, pp. 635–654, 1958.

[12] A. Dembo and T.M. Cover, "Information theoretic inequalities," *IEEE Trans. On Information Theory*, vol. 37, no. 6, pp. 1501–1518, 1991.