Infomax and Maximum Likelihood for Blind Source Separation

Jean-François Cardoso, *Member, IEEE*

***Abstract*—Algorithms for the blind separation of sources can be derived from several different principles. This letter shows that the recently proposed infomax principle is equivalent to maximum likelihood.**

I. Introduction

SOURCE separation consists in recovering a set of unobservable signals (sources) from a set of observed mixtures. In its simplest form, an $n\times 1$ vector $\mathbf{x}$ of observations (typically, the output of n sensors) is modeled as

$$\mathbf{x} = A_*\mathbf{s} \tag{1}$$

where the "mixing matrix" A_* is invertible and the $n\times 1$ vector $\mathbf{s} = [s_1, \cdots, s_n]^T$ has independent components: Its probability density function (pdf) $r(\mathbf{s})$ factors as

$$r(\mathbf{s}) = \prod_{i=1,\, n} r_i(s_i) \tag{2}$$

where $r_i(s_i)$ is the pdf of s_i.[1] Based on these assumptions and on realizations of $\mathbf{x}$, the aim is to estimate matrix A_* or, equivalently, to find a "separating matrix" B such that

$$\mathbf{y} = B\mathbf{x} = BA_*\mathbf{s}$$

is an estimate of the source signals.

A guiding principle for source separation is to optimize a function $\phi(B)$ called a *contrast* function, which is a function of the distribution of $\mathbf{y} = B\mathbf{x}$ [4]. Based on the *infomax principle*, an apparently new contrast function has recently been derived by Bell and Sejnowski [1], attracting a lot of interest. In this letter, we exhibit the contrast function associated to the well established maximum likelihood (ML) principle. By this device, we show that, regarding source separation, infomax principle boils down to ML.

II. Infomax

According to Bell and Sejnowski, application of the infomax principle [1] to source separation consists in maximizing an output entropy (see Fig. 1).

Manuscript received July 9, 1996. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. T. S. Durrani.

The author is with the Centre National de la Recherche Scientifique and École Nationale Supérieure des Télécommunications, 75634 Paris, France (e-mail: cardoso@sig.enst.fr).

Publisher Item Identifier S 1070-9908(97)02520-0.

[1] Throughout this letter, pdf's are with respect to Lebesgue measure. All distributions are assumed continuous with respect to it.

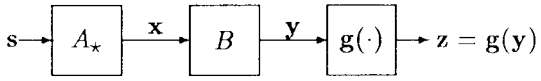

Fig. 1. Mixing, unmixing, and nonlinear transformation.

Plain maximization would be inappropriate because the entropy of $\mathbf{y} = B\mathbf{x}$ diverges to infinity for an arbitrarily large separating system B. Thus, the infomax principle is implemented by maximizing with respect to B the entropy of $\mathbf{z} = \mathbf{g}(\mathbf{y}) = \mathbf{g}(B\mathbf{x})$ where

$$[\mathbf{g}(\mathbf{y})]_i = g_i(y_i) \qquad 1 \le i \le n$$

is a $R^n \to R^n$ componentwise nonlinear function. Thus, the infomax contrast function is

$$\phi_I(B) \stackrel{\text{def}}{=} H(\mathbf{g}(B\mathbf{x})) \tag{3}$$

where $H(\cdot)$ is the (Shannon) differential entropy (11). Scalar functions $g_1, \cdots, g_n$ are taken to be "squashing functions," mapping the real line to the interval $(0, 1)$ and monotonously increasing. Thus, if g_i is differentiable, it is the cumulative distribution function (cdf) of some pdf q_i on the real line

$$q_i(s) = \frac{dg_i(s)}{ds}$$
$$g_i(s) = \int_{-\infty}^{s} q_i(y)\, dy.$$

Denote $\tilde{\mathbf{s}} = [\tilde{s}_1, \cdots, \tilde{s}_n]^T$ an $n\times 1$ random vector with pdf

$$q(\tilde{\mathbf{s}}) = \prod_{i=1}^{n} q_i(\tilde{s}_i). \tag{4}$$

Then, $g_i(\tilde{s}_i)$ is distributed *uniformly* on $(0, 1)$ since g_i is the cdf of q_i. Thus, $\mathbf{u} \stackrel{\text{def}}{=} \mathbf{g}(\tilde{\mathbf{s}})$ is distributed uniformly on $(0, 1)^n$ and the infomax contrast is rewritten as

$$\phi_I(B) = -K(\mathbf{g}(B\mathbf{x})\|\mathbf{u}) = -K(B\mathbf{x}\|\tilde{\mathbf{s}}). \tag{5}$$

where $K(\cdot\|\cdot)$ denotes the Kullback–Leibler (KL) divergence (12). The first equality is the combination of (3) and (14); the second results from (13). It shows that "infomaximization" is identical to minimization of the KL divergence between the distribution of the output vector $\mathbf{y} = B\mathbf{x}$ and the distribution (4) of $\tilde{\mathbf{s}}$.

III. Maximum Likelihood

We first recall how the ML principle is associated with a contrast function. This is then specialized to the source separation model.

Consider a sample of T independent realizations $\mathbf{x}_1, \cdots, \mathbf{x}_T$ of a random variable $\mathbf{x}$ distributed according to a common density $p_*(\mathbf{x})$. Let $\mathcal{P} = \{p_\theta(\mathbf{x}) | \theta \in \Theta\}$ be a parametric model for the density of $\mathbf{x}$. The likelihood that the sample is drawn with a particular distribution p_θ is the product $\prod_{t=1}^{T} p_\theta(\mathbf{x}_t)$. Taking the logarithm and dividing by the number of observations results in the normalized log-likelihood

$$L_T(\theta) \stackrel{\text{def}}{=} \frac{1}{T} \log \prod_{t=1,\,T} p_\theta(\mathbf{x}_t) = \frac{1}{T} \sum_{t=1,\,T} \log p_\theta(\mathbf{x}_t).$$

Since this is the sample average of $\log p_\theta(\mathbf{x})$, it converges in probability, by the law of large numbers, to its expectation

$$L_T(\theta) \stackrel{\mathcal{P}}{\rightarrow} L(\theta) \stackrel{\text{def}}{=} \mathrm{E} L_T(\theta) = \int p_*(\mathbf{x}) \log p_\theta(\mathbf{x})\, d\mathbf{x}.$$

Setting $p_\theta(\mathbf{x}) = [p_\theta(\mathbf{x})/p_*(\mathbf{x})]\, p_*(\mathbf{x})$ in the above yields

$$L(\theta) = -K(p_* || p_\theta) - H(p_*). \tag{6}$$

Note that this conclusion is reached *without* assuming an exact model, i.e., it is not assumed that $p_* = p_{\theta_*}$ for some $\theta_* \in \Theta$.

This result applies to source separation as follows. The distribution p_* is the true distribution of $\mathbf{x}$, i.e., the distribution of vector $A_*\mathbf{s}$; the parameter of the parametric model $\mathcal{P}$ is the unknown mixing matrix: $\theta = A$; the parameter set Θ is the set of all invertible $n \times n$ matrices; the pdf's in $\mathcal{P}$ are the distributions of vector $A\tilde{\mathbf{s}}$ where the source vector is assumed to be distributed as $\tilde{\mathbf{s}}$, i.e., according to distribution (4).

In this setting, (6) becomes $L(A) = -K(A_*\mathbf{s} || A\tilde{\mathbf{s}}) - H(A_*\mathbf{s})$: The contrast function associated with the likelihood appears to be

$$\phi_L(A) \stackrel{\text{def}}{=} -K(A_*\mathbf{s} || A\tilde{\mathbf{s}}) \tag{7}$$

because $H(A_*\mathbf{s})$ being an additive constant (not depending on the parameter A) can be discarded.

IV. Discussion

In view of (5) and (7), the function ϕ: $\mathrm{R}^{n\times n} \mapsto \mathrm{R}$

$$\phi(C) \stackrel{\text{def}}{=} -K(C\mathbf{s} || \tilde{\mathbf{s}}) \tag{8}$$

appears to play a central role. Indeed, we can write

$$\begin{aligned} \phi_I(B) &= \phi(BA_*) \\ \phi_L(A) &= \phi(A^{-1}A_*). \end{aligned}$$

The first equality stems from (5) and $B\mathbf{x} = BA_*\mathbf{s}$. The second equality is by applying property (13) to (7) with $\mathbf{f}(\mathbf{u}) = A^{-1}\mathbf{u}$.

Hence, it is found that the contrasts associated with infomax and with maximum likelihood coincide, provided B is identified to A^{-1}. Interpretation is straightforward in either case: Contrast optimization corresponds to a "Kullback matching" between the hypothesized source distribution $\tilde{\mathbf{s}}$ and the distribution of $C\mathbf{s}$ where C is either BA_* or $A^{-1}A_*$.

With the correct source model, i.e., when $\tilde{\mathbf{s}}$ is distributed as $\mathbf{s}$, the contrast reduces to $\phi(C) = -K(C\mathbf{s} || \mathbf{s})$. This is maximized at $C = I$ because $K(\mathbf{s} || \mathbf{s}) = 0$, which is the lowest possible value, the KL divergence being nonnegative.

What happens with a wrong source model or—equivalently—when the squashing functions are *not* the cdf's of the source distributions? We sketch an answer, based on the following functions:

$$f_i(s) \stackrel{\text{def}}{=} -\frac{q_i'(s)}{q_i(s)} = -\frac{d \log q_i(s)}{ds} \qquad 1 \le i \le n.$$

The stationary points of the likelihood/infomax contrast cancel the gradient of ϕ. Differentiating ϕ, it is easily found that these are the matrices C such that $\mathbf{y} = C\mathbf{s}$ verifies

$$\mathrm{E} f_i(y_i) y_j = \delta_{ij} \qquad 1 \le i, j \le n \tag{9}$$

$\delta_{ij} = 1$ if $i = j$ and 0 otherwise. Let λ_i be a solution of

$$\mathrm{E}\{f_i(\lambda_i s_i)\lambda_i s_i\} = 1 \qquad 1 \le i \le n. \tag{10}$$

Then, matrix $C = \mathrm{diag}\,(\lambda_1, \cdots, \lambda_n)$ is a stationary point of $\phi(C)$ if the source signals have zero mean: $\mathbf{E}\mathbf{s} = 0$. This is easily seen because the stationarity condition (9) is satisfied for $i = j$ thanks to (10) and for $i \neq j$ due to

$$\mathrm{E}\{f_i(\lambda_i s_i)\lambda_j s_j\} = \mathrm{E}\{f_i(\lambda_i s_i)\}\mathrm{E}\{\lambda_j s_j\} = 0.$$

The first equality is by independence of the source signals; the second one by the zero-mean condition. For a wrong source model: $q_i \neq r_i$, one generally has $\lambda_i \neq 1$ and thus $C = \mathrm{diag}\,(\lambda_1, \cdots, \lambda_n) \neq I$. However, such a C still is a satisfactory solution with respect to source separation, since source signals are recovered up to scaling factors. Unfortunately, one cannot conclude that the likelihood/infomax contrast is completely "robust" to misspecifying source distributions because too large a mismatch may turn $C = \mathrm{diag}\,(\lambda_1, \cdots, \lambda_n)$ into an *unstable* stationary point, as observed in [1].

V. Conclusion

The infomax principle was shown to coincide with the ML principle in the case of source separation. Both principles explicitly or implicitly assume given source distributions. They consist in minimizing the Kullback divergence between the distribution at the output of a separating matrix and the hypothesized source distribution. It thus appears important to match source distributions as closely as possible, as already noted in [1] and as is obvious from the ML standpoint. Even with a wrong source model, a stationary point of the contrast is still obtained for separated (but generally scaled) signals, even though the stability cannot be guaranteed for too large a mismatch. Both principles also hint at a more general approach involving joint estimation of the mixing matrix *and* of (some characteristics of) source distributions. First steps in this direction are taken in [1]; an elegant and well-developed approach is to be found in [5].

APPENDIX
INFORMATION THEORETIC QUANTITIES

- The *differential entropy* of a variable with pdf p is denoted $H(p)$; the *Kullback–Leibler divergence* between two pdf's p and q is denoted $K(p||q)$. Definitions are

$$H(p) = -\int_{\mathbf{y}} \log\left(p(\mathbf{y})\right) p(\mathbf{y})\, d\mathbf{y} \tag{11}$$

$$K(p||q) = \int_{\mathbf{y}} \log\left(\frac{p(\mathbf{y})}{q(\mathbf{y})}\right) p(\mathbf{y})\, d\mathbf{y} \tag{12}$$

  whenever the integrals exist. A convenient abuse of notation is

$$H(\mathbf{y}) = H(p)$$
$$K(\mathbf{y}||\mathbf{z}) = K(p||q)$$

  if p (resp. q) is the density of a random vector $\mathbf{y}$ (resp. $\mathbf{z}$).
- $K(p||q) \geq 0$ with equality iff p and q agree p-almost everywhere.
- The KL divergence is invariant under an invertible transformation $\mathbf{f}$ of the sample space, as follows:

$$K[(\mathbf{f}(\mathbf{u})||\mathbf{f}(\mathbf{v})) = K(\mathbf{u}||\mathbf{v}) = K(\mathbf{f}^{-1}(\mathbf{u})||\mathbf{f}^{-1}(\mathbf{v})). \tag{13}$$

- The differential entropy of a distribution with support $(0, 1)^n$ also is the KL divergence between this distribution and the uniform distribution $\mathbf{u}$ on $(0, 1)^n$. Clearly

$$H(\mathbf{z}) = -\int p_{\mathbf{z}}(\mathbf{z}) \log\left(\frac{p_{\mathbf{z}}(\mathbf{z})}{\prod_{i=1}^{n} \mathbf{1}_{(0,1)}(z_i)}\right) d\mathbf{z}$$
$$= -K(\mathbf{z}||\mathbf{u}). \tag{14}$$

ACKNOWLEDGMENT

This work was completed while the author was visiting S. I. Amari at Riken Institute of Japan.

REFERENCES

[1] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, no. 6, pp. 1004–1034.
[2] J.-F. Cardoso, "The equivariant approach to source separation," in *Proc. NOLTA*, 1995, pp. 55–60. Available as {ftp://sig.enst.fr/pub/jfc/Papers/nolta95.ps.gz}.
[3] T. M. Cover and J. A. Thomas, *Elements of Information Theory: Wiley Series in Telecommunications.* New York: Wiley, 1991.
[4] P. Comon, "Independent component analysis, a new concept?," *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1988.
[5] D.-T. Pham, P. Garrat, and C. Jutten, "Separation of a mixture of independent sources through a maximum likelihood approach," in *Proc. EUSIPCO*, 1992, pp. 771–774.