ESTIMATING EQUATIONS FOR SOURCE SEPARATION

Jean-François Cardoso

École Nationale Supérieure des Télécommunications (ENST), Département Signal
46 rue Barrault, 75634 Paris Cedex 13, France
WWW: http://sig.enst.fr/~cardoso/stuff.html

ABSTRACT

This paper proposes a unifying view of source separation via the concepts of 'estimating function' and 'estimating equation'. We exhibit the estimating functions corresponding to various known techniques like ICA, JADE, infomax, maximum likelihood, cumulant matching, etc... We also show how equivariant batch and adaptive algorithms stem from each particular estimating function and discuss their stability and asymptotic performance.

1. INTRODUCTION.

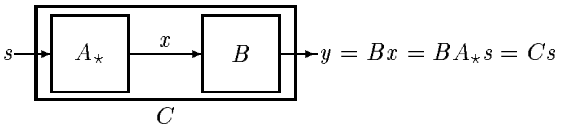
1.1. Source separation.

The simplest source separation model is that of an $n \times 1$ vector x of observations with structure

$$x = A_* s \qquad r(s) = \prod_{i=1}^{n} r_i(s_i) \tag{1}$$

where A_* is an invertible $n \times n$ unknown matrix and s is an unobserved $n \times 1$ vector. The second equation in (1) expresses that the probability density function (p.d.f) $r(s)$ (w.r.t. Lebesgue measure) of the source vector s is the product of the densities of its components, *i.e.* that s is a vector of *independent components*, the so-called 'sources'. The task is to recover the source signals and/or to identify matrix A_* using only the assumption of source independence. Only the case of real signals is considered here, but all the arguments carry over to the complex case.

Many source separation algorithms have been recently proposed, either adaptive [1, 2, 3, 4, 5, 6, 7] and many others, or batch, based on higher order criteria [8, 9, 10, 11] or on the likelihood [12, 13, 4]. In adaptive (on-line) approaches, one explicitly updates an $n \times n$ 'separating matrix' B which yields an 'output vector' $y = Bx$ estimating the source vector s.



In many instances, the stationary points of the learning algorithm may be characterized by an equation in the form

$$\mathrm{E}H(y) = 0 \quad \text{with} \quad y = Bx \tag{2}$$

where $H : \mathbf{R}^n \to \mathbf{R}^{n \times n}$ is an appropriately defined matrix-valued function. Function H precisely is an *estimating function* for source separation; it is the purpose of this contribution to show how this notion informs most of on-line and off-line approaches to source separation.

1.2. Estimating functions.

Consider an inference problem where the distribution of a random variable $X \in \mathcal{X}$ is parameterized by a parameter θ: $X \sim p(x;\theta)$, $\theta \in \Theta \subset \mathbf{R}^p$. In this generic parametric context, an *estimating function* is a function $h : \mathcal{X} \times \Theta \to \mathbf{R}^p$ such that $\mathrm{E}_\theta h(x;\theta) = 0$ for any $\theta \in \Theta$. If T independent realizations $x(1), \ldots, x(T)$ of X are available, the unknown parameter vector θ can be estimated as the solution $\hat{\theta}$ of the 'estimating equation':

$$\frac{1}{T} \sum_t h(x(t);\hat{\theta}) = 0 \tag{3}$$

which is just the sample counterpart of $\mathrm{E}_\theta h(x;\theta) = 0$. The resulting estimates are sometimes called 'M-estimates' and have been carefully studied in the statistical literature [14]. Note that M-estimation generalizes maximum likelihood estimation since the latter is obtained by taking $h(x;\theta) = \partial \log p(x;\theta)/\partial\theta$.

1.3. Equivariance

In the source separation problem, the unknown parameter is not an unstructured vector $\theta \in \mathbf{R}^p$ but an invertible $n \times n$ matrix: $\theta \equiv A \in \mathbf{R}^{n \times n}$. Thus the parameter set is the group, traditionally denoted $GL(n)$, of all the invertible linear transformations on $\mathbf{R}^n$. It is important to take this fact into account. An estimator of $A_\star$ which is compatible with the group structure is said to be *equivariant* [15]. This property means that if an estimate $\hat{A}$ is computed from a data set $x(1), \ldots, x(T)$, then the estimate computed from $Mx(1), \ldots, Mx(T)$ should be $M\hat{A}$ for any invertible matrix M.

It is easy to see that equivariant estimators have the desirable property of having uniform performance: their behavior in terms of source separation is independent of the particular value $A_\star$ of the mixing matrix [16]. It is also possible to design equivariant *adaptive* algorithms [7, 3, 4, 17]. In this paper, we will consider only equivariant estimating functions, defined in section 2.1..

Outline of the paper

In section 2., we introduce equivariant estimating functions for source separation and show how they can be derived by 'relative differentiation' of contrast functions. In section 3., we show how the theory is extended to source separation techniques which are only asymptotically equivalent

To appear in *Proc. ICASSP-97, Apr 21-24, Munich, Germany*

to solving estimating equations. The final section briefly describes adaptive and batch algorithms to solve estimating equations.

2. ESTIMATING FUNCTIONS FOR SOURCE SEPARATION.

2.1. Equivariant estimating functions

The equivariance principle suggests to focus on estimating functions for source separation having the following special structure [16]:

$$h(x;\theta) \equiv h(x;A) = H(A^{-1}x) = H(y) \tag{4}$$

where $H : \mathbf{R}^n \to \mathbf{R}^{n\times n}$ is a matrix-valued function of a vector-valued argument. The corresponding estimates $\widehat{A}$ of $A_\star$ are solutions of the estimating equation

$$\frac{1}{T}\sum_{t=1}^{T} H(y(t)) = 0 \quad \text{where} \quad y(t) = \widehat{A}^{-1}x(t). \tag{5}$$

This is to be related to eq. (2) with the identification of the mixing matrix B to the inverse of the parameter A. The condition $\mathrm{E}_\theta h(x;\theta) = 0$ which is characteristic of an estimating function in the general case now reads for source separation: $\mathrm{E}H(s) = 0$. In practice, it will be sufficient to find H verifying the weaker condition $\mathrm{E}H(Cs) = 0$ for C a non-mixing matrix, *i.e.* the components of Cs are the source signals possibly permuted and scaled.

2.2. Contrast functions and relative gradient

Some source separation techniques are based on the optimization of contrast functions: these are functions $c[y]$ of the distribution of vector $y = Bx$ taking their extremal values when B is a separating matrix. Typical instances are contrast functions measuring the independence of the components of y for instance by information-theoretic criteria or by using cross-cumulants (see below).

Stationary points of a contrast function $c[y]$ are characterized by the cancellation of the gradient of $c[y]$. For source separation, it is appropriate to use the relative [7, 17] or natural [4] gradient of $c[y]$. This is the $n \times n$ matrix denoted $\nabla c = \nabla c[y]$ such that:

$$\forall \mathcal{E} \in \mathbf{R}^{n\times n} \quad c[y + \mathcal{E}y] = c[y] + \mathrm{tr}\left\{\nabla c^\dagger \mathcal{E}\right\} + o(\mathcal{E}).$$

The relative gradient matrix ∇c characterizes the first-order variation of $c[y]$ when vector y is modified in $y + \mathcal{E}y$, *i.e.* when it is multiplied by $I + \mathcal{E}$.

It is often the case (exemples below) that the relative gradient of a contrast function $c[y]$ takes the form:

$$\nabla c[y] \;=\; \mathrm{E}H_c(y) \quad \text{or} \quad \nabla c[y] \;=\; \frac{1}{T}\sum_{t=1}^{T} H_c(y(t)) \tag{6}$$

for some function $H_c : \mathbf{R}^n \to \mathbf{R}^{n\times n}$. The first form is when $c[y]$ depends on the distribution of y; the second form when it depends on the sample distribution. The point here is that an estimating function H_c is derived from a contrast function c via relative differentiation. However, it is not necessarily the case that a valid estimating function derives from a contrast function.

2.3. Likelihood and related contrasts

If we believe that the source vector has a differentiable p.d.f. $q(s) = \prod_{i=1}^n q_i(s_i)$ and that T i.i.d. samples are available, it is a simple matter to write down the equation satisfied by the maximum likelihood (ML) estimator [13]. It turns out to be an equivariant estimating equation, associated to the estimating function

$$H_{ML}(y) = \phi(y)y^\dagger - I \tag{7}$$

where I is the $n \times n$ identity matrix and

$$\phi(y) = [\phi_1(y_1), \ldots, \phi_n(y_n)]^\dagger \qquad \phi_i = -\frac{q_i'}{q_i} \quad 1 \le i \le n. \tag{8}$$

This is also the form found in the infomax algorithm [6] showing that the latter actually is, in the case of source separation. Actually, any algorithm using non linear-functions ϕ as in (7) may be seen as a ML solver working under the assumption of i.i.d. source signals with pdf's related to ϕ as in eq. (8).

CMA-like criterion Maximizing the likelihood can be seen as minimizing the Kullback divergence between the hypothesized distribution of the sources and the empirical distribution of the output y [18]. One may also try to match other distributional properties. For instance, if the sources take only the values ± 1, the following criterion is of interest:

$$c_{\mathrm{CMA}}[y] = \sum_{i=1}^{n} \mathrm{E}(y_i^2 - 1)^2.$$

This is a simple-minded extension to real source separation of the well known CM criterion for the blind deconvolution of constant modulus signals. It is easily found to be associated to the estimating function

$$H_{\mathrm{CMA}}[y] = (y^3 - y)y^\dagger$$

where the i-th component y^3 is y_i^3.

2.4. Orthogonal contrasts

Orthogonal contrasts for source separation are to be optimized under the constraint that the output is spatially white: $\frac{1}{T}\sum_{t=1}^{T} y(t)y(t)^\dagger = I$. They can be implemented by first whitening the data and then constraining the separating matrix to be a rotation matrix. They give rise to estimating functions of a specific form, as shown below.

Orthogonal maximum likelihood Keeping the same setting as in sec. 2.3., but assuming (thanks to the whiteness constraint) that the mixing matrix is a rotation, one finds the estimating function to exist and to be equal to $\phi(y)y^\dagger - y\phi(y)^\dagger$. This can be combined with the whiteness condition to yield the estimating function

$$H_\phi^O(y) = yy^\dagger - I + \phi(y)y^\dagger - y\phi(y)^\dagger \tag{9}$$

whose symmetric part expresses the whiteness constraint and skew-symmetric part expresses the stationarity of the likelihood under the whiteness constraint. This particular form also appears in other instances below.

It is important to realize that the estimating function

$$H_\phi^O(y) = \alpha(yy^\dagger - I) + \beta(\phi(y)y^\dagger - y\phi(y)^\dagger) \tag{10}$$

yields the same estimates as (9) if α and β are two non zero scalars because the cancellation of a matrix is equivalent to the cancellation of both its symmetric and skew-symmetric parts. It follows that numeric factors affecting each of these parts do not affect the solutions of the estimation equations. However, they do affect the convergence of the *algorithms* which, as those described in section 4., try to solve the estimating equations by directly using the mean values of the H functions

Minimum kurtosis. If all the sources have negative kurtosis, separation may be achieved by minimizing under the whiteness constraint the contrast function $\sum_{i=1}^{n} \widehat{\mathrm{cum}}^{(4)}(y_i)$ where $\widehat{\mathrm{cum}}^{(4)}(y_i)$ denotes the sample kurtosis of y_i. It is not difficult to see that the resulting estimate also is the solution of the estimating equation using the estimating function

$$H^{O}_{\mathrm{mK}}(y) = yy^\dagger - I + y^3 y^\dagger - y y^{3\dagger}. \qquad (11)$$

Orthogonal cumulant matching. Denote k_i the kurtosis of the i-th source and define the matching criteria:

$$c_2[y] = \sum_{ij} |\widehat{\mathrm{cum}}(y_i, y_j) - \delta_{ij}|^2 \qquad (12)$$

$$c_4[y] = \sum_{ijkl} |\widehat{\mathrm{cum}}(y_i, y_j, y_k, y_l) - k_i \delta_{ijkl}|^2 \qquad (13)$$

Note that $c_2[y] = 0$ is equivalent to the whiteness constraint, while $c_4[y]$ measures the mismatch between all the (sample) 4th-order cumulants of y and the corresponding cumulants of the sources. Some algebra shows that if $c_2[y] = 0$, then $c_4[y] = \frac{1}{T}\sum_{t=1}^{T} h_4(y(t)) + cst$ where $h_4(y) = -2\sum_{i=1,n} k_i y_i^4$. From this, it is easily found that the source separation technique minimizing $c_4[y]$ under $c_2[y] = 0$ corresponds to an estimating function in the form (9) with $\phi = \phi_4$ defined by

$$\phi_4(y)_i = -k_i y_i^3 \qquad i = 1, \ldots, n. \qquad (14)$$

We note that a factor 8 actually appears in the computation of ϕ_4 but is discarded in (14), according to the remark of section 2.4.

3. ASYMPTOTIC ESTIMATING FUNCTIONS.

More sophisticated estimation techniques are not always *exactly* equivalent to solving estimating equations in the form (5). However, it usually exists an asymptotically equivalent form in the following sense. For a contrast function $c[y]$ estimated from T data samples, an asymptotic estimating function, if it exists, is a function H_c such that

$$\nabla c[y] = \frac{1}{T}\sum_{t=1}^{T} H_c(y(t)) + o(T^{-\frac{1}{2}}) \qquad (15)$$

for $y = (I + \mathcal{E})s$ with $\mathcal{E} = O(T^{-\frac{1}{2}})$ (this precision of order $O(T^{-\frac{1}{2}})$ indeed is the expected precision in regular estimation problems). In this case, one can show under standard regularity assumptions that the estimates obtained by optimizing $c[y]$ and those obtained as solutions of the estimating equation $\frac{1}{T}\sum_{t=1}^{T} H_c(y(t)) = 0$ differ only by a term of order $o(T^{-\frac{1}{2}})$. Hence, the asymptotic behavior of the estimates is essentially governed by the specification of H_c. In the next sections, we exhibit the functions H associated to known contrasts.

3.1. ICA and JADE

The ICA approach of Comon [9] consists in minimizing

$$c_{\mathrm{ICA}}[y] = \sum_{ijkl \neq iiii} |\widehat{\mathrm{cum}}(y_i, y_j, y_k, y_l)|^2$$

under the whiteness constraint. The relative gradient of this contrast cannot be put in the form $\mathrm{E}H_c(y)$, but it admits an asympotic form (15) which, after some calculations, is found to be: This is asymptotically equivalent to using the estimating function:

$$[H(y)]_{ij} = y_i y_j - \delta_{ij} - k_i y_i^3 y_j + k_j y_i y_j^3. \qquad (16)$$

We can also establish that the joint diagonalization criterion [8]

$$c_{\mathrm{JADE}}[y] = \sum_{ijkl \neq ijkk} |\widehat{\mathrm{cum}}(y_i, y_j, y_k, y_l)|^2$$

admits exactly the same asymptotic estimating function. This is no surprise, since we already know that these two criteria offer the same asymptotic performance [19]. We further note that the asymptotic estimating function (16) is identical to eq. (9) with ϕ defined in eq. (14). We conclude that, regarding off-line algorithms, identical performance are obtained using either ICA, JADE or the orthogonal 4th-order cumulant matching of section 2.4.. However, a fast optimization technique exists only for the JADE criterion.

3.2. Optimal cumulant matching

Another cumulant matching idea is to precompute the optimal weights to apply in matching the cumulants. This is described in [20] for the case of complex signals. We mention here, still without proof, a similar result for the real case. Interestingly enough, the optimal weights tend to simple numerical constants when the source distributions tend to normality. For nearly Gaussian signals, optimal (in an asymptotic MSE sense) weighting turns out to be very simple: the best matching criterion involving 2nd and 4th-order cumulants is

$$c^*_{24}[y] = 12\, c_2[y] + c_4[y].$$

The asymptotic estimating equation associated to this contrast is

$$H^*_{24}(y) = y\phi^*_{24}(y)^\dagger - I$$

where function ϕ^*_{24} is given by

$$\phi^*_{24}(y)_j = y_j - \frac{k_i}{6}(y_j^3 - 3y_j).$$

4. SOLVING ESTIMATING EQUATIONS.

4.1. Algorithms

The concept of estimating functions not only provides a unifying framework by which several source separation approaches can be compared: it is also straightforward to associate batch and/or adaptive algorithm to a particular estimating function $H(y)$. An adaptive algorithm for updating a separating matrix B_t upon reception of a new sample $x(t)$ is:

$$B_{t+1} = (I - \mu_t H(y(t)))B_t$$

where $y(t) = B_t x(t)$ and μ_t is a sequence of adaptation steps. Such an algorithm admits as a stationary any matrix B_* such that $\mathrm{E}H(y) = \mathrm{E}H(B_* x) = 0$. Adaptive algorithms based on an H function in the form (7) are described in [4]; those based on form (9) are studied in detail in [17].

A batch algorithm for the iterative solution of the estimating equation based on T samples is by setting $y(t) = x(t)$ for $1 \leq t \leq T$ and then by looping through the two steps

1 $\widehat{H} \leftarrow T^{-1} \sum_t H(y(t))$

2 $y(t) \leftarrow (I - \mu \widehat{H})\, y(t)$ for $t = 1, \ldots, T$

4.2. Performance.

The special form (4) of estimating function for source separation automatically ensures uniform performance of both the adaptive and batch versions of the algorithms outlined in the previous section. Here 'uniform' means 'independent of the mixing matrix A_*'. It follows that the (asymptotic) performance can be characterized uniquely in terms of the distribution of the input and of the estimating function $H(\cdot)$.

Because of their links with ML estimation, the particular forms (7) and (9) of estimating functions have already been studied in some detail. An asymptotic performance analysis of batch algorithms using estimating functions in the form (7) can be found in [13]. A similar study for adaptive algorithms based on form (9) can be found in [17].

CONCLUSION

We have informally presented the general framework of estimating functions for source separation which was shown to encompass many known techniques. Due to lack of space, calculations were omitted ans several issues have been left pending such as the unicity of estimating functions and a more serious treatment of asymptotic estimating functions. They will be addresed in a more formal study in preparation.

REFERENCES

[1] Christian Jutten and Jeanny Hérault. Blind separation of sources: an adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24:1–10, 1991.

[2] Eric Moreau and Odile Macchi. A one stage self-adaptive algorithm for source separation. In *Proc. ICASSP*, Adelaide, Australia., 1994.

[3] A. Cichocki, R. Unbehauen, and E. Rummert. Robust learning algorithm for blind separation of signals. *Electronic letters*, 30(17):1386–87, 1994.

[4] S. Amari, A. Cichocki, and H. Yang. A new learning algorithm for blind signal separation. In *Proc. NIPS*, Denver, 1995.

[5] N. Delfosse and P. Loubaton. Adaptative blind separation of independent sources: a deflation approach. *Signal Processing*, 45(1):59–83, 1995.

[6] Anthony J. Bell and Terrence Sejnowski. Blind separation and blind deconvolution: an information-theoretic approach. In *Proc. ICASSP*, Detroit, 1995.

[7] Beate Laheld and Jean-François Cardoso. Adaptive source separation with uniform performance. In *Proc. EUSIPCO*, pages 183–186, Edinburgh, September 1994.

[8] Jean-François Cardoso and Antoine Souloumiac. Blind beamforming for non Gaussian signals. *IEE Proceedings-F*, 140(6):362–370, December 1993.

[9] P. Comon. Independent component analysis, a new concept ? *Signal Processing, Elsevier*, 36(3):287–314, April 1994. Special issue on Higher-Order Statistics.

[10] Lang Tong, Yujiro Inouye, and Ruey-wen Liu. Waveform preserving blind estimation of multiple independent sources. *IEEE Tr. on SP*, 41(7):2461–2470, July 1993.

[11] N. Yuen and B. Friedlander. Asymptotic performance analysis of blind signal copy using fourth-order cumulants. *International Journal of Adaptive Control and Signal Processing*, pages 239–65, mar 1996.

[12] Michel Gaeta and Jean-Louis Lacoume. Source separation without a priori knowledge: the maximum likelihood solution. In *Proc. EUSIPCO*, pages 621–624, 1990.

[13] Dinh-Tuan Pham, Philippe Garrat, and Christian Jutten. Separation of a mixture of independent sources through a maximum likelihood approach. In *Proc. EUSIPCO*, pages 771–774, 1992.

[14] Robert J. Serfling. *Approximation theorems of mathematical statistics*. Wiley series in probability and mathematical statistics. John Wiley and sons, 1980.

[15] E. L. Lehmann. *Testing statistical hypothesis*. Wiley pub. in statistics. John Wiley, 1959.

[16] Jean-François Cardoso. The equivariant approach to source separation. In *Proc. NOLTA*, pages 55–60, 1995.

[17] Jean-François Cardoso and Beate Laheld. Equivariant adaptive source separation. *IEEE Trans. on S.P.*, 44(12), December 1996. To appear.

[18] Jean-François Cardoso. Infomax and maximum likelihood for source separation. Accepted for publication in IEEE Letters on S.P., 1997.

[19] Antoine Souloumiac and Jean-François Cardoso. Comparaison de méthodes de séparation de sources. In *Proc. GRETSI, Juan les Pins,France*, pages 661–664, 1991.

[20] Jean-François Cardoso, Sandip Bose, and Benjamin Friedlander. On optimal source separation based on second and fourth order cumulants. In *Proc. IEEE Workshop on SSAP, Corfou, Greece*, 1996.