THE MAXIMUM LIKELIHOOD APPROACH TO COMPLEX ICA

Jean-François Cardoso

ENST/TSI
75634 Paris, France

Tülay Adalı

University of Maryland Baltimore County
Baltimore, MD 21250, USA

ABSTRACT

We derive the form of the best non-linear functions for performing independent component analysis (ICA) by maximum likelihood estimation. We show that both the form of nonlinearity and the relative gradient update equations for likelihood maximization naturally generalize to the complex case, and that they coincide with the real case. We discuss several special cases for the score function as well as adaptive scores.

1. INTRODUCTION

There are a number of approaches for performing independent component analysis (ICA) of complex-valued sources (see *e.g.*, [1, 2, 3, 6, 7, 10]), some with specific assumptions on the source distributions such as their circular/noncircular nature. Maximum likelihood (ML) provides a desirable framework for many estimation problems, and in the case of ICA, it provides guidance for the choice of nonlinear functions to generate the higher-order statistics. The nonlinearity to achieve ICA should be chosen as close as possible to the score function of each source, which is given by $\psi_i = -{p'}_i/p_i$ where p_i is the probability density function of the ith source.

In this paper, using a simple transform from the complex n-dimensional space to the real $2n$-dimensional one, we derive the form of the best nonlinearity for complex-valued ML. We show that the score function coincides with the score for the real-valued case and discuss a number of special cases for the source distribution as well as adaptive scores as in [9]. The derivation we present is straightforward and eliminates the need to work with complex *partial* differential operators as in [8], where the form of the nonlinearity is derived for mutual information minimization, and, as expected, coincides with the form we derive. We show that the relative gradient algorithm naturally generalizes to the complex case as well.

2. COMPLEX PRELIMINARIES AND AN ORTHOGONAL DECOMPOSITION

A complex random variable X is defined as a random variable $X = X_r + jX_i$ where the real and imaginary parts, X_r and X_i are real-valued random variables. We consider the case where the joint probability density function (pdf) $p_X(x) \equiv p(x_r, x_i)$ exists and is log-differentiable as an $\mathbb{R}^2 \mapsto \mathbb{R}$ function.

We consider the traditional linear mixing model, *i.e.*, the vector of observed random variables $\mathbf{x} \in \mathbb{C}^n$ can be written as a linear mixture of the sources $\mathbf{s} \in \mathbb{C}^n$, such that $\mathbf{x} = \mathbf{A}\mathbf{s}$ where $\mathbf{A} \in \mathbb{C}^{n\times n}$. The ICA problem is then the recovery of the source estimates $\mathbf{y} = \mathbf{B}\mathbf{x}$ by optimizing a suitable cost.

We define the following linear invertible mappings from the complex space to the real one as in [11]:

$$\begin{aligned} \mathbf{x} \in \mathbb{C}^{n\times 1} &\mapsto \bar{\mathbf{x}} = \begin{bmatrix} \mathbf{x}_r \\ \mathbf{x}_i \end{bmatrix} \in \mathbb{R}^{2n\times 1} \\ \mathbf{A} \in \mathbb{C}^{n\times n} &\mapsto \bar{\mathbf{A}} = \begin{bmatrix} \mathbf{A}_r & -\mathbf{A}_i \\ \mathbf{A}_i & \mathbf{A}_r \end{bmatrix} \in \mathbb{R}^{2n\times 2n} \end{aligned} \quad (1)$$

and write the linear mixing model as $\bar{\mathbf{x}} = \bar{\mathbf{A}}\bar{\mathbf{s}}$. We note that the group of invertible complex $n \times n$ matrices is isomorphic to the group of invertible matrices obtained through the $\bar{\cdot}$ mapping. This isomorphism is the key in making our derivations work out nicely.

Next, we define an orthogonal decomposition of the space of $2n\times 2n$ matrices, which is instrumental in our development of the maximum likelihood formulation for the complex case. We write a $2n \times 2n$ matrix $\mathbf{M}$ in terms of 4 blocks of size $n \times n$ as $\mathbf{M} = \begin{bmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{bmatrix}$. The linear space of $2n \times 2n$ matrices can be decomposed into two orthogonal spaces: $\mathbb{R}^{2n\times 2n} = \mathcal{M}_n^+ \oplus \mathcal{M}_n^-$ where $\mathcal{M}^+$ (*resp.* $\mathcal{M}^-$) contains any matrix such that $\mathbf{M}_{11} = \mathbf{M}_{22}$ and $\mathbf{M}_{12} = -\mathbf{M}_{21}$ (*resp.* $\mathbf{M}_{11} = -\mathbf{M}_{22}$ and $\mathbf{M}_{12} = \mathbf{M}_{21}$). Hence a $2n \times 2n$ real matrix has the orthogonal decomposition $\mathbf{M} = \mathbf{M}^+ + \mathbf{M}^-$ where

$$\mathbf{M}^+ = \tfrac{1}{2}\begin{bmatrix} \mathbf{M}_{11}+\mathbf{M}_{22} & \mathbf{M}_{12}-\mathbf{M}_{21} \\ \mathbf{M}_{21}-\mathbf{M}_{12} & \mathbf{M}_{11}+\mathbf{M}_{22} \end{bmatrix} \in \mathcal{M}^+.$$

A similar expression can be written for $\mathbf{M}^-$. Note that the set of invertible matrices of $\mathbf{M}^+$ form a group for the usual multiplication of matrices and we have $\bar{\mathbf{A}} \in \mathcal{M}^+$, which is defined in (1).

The following are some useful properties of the mapping $\bar{\cdot}$ defined in (1) that can easily be verified.

P1: $\overline{\mathbf{A}\mathbf{x}} = \bar{\mathbf{A}}\bar{\mathbf{x}}$;

P2: $\overline{\mathbf{A}\mathbf{B}} = \bar{\mathbf{A}}\bar{\mathbf{B}}$ and thus $\overline{\mathbf{A}^{-1}} = \left(\bar{\mathbf{A}}\right)^{-1}$, and

P3: $\overline{\mathbf{x}\mathbf{y}^H} = 2\left(\bar{\mathbf{x}}\,\bar{\mathbf{y}}\right)^+$ where $\cdot^H$ denotes Hermitian transpo-

sition (transposition plus complex conjugation) and $\cdot^+$ is defined above.

3. LIKELIHOOD OF COMPLEX MIXTURES

We review the ML formulation for the real case and its maximization with relative gradients in Appendix A. For the complex case, we can use the $\bar{\cdot}$ mapping and write the log-likelihood function for T independent samples $\bar{\mathbf{x}}(t) \in \mathbb{R}^{2n}$ as in the real case: $\mathcal{L}(\bar{\mathbf{A}}) = \sum_{t=1}^{T} \ell(\bar{\mathbf{A}})$, where

$$\ell(\bar{\mathbf{A}}) = \log p(\bar{\mathbf{x}}(t)|\bar{\mathbf{A}}) = \log p_{\bar{s}}(\bar{\mathbf{A}}^{-1}\bar{\mathbf{x}}) - \log|\det \bar{\mathbf{A}}|.$$

We then let $p_{\bar{s}}(\bar{\mathbf{s}})$ represent the pdf of $\bar{\mathbf{s}}$ and $\bar{\psi} : \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ the associated score function, *i.e.*, $\bar{\psi}(\bar{\mathbf{y}}) = -\frac{\partial}{\partial \bar{\mathbf{y}}} \log p_{\bar{s}}(\bar{\mathbf{y}})$ where $\bar{\mathbf{y}} = \bar{\mathbf{A}}^{-1}\bar{\mathbf{x}}$.

We next find the stationarity condition for the likelihood and derive the relative gradient algorithm.

Relative variation of the likelihood. If $\mathbf{A}$ is changed into $\mathbf{A}(\mathbf{I}_n + \boldsymbol{\mathcal{E}})$, then $\bar{\mathbf{A}}$ is changed into $\overline{\mathbf{A}(\mathbf{I}_n + \boldsymbol{\mathcal{E}})} = \bar{\mathbf{A}}(\mathbf{I}_{2n} + \overline{\boldsymbol{\mathcal{E}}})$ by **P2**. Therefore, we can use the result obtained in the real case and write

$$\mathcal{L}(\bar{\mathbf{A}}(\mathbf{I}_{2n} + \overline{\boldsymbol{\mathcal{E}}})) = \mathcal{L}(\bar{\mathbf{A}}) + \langle \mathcal{H}_2(\bar{\mathbf{A}}), \overline{\boldsymbol{\mathcal{E}}} \rangle + O(\|\overline{\boldsymbol{\mathcal{E}}}\|^2)$$

using $\bar{\mathbf{A}}^{-1}\bar{\mathbf{x}} = \overline{\mathbf{A}^{-1}}\bar{\mathbf{x}} = \overline{\mathbf{A}^{-1}\mathbf{x}} = \bar{\mathbf{y}}$ and defining

$$\mathcal{H}_2(\bar{\mathbf{A}}) = \sum_{t=1}^{T} H_2(\bar{\mathbf{y}}(t)) \quad \text{with} \quad H_2(\bar{\mathbf{y}}) = \overline{\psi}(\bar{\mathbf{y}})\bar{\mathbf{y}}^T - \mathbf{I}_{2n}.$$

Here, $\langle \cdot, \cdot \rangle$ denotes the inner product between real matrices: $\langle \mathbf{M}, \mathbf{N} \rangle = \mathrm{trace}(\mathbf{M}^T\mathbf{N})$.

However $\mathcal{H}_2(\bar{\mathbf{A}})$ is not necessarily equal to $\overline{\mathbf{M}}$ for some matrix $\mathbf{M}$ of $\mathbb{C}^{n\times n}$ (because it does not have, in general, the required block structure). But, we have

$$\langle \mathcal{H}_2(\bar{\mathbf{A}}), \overline{\boldsymbol{\mathcal{E}}} \rangle = \langle \mathcal{H}_2(\bar{\mathbf{A}})^+, \overline{\boldsymbol{\mathcal{E}}} \rangle$$

because $\mathcal{H}_2(\bar{\mathbf{A}}) = \mathcal{H}_2(\bar{\mathbf{A}})^+ + \mathcal{H}_2(\bar{\mathbf{A}})^-$ and $\langle \mathcal{H}_2(\bar{\mathbf{A}})^-, \overline{\boldsymbol{\mathcal{E}}} \rangle = 0$ since $\mathcal{H}_2(\bar{\mathbf{A}})^- \in \mathcal{M}_n^-$ and $\overline{\boldsymbol{\mathcal{E}}} \in \mathcal{M}_n^+$, which are orthogonal subspaces. Now, $\mathcal{H}_2(\bar{\mathbf{A}})^+$ being in $\mathcal{M}_n^+$ is the image of some complex $n \times n$ matrix. Using the definition of $\mathcal{H}_2$ and property **P3** given in section 2, one easily finds that $\mathcal{H}_2(\bar{\mathbf{A}})^+ = \frac{1}{2}\overline{\mathcal{H}(\mathbf{A})}$ for

$$\mathcal{H}(\mathbf{A}) = \sum_{t=1}^{T} \psi(\mathbf{y}(t))\mathbf{y}^H(t) - \mathbf{I}_n, \tag{2}$$

where the associated score function $\psi_k(\cdot) : \mathbb{C} \mapsto \mathbb{C}$ is given by

$$\psi_k(y) = -\frac{1}{2}\left(\frac{\partial \log p_{s_k}(y_r, y_i)}{\partial y_r} + j\frac{\partial \log p_{s_k}(y_r, y_i)}{\partial y_i}\right) \tag{3}$$

for the kth source.

We can thus write the final result as:

$$\mathcal{L}(\mathbf{A}(\mathbf{I} + \boldsymbol{\mathcal{E}})) = \mathcal{L}(\mathbf{A}) + \frac{1}{2}\langle \overline{\mathcal{H}(\mathbf{A})}, \overline{\boldsymbol{\mathcal{E}}} \rangle + O(\|\overline{\boldsymbol{\mathcal{E}}}\|^2).$$

Maximum likelihood. Hence, a stationary point of the likelihood is characterized by $\langle \overline{\mathcal{H}(\mathbf{A})}, \overline{\boldsymbol{\mathcal{E}}} \rangle = 0$ for any $\boldsymbol{\mathcal{E}}$ in $\mathbb{C}^{n\times n}$ and therefore $\langle \overline{\mathcal{H}(\mathbf{A})}, \mathbf{M} \rangle = 0$ for any $\mathbf{M} \in \mathcal{M}_n^+$ which implies that $\overline{\mathcal{H}(\mathbf{A})}$ can only be zero since this is the only matrix of $\mathcal{M}_n^+$ which is orthogonal to all the other elements. The latter implies that the ML estimate of $\mathbf{A}$ is a solution of $\mathcal{H}(\mathbf{A}) = \mathbf{0}$.

Relative gradient algorithm. One can repeat here the same argument as in the real case: since the first order variation of the log-likelihood when $\mathbf{A}$ is changed into $\mathbf{A}(\mathbf{I}_n + \boldsymbol{\mathcal{E}})$ is $\langle \overline{\mathcal{H}(\mathbf{A})}, \overline{\boldsymbol{\mathcal{E}}} \rangle$, one is guaranteed to increase it by making in small in the direction of the relative gradient, *i.e.*, by taking $\overline{\boldsymbol{\mathcal{E}}} = \mu\overline{\mathcal{H}(\mathbf{A})}$ for a small enough step μ. Of course, this implies $\boldsymbol{\mathcal{E}} = \mu\mathcal{H}(\mathbf{A})$. As in the real case, it is again simpler to update the inverse of $\mathbf{A}$.

Hence the ML solutions have exactly the same structure as in the real case. The relative gradient algorithm [5] also generalizes naturally as:

1. Compute $\mathcal{H} = \frac{1}{T}\sum_t \psi(\mathbf{y}(t))\mathbf{y}^H(t) - \mathbf{I}_n$ with $\mathbf{y}(t) = \mathbf{B}\mathbf{x}(t)$. If $\|\mathcal{H}\|$ is small enough, stop.

2. Update $\mathbf{B}$ into $(\mathbf{I}_n - \mu\mathcal{H})\mathbf{B}$ and go to step 1.

Each step of this algorithm is guaranteed to increase the likelihood of $\mathbf{A} = \mathbf{B}^{-1}$ for small enough step size μ.

Note that $\mathcal{H}$ can be replaced with the instantaneous values $H(\mathbf{A}) = \psi(\mathbf{y}(t))\mathbf{y}^H(t) - \mathbf{I}_n$ in the updates for stochastic relative gradient optimization. This stochastic relative gradient update coincides with the updates used in [1, 4, 2, 10]. In [1], this is introduced in the context of ICA using nonlinear decorrelations with a set of complex functions from the trigonometric and hyperbolic family, in [4] for performing complex Infomax, and in [2, 10], using ML/Infomax assuming that the sources are circular.

As the results for the updates and the score function show, there is perfect parallelism between the complex and real cases and the development presented provides a straightforward way to derive the form of the nonlinearity as well as the form of the updates.

4. SOME SCORE FUNCTIONS

To recover independent sources through ML, the form of the score function ψ_k should be adapted to each individual source s_k. Next we consider a number of cases for the source distribution, discuss the form of corresponding score function, and define adaptive scores as in [9].

4.1. Gaussian variables

For a zero mean Gaussian-distributed source, the score function can be written as

$$\psi_g(s) = \frac{s\mathrm{E}\{|s|^2\} - s^*\mathrm{E}\{s^2\}}{2(\mathrm{E}\{|s|^2\}^2 - |\mathrm{E}\{s^2\}|^2)}. \quad (4)$$

The Gaussian source is *circular*, *i.e.*, its pdf is rotation invariant, when $E\{s^2\} = 0$ implying zero correlation between the real and imaginary parts and equal standard deviations, *i.e.*, $\sigma_r = \sigma_i \equiv \sigma$. Note that only when the source is circular, the score function is a linear function of s (given by $\psi_g(s) = s/4\sigma^2$) coinciding with the result for the real case, *i.e.*, the best nonlinearity for the gaussian random variable is the linear one. Otherwise, it is linear either in s_r or s_i but not in s.

The Gaussian score function given in (4) implies that $\mathrm{E}\{s_k s_l^*\} = 0$ and $\mathrm{E}\{s_k^* s_l^*\} = 0$ for $k \neq l$, since we have $\mathrm{E}\{\psi_g(s_k)s_l^*\} = 0$ for $k \neq l$. Hence, it satisfies the condition for the strong uncorrelating transform defined in [7], which requires that both the covariance matrix $\mathrm{E}\{\mathbf{s}\mathbf{s}^H\}$ and the pseudo-covariance matrix $\mathrm{E}\{\mathbf{s}\mathbf{s}^T\}$ to be diagonal for s_k and s_l to be *strongly uncorrelated*. It is shown that this transform can be used to achieve ICA of noncircular sources as long as they all have distinct spectral coefficients, which are defined as $\sigma_r^2 - \sigma_i^2$ for each source.

4.2. Circular variables

If a given source has a circular distribution, *i.e.*, $p_c(s) = g(|s|)$. Then, it is a simple matter to find that the score function has the form

$$\psi_c(s) = \frac{s}{2|s|} f(|s|) \quad \text{for} \quad f = -\frac{g'}{g}.$$

In other words, the score function always has the same phase as its argument. This is the form of the score function proposed in [2] where all sources are assumed to be circular.

4.3. Adaptive scores

We consider the issue of adapting the non-linear functions to the sources distributions. From ML theory, it is clear that for each source, the non-linear function $\psi(s)$ should be as close as possible to the "true" score function $\psi_o(s)$ for the corresponding source.

We extend the elegant method of Pham and Garat [9] to the complex case. It is based on using a set $\mathcal{F}$ of M predefined functions $f_1, \ldots, f_M : \mathbb{R}^2 \to \mathbb{R}^2$. Each non-linear function $\psi(\bar{\mathbf{s}}) : \mathbb{R}^2 \to \mathbb{R}^2$ is taken to be a linear combination: $\psi(\bar{\mathbf{s}}) = \sum_{m=1}^{M} \alpha_m f_m(\bar{\mathbf{s}}) = F(\bar{\mathbf{s}})\boldsymbol{\alpha}$ with $F(\bar{\mathbf{s}}) = [f_1(\bar{\mathbf{s}}), \ldots, f_M(\bar{\mathbf{s}})]$ and $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_M]^T$. For a given source, the coefficients are chosen to minimize $E\|\psi_o(\bar{\mathbf{s}}) - F(\bar{\mathbf{s}})\boldsymbol{\alpha}\|^2$ where $\psi_o(\bar{\mathbf{s}})$ denotes the score function of the true distribution of $\bar{\mathbf{s}}$. Since this score-matching criterion is quadratic in $\boldsymbol{\alpha}$, the optimal value of $\boldsymbol{\alpha}$ is easily found to be $\boldsymbol{\alpha}_{\mathrm{opt}} = \mathbf{C}^{-1}\mathbf{c}$ where $C_{lm} = \mathrm{E}\{f_l(\bar{\mathbf{s}})^T f_m(\bar{\mathbf{s}})\}$ and $c_m = \mathrm{E}\{f_m(\bar{\mathbf{s}})^T \psi_o(\bar{\mathbf{s}})\}$. Matrix $\mathbf{C}$ is easily estimated by replacing the expectation operator by a sample average. In order to estimate vector $\mathbf{c}$, we need the true score function $\psi_o(\bar{\mathbf{s}})$, which is unknown. The nice observation given in [9] to overcome this problem (as extended here to the complex case) is to recognize that, for any well-behaved function $f(\bar{\mathbf{s}}) = [f_r(s), f_i(s)]^T$, we can use integration by parts to write

$$\mathrm{E}\{f(\bar{\mathbf{s}})^T \psi_o(\bar{\mathbf{s}})\} = E\left\{\frac{\partial f_r(s)}{\partial s_r} + \frac{\partial f_i(s)}{\partial s_i}\right\} \quad (5)$$

provided that $f_r(s)p_s(s_r, s_i)$ vanishes at infinity—and similarly for $f_i(s)p_s(s_r, s_i)$. Hence, even if $\psi_o(\bar{\mathbf{s}})$ is unknown, its correlation with any function (hence vector $\mathbf{c}$) is easily estimated as the sample average of derivatives as suggested by equation (5). In the real case, it is shown that if the set $\mathcal{F}$ contains at least the identity function plus some other non-linear function, then the stability of the separation is guaranteed [9].

Consider the Gaussian score function $\psi_g(s)$ defined in (4). Its mapping is

$$\overline{\psi_g(s)} = \frac{1}{4(1-\rho^2)} \begin{bmatrix} s_r/\sigma_r^2 - \rho s_i/\sigma_r\sigma_i \\ s_i/\sigma_i^2 - \rho s_r/\sigma_r\sigma_i \end{bmatrix}$$

i.e., $\overline{\psi_g(s)} = F_g(\bar{\mathbf{s}})\,\boldsymbol{\alpha}_g$ for $F_g(\bar{\mathbf{s}}) = \begin{bmatrix} s_r & 0 & s_i \\ 0 & s_i & s_r \end{bmatrix}$ and $\boldsymbol{\alpha}_g = \frac{1}{4(1-\rho^2)}[\frac{1}{\sigma_r^2}, \frac{1}{\sigma_i^2}, \frac{-\rho}{\sigma_i\sigma_r}]^T$ where $\rho = \mathrm{E}\{s_r s_i\}$. That is, we need $M = 3$ and the basis functions: $f_1(s) = \mathrm{Re}(s) = s_r$, $f_2(s) = j\mathrm{Im}(s) = js_i$, and $f_3(s) = s_i + js_r = js^*$) (or another equivalent choice) to represent the score function of a non-circular Gaussian complex variable. Hence, if we take $F_g(\bar{\mathbf{s}})$ as the basis functions, we do not need to use (5) explicitly because the best coefficients are already apparent: they are the $\boldsymbol{\alpha}_g$ defined above and they are optimal for Gaussian variables since $F_g(\bar{\mathbf{s}})\boldsymbol{\alpha}_g$ is the Gaussian score.

Normalization. Note that in the blind separation problem, the phase of each source cannot be identified. Thus, we can always add an arbitrary phase to each source in such a way that $\rho = 0$. Similarly, a global scale can be imposed on each source. We choose to normalize the separation by imposing $\sigma_r^2 + \sigma_i^2 = 2$. One can also add an extra $\pi/2$ phase shift to ensure that $\sigma_r^2 \geq \sigma_i^2$. Then the Gaussian score reduces to $\psi_g(s) = (s_r/\sigma_r^2 + js_i/\sigma_i^2)/4$.

We have then the following algorithm for non-circular Gaussian sources: start with $\mathbf{y}(t) = \mathbf{x}(t)$ and iterate through

1. Normalize each source $y_m(t)$ as described above.
2. Compute σ_r^2 and σ_i^2 for each source and apply the Gaussian score $\psi_g(s) = (s_r/\sigma_r^2 + js_i/\sigma_i^2)/4$ to it.
3. Compute $\mathcal{H}$ as before. Exit if it is small enough. Otherwise, update the sources by $\mathbf{y}(t) = (\mathbf{I}_n - \mu\mathcal{H})\mathbf{y}(t)$

Since this algorithm computes the ML estimate under a Gaussian model, it does perform strong uncorrelating transform as discussed earlier. However, SUT algorithm given in [7] performs a second-order direct implementation to achieve independence and hence is a much more efficient way to perform ML in the case of Gaussian sources.

In the case of non-Gaussian sources, it is desirable to include extra information in addition to second-order statistics. This is easily done by complementing $F_G(\bar{s})$ with non-linear functions. The simplest possibility is to include a single 'circular score' of the form $f_c(s) = g(|s|^2)s/|s|^2$ for some function $g : \mathbb{C} \to \mathbb{R}$. Then, taking into account the simplification introduced by our normalization convention, we would use the basis $F(\bar{\mathbf{s}}) = \begin{bmatrix} s_r & 0 & g(|s|^2)s_r/|s|^2 \\ 0 & s_i & g(|s|^2)s_i/|s|^2 \end{bmatrix}$ and we would need to use equation (5) explicitly to obtain the best coefficients.

Remark: in this simple approach, all the information carried by the non-circularity is exploited by the linear part of the score *i.e.*, using second-order information and, conversely, the other-than-second-order information is exploited 'circularly' via the circular function $g(\cdot)$. This approach is much simpler than the two-step algorithm in [8].

A. APPENDIX

Summary of ML estimation in the real case

If $\mathbf{s}$ is an $n \times 1$ real random vector with a probability density $p_s(\mathbf{s})$ on $\mathbb{R}^n$ and one observes $\mathbf{x} = \mathbf{A}\mathbf{s}$ for some $n \times n$ matrix $\mathbf{A}$, then the pdf of $\mathbf{x}$ is $p(\mathbf{x}|\mathbf{A}) = p_s(\mathbf{A}^{-1}\mathbf{x})/|\det \mathbf{A}|$. The *score function* associated with a (differentiable) pdf p_s is the function $\psi : \mathbb{R}^n \to \mathbb{R}^n$ defined as $\psi(s) = -\frac{\partial}{\partial s} \log p(s)$.

Relative variations. Let $\ell(\mathbf{A})$ denote the log-likelihood of $\mathbf{A}$, *i.e.*, $\ell(\mathbf{A}) = \log p(\mathbf{x}|\mathbf{A})$. The first order variations of $\ell(\mathbf{A})$ around a given point $\mathbf{A}$ are conveniently expressed in terms of relative variations. A classic derivation yields

$$\ell(\mathbf{A}(\mathbf{I} + \boldsymbol{\mathcal{E}})) = \ell(\mathbf{A}) + \langle H(\mathbf{A}^{-1}\mathbf{x}), \boldsymbol{\mathcal{E}} \rangle + O(\|\boldsymbol{\mathcal{E}}\|^2)$$

where $H : \mathbb{R}^n \to \mathbb{R}^{n \times n}$ is the function $H(s) = \psi(\mathbf{s})\mathbf{s}^T - \mathbf{I}$.

Denote $\mathcal{L}(\mathbf{A}) = \log p(x(1), \ldots, x(T)|\mathbf{A})$ the log-likelihood of $\mathbf{A}$ in front of T samples $x(1), \ldots, x(T)$. If they are modeled as independent, then $\mathcal{L}(\mathbf{A}) = \sum_{t=1}^{T} \log p(\mathbf{x}(t)|\mathbf{A})$ and the (relative) variation of $\mathcal{L}(\mathbf{A})$ is

$$\mathcal{L}(\mathbf{A}(\mathbf{I} + \boldsymbol{\mathcal{E}})) = \mathcal{L}(\mathbf{A}) + \langle \mathcal{H}(\mathbf{A}), \boldsymbol{\mathcal{E}} \rangle + O(\|\boldsymbol{\mathcal{E}}\|^2)$$

where $\mathcal{H}(\mathbf{A}) = \sum_{t=1}^{T} H\left(\mathbf{A}^{-1}\mathbf{x}(t)\right)$.

Maximum likelihood estimate. If $\mathbf{A}$ is a maximum of the likelihood, then it must be stationary at first order against all variations of $\mathbf{A}$. Therefore, it must verify $\langle \mathcal{H}(\mathbf{A}), \boldsymbol{\mathcal{E}} \rangle = 0$ for all $\boldsymbol{\mathcal{E}}$ in $\mathbb{R}^{n \times n}$. This, of course, implies that a stationary point of the likelihood is characterized by $\mathcal{H}(\mathbf{A}) = 0$.

Maximization of the likelihood with the relative gradient. Let $\mathbf{A}$ be a matrix such that $\mathcal{H}(\mathbf{A}) \neq 0$. Let us change $\mathbf{A}$ to increase the likelihood. Consider a small step in the (relative) direction $\mathbf{D} \in \mathbb{R}^{n \times n}$, *i.e.*, $\mathbf{A}$ is replaced by $\mathbf{A}(\mathbf{I} + \mu\mathbf{D})$ for some small step μ. Then the log-likelihood is increased by $\langle \mathcal{H}(\mathbf{A}), \mu\mathbf{D} \rangle + O(\mu^2)$. Taking $\mathbf{D} = \mathcal{H}(\mathbf{A})$ (*i.e.* moving in the direction of the relative gradient) and μ a small enough positive step guarantees a positive increase since $\langle \mathcal{H}(\mathbf{A}), \mu\mathcal{H}(\mathbf{A}) \rangle = \mu\|\mathcal{H}(\mathbf{A})\|^2$.

Hence the relative gradient for maximizing the likelihood of $\mathbf{A}$ in front of T samples assumed to be independent and generated by $\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t)$ is: while $\mathcal{H}(\mathbf{A})$ is not small enough, change $\mathbf{A}$ into $\mathbf{A}(\mathbf{I} + \mu\mathcal{H}(\mathbf{A}))$.

It is more natural (and cheaper) to implement this idea by updating the inverse $\mathbf{B}$ of $\mathbf{A}$ rather than $\mathbf{A}$ itself. A small relative change of $\mathbf{A}$ into $\mathbf{A}(\mathbf{I} + \boldsymbol{\mathcal{E}})$ induces a small change of $\mathbf{B} = \mathbf{A}^{-1}$ into $(\mathbf{A}(\mathbf{I} + \boldsymbol{\mathcal{E}}))^{-1} = (\mathbf{I} + \boldsymbol{\mathcal{E}})^{-1}\mathbf{A}^{-1} \approx (\mathbf{I} - \boldsymbol{\mathcal{E}})\mathbf{B}$. Hence, we never need to invert $\mathbf{A}$. More explicitly the algorithm for a given learning step μ starts with some matrix $\mathbf{B}$ and iterates through:

1. Compute $\mathbf{y}(t) = \mathbf{B}\mathbf{x}(t)$ and $\mathcal{H} = \frac{1}{T}\sum_t H(\mathbf{y}(t))$. If $\|\mathcal{H}\| \approx 0$, stop.
2. Update $\mathbf{B}$ into $(\mathbf{I}_n - \mu\mathcal{H})\mathbf{B}$ and go to step 1.

B. REFERENCES

[1] T. Adalı, T. Kim, and V. Calhoun, "Independent component analysis by complex nonlinearities," in *Proc. ICASSP*, Montreal, Canada, May 2004, vol. 5, pp. 525–528.

[2] J. Anemüller, T. J. Sejnowski, and S. Makeig, "Complex independent component analysis of frequency-domain EEG data," *Neural Networks*, vol. 16, pp. 1311–23, 2003.

[3] E. Bingham and A. Hyvärinen, "A fast fixed-point algorithm for independent component analysis of complex-valued signals," *Int. J. Neural Systems*, 10(1):1, pp. 1–8, 2000.

[4] V. Calhoun and T. Adalı, "Complex Infomax: Convergence and approximation of Infomax with complex nonlinearities," in *Proc. NNSP*, Martigny, Switzerland, Sep. 2002.

[5] J.-F. Cardoso and B. H. Laheld, "Equivariant adaptive source separation," *IEEE Trans. Signal Processing*, vol. 44, no. 12, pp. 3017–3030, Dec. 1996.

[6] J.-F. Cardoso, and A. Souloumiac, "Blind beamforming for non-Gaussian signals," *IEE Proc. Radar Signal Proc.*, vol. 140, no., 6, pp. 362–370, 1993.

[7] J. Eriksson and V. Koivunen, "Complex-valued ICA using second order statistics," in *Proc. MLSP*, Saõ Luis, Brazil, 2004.

[8] J. Eriksson, A.-M. Seppola, and V. Koivunen, "Complex ICA for circular and non-circular sources," in *Proc. EUSIPCO*, Antalya, Turkey, 2005.

[9] D.-T. Pham and P. Garat, "Blind separation of mixture of independent sources through a quasimaximum likelihood approach," *IEEE Trans. Signal Processing*, vol. 45, pp. 1712–1725, July 1997.

[10] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, pp. 21-34, 1998.

[11] N. N. Vakhania and N. P. Kandelaki, "Random vectors with values in complex Hilbert spaces," *Theory Probability Appl.*, vol. 41, no. 1, pp. 116–131, Feb. 1996.