# Independent Component Analysis by Wavelets

Pascal Barbedor

# Independent Component Analysis by Wavelets

**Pascal Barbedor**
*Laboratoire de probabilités et modéles aléatoires*
*CNRS-UMR 7599, Université Paris VI et Université Paris VII*
*PO box 7012, 75251 PARIS Cedex 05 (FRANCE)*
*barbedor@proba.jussieu.fr*

**Abstract**

We propose an ICA contrast based on density estimation of the observed signal and its marginal distributions through wavelets. The statistical risk of the wavelet contrast is linked with approximation properties in Besov spaces. Follows a discussion on computational issues; in particular, we resort to dyadic rational approximations to compute wavelet coefficients, instead of the usual histogram and filter scheme generally used in density estimation. The implemented wavelet contrast has linear complexity in $n$; numerical simulations give results as good as those of existing methods, if no better. The wavelet contrast also admits explicit differentials; using a simple jackknife, we give filter aware and computationally tractable formulations for the gradient and hessian of the contrast estimator.
**Keywords**: ICA, wavelets, Besov spaces, nonparametric density estimation

## 1. Introduction

In signal processing, blind source separation consists in the identification of analogical, independent signals mixed by a black-box device. In psychometry, one has the notion of structural latent variable whose mixed effects are only measurable through series of tests; an example are the Big Five (components of personality) identified from factorial analysis by researchers in the domain of personality evaluation (Roch, 1995). Other application fields such as digital imaging, biomedicine, finance and econometrics also use models aiming to recover hidden independent factors from observation. Independent component analysis (ICA) is one such tool; it can be seen as an extension of principal component analysis, in that it goes beyond a simple linear decorrelation only satisfactory for a normal distribution; or as a complement, since its application is precisely pointless with the assumption of normality.

Articles on ICA are found in the research fields of signal processing, artificial neural networks, statistics and information theory. Comon (1994) defined the concept of independent component analysis (ICA) as maximizing the degree of statistical independence among outputs using contrast functions approximated by the Edgeworth expansion of the Kullback-Leibler divergence.

The model is usually stated as follows: let $x$ be a random variable on $\mathbb{R}^d$, $d \geq 2$, one tries to find couples $(A, s)$, such that $x = As$, where $A$ is a square invertible matrix and $s$ a latent random variable whose components are mutually independent. This is usually done through some contrast function that cancels out if and only if the components of $Wx$ are independent, where $W$ is a candidate for the inversion of $A$.

Maximum-likelihood methods and contrast functions based on mutual information or other divergence measures between densities are commonly employed. Cardoso (1999) used higher-order cumulant tensors, which led to the Jade algorithm, Bell and Snejowski (1990s) published an approach based on the Infomax principle. Hyvarinen, Karhunen and Oja (1997) presented the fast ICA algorithm.

In the semi-parametric case, where the latent variable density is left unspecified, Bach and Jordan (2002) proposed a contrast function based on canonical correlations in a reproducing kernel hilbert space. Similarly, Gretton, Herbrich and Smola (2003) proposed kernel covariance and kernel mutual information contrast functions.

The density model assumes that the observed random variable $X$ has the density $f_A$ given by,

$$\begin{aligned} f_A(x) &= |\det A^{-1}| f(A^{-1}x) \\ &= |\det B| f^1(b_1 x) \dots f^d(b_d x), \end{aligned}$$

where $b_\ell$ is the $\ell^{th}$ row of the matrix $B = A^{-1}$; which results from a change of variable if the latent density $f$ is equal to the product of its marginal distributions $f^1 \dots f^d$. In this regard, latent variable $s = (s^1, \dots, s^d)$ having independent components means that the variables $s^\ell \circ \pi^\ell$ are independent random variables defined on some product probability space $\Omega = \prod \Omega^\ell$, with $\pi^\ell$ the canonical projections. So $s$ can be defined as the compound of the unrelated $s^1$,$\dots, s^d$ sources.

Tsybakov and Samarov (2002) proposed a method of simultaneous estimation of the directions $b_i$, based on nonparametric estimates of matrix functionals using in particular the gradient of $f_A$.

In this paper, we propose a wavelet based ICA contrast. Wavelets are orthonormal bases of $L_2$ with remarkable approximation and statistical properties. As the usual sine in Fourier analysis, wavelets split up a signal into frequency components, thus allowing a fine study at different scales, all within the framework of a so-called multiresolution analysis. An advantage of wavelets over the sine wave lies in the double localization property, both in frequency and time domain, which makes them well-suited to approximate data with sharp spikes.

The proposed contrast $C_j$ compares the mixed density $f_A$ and its marginal distributions through their projections on a multiresolution analysis at level $j$. It thus heavily relates on the procedures of wavelet density estimation which are found in a series of articles from Kerkyacharian and Picard (1992) and Donoho, Johnstone, Kerkyacharian and Picard (1996), among others. See also the book from Härdle, Kerkyacharian, Picard and Tsybakov (1998) for a self-contained presentation of wavelets linked with Sobolev and Besov approximation theorems and statistical applications.

As will be shown, the wavelet contrast has the property to be zero only on a projected density with independent components. The key parameter of the method lies in the choice of a resolution $j$, so that minimizing the contrast at that resolution gives a satisfactory approximate solution to the ICA problem.

Besov spaces are a general tool in describing smoothness properties of functions; they also constitute the natural choice when dealing with projections on a multiresolution analysis. We first show that a linear mixing operation is conservative as to Besov membership; after what we are in position to derive a statistical risk that will hold for the entire ICA minimization procedure.

Under its simplest form, the wavelet contrast is a linear function of the empirical measure on the observation. We give the rule for the choice of the resolution level $j$ that minimizes the risk, assuming a known regularity for the latent signal.

The rest of the article is about computation. In particular, we resort to the moment estimator of wavelet coefficients with approximation at dyadic rationals, instead of the widely used histogram followed by smoothing, typical of density estimation. We have implemented this way a fast wavelet contrast that meets the needs of the minimization procedure in ICA.

Using a plain steepest descent algorithm with empirical gradient, the efficacy of the wavelet contrast for demixing seems as good as existing algorithms, if not challenging.

We end by a differential calculus applied to the wavelet contrast, and give formulations based on simple jackknifes for the gradient and hessian estimators. We show that these formulations can be implemented in a tractable way and thus constitute a potential improvement in accuracy or convergence rate for the minimization stage.

## 2. Wavelet contrast, Besov membership

Let $(V_j)_{j\in\mathbb{Z}}$ be a multiresolution analysis of $L_2(\mathbb{R})$, with $V_j$ spanned by $\{\varphi_{jk} = 2^{j/2}\varphi(2^j . - k), k \in \mathbb{Z}\}$ and $W_j$ the complement of $V_j$ in $V_{j+1}$. A function expandable on $(\varphi, \psi)$ is written,

$$f(x) = \sum_k \alpha_{j_0 k}\varphi_{jk}(x) + \sum_{j=j_0}^{\infty}\sum_k \beta_{jk}\psi_{jk}(x).$$

Define $V_j^d$ as the tensorial product of $d$ copies of $V_j$. The increasing sequence $(V_j^d)_{j\in\mathbb{Z}}$ defines a multiresolution analysis of $L_2(\mathbb{R}^d)$ (Meyer, 1997):

for $(i^1 \dots, i^d) \in \{0,1\}^d$ and $(i^1 \dots, i^d) \neq (0 \dots, 0)$, define $\Psi(x)_{i^1\dots,i^d} = \prod_{\ell=1}^d \psi^{(i^\ell)}(x^\ell)$, with $\psi^{(0)} = \varphi$, $\psi^{(1)} = \psi$, so that $\psi$ appears at least once in the product $\Psi(x)$ (we now on omit $i^1 \dots, i^d$ in the notation for $\Psi$, and in (1), although it is present each time);

for $(i^1 \dots, i^d) = (0 \dots, 0)$, define $\Phi(x) = \prod_{\ell=1}^d \varphi(x^\ell)$;

for $j \in \mathbb{Z}$, $k \in \mathbb{Z}^d$, $x \in \mathbb{R}^d$, let $\Psi_{jk}(x) = 2^{\frac{jd}{2}}\Psi(2^j x - k)$ and $\Phi_{jk}(x) = 2^{\frac{jd}{2}}\Phi(2^j x - k)$;

define $W_j^d$ as the orthogonal complement of $V_j^d$ in $V_{j+1}^d$; it is an orthogonal sum of $2^d - 1$ spaces having the form $U_{1j} \dots \otimes U_{dj}$, where $U$ is a placeholder for $V$ or $W$; $V$ or $W$ are thus placed using up all permutations, but with $W$ represented at least once, so that a fraction of the overall innovation brought by the finer resolution $j+1$ is always present in the tensorial product.

A function expandable on the basis $(\Phi, \Psi)$ can be written,

$$f(x) = \sum_k \alpha_{j_0 k}\Phi_{j_0 k}(x) + \sum_{j=j_0}^{\infty}\sum_k \beta_{jk}\Psi(x), \tag{1}$$

with implicit multidimensional notation for $k$.

In what follows, we assume that $f$ is a density function with compact support, expandable on a compactly supported tensorial wavelet basis $\Phi, \Psi \in L^2(\mathbb{R}^d)$.

We define $f^{\star\ell}$ to be the marginal distribution in dimension $\ell$,

$$f^{\star\ell}: x^\ell \mapsto \int_{\mathbb{R}^{d-1}} f(x^1 \dots, x^d)\, dx^1 \dots dx^{\ell-1} dx^{\ell+1} \dots dx^d,$$

and assume it is expandable on the basis $\varphi, \psi \in L^2(\mathbb{R})$, used to build $\Phi, \Psi$.

The wavelet expansions up to order $j$, $P_j f$ and $P_j^\ell f^{\star\ell}$, that is to say the projections of $f$ and $f^{\star\ell}$ on $V_j^d$ and $V_j$ respectively can be written,

$$P_j f(x) = \sum_{k \in \mathbb{Z}^d} \alpha_{jk} \Phi_{jk}(x) \quad \text{and} \quad P_j^\ell f^{\star\ell}(x^\ell) = \sum_{k^\ell \in \mathbb{Z}} \alpha_{jk^\ell} \varphi_{jk^\ell}(x^\ell),$$

where $\alpha_{jk^\ell} = \int f^{\star\ell}(x^\ell) \varphi_{jk^\ell}(x^\ell)\, dx^\ell$ and $\alpha_{jk} = \alpha_{jk^1 \dots, k^d} = \int f(x) \Phi_{jk}(x)\, dx$.

**Proposition 2.1 (wavelet contrast)**

*Define the contrast function,*

$$C_j(f) = \sum_{k^1 \dots, k^d} (\alpha_{jk^1 \dots, k^d} - \alpha_{jk^1} \dots \alpha_{jk^d})^2,$$

*with* $\alpha_{jk^\ell} = \int_{\mathbb{R}} f^{\star\ell}(x^\ell) \varphi_{jk^\ell}(x^\ell)\, dx^\ell$ *and* $\alpha_{jk^1 \dots, k^d} = \int_{\mathbb{R}^d} f(x) \Phi_{jk^1, \dots, k^d}(x)\, dx$.

$C_j$ *is equal to the square of the* $L_2$ *norm of* $P_j f - P_j^1 f^{\star 1} \dots P_j^d f^{\star d}$ *and,*

$$f \text{ factorizable} \implies C_j(f) = 0,$$
$$C_j(f) = 0 \implies P_j f = \prod_{\ell=1}^d P_j^\ell f^{\star\ell}.$$

**Proof**

As $V_j^d = V_j \otimes \dots V_j$, one has $P_j f = (P_j^d \circ \dots P_j^1) f$, and the same for any other permutation of the order of projections.

As for the first assertion, with $f = f^1 \dots f^d$, one has $f^{\star\ell} = f^\ell$, $\ell = 1, \dots d$.

By linearity of projection operators, and seeing that the product $f^1 \dots f^{\ell-1}\ f^{\ell+1} \dots f^d$ is a multiplicative constant to $P_j^\ell$, one has $P_j f = P_j^1 f^1 \dots P_j^d f^d$.

Since the density $f$ and the wavelet $\varphi$ are compactly supported, there are finitely many k needed for complete expansion of each $P_j^\ell f^\ell$, and for all $x \in \mathbb{R}^d$ their product is written,

$$\left( \sum_{k^1} \alpha_{jk^1} \varphi_{jk^1}(x^1) \right) \dots \left( \sum_{k^d} \alpha_{jk^d} \varphi_{jk^d}(x^d) \right) = \sum_{k^1 \dots, k^d} \alpha_{jk^1} \dots \alpha_{jk^d}\, \varphi_{jk^1}(x^1) \dots \varphi_{jk^d}(x^d).$$

The finite set $K(f) = \{(k^1 \dots, k^d) \in \mathbb{Z}^d : \alpha_{jk^1 \dots, k^d} \neq 0\}$ needed to express $P_j f$ is always a subset of the one needed for $P_j^1 f^{\star 1} \dots P_j^d f^{\star d}$, the latter being the tensorial product of the canonical projections of $K(f)$; both indice sets are equal when $f = f^{\star 1} \dots f^{\star d}$.

Therefore one can write,

$$0 = \left(P_j f - P_j^1 f^1 \dots P_j^d f^d\right) = \sum_{k^1 \dots, k^d} \left(\alpha_{jk^1 \dots, k^d} - \alpha_{jk^1} \dots \alpha_{jk^d}\right) \varphi_{jk^1} \dots \varphi_{jk^d},$$

which implies $\alpha_{jk^1 \dots, k^d} = \alpha_{jk^1} \dots \alpha_{jk^d}$, $\forall (k^1 \dots, k^d) \in \mathbb{Z}^d$, since the $\varphi_{jk^1} \dots \varphi_{jk^d}$ form an orthonormal system; this proves the first assertion.

For the second assertion, $C_j = 0$ means exactly that $P_j f - P_j^1 f^{\star 1} \dots P_j^d f^{\star d}$ is equal to zero. □

For the zero contrast to give any clue as to whether the non projected difference $f - f^{\star 1} \dots f^{\star d}$ is itself close to zero, a key parameter is the order of projection $j$. Under the notations of the preceding proposition, when $C_j(f) = 0$, for any $L_p$ norm one has $\|P_j f - P_j^1 f^{\star 1} \dots P_j^d f^{\star d}\|_p = 0$, and so,

$$\begin{aligned} \| f - f^{\star 1} \dots f^{\star d} \|_p &\leq \|f - P_j f\|_p + \|P_j^1 f^{\star 1} \dots P_j^d f^{\star d} - f^{\star 1} \dots f^{\star d}\|_p \\ &= \|f - P_j f\|_p + \|P_j(f^{\star 1} \dots f^{\star d}) - f^{\star 1} \dots f^{\star d}\|_p. \end{aligned}$$

At least for $p = 2$, when $j \to +\infty$ and $f$, $f^{\star \ell}$ in $L_2$, the two terms above are as small as desired since $\overline{\bigcup_j V_j^d} = L_2(\mathbb{R}^d)$.

For other values of $p$, we suppose that $f$ belongs to the (inhomogeneous) Besov space $B_{spq}(\mathbb{R}^d)$, *i.e.*

$$J_{spq}(f) = \|\alpha_{0.}\|_{\ell_p} + \left[\sum_{j \geq 0} \left[2^{js} 2^{dj(\frac{1}{2} - \frac{1}{p})} \|\beta_{j.}\|_{\ell_p}\right]^q\right]^{\frac{1}{q}} < \infty,$$

with $s > 0$, $1 \leq p \leq \infty$, $1 \leq q \leq \infty$, and $\varphi, \psi$ $C^r, r > s$, and $d$ the dimension (Meyer, 1997).

With a $r$-regular wavelet $\varphi$, $r > s$, the very definition of Besov spaces implies for $f$ that (Meyer, 1997),

$$\|f - P_j f\|_p = 2^{-js}\, \epsilon_j, \quad \{\epsilon_j\} \in \ell_q(\mathbb{N}^d)$$

Assuming that the product $f^{\star 1} \dots f^{\star d}$ of marginal distributions also belong to another Besov space with same $p$ parameter, and using the Sobolev injection,

$$B_{s'pq'} \subset B_{spq} \text{ for } s' \geq s,\ q' \leq q,$$

we are through since both $f$ and $f^{\star 1} \dots f^{\star d}$ will belong to the enclosing Besov space $B_{s''pq''}$ with $s'' = \min(s, s')$ and $q'' = \max(q, q')$.

We are then left with the study of the Besov membership of a density $f$ compared with the membership of the product of its marginal distributions. In an ICA context, the function $f$ has the additional property to be of the type,

$$f_A(x) = |\det A^{-1}|\, f(A^{-1}x) = |\det B|\, f^1(b_1 x) \dots f^d(b_d x),$$

with $B = A^{-1}$, $b_\ell$ line $\ell$ of $B$, and $f = f^1 \dots f^d$.

**Proposition 2.2 (Besov membership of marginal distributions)**

*Let $f$ be a density function belonging to $B_{spq}(\mathbb{R}^d)$.*

*Each of the marginal distributions of $f$ belongs to $B_{spq}(\mathbb{R})$.*

**proof**

Let us first check the $L_p$ membership of the marginal distribution. For $p \geq 1$, by convexity one has,

$$\int_{\mathbb{R}^d} |f_A|^p \, dx = \int_{\mathbb{R}} \int_{\mathbb{R}^{d-1}} |f_A|^p \, dx^{\star\ell} dx^\ell \geq \int_{\mathbb{R}} \left| \int_{\mathbb{R}^{d-1}} f_A \, dx^{\star\ell} \right|^p dx^\ell = \int_{\mathbb{R}} |f_A^{\star l}|^p \, dx^\ell;$$

that is to say $\|f_A^{\star l}\|_p \leq \|f_A\|_p$.

For the Besov membership, we resort to a norm on $B_{spq}(\mathbb{R}^d)$ defined through the modulus of smoothness (Devore, Lorentz, 1993),

$$J''_{spq}(f) = \|f\|_p + \|\omega_r(f,.)\|_{s,q},$$

where, for $1 \leq p, q \leq \infty$ and $r = [s] + 1$,

$$\begin{aligned} \|\omega_r(f,.)\|_{s,q} &= \left( \int_{\mathbb{R}^d} \left[ \frac{\omega_r(f,t)_p}{|h|^s} \right]^q \frac{dh}{|h|} \right)^{1/q} \qquad 0 < q < \infty \\ &= \sup_{t \in \mathbb{R}^d > 0} |t|^{-s} \omega_r(f,t)_p \qquad\qquad q = \infty; \end{aligned}$$

and where the $r^{th}$ modulus of smoothness is defined by $\omega_r(f,t)_p = \sup_{0<h\leq t} \|\Delta_h^r(f,.)\|_p$, with

$$\Delta_h^r(f,x) = \sum_{k=0}^{r} C_r^k (-1)^{r-k} f(x + kh),$$

with implied multidimensional notation when $f \in L_p(\mathbb{R}^d)$.

With $f^{\star\ell}$ the marginal distribution $\ell$ of $f$, $e^\ell$ designating a vector of the canonical basis of $\mathbb{R}^d$, $h \in \mathbb{R}$ and $x = (x^1, \ldots, x^d)$,

$$\tau_h f^{\star\ell} = f^{\star\ell}(x^\ell + h) = \int_{\mathbb{R}^{d-1}} f(x + he^\ell) \, dx^{\star\ell};$$

so that by convexity,

$$\begin{aligned} \|\tau_h f^{\star\ell} - f^{\star\ell}\|_p^p &= \int \left| \int_{\mathbb{R}^{d-1}} f(x + he^\ell) - f(x) dx^{\star\ell} \right|^p dx^\ell \\ &\leq \int_{\mathbb{R}^d} |f(x + he^\ell) - f(x)|^p \, dx \\ &= \|\tau_{he^\ell} f - f\|_p^p. \end{aligned}$$

Since $\Delta_h^r(f^{\star\ell}) = \sum_{k=0}^r C_r^k (-1)^{r-k} \int_{\mathbb{R}^{d-1}} f(x + khe^\ell)\, dx^{\star\ell}$, one has in the same way,

$$\|\Delta_h^r(f^{\star\ell}, .)\|_{Lp(\mathbb{R})} \leq \|\Delta_{he^\ell}^r(f, .)\|_{Lp(\mathbb{R}^d)}.$$

So that for $h = (h^1, \ldots, h^d)$ and $t = (t^1, \ldots, t^d)$,

$$\begin{aligned}\omega_r(f^{\star\ell}, t^\ell)_p &= \sup_{0<h^\ell\leq t^\ell} \|\Delta_{h^\ell}^r(f^{\star\ell}, .)\|_{Lp(\mathbb{R})} \leq \sup_{0<h^\ell\leq t^\ell} \|\Delta_{h^\ell e^\ell}^r(f, .)\|_{Lp(\mathbb{R}^d)} \\ &= \omega_r(f, te^\ell)_p \leq \sup_{0<h\leq t} \|\Delta_h^r(f, .)\|_{Lp(\mathbb{R}^d)} = \omega_r(f, t)_p.\end{aligned}$$

Next, since $|t^\ell| = |te^\ell|$, one can write,

$$\|\omega_r(f^{\star\ell}, .)\|_{s,q} \leq \left[\int \left[\frac{\omega_r(f, t^\ell)_p}{|te^\ell|^s}\right]^q \frac{dt^\ell}{|te^\ell|}\right]^{\frac{1}{q}};$$

the right member is finite since it is the restricted semi-norm of $f$ in the direction $e^\ell$, which cannot be greater than the unrestricted one.
□

Next, we check that the mixed density $f_A$ belongs to the same Besov space than the original density $f$.

**Proposition 2.3 (Besov membership of the mixed density)**

*Let $f = f^1 \ldots f^d$ and $f_A(x) = |\det A^{-1}| f(A^{-1}x)$.*

*(a) Assume that $f$ belongs to $L_p(\mathbb{R}^d)$, or that each $f^\ell$ belongs to $L_p(\mathbb{R})$, then $f_A$ and the product of its marginal distributions $\prod f_A^{\star\ell}$ also belong to $L_p(\mathbb{R}^d)$.*

*(b) $f$ and $f_A$ have same $\gamma_{spq}$ semi-norm up to a constant; that is to say, together with (a), they belong to the same $B_{spq}$ inhomogeneous Besov space.*

**proof**

Another norm equivalent to $J_{spq}$ using the modulus of continuity is defined by (Bergh and Löstrom , 1976),

$$J'_{spq}(f) = \|f\|_p + \sum_{|m|\leq[s]} \gamma_{\alpha pq}(D^m f),$$

where $[s]$ is the integer part of $s$, $\alpha = s - [s]$ is the fractional part, $m$ is in multi-index notation and,

$$\gamma_{\alpha pq}(f) = \left[\int_{\mathbb{R}^d} \left[\frac{\|f(x-h) - f(x)\|_p}{|h|^\alpha}\right]^q \frac{dh}{|h|}\right]^{\frac{1}{q}} \quad \gamma_{\alpha p\infty}(f) = \sup_{h\in\mathbb{R}^{d\star}} \frac{\|f(x-h) - f(x)\|_p}{|h|^\alpha};$$

For (a), with $p \geq 1$, as in Prop. 2.2 above, one has $\|f_A^{\star\ell}\|_p \leq \|f_A\|_p$. Also,

$$\|f_A\|_p = |A|^{-p} \int |f(A^{-1}x)|^p\, dx = |A|^{-p} \int |f(x)|^p\, |A|\, dx = |A|^{1-p}\, \|f\|_p.$$

And finally by Fubini theorem, $\|f\|_{L_p(\mathbb{R}^d)} = \|f^1\|_{L_p(\mathbb{R})} \dots \|f^d\|_{L_p(\mathbb{R})}$, so that $f \in L_p(\mathbb{R}^d) \iff f^\ell \in L_p(\mathbb{R}), \ell = 1 \dots d$ .

For (b), we use the $\gamma_{\alpha pq}$ semi-norm.

With $\tau_h f = f(. - h)$ and a change of variable in the integral one has,

$$\|\tau_h f_A - f_A\|_p = |A|^{-1+\frac{1}{p}} \|\tau_{A^{-1}h} f - f\|_p ;$$

next, using the fact that $|A^{-1}h| \leq \|A^{-1}\| \, |h|$, *i.e.* $|h|^{-\alpha} \leq |A^{-1}h|^{-\alpha} \|A^{-1}\|^\alpha$, one has,

$$\begin{aligned} \frac{\|\tau_h f_A - f_A\|_p}{|h|^\alpha} &= \frac{|A|^{-1+\frac{1}{p}} \|\tau_{A^{-1}h} f - f\|_p}{|h|^\alpha} \\ &\leq |A|^{-1+\frac{1}{p}} \|A^{-1}\|^\alpha \frac{\|\tau_{A^{-1}h} f - f\|_p}{|A^{-1}h|^\alpha}. \end{aligned}$$

So that,

$$\begin{aligned} \gamma_{\alpha pq}(f_A) &\leq |A|^{-1+\frac{1}{p}} \|A^{-1}\|^{\alpha+\frac{1}{q}} \left[ \int \left[ \frac{\|\tau_{A^{-1}h} f - f\|_p}{|A^{-1}h|^\alpha} \right]^q \frac{dh}{|A^{-1}h|} \right]^{\frac{1}{q}} \\ &= |A|^{\frac{1}{p}} \|A^{-1}\|^{\alpha+\frac{1}{q}} \left[ \int \left[ \frac{\|\tau_\ell f - f\|_p}{|\ell|^\alpha} \right]^q \frac{d\ell}{|\ell|} \right]^{\frac{1}{q}} \\ &= |A|^{\frac{1}{p}} \|A^{-1}\|^{\alpha+\frac{1}{q}} \gamma_{\alpha pq}(f) \\ &= \lambda_1^{\frac{1}{p}-\alpha-\frac{1}{q}} (\lambda_2 \dots \lambda_d)^{\frac{1}{p}} \gamma_{\alpha pq}(f), \end{aligned}$$

with $\lambda_1, \dots, \lambda_d$ the eigenvalues of $A$ in decreasing order.

About the $s$ parameter, since $df_A(h) = |A^{-1}| \, df(A^{-1}h) \circ A^{-1}$ one can see that the $\gamma_{\alpha pq}$ semi-norms of the partial derivatives of $f_A$ are bounded the same way by the ones of $f$. So that both functions have the same $\gamma_{spq}$ semi-norm, up to a constant. Finally, $f$ and $f_A$ belong to the same Besov space. □

## 3. Risk of the wavelet contrast estimator

Define the experiment $\mathcal{E}^n = (\mathcal{X}^{\otimes n}, \mathcal{A}^{\otimes n}, (X_1, \dots, X_n), P^n_{f_A}, f \in V)$, where $X_1, \dots, X_n$ is an iid sample of $X = AS$, and $P^n_{f_A} = P_{f_A} \dots \otimes P_{f_A}$ is the joint distribution of $(X_1 \dots, X_n)$. Likewise, define $P^n_f$ as the joint distribution of $(S_1 \dots, S_n)$.

The coordinates $\alpha_{jk}$ in the wavelet contrast are estimated as usual by,

$$\hat{\alpha}_{jk^1 \dots, k^d} = \frac{1}{n} \sum_{i=1}^n \varphi_{jk^1}(X_i^1) \dots \varphi_{jk^d}(X_i^d) \text{ and } \hat{\alpha}_{jk^\ell} = \frac{1}{n} \sum_{i=1}^n \varphi_{jk^\ell}(X_i^\ell), \ \ell = 1, \dots d.$$

And the linear wavelet contrast estimator is given by,

$$\hat{C}_j(x_1 \dots, x_n) = \sum_{k^1 \dots, k^d} (\hat{\alpha}_{jk^1 \dots, k^d} - \hat{\alpha}_{jk^1} \dots \hat{\alpha}_{jk^d})^2 = \sum_{k \in \mathbb{Z}^d} \hat{\delta}^2_{jk},$$

where we define $\hat{\delta}_{jk}$ as the difference $\hat{\alpha}_{jk^1\ldots,k^d} - \hat{\alpha}_{jk^1} \ldots \hat{\alpha}_{jk^d}$.

The sum in $k$ is finite for a compactly supported wavelet and a compactly supported density. Moreover we take the convention that the whole signal is contained in a cube $[0,1]^{\otimes d}$ after possible rescaling. For the compactly supported Daubechies wavelets (Daubechies, 1992), $D2N, N = 1, 2, \ldots$, whose support is $[0, 2N-1]$, the maximum number of $k$ intersecting with an observation lying in the cube is $(2^j + 2N - 2)^d$.

Using a quadratic loss function, the risk associated to $\hat{C}_j$ in estimating $C_j$, $R(\hat{C}_j, f_A)$, is given by $R(\hat{C}_j, f_A) = E^n_{f_A}(\hat{C}_j - C_j)^2 = b^2 + \sigma^2$ with $b = E^n_{f_A}\hat{C}_j - C_j$ and $\sigma^2 = E^n_{f_A}(\hat{C}_j - E^n_{f_A}\hat{C}_j)^2$.

We are in fact specially interested by the risk relative to zero, $E^n_{f_{WA}}\hat{C}_j^2$, since we estimate a contrast function $C_j$ that should be zero after we applied to $X = AS$ a transformation $W$, found by some procedure, and supposed to reverse the mixing by $A$.

In this part we come across terms of the form,

$$V_n = \frac{1}{n^m} \sum_{i_1,\ldots,i_m} g^1_{jk}(X_{i_1}) \ldots g^m_{jk}(X_{i_m}),$$

with $(X_1 \ldots, X_n)$ the independent identically distributed sample. This is the Von Mises statistic associated with the kernel function $h(x^1, \ldots, x^m) = g^1_{jk}(x^1) \ldots g^m_{jk}(x^m)$, symmetrized if necessary by $h_s = \frac{1}{m!}\sum_p h(x^1, \ldots, x^m)$, where $\sum_p$ denotes summation over the $m!$ permutation of $(1, \ldots, m)$.

The associated unbiased $U$-statistic,

$$U_n = [C^m_n]^{-1} \sum_{i_1 < \ldots < i_m} g^1_{jk}(X_{i_1}) \ldots g^m_{jk}(X_{i_m})$$

has expectation $Eg^1_{jk}(X) \ldots Eg^m_{jk}(X)$.

We then use the following lemma which connects a $V$-statistic with its associated $U$-statistic (Serfling, 1980),

**Lemma 3.1** **($U$-$V$ statistic connection )**

*if $r$ is a positive integer, and $E|h(X_{i_1}, \ldots, X_{i_m})|^r < \infty$ for all $i_1, \ldots i_m \in \{1, \ldots n\}$, then*

$$E|U_n - V_n|^r = O(n^{-r}).$$

For our purpose, the kernel function will depend on $j$, that we need to pull out of the $O(n^{-r})$ term; from the proof of Serfling's lemma we see that,

$$\begin{aligned}
E^n_{f_A}|U_n - V_n|^r &= n^{-mr}\, E^n_{f_A}|n^m(U_n - V_n)|^r \\
&= n^{-mr}\, E^n_{f_A}|(n^m - C^m_n)(U_n - W_n)|^r \\
&= n^{-mr}(n^m - C^m_n)^r\, E^n_{f_A}|U_n - W_n|^r \\
&\leq n^{-mr}(n^m - C^m_n)^r \left(E^n_{f_A}|U_n|^r + E^n_{f_A}|W_n|^r\right) \\
&\leq O(n^{-r}) \left(E^n_{f_A}|U_n|^r + E^n_{f_A}|W_n|^r\right),
\end{aligned}$$

with $W_n$ the average of all terms $h(X_{i_1},\dots,X_{i_m})$ with at least one equality $i_a = i_b$, $a \neq b$, and using the fact that $n^m(U_n - V_n) = (n^m - C_n^m)(U_n - W_n)$, and the other fact that $(n^m - C_n^m) = O(n^{m-1})$ is positive.

The $g_{jk}$ that we are going to deal with, will be either $\Phi$ either $\varphi \circ \pi^\ell$, where $\pi^\ell$ designates the canonical projection on component $\ell$; assuming there are $m_d$ $\Phi$ and $m_1$ $\varphi \circ \pi^\ell$, with $m_d + m_1 = m$,

$$E^n_{f_A}|U_n| \leq [C_n^m]^{-1} \sum_{i_1<\dots<i_m} \left(2^{\frac{j}{2}}\|\varphi\|_\infty\right)^{dm_d+m_1} = \left(2^{\frac{j}{2}}\|\varphi\|_\infty\right)^{dm_d+m_1},$$

and by the same means, $E^n_{f_A}|W_n| \leq \left(2^{\frac{j}{2}}\|\varphi\|_\infty\right)^{dm_d+m_1}$.

So that we can finally write,

$$E\frac{1}{n^m}\sum_{i_1,\dots,i_m} g^1_{jk}(X_{i_1})\dots g^m_{jk}(X_{i_m}) = Eg^1_{jk}(X)\dots Eg^m_{jk}(X) + 2^{\frac{j}{2}(dm_d+m_1)}O(n^{-1}), \tag{2}$$

with $m_d + m_1 = m$, $0 \leq m_d \leq m$, $0 \leq m_1 \leq m$.

**Proposition 3.4 (Risk of $\hat{C}_j$)**

*a) A convergence rate for the bias of $\hat{C}_j$ relative to zero is given by $E^n_{f_A}\hat{C}_j = C_j + 2^{2jd}O(1/n)$; incidentally, the plain bias is then $2^{2jd}O(1/n)$.*

*b) A convergence rate for the variance of $\hat{C}_j$ is given by $E^n_{f_A}\left(\hat{C}_j - E^n_{f_A}\hat{C}_j\right)^2 = 2^{4jd}O(1/n)$.*

*c) The risk of $\hat{C}_j$ relative to zero is $E^n_{f_A}\hat{C}_j^2 = C_j^2 + 2^{4jd}O(1/n)$.*

**proof**

As for the bias, one has,

$$\begin{aligned}\hat{\delta}^2_{jk} &= (\hat{\alpha}_{jk} - \hat{\alpha}_{jk^1}\dots\hat{\alpha}_{jk^d})^2 \\ &= \hat{\alpha}^2_{jk} - 2\hat{\alpha}_{jk}\hat{\alpha}_{jk^1}\dots\hat{\alpha}_{jk^d} + (\hat{\alpha}_{jk^1}\dots\hat{\alpha}_{jk^d})^2.\end{aligned}$$

For the first term, apply (2) to $\frac{1}{n}\sum_{i_1}\Phi(X_{i_1})\frac{1}{n}\sum_{i_2}\Phi(X_{i_2})$, with $g_1 = g_2 = \Phi$, $m_d = m = 2$ and $m_1 = 0$. So,

$$\begin{aligned}E^n_{f_A}\hat{\alpha}^2_{jk} &= E^n_{f_A}\frac{1}{n^2}\sum_{i_1,i_2}\Phi(X_{i_1})\Phi(X_{i_2}) \\ &= E^n_{f_A}\Phi(X_1)E^n_{f_A}\Phi(X_1) + 2^{jd}O(\frac{1}{n}) \\ &= \alpha^2_{jk} + 2^{jd}O(\frac{1}{n}).\end{aligned}$$

For the central term, apply (2) to $\frac{1}{n}\sum_{i_0}\Phi(X_{i_0})\frac{1}{n}\sum_{i_1}\varphi(X^1_{i_1})\dots\frac{1}{n}\sum_{i_d}\varphi(X^d_{i_d})$, with $g_0 = \Phi$ and $g_\ell = \varphi\circ\pi^\ell$, $\ell = 1,\dots,d$, $m_d = 1$, $m_1 = d$, $m = d+1$.

Therefore,

$$\begin{aligned} E^n_{f_A}[\hat{\alpha}_{jk}\hat{\alpha}_{jk^1}\dots\hat{\alpha}_{jk^d}] &= E^n_{f_A}\left[\frac{1}{n}\sum_{i_0}\Phi(X_{i_0})\frac{1}{n}\sum_{i_1}\varphi(X^1_{i_1})\dots\frac{1}{n}\sum_{i_d}\varphi(X^d_{i_d})\right] \\ &= E^n_{f_A}\Phi(X_1)E^n_{f_A}\varphi(X^1_1)\dots E^n_{f_A}\varphi(X^d_1) + 2^{jd}O(\frac{1}{n}) \\ &= \alpha_{jk}\alpha_{jk^1}\dots\alpha_{jk^d} + 2^{jd}O(\frac{1}{n}). \end{aligned}$$

For the last term, apply (2) to,

$$\frac{1}{n}\sum_{i_1}\varphi(X^1_{i_1})\frac{1}{n}\sum_{i_{d+1}}\varphi(X^1_{i_{d+1}})\dots\frac{1}{n}\sum_{i_d}\varphi(X^d_{i_d})\frac{1}{n}\sum_{i_{2d}}\varphi(X^d_{i_{2d}})$$

with $g_\ell = g_{\ell+d} = \varphi\circ\pi^\ell$, $\ell = 1,\dots,d$; $m_d = 0$, $m_1 = m = 2d$.

Finally, $E^n_{f_A}\hat{\delta}^2_{jk} = \delta^2_{jk} + 2^{jd}O(\frac{1}{n})$. So for the bias relative to zero,

$$E^n_{f_A}\sum_k\hat{\delta}^2_{jk} = \sum_k\Big(\delta^2_{jk} + 2^{jd}O(\frac{1}{n})\Big) = C_j + 2^{2jd}O(1/n).$$

On the other hand for the variance term,

$$\begin{aligned} \sigma^2 &= E^n_{f_A}\left(\hat{C}_j - E^n_{f_A}\hat{C}_j\right)^2 \\ &= E^n_{f_A}\left(\sum_k\hat{\delta}^2_{jk} - E^n_{f_A}\sum_k\hat{\delta}^2_{jk}\right)^2 \\ &= E^n_{f_A}\left(\sum_k\left[\hat{\delta}^2_{jk} - E^n_{f_A}\hat{\delta}^2_{jk}\right]\right)^2 \\ &= \sum_{k,\ell}E^n_{f_A}\left(\hat{\delta}^2_{jk} - E^n_{f_A}\hat{\delta}^2_{jk}\right)\left(\hat{\delta}^2_{j\ell} - E^n_{f_A}\hat{\delta}^2_{j\ell}\right) \\ &= \sum_{k,\ell}E^n_{f_A}\hat{\delta}^2_{jk}\hat{\delta}^2_{j\ell} - E^n_{f_A}\hat{\delta}^2_{jk}E^n_{f_A}\hat{\delta}^2_{j\ell}, \end{aligned}$$

with $\hat{\delta}^2_{jk}\hat{\delta}^2_{j\ell}$ given by,

$$\begin{aligned} \hat{\delta}^2_{jk}\hat{\delta}^2_{j\ell} &= (\hat{\alpha}_{jk} - \hat{\alpha}_{jk^1}\dots\hat{\alpha}_{jk^d})^2(\hat{\alpha}_{j\ell} - \hat{\alpha}_{j\ell^1}\dots\hat{\alpha}_{j\ell^d})^2 \\ &= \hat{\alpha}^2_{jk}\hat{\alpha}^2_{j\ell} - 2\hat{\alpha}^2_{jk}\hat{\alpha}_{j\ell}\hat{\alpha}_{j\ell^1}\dots\hat{\alpha}_{j\ell^d} + \hat{\alpha}^2_{jk}\hat{\alpha}^2_{jl^1}\dots\hat{\alpha}^2_{jl^d} - 2\hat{\alpha}_{jk}\hat{\alpha}_{jk^1}\dots\hat{\alpha}_{jk^d}\hat{\alpha}^2_{j\ell} \\ &+ 4\hat{\alpha}_{jk}\hat{\alpha}_{jk^1}\dots\hat{\alpha}_{jk^d}\hat{\alpha}_{j\ell}\hat{\alpha}_{j\ell^1}\dots\hat{\alpha}_{j\ell^d} - 2\hat{\alpha}_{jk}\hat{\alpha}_{jk^1}\dots\hat{\alpha}_{jk^d}\hat{\alpha}^2_{jl^1}\dots\hat{\alpha}^2_{jl^d} \\ &+ \hat{\alpha}^2_{jk^1}\dots\hat{\alpha}^2_{jk^d}\hat{\alpha}^2_{j\ell} - 2\hat{\alpha}^2_{jk^1}\dots\hat{\alpha}^2_{jk^d}\hat{\alpha}_{j\ell}\hat{\alpha}_{j\ell^1}\dots\hat{\alpha}_{j\ell^d} + \hat{\alpha}^2_{jk^1}\dots\hat{\alpha}^2_{jk^d}\hat{\alpha}^2_{jl^1}\dots\hat{\alpha}^2_{jl^d}. \end{aligned}$$

The decoupling equation (2) applies to all the terms of $\hat{\delta}^2_{jk}\hat{\delta}^2_{j\ell}$ expansion. Moreover, each time the numbers $m_d$ and $m_1$ are such that $dm_d + m_1$ is equal to $4d$.

In sum, after re-factorization, one has $E^n_{f_A}\hat{\delta}^2_{jk}\hat{\delta}^2_{j\ell} = \delta^2_{jk}\delta^2_{j\ell} + 2^{2jd}O(\frac{1}{n})$; and also $E^n_{f_A}\hat{\delta}^2_{jk} = \delta^2_{jk} + 2^{jd}O(\frac{1}{n})$. So the variance of $\hat{C}_j$ is in $2^{4jd}O(1/n)$.

Finally the risk of $\hat{C}_j$ relative to zero is given by,

$$\begin{aligned} R(\hat{C}_j, f_A) &= E^n_{f_A}\left(\hat{C}_j - E^n_{f_A}\hat{C}_j\right)^2 + \left(E^n_{f_A}\hat{C}_j\right)^2 = E^n_{f_A}\hat{C}_j^2 \\ &= E^n_{f_A}\sum_{k,\ell}\hat{\delta}^2_{jk}\hat{\delta}^2_{j\ell} \\ &= \sum_{k,\ell}\left(\delta^2_{jk}\delta^2_{j\ell} + 2^{2jd}O(1/n)\right) \\ &= C_j^2 + 2^{4jd}O(1/n). \end{aligned} \tag{3}$$

□

We now give a rule for choosing the resolution $j$ that minimizes the risk of the estimator $\hat{C}_j$ relative to zero. This rule, obtained as usual by balancing bias and variance, depends on $s$, the unknown regularity of the density that is assumed to belong to some Besov space $B_{spq}$.

**Proposition 3.5 (resolution $j$ minimizing the risk)**

*Assume that $f$ belongs to $B_{spq}(\mathbb{R}^d)$ and $C_j$ is based on a $r$-regular wavelet $\varphi$, $r > s'$.*

*The risk of $\hat{C}_j$ relative to zero is minimized by choosing $j$ such that $2^j \approx n^{\frac{1}{4}\frac{1}{s'+d}}$, with $s' = s$ if $p > 2$ and $s' = s + d/2 - d/p$ if $1 \le p \le 2$. The corresponding convergence rate under independence is $n^{\frac{-s'}{s'+d}}$.*

*This choice ensures a risk of $\hat{C}_j$ relative to $C_j$ in $n^{\frac{-s'}{s'+d}}$.*

**proof**

It was seen that,

$$\begin{aligned} C_j(f_A)^{\frac{1}{2}} &= \|P_j(f_A^{\star 1}\dots f_A^{\star d}) - P_j f_A\|_2 \\ &\le \|f_A - P_j f_A\|_2 + \|f_A - f_A^{\star 1}\dots f_A^{\star d}\|_2 + \|P_j(f_A^{\star 1}\dots f_A^{\star d}) - f_A^{\star 1}\dots f_A^{\star d}\|_2; \end{aligned}$$

in the same way we have,

$$\|f_A - f_A^{\star 1}\dots f_A^{\star d}\|_2 \le \|f_A - P_j f_A\|_2 + \|P_j(f_A^{\star 1}\dots f_A^{\star d}) - f_A^{\star 1}\dots f_A^{\star d}\|_2 + C_j(f_A)^{\frac{1}{2}}.$$

Let $K_\star(A, f) = \|f_A - f_A^{\star 1}\dots f_A^{\star d}\|_2$; $K_\star(A, f)$ is constant relatively to $j$ and $n$ just like $C_j$ and from the two inequations above one has,

$$|K_\star(A, f) - C_j(f_A)^{\frac{1}{2}}| \le \|f_A - P_j f_A\|_2 + \|P_j(f_A^{\star 1}\dots f_A^{\star d}) - f_A^{\star 1}\dots f_A^{\star d}\|_2$$

By Prop. 2.2 and 2.3 we know that $f_A$ and the product of its marginal distributions belong to the same $B_{spq}$ than the original $f$.

If $1 \le p \le 2$, using the Sobolev embedding $B_{spq} \subset B_{s'p'q}$ for $p \le p'$ and $s' = s + d/p' - d/p$, one can see that $f_A$ belongs to $B_{s'2q}$ with $s' = s + d/2 - d/p$, and so by definition,

$$\|f_A - P_j f_A\|_2 \le \epsilon_j 2^{-j(s+d/2-d/p)},$$

with $\epsilon_j \in \ell_q$.

if $p > 2$, since we consider compactly supported densities, one can write,

$$\|f_A - P_j f_A\|_2 \leq \|f_A - P_j f_A\|_p \leq \epsilon_j 2^{-js}.$$

Finally with $s$ modified to $s + d/2 - d/p$ if $p < 2$, one has,

$$|K_\star(A, f) - C_j(f_A)^{\frac{1}{2}}| \leq K 2^{-js}; \tag{4}$$

with $K$ a constant.

When $A = I$, under independence, $K_\star(A, f)$ vanish and $C_j(f_A)$ is convergent to zero with $j$; when $A \neq I$, $K_\star(A, f)$ is strictly positive and $C_j(f_A)$ cannot reach zero.

Taking power 4 of (4) and using (3),

$$R(\hat{C}_j, f_A) + K^\star Q(C_j, K^\star) \leq K 2^{-4js} + 2^{4jd} K n^{-1},$$

with $K$ a placeholder for an unspecified constant, and $Q(a, b) = -4a^3 + 6a^2 b - 4ab^2 + b^3$.

When $A$ is far from $I$, the constant $K_\star$ is strictly positive and the risk relative to zero has no useful upper bound. Although the risk relative to $C_j$ is always in $2^{4jd} K n^{-1}$.

With $A$ getting closer to $I$, $K_\star$ is brought down to zero and, the risk is then minimized when, constants appart, we balance $2^{-4js}$ with $2^{4jd} n^{-1}$, or $2^{4j(d+s)}$ with $n$.

This gives, $2^j = O(n^{\frac{1}{4}\frac{1}{s+d}})$, and a convergence rate of $n^{\frac{-s}{s+d}}$ for the risk relative to zero under independence and for the risk relative to $C_j$.
□

## 4. Computation of the estimator $\hat{C}_j$

The estimator can be computed with any Daubechies wavelet, including Haar.

For a regular wavelet ($D2N, N > 1$), it is known how to compute the values $\varphi_{jk}(x)$ (and any derivative) at dyadic rationals, see for instance the book of Nguyen and Strang (1996); this is the approach we used in this paper.

Alternately, using the usual filtering scheme, one can compute the Haar projection at high $j$ and use a discrete wavelet transform (DWT) by a $D2N$ to synthetize the coefficients at a lower, more desirable resolution before computing the contrast. This avoids the need to precompute any value at dyadics, because the Haar projection is like a histogram, but adds the time of the DWT.

While this second approach is almost exclusively used in density estimation, in the ICA context it leads to either an inflation of computational resources, or a possibly inoperative contrast at minimization stage. Indeed, for the Haar contrast to show any elasticity under a small perturbation, $j$ must be very high regardless of what would be required by the

signal regularity and the number of observations; whereas for a D4 and above, we just need to set high the precision of dyadic rational approximation, which present no inconvenience and can be viewed as a memory friendly refined binning inside the binning in $j$.

We now review some useful points for practical computation of the contrast estimator.

**Direct evaluation of $\varphi_{jk}(x)$ at floating point numbers**

Consider $x \in \mathbb{R}$, define $x_L = 2^{-L}\lfloor 2^L x \rfloor$ as the closest dyadic at approximation level $L$, where $\lfloor . \rfloor$ is the integer part or floor rounding.

To compute $\varphi_{jk}(x)$, one can evaluate $\varphi((2^j x - k)_L)$ or else $\varphi(2^j x_L - k)$:

$$\begin{aligned}(2^j x - k)_L &= 2^{-L}\lfloor 2^L 2^j x - 2^L k \rfloor \\ &= 2^{-L}(\lfloor 2^L 2^j x \rfloor - 2^L k) \\ &= 2^{-L}\lfloor 2^{L+j} x \rfloor - k \\ &= 2^j x - 2^{-L} F(2^{L+j} x) - k,\end{aligned}$$

where $F(x)$ is the fractional part of $x$; in this case the error in $x$ is less than $2^{-L}$ in absolute value. This is the approximation method we used. For the case $\varphi(2^j x_L - k)$, the error in $x$ is less than $2^{j-L}$ in absolute value and boils down to raising $L$.

When computing $\varphi((2^j x - k)_L) = \varphi(2^{-L}(\lfloor 2^L 2^j x \rfloor - 2^L k))$, the evaluation can only take $2^L(2N-1)$ different values; this is the needed size of an array designed to hold the precomputed values.

Suppose `tab` is such an array sized $2^L(2N-1)$, and containing the values of the $D2N$ at dyadics $\{k + i2^{-L}, i = 0, \ldots, 2^L - 1,\ k = 0, \ldots 2N-2\}$ with $\varphi(0) = \varphi(2N-1) = 0$, except for Haar where $\varphi(0) = 1$.

Given an observation $x$ there is exactly $2N-1$ functions $\varphi_{jk}$ whose support contains $x$, namely $\varphi_{je_j} \ldots \varphi_{je_j - 2N+2}$, with $e_j = \lfloor 2^j x \rfloor$ the integer part of $2^j x$ (if $e_j - 1, ...e_j - 2N + 2$ goes out of bound for the array containing the projection coordinates $\alpha_{jk}$, we use circular shifting).

So one needs to compute for each $x$, $2^{\frac{j}{2}}\varphi(f_j), \ldots 2^{\frac{j}{2}}\varphi(f_j + 2N - 2)$, with $f_j$ the fractional part of $2^j x$. If the index of $\varphi(f_j)$ in `tab` is $i = \lfloor 2^L f_j \rfloor$, we can safely retrieve `tab`$[i], \ldots,$ `tab`$[i + (2N-2)2^L]$ provided the shift stays below `tab`'s upper bound (*i.e.* within the support of the $D2N$), because $\lfloor 2^L(f_j + k) \rfloor = \lfloor 2^L f_j \rfloor + k2^L$.

Note that precomputed values at octave $L+1$ are those computed at octave $L$, interleaved with new values.

Note also that when $j + L$ has passed the machine floating point precision, (24 in single precision, for IEEE 754 on 32 bit), all machine numbers smaller than 1 in absolute value are covered and the evaluation $\varphi((2^j x - k)_L)$ may give no new value.

In effect, with $L + j \geq p$, the machine precision, $\lfloor 2^{L+j+b} x \rfloor = 2^b \lfloor 2^{L+j} x \rfloor \, \forall x$, and so if the added $b$ was added precision in $L$ we have, with $k \in \{\lfloor 2^j x \rfloor - k', \quad k' = 0 \ldots 2N-2\}$

$$\begin{aligned}\lfloor 2^{j+L+b} x - 2^{L+b} k \rfloor &= 2^b \lfloor 2^{j+L} x \rfloor - 2^{L+b} k \\ &= 2^b \big( \lfloor 2^{j+L} x \rfloor - 2^L \lfloor 2^j x - k' \rfloor \big), \quad k' = 0 \ldots 2N-2\end{aligned}$$

and the index will point to the exact same values in the $2^b$ times larger table of precomputed $\varphi$ values;

or if $b$ was added resolution in $j$, we have, with $k \in \{\lfloor 2^{j+b}x \rfloor - k', \quad k' = 0 \dots 2N-2\}$

$$\begin{aligned}\lfloor 2^{j+L+b}x - 2^L k \rfloor &= \lfloor 2^{j+L+b}x \rfloor - 2^L k \\ &= 2^b \lfloor 2^{j+L}x \rfloor - 2^L \lfloor 2^{j+b}x - k' \rfloor \quad k' = 0 \dots 2N-2\end{aligned}$$

with no new value if $b \geq L$.

**Relocation by affine or linear transform**

Relocate the observation so that it fits in a $d$-cube of volume 1; this does not change the ICA problem.

Let $Y = WX$ be a particular mixing of the observation at hand; with $b = \min_{i,j} y_i^j$ and $a = \max_{i,j} y_i^j - b$, $Y_a = \frac{1}{a}(Y - b)$ is entirely contained in the $d$-cube placed at zero; $Y$ components are independent if and only if $Y_a$ components are independent. So $\underset{W}{\operatorname{argmin}} C(Y) = \underset{W}{\operatorname{argmin}} C(Y_a)$.

$Y_a$ is not anymore whitened nor centered if $Y$ was; if we need to keep a centered or whitened observation we use the linear scaling,

$$Y_c = \frac{1}{4} \frac{\bar{Y}}{\max\limits_{i=1,\dots n} \|\bar{y}_i\|},$$

with $\bar{Y}$ the centered (possibly whitened) version of $Y$ ; $Y_c$ is centered and included in a cube of volume 1, namely $[-\frac{1}{2}, \frac{1}{2}]^{\otimes d}$.

Next, if we restrict to orthogonal $W$ and $Y$ is whitened, the element $(i, \ell)$ of the transform by $W$ satisfies,

$$|(WY_c)_i^\ell|^2 = | < {}^tW^\ell, (Y_c)_i > |^2 \leq \|(Y_c)_i\| \leq \frac{1}{4},$$

with $W^\ell$ the (normed) line vector $\ell$ of $W$, and $(Y_c)_i$ the column vector $i$ of $Y_c$.

This way an initial relocation covers all the minimization process.

**Whitening step**

Let $X$ be a $d \times n$ matrix with empirical dispersion $S_n = \frac{1}{n-1} X(I_n - \frac{1}{n} 1_n \, {}^t1_n) \, {}^tX$, with $\frac{1}{n} 1_n \, {}^t1_n$ an optional term needed for X centering; set,

$$\tilde{X} = (n-1)^{\frac{-1}{2}} D^{\frac{-1}{2}} \, {}^tPX(I_n - \frac{1}{n} 1_n \, {}^t1_n),$$

with $D$ the diagonal matrix of the eigenvalues of $S_n$, and $P$ the matrix $(v_1 \quad \dots \quad v_p)$ of associated eigenvectors, *i.e.* ${}^tPS_nP = D$. $\tilde{X}$ is the whitened version of $X$, after principal component transformation.

Note that ${}^t\tilde{X}$ is an element of $S_{n,d}$ the Stiefel manifold of $n \times d$ matrices $M$ such that ${}^tMM = I_d$, which is equivalent to say that the empirical dispersion of $\tilde{X}$ is equal to $\frac{1}{n-1} I_d$.

With $X = AS$, $X\,{}^tX = I_d$ and $S\,{}^tS = I_d$ implies that $A$ is orthogonal; or also $X\,{}^tX = I_d$ and $A$ orthogonal implies $S\,{}^tS = I_d$. Indeed, $X\,{}^tX = I_d$ implies $I_d = AS\,{}^tX$, *i.e.* $S\,{}^tX$ is a right inverse of $A$; it is also a left-inverse since $A$ is supposed invertible, so we must have $S\,{}^tXA = S\,{}^t(AS)A = S\,{}^tS\,{}^tAA = I_d$; verified in particular if A is orthogonal and if ${}^tS \in S_{n,d}$.

When $X$ is centered, $\tilde{X}$ is a left-side transformation of $X$, so we have $\tilde{X} = NAS = \tilde{A}S$, with $\tilde{A}$ orthogonal, seeing that $\frac{1}{n-1}I_d = \mathrm{var}(\tilde{X}) = \tilde{A}\mathrm{var}(S)\,{}^t\tilde{A} = \frac{1}{n-1}\tilde{A}\,{}^t\tilde{A}$.

Since $A^{-1} = \tilde{A}^{-1}N = {}^t\tilde{A}N$, the inverse of $A$ can be written as the product of an orthogonal matrix ${}^t\tilde{A}$ and a matrix already known $N$; so that it remains only to estimate the couple $(\tilde{A}, s)$, with the constraint $\tilde{A}$ orthogonal.

If $X$ is not centered, $\tilde{X}$ is not a left side transformation of $X$, so one cannot recover $A$ from $\tilde{A}$ ($1_n\,{}^t1_n$ does not in general commute with $A$). But centering $X$ shifts to a new ICA problem with A unchanged. $EX = AES \Rightarrow X - EX = AS - AES = A(S - ES)$.

**First simulations**

In this section, we compare independent and mixed D2 to D8 contrasts on a uniform whitened signal, in dimension 2 with 100000 observations, and in dimension 4 with 50000 observations. According to proposition 3.5, for $s = +\infty$ the best choice is $j = 0$, to be interpreted as the smallest of technically working $j$, *i.e.* satisfying $2^j > 2N - 1$, to ensure that the wavelet support is contained in the observation support. For $j = 0$, there is only one cell in the cube and the contrast is unable to detect any mixing effect: for Haar it is identically zero, and for the others D2N, it is a constant (quasi for round-off errors) because we use circular shifting if the wavelet support goes out of the observation support. At small $j$ such that $2 \leq 2^j \leq 2N-1$, D2N wavelets behave more or less like the Haar wavelet, except they are more responsive to a small perturbation. We use the Amari distance as defined in Amari (1996) rescaled from 0 to 100.

In this example, we have deliberately chosen an orthogonal matrix producing a small Amari error (less than 1 on a scale from 0 to 100), pushing the contrast to its limits.

| j | D2 indep | D2 mixed | cpu |
|---|---|---|---|
| 0 | 0.000E+00 | 0.000E+00 | 0.12 |
| 1 | 0.184E-06 | 0.102E-10 | 0.06 |
| 2 | 0.872E-04 | 0.199E-04 | 0.06 |
| 3 | 0.585E-03 | 0.294E-03 | 0.06 |
| 4 | 0.245E-02 | 0.285E-02 | 0.06 |
| 5* | 0.926E-02 | 0.110E-01 | 0.07 |
| 6 | 0.395E-01 | 0.387E-01 | 0.07 |
| 7 | 0.162E+00 | 0.162E+00 | 0.07 |
| 8 | 0.651E+00 | 0.661E+00 | 0.08 |
| 9 | 0.262E+01 | 0.262E+01 | 0.12 |
| 10 | 0.105E+02 | 0.105E+02 | 0.23 |
| 11 | 0.419E+02 | 0.419E+02 | 0.69 |
| 12 | 0.168E+03 | 0.168E+03 | 2.48 |

| j | D4 indep | D4 mixed | cpu |
|---|---|---|---|
| 0 | 0.250E+00 | 0.250E+00 | 0.21 |
| 1* | 0.239E+00 | 0.522E+00 | 0.17 |
| 2 | 0.198E-04 | 0.209E-04 | 0.17 |
| 3 | 0.127E-03 | 0.159E-03 | 0.17 |
| 4 | 0.635E-03 | 0.714E-03 | 0.17 |
| 5 | 0.235E-02 | 0.282E-02 | 0.17 |
| 6 | 0.988E-02 | 0.105E-01 | 0.17 |
| 7 | 0.405E-01 | 0.419E-01 | 0.17 |
| 8 | 0.163E+00 | 0.165E+00 | 0.21 |
| 9 | 0.653E+00 | 0.653E+00 | 0.26 |
| 10 | 0.261E+01 | 0.262E+01 | 0.39 |
| 11 | 0.104E+02 | 0.105E+02 | 0.87 |
| 12 | 0.419E+02 | 0.420E+02 | 2.67 |

Table 1a. D2, D4 uniform, dim=2 , nobs=100000, L=10, half degree rotation (Amari error $\approx .8$)

| j | D6 indep | D6 mixed | cpu |
|---|---|---|---|
| 0 | 0.304E+00 | 0.304E+00 | 0.37 |
| 1 | 0.304E+00 | 0.305E+00 | 0.37 |
| 2* | 0.215E+00 | 0.666E+00 | 0.37 |
| 3 | 0.132E-03 | 0.188E-03 | 0.36 |
| 4 | 0.641E-03 | 0.717E-03 | 0.36 |
| 5 | 0.295E-02 | 0.335E-02 | 0.35 |
| 6 | 0.123E-01 | 0.126E-01 | 0.37 |
| 7 | 0.495E-01 | 0.518E-01 | 0.36 |
| 8 | 0.198E+00 | 0.200E+00 | 0.41 |
| 9 | 0.796E+00 | 0.791E+00 | 0.49 |
| 10 | 0.319E+01 | 0.319E+01 | 0.64 |
| 11 | 0.127E+02 | 0.128E+02 | 1.13 |
| 12 | 0.509E+02 | 0.511E+02 | 2.97 |

| j | D8 indep | D8 mixed | cpu |
|---|---|---|---|
| 0 | 0.966E+00 | 0.966E+00 | 0.65 |
| 1 | 0.966E+00 | 0.197E+01 | 0.64 |
| 2* | 0.914E+00 | 0.333E+01 | 0.65 |
| 3 | 0.446E-03 | 0.409E-03 | 0.64 |
| 4 | 0.220E-02 | 0.214E-02 | 0.64 |
| 5 | 0.932E-02 | 0.104E-01 | 0.63 |
| 6 | 0.388E-01 | 0.383E-01 | 0.63 |
| 7 | 0.157E+00 | 0.160E+00 | 0.64 |
| 8 | 0.628E+00 | 0.630E+00 | 0.71 |
| 9 | 0.253E+01 | 0.252E+01 | 0.84 |
| 10 | 0.101E+02 | 0.101E+02 | 1.03 |
| 11 | 0.405E+02 | 0.406E+02 | 1.53 |
| 12 | 0.162E+03 | 0.162E+03 | 3.37 |

Table 1b. D6, D8 uniform, dim=2 , nobs=100000, L=10, half degree rotation (Amari error $\approx .8$)

Firstly, the Haar contrast is out of touch; at low resolution, the mixing passes unnoticed because the observations stay in their original bins, and at high resolution, like for the other wavelets, things become impossible because the ratio $2^{jd}/n$ gets too big, and clearly wanders from the optimal rule of Prop. 3.5.

Had we chosen a mixing with bigger Amari error, say 10, the Haar contrast would have worked at many more resolutions (this can be checked using the program `icalette1`); still, the Haar contrast is less likely to reach small Amari errors in a minimization procedure.

For wavelets D4 and above, the contrast is able to capture the mixing effect especially at low resolution (best observed resolution marked) and up to $j = 8$. Also, the wavelet support technical constraint is apparent between D4 and D6 or D8.

Finally we observe that the difference in computing time between Haar and a D8 is not significative in small dimension; it gets important starting from dimension 4 (Table 2). Note that the relatively longer cpu time for $2^j < 2N - 1$ is caused by the need to compute a circular shift for practically all points instead of only at borders.

| j | D2 indep | D2 mixed | cpu |
|---|---|---|---|
| 0 | 0.000E+00 | 0.000E+00 | 0.08 |
| 1 | 0.100E-03 | 0.155E-06 | 0.05 |
| 2 | 0.411E-02 | 0.221E-02 | 0.05 |
| 3 | 0.831E-01 | 0.684E-01 | 0.05 |
| 4 | 0.132E+01 | 0.129E+01 | 0.08 |
| 5 | 0.210E+02 | 0.210E+02 | 0.29 |
| 6 | 0.336E+03 | 0.335E+03 | 3.62 |

| j | D4 indep | D4 mixed | cpu |
|---|---|---|---|
| 0 | 0.625E-01 | 0.625E-01 | 0.85 |
| 1 | 0.624E-01 | 0.304E+00 | 0.83 |
| 2 | 0.283E-03 | 0.331E-03 | 0.82 |
| 3 | 0.503E-02 | 0.453E-02 | 0.83 |
| 4 | 0.818E-01 | 0.824E-01 | 0.92 |
| 5 | 0.130E+01 | 0.133E+01 | 1.30 |
| 6 | 0.211E+02 | 0.211E+02 | 4.68 |

| j | D6 indep | D6 mixed | cpu |
|---|---|---|---|
| 0 | 0.926E-01 | 0.926E-01 | 6.03 |
| 1 | 0.927E-01 | 0.929E-01 | 6.01 |
| 2 | 0.884E-01 | 0.825E+00 | 6.01 |
| 3 | 0.725E-02 | 0.744E-02 | 6.07 |
| 4 | 0.122E+00 | 0.117E+00 | 6.40 |
| 5 | 0.193E+01 | 0.195E+01 | 7.51 |
| 6 | 0.311E+02 | 0.311E+02 | 11.0 |

| j | D8 indep | D8 mixed | cpu |
|---|---|---|---|
| 0 | 0.934E+00 | 0.934E+00 | 22.8 |
| 1 | 0.934E+00 | 0.364E+01 | 22.8 |
| 2 | 0.937E+00 | 0.111E+02 | 22.8 |
| 3 | 0.751E-01 | 0.751E-01 | 22.9 |
| 4 | 0.124E+01 | 0.117E+01 | 24.1 |
| 5 | 0.196E+02 | 0.196E+02 | 27.0 |
| 6 | 0.313E+03 | 0.313E+03 | 30.8 |

Table 2. uniform, dim=4 , nobs=50000, L=10, Amari error $\approx .5$

Computations use double precision, but single precision works just as well. There is no guard against inaccurate sums that occur about 10% of the time for D4 and above, because

it does not prevent a minimum contrast from detecting independence. Dyadic approximation parameter $L$ is set at octave 10, about three exact decimals, and shows enough. Cpu times, in seconds, correspond to the total of the projection time on $V_j^d$ and on the $d$ $V_j$, plus the wavelet contrast time; machine used for simulations is a G4 1,5Mhz, with 1Go ram; programs are written in fortran and compiled with IBM xlf (program `icalette1` to be found in Appendix).

**Contrast complexity**

By complexity we mean the length of do-loops.

The projection on the tensorial space $V_j^d$ and the $d$ margins for a Db(2N), based on $n$ observations, is in $O(n(2N-1)^d)$. This is $O(n)$ for a Haar wavelet (2N=2) which correspond to making a histogram. The projection is almost independent of $j$ except for memory allocation. Once the projection at level $j$ is known, the contrast is computed in $O(2^{jd})$.

On the other hand, the complexity to apply one discrete wavelet transform at level $j$ is in $O(2^{jd}(2N-1)^d)$; so we see that the filtering approach consisting in taking the Haar projection for a high $j_1$ (typically $2^{j_1 d} \approx \frac{n}{log\, n}$) and filter down to a lower $j_0$, as a shortcut to direct D2N moment approximation at level $j_0$, is definitely a shortcut, except that the Haar wavelet is intrinsically immune to small perturbations, which is a problem for empirical gradient evaluation or the detection of a small departure from independence.

### 4.1 Dimension 2 example

In dimension 2, we are exempted from any further complication brought by a gradient descent and minimization on the Stiefel manifold, and at the same time this particular case is already illustrative of higher dimensions, at least concerning the needed resolution $j$ function of the density regularity.

After whitening, the inverse of $A$ is an orthogonal matrix, whose membership can be restricted to $SO(2)$ because reflections are not relevant to ICA. So there is only one parameter $\theta$ to find to achieve reverse mixing. Since permutations of axes are also void operations in ICA, angles in the range 0 to $\pi/2$ are enough to find out the minimum $W_0$ which, right multiplied by $N$, will recover the ICA inverse of $A$. And $A$ can be set to the identity matrix, because what changes when $A$ is not the identity, but any invertible matrix, is completely contained in $N$.

Figures below show the wavelet contrast in $W$ and the amari distance $d(A, WN)$ (where $N$ is the matrix computed after whitening), in function of the angle of rotation of the matrix $W$ restricted to one period, $[\,0, \pi/2\,]$. The minimums are not necessarily at a zero angle, because whitening leaves the signal in a random rotated position (to reproduce the following results run the program `icalette2`).

We see that, provided the contrast curve has a minimum that coincides with Amari minimum, any line search algorithm will find the angle to reverse the mixing effect. The only remaining question at this stage is with what precision.

Note that D4 needs at least $j = 2$, D6 or D8 need at least $j = 3$, for the wavelets to fully deploy inside the observation support; otherwise we use circular shifting, so that the

computation is more or less equivalent to that of Haar. Haar contrast curves are usable if performing a line-search with no gradient computation.

For an exponential signal, for D4 the minimum is localized, though loosely, at $j = 4$ (Fig. 1) and below, at $j = 3$ and 2 (not shown), and very well localized at $j = 6$ (Fig. 2) and 7 to 10 (not shown). For Haar, good localization also starts at $j = 5$, but with an annoying local minimum (Fig. 3), disappearing at $j = 8$ and above (not shown).

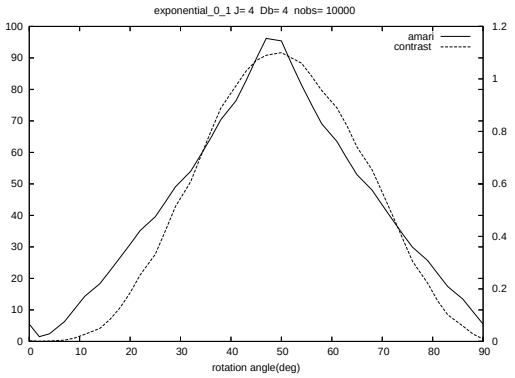


Fig.1. Exponential, D4, j=4, n=10000

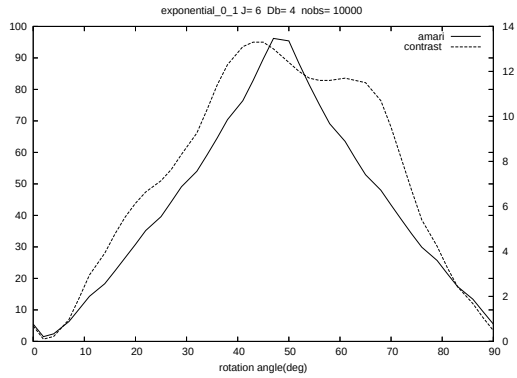


Fig.2. Exponential, D4, j=6, n=10000

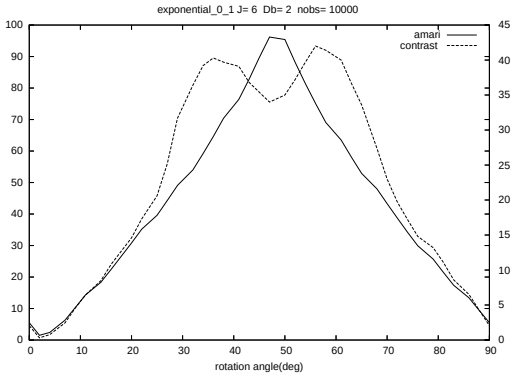


Fig.3. Exponential, D2, j=6, n=10000

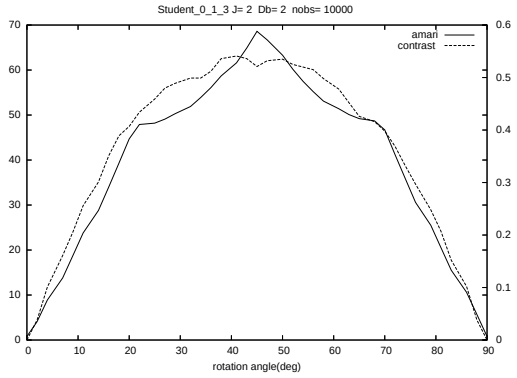


Fig.4. Student, D2, j=2, n=10000

For a Student, both Haar and D4 are very effective at $j = 2$ (Fig. 4, 5).

For a uniform, the minimum is accuratly localized from $j = 3$ to 6 (not shown), after that, the contrast variation is more chaotic, although the minimum still shows (Fig. 6).

For a semi-circular density, the minimum is accuratly localized from $j = 3$ (Fig. 7) to 5

(not shown).

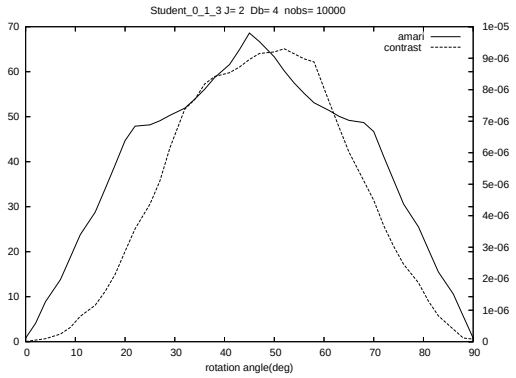


Fig.5. Student, D4, j=2, n=10000

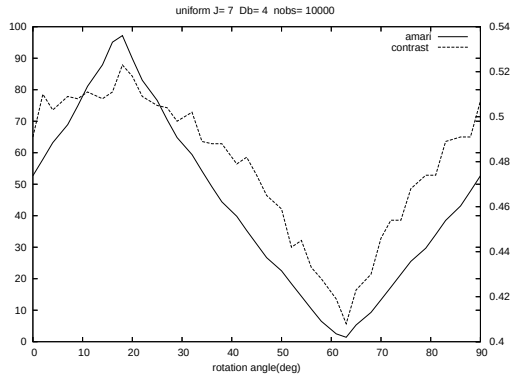


Fig.6. Uniform, D4, j=7, n=10000

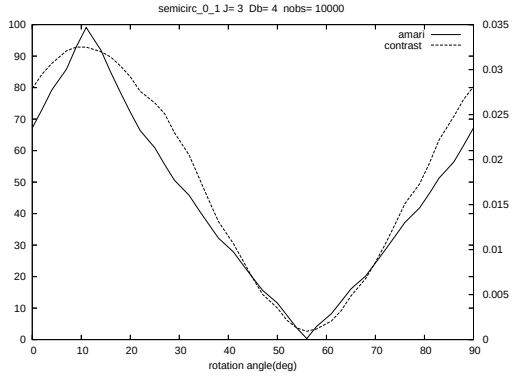


Fig.7. Semi-circular, D4, j=3, n=10000

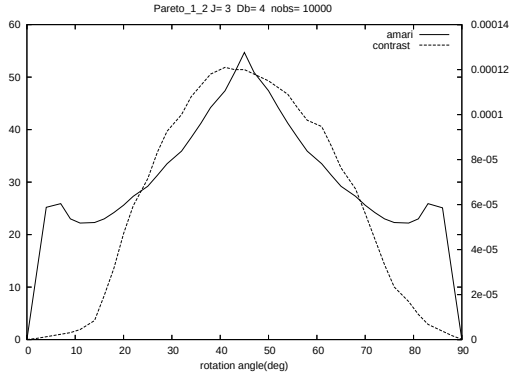


Fig.8. Pareto, D4, j=3, n=10000

Idem for a Pareto, at all resolutions from 3 to 9 (Fig. 8, $j = 3$ only shown).

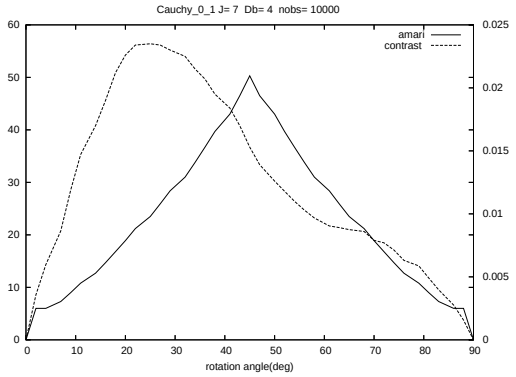


Fig.9. Cauchy, D4, j=7, n=10000

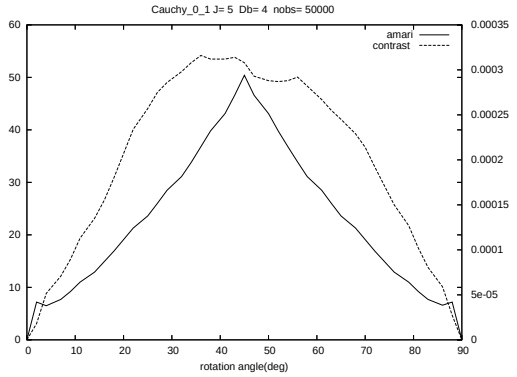


Fig.10. Cauchy, D4, j=5, n=50000

Finally for a Cauchy distribution, with a D4, the minimum is accuratly localized without interfering local minimums starting from $j = 7$ for 10000 observations (Fig. 9), and starting from $j = 5$ for 50000 observations (Fig. 10), which matches the fact that this density has a lower regularity than previous ones, and needs higher $j$.

Through these examples, it looks like the theoretical rule is pessimistic in the sense that with a ratio $2^{4jd}/n$ rather big, the ICA minimum is occasionaly well localized, even without a necessarily well estimated density.

## 5. Contrast minimization

The natural way to minimize the ICA contrast as a function of a demixing matrix $W$, is to whiten the signal and then perform a steepest descent algorithm given the constraint ${}^tWW = I_d$, corresponding to $W$ lying on the the Stiefel manifold $S(d,d) = O(d)$. In the ICA context, as in dimension 2, we can restrict to $SO(d) \subset O(d)$.

Needed material for minimization on the Stiefel manifold can be found in the paper of Smith, Edelman and Arias (1998). Another very close method uses the the Lie group structure of $SO(d)$ and the corresponding Lie algebra $so(d)$ mapped together by the matrix logarithm and exponential (Plumbley, 2004). For convenience we reproduce here the algorithm in question, which is equivalent to a line search in the steepest descent direction in $so(d)$:

- start at $O \in so(d)$, equivalent to $I \in SO(d)$;
- move about in $so(d)$ from 0 to $-\eta\nabla_B J$, where $\eta \in \mathbb{R}^+$ corresponds to the minimum in direction $\nabla_B J$ found by a line search algorithm, where $\nabla_B J = \nabla J\, {}^tW - W\, {}^t\nabla J$ is the gradient of $J$ in $so(d)$, and where $\nabla J$ is the gradient of $J$ in $SO(d)$;
- use the matrix exponential to map back into $SO(d)$, giving $R = \exp(-\eta\nabla_B J)$;
- calculate $W' = RW \in SO(d)$ and iterate.

We reproduce below some typical runs (program `icalette3`), with a D4 and $L = 10$. Note that on example 2, the contrast cannot be usefully minimized because of a wrong resolution.

| d=3, j=3, n=30000 uniform | | |
|---|---|---|
| it | contrast | amari |
| 0 | 0.127722 | 65.842 |
| 1 | 0.029765 | 15.784 |
| 2 | 0.002600 | 2.129 |
| 3 | 0.001939 | 0.288 |
| 4 | - | - |
| 5 | - | - |

| d=3, j=5, n=30000 uniform | | |
|---|---|---|
| it | contrast | amari |
| 0 | 0.321970 | 65.842 |
| 1 | 0.321948 | 65.845 |
| 2 | 0.321722 | 65.999 |
| 3 | 0.321721 | 65.999 |
| 4 | - | - |
| 5 | - | - |

| d=3, j=3, n=10000 uniform | | |
|---|---|---|
| it | contrast | amari |
| 0 | 0.092920 | 42.108 |
| 1 | 0.035336 | 14.428 |
| 2 | 0.007458 | 3.392 |
| 3 | 0.006345 | 1.684 |
| 4 | 0.006122 | 1.109 |
| 5 | 0.006008 | 0.675 |

| d=4, j=2, n=10000 uniform | | |
|---|---|---|
| it | contrast | amari |
| 0 | 0.025193 | 22.170 |
| 1 | 0.010792 | 9.808 |
| 2 | 0.003557 | 4.672 |
| 3 | 0.001272 | 1.167 |
| 4 | 0.001033 | 0.502 |
| 5 | 0.000999 | 0.778 |

| d=3, j=4, n=30000 expone. | | |
|---|---|---|
| it | contrast | amari |
| 0 | 8.609670 | 52.973 |
| 1 | 5.101633 | 48.744 |
| 2 | 0.778619 | 16.043 |
| 3 | 0.017585 | 3.691 |
| 4 | 0.008027 | 2.262 |
| 5 | 0.006306 | 1.542 |

| d=3, j=3, n=10000 semici. | | |
|---|---|---|
| it | contrast | amari |
| 0 | 0.041392 | 35.080 |
| 1 | 0.029563 | 22.189 |
| 2 | 0.007775 | 5.601 |
| 3 | 0.006055 | 3.058 |
| 4 | 0.005387 | 2.261 |
| 5 | 0.005355 | 1.541 |

Table 3. Minimization examples at various j, d and n with D4 and L=10

In our simulations, $\nabla J$ is computed by first differences; in doing so we cannot keep the orthogonality of the perturbed $W$, and we actually compute a plain gradient in $\mathbb{R}^{dd}$.

Note that a Haar contrast empirical gradient is tricky to obtain because a small perturbation in $W$ will likely result in an unchanged histogram at small $j$, whereas, for contrasts D4 and above, the response to perturbation is practically automatic and is anyway adjustable by the dyadic approximation parameter $L$.

Below is the average of 100 runs in dimension 2 with 10000 observations, D4, $j = 3$ and $L = 10$ for different densities; the start columns indicate Amari distance (on the scale 0 to 100) and the wavelet contrast on entry; `it` is the average number of iterations. Note that for some densities after whitening we are already close to the minimum, but the contrast still detects a departure from independence; the minimization routine exits on entry if the contrast or the gradient are too small, and this practically always correspond to an Amari distance less than 1 in our simulations.

| density | Amari start | Amari end | cont. start | cont. end | it. |
|---|---|---|---|---|---|
| uniform | 53.193 | 0.612 | 0.509E-01 | 0.104E-02 | 1.7 |
| exponential | 32.374 | 0.583 | 0.616E-01 | 0.150E-03 | 1.4 |
| Student | 2.078 | 1.189 | 0.534E-04 | 0.188E-04 | 0.1 |
| semi-circ | 51.401 | 2.760 | 0.222E-01 | 0.165E-02 | 1.8 |
| Pareto | 4.123 | 0.934 | 0.716E-03 | 0.415E-05 | 0.3 |
| triangular | 46.033 | 7.333 | 0.412E-02 | 0.109E-02 | 1.6 |
| normal | 45.610 | 45.755 | 0.748E-03 | 0.408E-03 | 1.4 |
| Cauchy | 1.085 | 0.120 | 0.261E-04 | 0.596E-06 | 0.1 |

Table 4. Average results of 100 runs in dimension 2, j=3 with a D4 at L=10

These first results compare rather favourably with the performance of existing ICA algorithms, as presented for instance in the paper of Bach and Jordan (2002). We have found equivalent results in higher dimension, with suitable $j$ and $n$ adjustment.

### 5.1 Wavelet contrast differential and implementation issues

Recall the notation,

$$C_j(y_1, \dots, y_n) = \sum_{k^1 \dots, k^d} \left(\hat{\alpha}_{jk^1 \dots, k^d} - \hat{\alpha}_{jk^1} \dots \hat{\alpha}_{jk^d}\right)^2 = \sum_{k^1 \dots, k^d} \hat{\delta}^2_{jk},$$

with $\hat{\alpha}_{jk^\ell} = \frac{1}{n} \sum_i \varphi_{jk^\ell}(y_i^\ell)$ and $\hat{\alpha}_{jk^1 \dots, k^d} = \hat{\alpha}_{jk} = \frac{1}{n} \sum_i \Phi_{jk}(y_i)$.

**Proposition 5.6** **($C_j$ differential in $Y$)**

*If the wavelet $\varphi$ is a $C^1$ function the differential of $C_j$ is well defined and, when $Y$ shaped $d \times n$ is in row major order, $dC_j(Y) \in \mathcal{L}(\mathbb{R}^{nd}; \mathbb{R})$ is given by the matrix,*

$$dC_j(Y) = 2 \sum_k \hat{\delta}_{jk} \left( D_1^1 \hat{\delta}_{jk} \quad D_2^1 \hat{\delta}_{jk} \quad \dots \quad D_n^1 \hat{\delta}_{jk} \quad D_1^2 \hat{\delta}_{jk} \quad \dots \quad D_n^d \hat{\delta}_{jk} \right),$$

*or, when $Y \in \mathbb{R}^{nd}$ is in column major order, by the matrix,*

$$dC_j(Y) = 2 \sum_k \hat{\delta}_{jk} \left( D_1^1 \hat{\delta}_{jk} \quad D_1^2 \hat{\delta}_{jk} \quad \dots \quad D_1^d \hat{\delta}_{jk} \quad D_n^1 \hat{\delta}_{jk} \quad \dots \quad D_n^d \hat{\delta}_{jk} \right),$$

*with $D_i^\ell \hat{\delta}_{jk}$ the partial derivative in direction $i, \ell$ given by,*

$$D_i^\ell \hat{\delta}_{jk} = \varphi'_{jk_\ell}(y_i^\ell) \left( \frac{1}{n} \prod_{h \neq \ell} \varphi_{jk_h}(y_i^h) - \frac{1}{n^d} \prod_{h \neq \ell} \sum_j \varphi_{jk_h}(y_j^h) \right).$$

*Incidentally, let $D$ be the $n \times d$ matrix whose $(i,j)$ element is $2 \sum_k \hat{\delta}_{jk} D_i^j \hat{\delta}_{jk}$, and $H \in \mathbb{R}^{nd}$ identified with a $d \times n$ matrix, one has*

$$dC_j(Y)(H) = \text{trace}[HD]$$

**proof**

The $C_j$ partial derivative in direction $r_i^\ell$, $1 \leq i \leq n$, $1 \leq \ell \leq d$, is given by $D_i^\ell C_j = \sum_k 2\hat{\delta}_{jk} D_i^\ell \hat{\delta}_{jk}$, where,

$$\begin{aligned}
D_i^\ell \hat{\delta}_{jk} =& D_i^\ell \left[ \frac{1}{n} \sum_i \Phi_{jk}(y_i) - \prod_h \frac{1}{n} \sum_j \varphi_{jk_h}(y_j^h) \right] \\
=& D_i^\ell \left[ \frac{1}{n} \sum_i \prod_h \varphi_{jk_h}(y_i^h) - \prod_h \frac{1}{n} \sum_j \varphi_{jk_h}(y_j^h) \right] \\
=& \frac{1}{n} \varphi'_{jk_\ell}(y_i^\ell) \prod_{h \neq \ell} \varphi_{jk_h}(y_i^h) - \frac{1}{n^d} \varphi'_{jk_\ell}(y_i^\ell) \prod_{h \neq \ell} \sum_j \varphi_{jk_h}(y_j^h) \\
=& \varphi'_{jk_\ell}(y_i^\ell) \left( \frac{1}{n} \prod_{h \neq \ell} \varphi_{jk_h}(y_i^h) - \frac{1}{n^d} \prod_{h \neq \ell} \sum_j \varphi_{jk_h}(y_j^h) \right),
\end{aligned} \tag{5}$$

since, in the different sums appearing in $\hat{\delta}_{jk}$, the derivative in $(i, \ell)$ is zero if the observation $j$ is not equal to $i$ or if the component $h$ is not equal to $\ell$.

$D_i^\ell \hat{\delta}_{jk}$ is continuous and so the differential exists.

Considered as a matrix, $Y$ was shaped $d \times n$ by convention; as an element of $\mathbb{R}^{nd}$ we adopt the row major order, *i.e.* the first $n$ elements are the first line of the matrix, the $n$ following, the second line, etc. Thus $dC_j(Y) \in \mathcal{L}(\mathbb{R}^{nd}; \mathbb{R})$ is given in the order,

$$dC_j(Y) = 2 \sum_k \hat{\delta}_{jk} \left( D_1^1 \hat{\delta}_{jk} \quad D_2^1 \hat{\delta}_{jk} \quad \dots \quad D_n^1 \hat{\delta}_{jk} \quad D_1^2 \hat{\delta}_{jk} \quad \dots \quad D_n^d \hat{\delta}_{jk} \right).$$

So for $H \in \mathbb{R}^{nd}$, and $D_i^\ell \hat{\delta}_{jk}$ shortened in $D_i^\ell$,

$$dC_j(Y)(H) = 2 \sum_k \hat{\delta}_{jk} \left[ (D_1^1 \dots D_n^1) \begin{pmatrix} h_{11} \\ \vdots \\ h_{1n} \end{pmatrix} + (D_1^2 \dots D_n^2) \begin{pmatrix} h_{21} \\ \vdots \\ h_{2n} \end{pmatrix} + \dots \right]$$

which is exactly the sum of the diagonal elements of the matrix $HD$. For the column major order, the gradient is reversed as announced and still equals the trace of $HD$ when applied to $H$. □

The differential in $W$ represents a weighted sum of the differential in $Y$ as we see now.

**Proposition 5.7** **($C_j$ differential in $W$)**

*Under the notation of Prop. 5.6, with $C_j(W) \equiv C_j(Wx_1, \ldots, Wx_n)$ and $X = (\, x_1 \quad \ldots \quad x_n \,)$, the $d \times n$ matrix arrangement of the observation $\{x_1 \ldots, x_n\}$, the differential of $C_j$ is well defined and,*

*Under the row major order, $dC_j(W) \in \mathcal{L}(\mathbb{R}^{dd}; \mathbb{R})$ is given by,*

$$dC_j(W) = 2 \sum_k \delta_{jk} \, (\, (D_1^1 \ldots D_n^1)^t X \quad (D_1^2 \ldots D_n^2)^t X \quad \ldots \quad (D_1^d \ldots D_n^d)^t X \,) \in \mathcal{L}(\mathbb{R}^{dd}; \mathbb{R})$$

*with $D_i^\ell$, the abridged notation for $D_i^j \hat{\delta}_{jk}$ given in Prop. 5.6.*

*Under column major order,*

$$dC_j(W) = 2 \sum_i X_i \otimes \sum_k \hat{\delta}_{jk} D_i \tag{6}$$

*with $D_i = (D_i^1, \ldots, D_i^d)$, $X_i = (X_i^1, \ldots, X_i^d)$ and $\otimes$ denoting the inlined kronecker product, i.e. $X_i \otimes D_i = (\, X_i^1 D_i^1 \quad \ldots \quad X_i^1 D_i^d \quad X_i^2 D_i^1 \quad \ldots \quad X_i^2 D_i^d \quad \ldots \,) \in \mathbb{R}^{dd}$.*

*Incidentally, let $D$ the matrix $n \times d$ whose $(i, j)$ element is $2 \sum_k \hat{\delta}_{jk} D_i^j \hat{\delta}_{jk}$, and $Z \in \mathbb{R}^{dd}$ identified with a $d \times d$ matrix in row major order, one has*

$$dC_j(W)(Z) = \text{trace}[ZXD]$$

**proof**

Consider the linear function $T : W \in \mathbb{R}^{dd} \mapsto Wx = y \in \mathbb{R}^{np}$ where $y_i = Wx_i$ and $x_i \in \{x_1, \ldots, x_n\}$, the given sample; one has

$$d(C_j \circ T)(W) = dC_j(y_1, \ldots, y_n) \circ dT(W) = dC_j(y_1, \ldots, y_n) \circ T$$

Let's find the expression of $dT(W) \in \mathcal{L}(\mathbb{R}^{dd} \, ; \, \mathbb{R}^{nd})$.

$(Wx)_{ij} = \sum_{k=1\ldots,d} W_{ik} x_j^k$, so $\frac{d}{dw_{rs}}(Wx)_{ij} = 0$ if $r \neq i$ and $\frac{d}{dw_{rs}}(WX)_{rj} = x_j^s$. Finally, when adopting the row major order, $dT(W)$ has the form,

$$dT(W) = dT = \begin{pmatrix} {}^tX & & & \\ & {}^tX & & \\ & & \ddots & \\ & & & {}^tX \end{pmatrix} \tag{7}$$

which boils down to say that for a $d \times d$ matrix $F$, $dT(W)(F) = FX$ in line with the fact that $T$ is linear and so $dT(W) \equiv T$.

Under column major order, $dT(W)$ has the same form as above with rows and columns of zero permuted with rows and columns of ${}^tX$.

$dC_j(y_1^1, y_2^1, \ldots, y_n^d) \in \mathcal{L}(\mathbb{R}^{nd}; \mathbb{R})$ was found in the preceding proposition to be,

$$dC_j(y_1^1, y_2^1, \ldots, y_n^d) = 2 \sum_k \hat{\delta}_{jk} \, (\, D_1^1 \hat{\delta}_{jk} \quad D_2^1 \hat{\delta}_{jk} \quad \ldots \quad D_n^1 \hat{\delta}_{jk} \quad D_1^2 \hat{\delta}_{jk} \quad \ldots \quad D_n^d \hat{\delta}_{jk} \,)$$

Finally, with the abuse of notation $(C_j \circ T)(W) = C_j(W)$, the expression of the differential $d(C_j \circ T)(W) = dC_j(y_1^1, y_1^2, \ldots, y_n^d) \circ dT(W)$ is the matrix,

$$dC_j(W) = 2\sum_k \delta_{jk} \left( (D_1^1 \ldots D_n^1)^t X \quad (D_1^2 \ldots D_n^2)^t X \quad \ldots \quad (D_1^d \ldots D_n^d)^t X \right) \in \mathcal{L}(\mathbb{R}^{dd}; \mathbb{R})$$

And,

$$dC_j(W)(Z) = 2\sum_k \hat{\delta}_{jk} \left[ (D_1^1 \ldots D_n^1)\, {}^t X \begin{pmatrix} z_{11} \\ \vdots \\ z_{1d} \end{pmatrix} + (D_1^2 \ldots D_n^2)\, {}^t X \begin{pmatrix} z_{21} \\ \vdots \\ z_{2d} \end{pmatrix} + \ldots \right]$$

which is written also $dC_j(W)(Z) = \text{trace}[ZXD]$.

Next, $dC_j(W)(z)$ is also written,

$$\begin{aligned} &dC_j(W)(z) = \\ &= 2\sum_k \hat{\delta}_{jk} \left[ (\sum_i D_i^1 X_i^1 \ldots \sum_i D_i^1 X_i^d) \begin{pmatrix} z_{11} \\ \vdots \\ z_{1d} \end{pmatrix} + (\sum_i D_i^2 X_i^1 \ldots \sum_i D_i^2 X_i^d) \begin{pmatrix} z_{21} \\ \vdots \\ z_{2d} \end{pmatrix} + \ldots \right] \\ &= 2\sum_k \hat{\delta}_{jk} \left[ (\sum_i D_i^1 X_i^1 \ldots \sum_i D_i^d X_i^1) \begin{pmatrix} z_{11} \\ \vdots \\ z_{d1} \end{pmatrix} + \ldots + (\sum_i D_i^1 X_i^d \ldots \sum_i D_i^d X_i^d) \begin{pmatrix} z_{1d} \\ \vdots \\ z_{dd} \end{pmatrix} \right] \end{aligned}$$

where $z$ can now be considered under column major order.

This simplifies to,

$$\begin{aligned} dC_j(W)(z) &= 2\sum_k \hat{\delta}_{jk} \sum_i (X_i \otimes D_i) \quad z \\ &= 2\sum_i X_i \otimes \sum_k \hat{\delta}_{jk} D_i \quad z \end{aligned}$$

with notations given above.

□

**Proposition 5.8 ($C_j$ second derivatives)**

*Under the notations of the preceding propositions, and assuming the wavelet is $C^2$,*

$$D_{i'}^{\ell'} D_i^{\ell} C_j = 2\sum_k D_{i'}^{\ell'} \hat{\delta}_{jk} D_i^{\ell} \hat{\delta}_{jk} + \hat{\delta}_{jk} D_{i'}^{\ell'} D_i^{\ell} \hat{\delta}_{jk}$$

*For $(i, \ell) = (i', \ell')$,*

$$\begin{aligned} D_i^{\ell} D_i^{\ell} \hat{\delta}_{jk} &= \varphi''_{jk^{\ell}}(y_i^{\ell}) \left( \frac{1}{n} \prod_{h \neq \ell} \varphi_{jk_h}(y_i^h) - \frac{1}{n^d} \prod_{h \neq \ell} \sum_j \varphi_{jk_h}(y_j^h) \right) \\ &= \varphi''_{jk^{\ell}}(y_i^{\ell}) Q_1(i, \ell) \end{aligned} \tag{8}$$

*For $i \neq i'$ and $\ell = \ell'$, $D^{\ell'}_{i'} D^{\ell}_{i} \hat{\delta}_{jk} = 0$.*

*For $i = i'$ and $\ell \neq \ell'$,*

$$\begin{aligned} D^{\ell'}_{i'} D^{\ell}_{i} \hat{\delta}_{jk} &= \varphi'_{jk^{\ell}}(y^{\ell}_{i}) \varphi'_{jk^{\ell'}}(y^{\ell'}_{i}) \left[ \frac{1}{n} \prod_{h \neq \ell, h \neq \ell'} \varphi_{jk^{h}}(y^{h}_{i}) - \frac{1}{n^{d}} \prod_{h \neq \ell, h \neq \ell'} \sum_{j} \varphi_{jk^{h}}(y^{h}_{j}) \right] \\ &= \varphi'_{jk^{\ell}}(y^{\ell}_{i}) \varphi'_{jk^{\ell'}}(y^{\ell'}_{i}) Q_2(i, \ell, \ell') \end{aligned} \tag{9}$$

*For $i \neq i'$ and $l \neq \ell'$,*

$$\begin{aligned} D^{\ell'}_{i'} D^{\ell}_{i} \hat{\delta}_{jk} &= \varphi'_{jk^{\ell}}(y^{\ell}_{i}) \varphi'_{jk^{\ell'}}(y^{\ell'}_{i'}) \left[ -\frac{1}{n^{d}} \prod_{h \neq \ell, h \neq \ell'} \sum_{j} \varphi_{jk^{h}}(y^{h}_{j}) \right] \\ &= \varphi'_{jk^{\ell}}(y^{\ell}_{i}) \varphi'_{jk^{\ell'}}(y^{\ell'}_{i'}) Q_3(\ell, \ell') \end{aligned} \tag{10}$$

**proof**

Nothing more than calculus. Note that $Q_3$ and $Q_2$ are symmetric in $(\ell, \ell')$.
□

**Proposition 5.9** **($C_j$ hessian in $Y$)**

*Under the assumption of the preceding proposition, and assuming the wavelet is $C^2$, the hessian of $C_j$ is a $nd \times nd$ matrix given by,*

$$\nabla^2 C_j(Y) = 2 \sum_{k} (D^{\ell}_{i} \hat{\delta}_{jk})_{(i\ell)} \, {}^{t}(D^{\ell}_{i} \hat{\delta}_{jk})_{(i\ell)} + \hat{\delta}_{jk} \begin{pmatrix} B^{11} & \dots & B^{1d} \\ & \ddots & \\ B^{d1} & \dots & B^{dd} \end{pmatrix}$$

*with, $B^{\ell\ell}$ diagonal whose term $(ii) = \varphi''_{jk^{\ell}}(y^{\ell}_{i}) Q_1(i, \ell)$ is given by (8);*

*and $B^{\ell\ell'}$ symmetric whose terms $(ii) = \varphi'_{jk^{\ell}}(y^{\ell}_{i}) \varphi'_{jk^{\ell'}}(y^{\ell'}_{i}) Q_2(i, \ell, \ell')$ are given by (9), and whose terms $(ii') = \varphi'_{jk^{\ell}}(y^{\ell}_{i}) \varphi'_{jk^{\ell'}}(y^{\ell'}_{i'}) Q_3(\ell, \ell')$ are given by (10); and with $(D^{\ell}_{i} \hat{\delta}_{jk})_{(i\ell)}$ the $nd \times 1$ vector given in (5).*

*The number of free terms of the matrix on the right is $\frac{d^2-d}{2} \frac{n^2+n}{2} + nd = \frac{nd}{4}(nd + d - n + 3)$.*

*With the notation $D^{\ell} = {}^{t}(D^{\ell}_{1} \dots D^{\ell}_{n}) \equiv {}^{t}(D^{\ell}_{1} \hat{\delta}_{jk} \dots D^{\ell}_{n} \hat{\delta}_{jk})$, one has the other expression,*

$$\nabla^2 C_j(Y) = 2 \sum_{k} \begin{pmatrix} D^1 \, {}^{t}D^1 & \dots & D^1 \, {}^{t}D^d \\ & \ddots & \\ D^d \, {}^{t}D^1 & \dots & D^d \, {}^{t}D^d \end{pmatrix} + \hat{\delta}_{jk} \begin{pmatrix} B^{11} & \dots & B^{1d} \\ & \ddots & \\ B^{d1} & \dots & B^{dd} \end{pmatrix} \tag{11}$$

**proof**

This is the definition of the hessian matrix.

The number of free terms of the matrix on the right is at first sight $(n^2d^2 - nd)/2 + nd = (n^2d^2 + nd)/2$, but since each of sub-diagonal blocks is also symmetric, the number of free terms is in fact $\frac{d^2-d}{2}\frac{n^2+n}{2} + nd = \frac{nd}{4}(nd+d-n+3)$ ; this is a diminution of $\frac{nd}{4}(nd-d+n-1)$. □

The left part of (11) is the Gauss-Newton matrix appearing in the second derivatives of any least square type function, and sometimes used as a hessian approximate (see for instance Lemarechal et al, 1997).

**Proposition 5.10** **($C_j$ hessian in $W$)**

*The hessian of $C_j$ in $W$ is given by,*

$$\nabla^2 C_j(W) = {}^t dT \nabla^2 C_j(Y) dT$$

*with $T : W \in \mathbb{R}^{dd} \mapsto Wx = y \in \mathbb{R}^{np}$.*

*The differential is given by $d^2C_j(W)(H_1, H_2) = d^2C_j(Y)(H_1X, H_2X)$.*

**proof**

With the abuse of notation $d^2C_j(W) \equiv d^2(C_j \circ T)(W)$,

$$\begin{aligned} d^2(C_j \circ T)(W)(H_1, H_2) &= d\Big[d(C_j \circ T)(W)(H_1)\Big](H_2) \\ &= d\Big[\big[dC_j(Y) \circ dT(W)\big](H_1)\Big](H_2) \\ &= d\Big[\big[dC_j(Y) \circ T\big](H_1)\Big](H_2) \\ &= \Big[d^2C_j(Y)(T(H1)) \circ dT\Big](H_2), \end{aligned}$$

where $dT$ is the matrix given in (7) and $Y = WX$.

After identification of $\mathcal{L}(\mathbb{R}^{dd}; \mathcal{L}(\mathbb{R}^{dd}; \mathbb{R}))$ with $\mathcal{L}(\mathbb{R}^{dd}, \mathbb{R}^{dd}; \mathbb{R})$ this is also written as,

$$d^2C_j(W)(H_1, H_2) = d^2C_j(Y)(dT(H_1), dT(H_2)) = d^2C_j(Y)(H_1X, H_2X).$$

So we have,

$${}^t H_1 \nabla^2 C_j(W) H_2 = d^2C_j(W)(H_1, H_2) = d^2C_j(Y)(H_1X, H_2X) = {}^t X {}^t H_1 \nabla^2 C_j(WX)\ H_2X$$

and by identification $\nabla^2 C_j(W) = {}^t dT \nabla^2 C_j(Y) dT$. □

**Filter aware formulations for the gradient and the hessian**

We now give a formulation of the gradient and the hessian that will be very helpful for practical computations, and accomodate well to possible subsequent filtering operations.

A Daubechies wavelet, $D2N$, satisfy the usual equation $\varphi(t) = \sqrt{2}\sum_k c_k \varphi(2t-k)$ with $c_0 \ldots, c_{2N-1}$ the only non zero coefficients; we have as well $\varphi'(t) = 2\sqrt{2}\sum_k c_k \varphi'(2t-k)$.

With $\varphi_{jk}(t) = 2^{j/2}\varphi(2^j t - k)$, we thus have the relation,

$$\begin{aligned}\varphi_{jk}(t) &= 2^{\frac{j}{2}}\varphi(2^j t - k) \\ &= 2^{\frac{j}{2}}\sqrt{2}\sum_\ell c_\ell \varphi(2(2^j t - k) - \ell) \\ &= \sum_\ell c_\ell \varphi_{j+1,2k+\ell}(t) = \sum_\ell c_{\ell-2k}\varphi_{j+1,\ell}(t)\end{aligned} \tag{12}$$

which gives directly $\hat{\alpha}_{jk} = \frac{1}{n}\sum_i \varphi_{jk}(X_i) = \sum_\ell c_{\ell-2k}\hat{\alpha}_{j+1,\ell} = (\hat{\alpha}_{j+1} * \bar{c})_{2k}$, where $\bar{c}_k = c_{-k}$, and $*$ designates convolution; this is the discrete wavelet transform algorithm (DWT) (Mallat, 2000).

This algorithm extends to the multidimensional $\alpha_{jk}$, and to $\hat{\delta}_{jk} = \hat{\alpha}_{jk} - \hat{\alpha}_{jk^1}\ldots\hat{\alpha}_{jk^d}$:

$$\begin{aligned}\hat{\delta}_{jk} &= \hat{\alpha}_{jk} - \hat{\alpha}_{jk^1}\ldots\hat{\alpha}_{jk^d} \\ &= \frac{1}{n}\sum_i \varphi_{jk^1}\ldots\varphi_{jk^1} - \frac{1}{n}\sum_i \varphi_{jk^1}\ldots\frac{1}{n}\sum_i \varphi_{jk^d} \\ &= \sum_{\ell^1\ldots,\ell^d} c_{\ell^1-2k^1}\ldots c_{\ell^d-2k^d}\hat{\alpha}_{j+1,\ell} - \sum_{\ell^1} c_{\ell^1-2k^1}\hat{\alpha}_{j+1,\ell^1}\ldots\sum_{\ell^d} c_{\ell^d-2k^d}\hat{\alpha}_{j+1,\ell^d} \\ &= \sum_{\ell^1\ldots,\ell^d} c_{\ell^1-2k^1}\ldots c_{\ell^d-2k^d}(\hat{\alpha}_{j+1,\ell} - \hat{\alpha}_{j+1,\ell^1}\ldots\hat{\alpha}_{j+1,\ell^d}) \\ &= \sum_\ell c_{\ell-2k}\hat{\delta}_{j+1,\ell},\end{aligned}$$

where the last line makes use of a condensed notation; and where we used the fact that there exists no index on the $\ell$ margin that does not exist also on the dimension $\ell$ of the cube.

Let us introduce the jackknife estimator $\hat{\alpha}_{jk}^{(i)} = \frac{1}{n-1}\sum_{j\neq i}\Phi_{jk}(X_j)$.

**Proposition 5.11 (filter aware gradient formulation)**

*$D_i^\ell\hat{\delta}_{jk}$ is a function of Jackknife of the original $\hat{\alpha}_{jk}$ coefficients; the following relation holds,*

$$D_i^\ell\hat{\delta}_{jk} = \frac{\varphi'_{jk_\ell}(X_i^\ell)}{\varphi_{jk_\ell}(X_i^\ell)}\left[\hat{\delta}_{jk} - \frac{n-1}{n}\left[\hat{\alpha}_{jk}^{(i)} - \hat{\alpha}_{jk^\ell}^{(i)}\prod_{h\neq\ell}\hat{\alpha}_{jk^h}\right]\right]$$

*It follows than the partial derivative of $D_i^\ell\hat{\delta}_{jk}$ is computable from the same elements than $D_i^\ell\hat{\delta}_{j+1k}$ is computed, with one DWT filtering pass.*

**proof**

The following relation holds,

$$\frac{1}{n}\Phi_{jk}(X_i) = \hat{\alpha}_{jk} - \frac{n-1}{n}\hat{\alpha}_{jk}^{(i)},$$

and starting from (5), the partial derivative can be expressed by,

$$\begin{aligned} D_i^\ell \hat{\delta}_{jk} =& \frac{\varphi'_{jk_\ell}(X_i^\ell)}{\varphi_{jk_\ell}(X_i^\ell)} \left[ \hat{\alpha}_{jk} - \frac{n-1}{n} \hat{\alpha}_{jk}^{(i)} - (\hat{\alpha}_{jk^\ell} - \frac{n-1}{n} \hat{\alpha}_{jk^\ell}^{(i)}) \prod_{h \neq \ell} \hat{\alpha}_{jk^h} \right] \\ =& \frac{\varphi'_{jk_\ell}(X_i^\ell)}{\varphi_{jk_\ell}(X_i^\ell)} \left[ \hat{\delta}_{jk} - \frac{n-1}{n} \left[ \hat{\alpha}_{jk}^{(i)} - \hat{\alpha}_{jk^\ell}^{(i)} \prod_{h \neq \ell} \hat{\alpha}_{jk^h} \right] \right]. \end{aligned} \tag{13}$$

□

There is theoretically an indetermination when the denominator of $\frac{\phi'_{jk^\ell}(X_i^\ell)}{\phi_{jk^\ell}(X_i^\ell)}$ is equal to zero. But in dyadic approximation, one always find points where the derivative is zero if the value at the point is zero.

The transition to a jackknifed estimator is,

$$\hat{\alpha}_{jk}^{(i)} = \frac{1}{n-1} \left[ n\hat{\alpha}_{jk} - \Phi_{jk}(X_i) \right] = \frac{n}{n-1} \hat{\alpha}_{jk} - \frac{2^{jd/2}}{n-1} \Phi(2^j X_i - k),$$

with implicit extended multidimensional notation.

This affects only $(2n-1)^d$ cells in the cube *i.e.* the power $d$ of the number of integers contained in the wavelet support ($\varphi(2N-1) = 0$). So once the full projection on the cube is known, transition to a Jackknife is a $O((2N-1)^d)$ operation.

Also in (13), the product of margins differs from the non jackknife product of margins only on a band with volume $2^{j(d-1)}(2N-1)$, and finally everything outside this band can be ignored since the premultiplication by $\varphi'/\varphi$ will produce zero, and from expression (6) the whole computations ends up with a sum in $k$.

For the transition between jackknifes, one has, $\hat{\alpha}_{jk}^{(j)} = \hat{\alpha}_{jk}^{(i)} + \frac{1}{n-1} \left[ \varphi_{jk}(X_i) - \varphi_{jk}(X_j) \right]$ which is true for all $j$, by linearity of the convolution, and translates in the Haar case in

$$\hat{\alpha}_{jk}^{(j)} = \hat{\alpha}_{jk}^{(i)} + \frac{2^{jd/2}}{n-1} \left[ I_{(2^j X_i \in A_{jk})} - I_{(2^j X_j \in A_{jk})} \right].$$

The hessian also possess a filter aware formulation, and besides, as compared to the gradient, the only additional quantities to compute are the $\varphi''_{jk}$, as we see next.

**Proposition 5.12 (filter aware hessian formulation)**

*The matrices $B^{\ell\ell}$ composing the hessian can be written as a function of Jackknife of the original $\alpha_{jk}$ coefficients ; one has the relations,*

$$B_{(ii)}^{\ell\ell} = \frac{\varphi''_{jk_\ell}(X_i^\ell)}{\varphi_{jk_\ell}(X_i^\ell)} \left[ \hat{\delta}_{jk} - \frac{n-1}{n} \left[ \hat{\alpha}_{jk}^{(i)} - \hat{\alpha}_{jk^\ell}^{(i)} \prod_{h \neq \ell} \hat{\alpha}_{jk^h} \right] \right]$$

*For matrices $B^{\ell\ell'}$, one has,*

$$B^{\ell\ell'}_{(ii)} = \frac{\varphi'_{jk^\ell}(y_i^\ell)\varphi'_{jk^{\ell'}}(y_i^{\ell'})}{\varphi_{jk^\ell}(y_i^\ell)\varphi_{jk^{\ell'}}(y_i^{\ell'})}\left[\hat{\alpha}_{jk} - \frac{n-1}{n}\hat{\alpha}^{(i)}_{jk} - (\hat{\alpha}_{jk^\ell} - \frac{n-1}{n}\hat{\alpha}^{(i)}_{jk^\ell})(\hat{\alpha}_{jk^{\ell'}} - \frac{n-1}{n}\hat{\alpha}^{(i)}_{jk^{\ell'}})\prod_{\substack{h\neq\ell\\ h\neq\ell'}}\hat{\alpha}_{jk^h}\right]$$

*with the bracket also written as,*

$$\left[\hat{\delta}_{jk} - \frac{n-1}{n}\left[\hat{\alpha}^{(i)}_{jk} + \hat{\alpha}^{(i)}_{jk^{\ell'}}\prod_{h\neq\ell'}\hat{\alpha}_{jk^h} + \hat{\alpha}^{(i)}_{jk^\ell}\prod_{h\neq\ell}\hat{\alpha}_{jk^h} - \frac{n-1}{n}\hat{\alpha}^{(i)}_{jk^\ell}\hat{\alpha}^{(i)}_{jk^{\ell'}}\prod_{h\neq\ell,h\neq\ell'}\hat{\alpha}_{jk^h}\right]\right]$$

*And,*

$$B^{\ell\ell'}_{(ii')} = \varphi'_{jk^\ell}(y_i^\ell)\varphi'_{jk^{\ell'}}(y_{i'}^{\ell'})\left[-\frac{1}{n^2}\prod_{h\neq\ell,h\neq\ell'}\hat{\alpha}_{jk^h}\right]$$

**proof**

Use the jackknife substitution. □

**References**


[1] E. Oja A. Hyvarinen, J. Karhunen. *Independent component analysis.* Inter Wiley Science, 2001.

[2] T.J. Sejnowski A. J. Bell. A non linear information maximization algorithm that performs blind separation. *Advances in neural information processing systems*, 1995.

[3] A. Tsybakov A. Samarov. Nonparametric independent component analysis. *Bernouilli*, 10:565–582, 2004.

[4] Steven T. Smith Alan Edelman, Tomas Arias. *The geometry of algorithms with orthogonality constraints.* SIAM, 1998.

[5] Alex Smola Arthur Gretton, Ralf Herbrich. The kernel mutual information. Technical report, Max Planck Institute for Biological Cybernetics, April 2003.

[6] J. Bergh and J. Löstrom. *Interpolation spaces.* Springer, Berlin, 1976.

[7] J.F. Cardoso. High-order contrasts for independent component analysis. *Neural computations 11*, pages 157–192, 1999.

[8] P. Comon. Independent component analysis, a new concept ? *Signal processing*, 1994.

[9] Ingrid Daubechies. *Ten lectures on wavelets.* SIAM, 1992.

[10] R. Devore and G. Lorentz. *Constructive approximation.* Springer-Verlag, 1993.

[11] G. Kerkyacharian D.L. Donoho, I.M. Johnstone and D. Picard. Density estimation by wavelet thresholding. *Annals of statistics*, 1996.

[12] Gérard Kerkyacharian Dominique Picard. Density estimation in Besov spaces. *Statistics and Probability Letters*, 13:15–24, 1992.

[13] M. I. Jordan F. R. Bach. Kernel independent component analysis. *J. of Machine Learning Research*, 3:1–48, 2002.

[14] Alexander Tsybakov Gérard Kerkyacharian, Dominique Picard and Wolfgang Härdle. *Wavelets, approximation and statistical applications*. Springer, 1998.

[15] Truong Nguyen Gilbert Strang. *Wavelets and filter banks*. Wellesley-Cambridge Press, 1996.

[16] A. Hyvarinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural computation*, 1997.

[17] Claude Lemarechal J.Frederic Bonnans, J.Charles Gilbert et al. *Numerical Optimization*. Springer, 1997.

[18] Stephane Mallat. *Une exploration des signaux en ondelettes*. Les éditions de l'ecole polytechnique, 2000.

[19] Yves Meyer. *Ondelettes et opérateurs*. Hermann, 1997.

[20] Mark D. Plumbley. Lie group methods for optimization with orthogonality constraints. In *ICA*, pages 1245–1252, 2004.

[21] Jean Roch. Le modèle factoriel des cinq grandes dimensions de personnalité : les big five. Technical report, AFPA, DO/DEM, March 1995.

[22] A. Cichocki S. Amari and H. Yang. A new algorithm for blind signal separation. *Advances in Neural Information Processing Systems*, 8:757–763, 1996.

[23] Robert J. Serfling. *Approximation theorems of mathematical statistics*. Wiley, 1980.

[24] R. B. Sidje. Expokit. A Software Package for Computing Matrix Exponentials. *ACM Trans. Math. Softw.*, 24(1):130–156, 1998.

**Programs available at** `http://www.proba.jussieu.fr/pageperso/barbedor`