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 the density estimation of the observed signal and its marginals by means of wavelets. The risk of the associated moment estimator is linked with approximation properties in Besov spaces. It is shown to converge faster than the at least expected minimax rate carried over from the underlying density estimations. Numerical simulations performed on some common types of densities yield very competitive results, with a high sensitivity to small departures from independence.

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 under the assumption of normality.

Papers on ICA are found in the research fields of signal processing, neural networks, statistics and information theory. Comon (1994) defined the concept of 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 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 et al (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_ℓ is the ℓ^{th} row of the matrix $B = A^{-1}$; this resulting from a change of variable if the latent density f is equal to the product of its marginals $f^1 \dots f^d$. In this regard, latent variable $s = (s^1, \dots, s^d)$ having independent components means the indepence of the random variables $s^\ell \circ \pi^\ell$ defined on some product probability space $\Omega = \prod \Omega^\ell$, with π^ℓ 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 the gradient of f_A.

In this paper, we propose a wavelet based ICA contrast. The wavelet 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 relies upon the procedures of wavelet density estimation which are found in a series of articles from Kerkyacharian and Picard (1992) and Donoho et al. (1996).

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 yields a satisfactory approximate solution to the ICA problem.

The wavelet contrast can be seen as a special case of quadratic dependence measure, as presented in Achard et al. (2003), which is equal to zero under independence. But in our case, the resolution parameter j allows more flexibility in controlling the reverse implication. Let's mention also that the idea of comparing in the L_2 norm a joint density with the product of its marginals, can be traced back to Rosenblatt (1975).

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 risk bound that will hold for the entire ICA minimization procedure.

Under its simplest form, the wavelet contrast estimator is a linear function of the empirical measure on the observation. We give the rule for the choice of a resolution level j minimizing the risk, assuming a known regularity s for a latent signal in some Besov space B_{spq}.

The estimator complexity is linear in the sample size but exponential in the dimension d of the problem; this is on account of an implicit multivariate density estimation. In compensation to this computational load, the wavelet contrast shows a very good sensitivity to small departures from independence, and encapsulates all practical tuning in a single parameter j.

2. Notations

We set here the main notations and recall some definitions for the convenience of ICA specialists. The reader already familiar with wavelets and Besov spaces can skip this part.

- *Wavelets*

Let φ be some function of $L_2(\mathbb{R})$ such that the family of translates $\{\varphi(.-k),\, k \in \mathbb{Z}\}$ is an orthonormal system; let $V_j \subset L_2(\mathbb{R})$ be the subspace spanned by $\{\varphi_{jk} = 2^{j/2}\varphi(2^j.-k), k \in \mathbb{Z}\}$.

By definition, the sequence of spaces $(V_j), j \in \mathbb{Z}$, is called a multiresolution analysis (MRA) of $L_2(\mathbb{R})$ if $V_j \subset V_{j+1}$ and $\bigcup_{j\geq 0} V_j$ is dense in $L_2(\mathbb{R})$; φ is called the father wavelet or scaling function.

Let $(V_j)_{j\in\mathbb{Z}}$ be a multiresolution analysis of $L_2(\mathbb{R})$, with V_j spanned by $\{\varphi_{jk} = 2^{j/2}\varphi(2^j.-k), k \in \mathbb{Z}\}$. Define W_j as the complement of V_j in V_{j+1}, and let the families $\{\psi_{jk}, k \in \mathbb{Z}\}$ be a basis for W_j, with $\psi_{jk}(x) = 2^{j/2}\psi(2^jx-k)$. Let $\alpha_{jk}(f) =< f, \varphi_{jk} >$ and $\beta_{jk}(f) =< f, \psi_{jk} >$.

A function $f \in L_2(\mathbb{R})$ admits a wavelet expansion on $(V_j)_{j\in\mathbb{Z}}$ if the series

$$\sum_k \alpha_{j_0k}(f)\varphi_{jk} + \sum_{j=j_0}^{\infty}\sum_k \beta_{jk}(f)\psi_{jk}$$

is convergent to f in $L_2(\mathbb{R})$; ψ is called a mother wavelet.

The definition of a multiresolution analysis on $L_2(\mathbb{R}^d)$ follows the same pattern. But an MRA in dimension one also induces an associated MRA in dimension d, using the tensorial product procedure below.

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

for $(i^1\ldots,i^d) = (0\ldots,0)$, 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 $\Psi_{jk}(x) = 2^{\frac{jd}{2}}\Psi(2^jx-k)$ and $\Phi_{jk}(x) = 2^{\frac{jd}{2}}\Phi(2^jx-k)$;

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 $U_{1j}\ldots\otimes U_{dj}$, 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 f admits a wavelet expansion on the basis (Φ, Ψ) if the series

$$\sum_{k\in\mathbb{Z}^d} \alpha_{j_0 k}(f)\Phi_{j_0 k} + \sum_{j=j_0}^{\infty} \sum_{k\in\mathbb{Z}^d} \beta_{jk}(f)\Psi_{jk}, \tag{1}$$

is convergent to f in $L_2(\mathbb{R}^d)$.

In fact, with the concentration condition

$$\sum_k |\varphi(x+k)| \leq C \; a.s., \tag{2}$$

verified in particular for a compactly supported wavelet, any function in $L_1(\mathbb{R}^d)$ admits a wavelet expansion. Otherwise any function in a Besov space $B_{spq}(\mathbb{R}^d)$ admits a wavelet expansion.

In connection with function approximation, wavelets can be viewed as falling in the category of orthogonal series methods, or also in the category of kernel methods.

The approximation at level j of a funtion f that admits a multiresolution expansion is the orthogonal projection $P_j f$ of f onto $V_j \subset L_2(\mathbb{R}^d)$ defined by:

$$(P_j f)(x) = \sum_{k\in\mathbb{Z}^d} \alpha_{jk}\Phi_{jk}(x),$$

where $\alpha_{jk} = \alpha_{jk^1\ldots,k^d} = \int f(x)\Phi_{jk}(x)\,dx$.

With the concentration condition above, the projection operator can also be written

$$(P_j f)(x) = \int_{\mathbb{R}^d} K_j(x,y) f(y) d(y),$$

with $K_j(x,y) = 2^{jd}\sum_{k\in\mathbb{Z}^d} \Phi_{jk}(x-k)\overline{\Phi_{jk}(y-k)}$. K_j is an orthogonal projection kernel with window 2^{-jd} (which is not translation invariant).

- *Besov spaces*

Let f be a function in $L_p(\mathbb{R}^d)$ and $h \in \mathbb{R}^d$. Define the first order difference $\Delta_h f$ by $\Delta_h f(x) = f(x+h) - f(x)$ and the k^{th} order difference $\Delta_h^k f = \Delta_h \Delta_h^{k-1} f$ ($k = 1, 2, \ldots$ with $\Delta_h^0 f = f$, $\Delta_h^1 f = \Delta_h f$).

The modulus of continuity of order k of f in the metric of L_p, along direction h, is defined by (Nikol'skiĭ, 1975, p.145-160)

$$\omega_h^k(f,\delta)_p = \sup_{|t|\leq\delta} \|\Delta_{th}^k f(x)\|_p, \qquad \delta \geq 0, \quad |h| = 1.$$

The modulus of continuity of order k of f in the direction of the subspace $\mathbb{R}^m \subset \mathbb{R}^d$ is defined by

$$\Omega^k_{\mathbb{R}^m}(f,\delta)_p = \sup_{|h|=1, h\in\mathbb{R}^m} \omega^k_h(f,\delta)_p.$$

If the function f has arbitrary derivatives of order ϱ relative to the first m coordinates, one can define, for $h \in \mathbb{R}^m$,

$$f^{(\varrho)}_h = \sum_{|n|=\varrho} f^{(n)} h^n,$$

with $h = (h_1, \ldots, h_m, 0, \ldots, 0)$, $|h| = 1$, $|n| = \sum_1^m n_i$ and $h^n = h_1^{n_1} \ldots h_m^{n_m} = h_1^{n_1} \ldots h_m^{n_m} 0^0 \ldots 0^0$.

The modulus of continuity of order k of the derivatives of order ϱ of f is then defined by

$$\Omega^k_{\mathbb{R}^m}(f^{(\varrho)},\delta)_p = \sup_{|h|=1, h\in\mathbb{R}^m} \omega^k_h(f^{(\varrho)}_h,\delta)_p = \sum_{|n|=\varrho} \Omega^k_{\mathbb{R}^m}(f^{(n)},\delta)_p.$$

Let $s = [s] + \alpha$; the Hölder space $H^s_p(\mathbb{R}^d)$ is defined as the collection of functions in $L_p(\mathbb{R}^d)$ such that

$$\|\Delta_h f^{(n)}\|_p \leq M|h|^\alpha, \quad \forall n = (n^1, \ldots, n^d), \text{ with } |n| = \sum_1^d n_i = [s],$$
$$\text{or equivalently, } \Omega_{\mathbb{R}^d}(f^{([s])},\delta)_p = \sup_{h\in\mathbb{R}^d} \omega_h(f^{([s])},\delta)_p \leq M\delta^\alpha,$$

where M does not depend on h.

Besov spaces introduce a finer scale of smoothness than is provided by Hölder spaces. For each $\alpha > 0$ this can be accomplished by introducing a second parameter q and applying (α, q) quasi-norms (rather than (α, ∞)) to the modulus of continuity of order k.

Let $s > 0$ and (ϱ, k) forming an admissible pair of nonnegative integers satisfying the inequalities $k > s - \varrho > 0$. By definition, $f \in L_p(\mathbb{R}^d)$ belongs to the class $B_{spq}(\mathbb{R}^d)$ if there exist generalized partial derivatives of f of order $n = (n^1, \ldots, n^d)$, $|n| \leq \varrho$, and one of the following semi-norms is finite:

$$\begin{aligned} J'_{spq}(f) &= \sum_{|n|=\varrho} \left(\int_0^\infty |t^{-(s-\varrho)} \Omega^k_{\mathbb{R}^d}(f^{(n)},t)_p|^q \frac{dt}{t} \right)^{\frac{1}{q}}, \\ J''_{spq}(f) &= \left(\int_0^\infty |t^{-(s-\varrho)} \Omega^k_{\mathbb{R}^d}(f^{(\varrho)},t)_p|^q \frac{dt}{t} \right)^{\frac{1}{q}}. \end{aligned} \tag{3}$$

For fixed s and p, the space B_{spq} gets larger with increasing q. In particular, for $q = \infty$, $B_{spq}(\mathbb{R}) = H^s_p(\mathbb{R})$; various other embeddings exist since Besov spaces cover many well known classical concrete function spaces having their own history.

Finally, Besov spaces also admit a characterization in terms of wavelet coefficients, which makes them intrinsically connected to the analysis of curves via wavelet techniques.

f belongs to the (inhomogeneous) Besov space $B_{spq}(\mathbb{R}^d)$ if

$$J_{spq}(f) = \|\alpha_{0.}\|_{\ell_p} + \left[\sum_{j \geq 0} \left[2^{js} 2^{dj(\frac{1}{2} - \frac{1}{p})} \|\beta_{j.}\|_{\ell_p} \right]^q \right]^{\frac{1}{q}} < \infty, \tag{4}$$

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

A more complete presentation of wavelets linked with Sobolev and Besov approximation theorems and statistical applications can be found in the book from Härdle et al. (1998). General references about Besov spaces are Peetre (1975), Bergh & Löfström (1976), Triebel (1992), DeVore & Lorentz (1993).

3. Wavelet contrast, Besov membership

Let f be a density function with marginal distribution in dimension ℓ,

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

denoted by $f^{\star\ell}$.

As integrable positive functions, f and the $f^{\star\ell}$ admit a wavelet expansion on a basis (φ, ψ) verifying the concentration condition (2). One can then consider the projections up to order j, that is to say the projections of f and $f^{\star\ell}$ on V_j^d and V_j respectively, namely

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

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

Proposition 3.1 (wavelet contrast)

Let f be a density function on $\mathbb{R}^d$ and let φ be the scaling function of a multiresolution analysis verifying the concentration condition (2).

Define the contrast function

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

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

The following relation hold:

$$f \text{ factorizable} \implies C_j(f) = 0.$$

If f and φ are compactly supported or else if $f \in L_2(\mathbb{R}^d)$, the following relation hold:

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

Proof

As for the first assertion, with $f = f^1 \dots f^d$, one has $f^{\star \ell} = f^\ell$, $\ell = 1, \dots d$. Whence for $k = (k^1, \dots, k^d) \in \mathbb{Z}^d$, one has by Fubini theorem,

$$\begin{aligned}\alpha_{jk}(f) = \alpha_{jk}(f^{\star 1} \dots f^{\star d}) &= \int_{\mathbb{R}^d} f^{\star 1} \dots f^{\star d} \Phi_{jk}(x) dx \\ &= \int_{\mathbb{R}} f^{\star 1} \varphi_{jk^1}(x^1) dx^1 \dots \int_{\mathbb{R}} f^{\star d} \varphi_{jk^d}(x^d) dx = \alpha_{jk^1}(f^{\star 1}) \dots \alpha_{jk^d}(f^{\star d}).\end{aligned}$$

For the second assertion, $C_j = 0$ entails $\alpha_{jk}(f) = \alpha_{jk^1}(f^{\star 1}) \dots \alpha_{jk^d}(f^{\star d})$ for all $k \in \mathbb{Z}^d$. So that for $P_j f \in L_p(\mathbb{R}^d)$,

$$\begin{aligned}P_j f &= \sum_k \alpha_{jk}(f) \Phi_{jk} = \sum_k \alpha_{jk^1}(f^{\star 1}) \varphi_{jk^1} \dots \alpha_{jk^d}(f^{\star d}) \varphi_{jk^d} \\ &= \sum_{k^1} \alpha_{jk^1}(f^{\star 1}) \varphi_{jk^1} \dots \sum_{k^d} \alpha_{jk^d}(f^{\star d}) \varphi_{jk^d} \\ &= P_j^1 f^{\star 1} \dots P_j^d f^{\star d},\end{aligned}$$

with passage to line 2 justified by the fact that $(\alpha_{jk}(f) \Phi_{jk})_{k \in \mathbb{Z}^d}$ is a summable family of $L_2(\mathbb{R}^d)$ or else is a finite sum for a compactly supported density and a compactly supported wavelet.
$\square$

For the zero wavelet 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 lies in the order of projection j.

Under the notations of the preceding proposition, with a zero wavelet contrast and assuming existence in L_p, one has $\| P_j f - P_j^1 f^{\star 1} \dots P_j^d f^{\star d} \|_p = 0$, and so

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

If we now impose some regularity conditions on the densities, in our case if we now require that f and the product of its marginals belong to the (inhomogeneous) Besov space $B_{spq}(\mathbb{R}^d)$, the approximation error can be evaluated precisely. With a r-regular wavelet φ, $r > s$, the very definition of Besov spaces implies for any member f that (Meyer, 1997)

$$\| f - P_j f \|_p = 2^{-js} \epsilon_j, \quad \{\epsilon_j\} \in \ell_q(\mathbb{N}^d). \tag{5}$$

Remark

In the special case where f_A and the product of its marginals belong to $L_2(\mathbb{R}^d)$, Parseval equality implies that C_j is equal to the square of the L_2 norm of $P_j f_A - P_j^1 f_A^{\star 1} \dots P_j^d f_A^{\star d}$. And one can write,

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

hence with notation $K_\star(A,f) = \|f_A - f_A^{\star 1} \dots f_A^{\star d}\|_2$,

$$|K_\star(A,f) - C_j(f_A)^{\frac{1}{2}}| \leq \|f_A - P_j f_A\|_2 + \|P_j(f_A^{\star 1} \dots f_A^{\star d}) - f_A^{\star 1} \dots f_A^{\star d}\|_2, \qquad (6)$$

which gives another illustration of the shrinking with j distance between the wavelet contrast and the true norm evaluated at f_A. In particular when $A \neq I$, $C_j(f_A)$ cannot be small and for $A = I$, C_j must be small, for j big enough.

Continuing on the special case $p = 2$, the wavelet contrast can be viewed as an example of quadratic dependence measure as presented in the paper from Achard et al (2003).

Using the orthogonal projection kernel associated to the function φ, one has the writing

$$C_j(f_A) = \int_{\mathbb{R}^d} \left(E_{f_A}^n K_j(x,Y) - \prod_{i=1}^d E_{f_A}^n K_j^i(x^i, Y^i) \right)^2 dx,$$

with $K_j(x,y) = 2^{jd} \sum_{k \in \mathbb{Z}^d} \Phi_{jk}(x-k)\Phi_{jk}(y-k)$ and $K_j^i(x,y) = 2^j \sum_{k \in \mathbb{Z}} \varphi_{jk}(x^i - k^i)\varphi_{jk}(y^i - k^i)$.

This is the form of the contrast in the paper from Achard et al. (2003), except that in our case the kernel is not scale invariant; but the ICA context is scale invariant by feature, since the inverse of A is conventionally determined up to a scaling diagonal or permutation matrix, after a whitening step.

▪

To take advantage of relation (5) in the ICA context, we need a fixed Besov space containing the mixed density f_A and the product of its marginals, for any invertible matrix A.

The two following propositions check that the mixing by A is conservative as to Besov membership, and that the product of the marginals of a density f belongs to the same Besov space than f. It is equivalent to assume that f is in $B_{spq}(\mathbb{R}^d)$ or that the factors f^i are in $B_{spq}(\mathbb{R})$. If the factors have different Besov parameters, one can theoretically always find a bigger Besov space using Sobolev inclusions

$$\begin{aligned} &B_{s'pq'} \subset B_{spq} && \text{for } s' \geq s, \quad q' \leq q; \\ &B_{spq} \subset B_{s'p'q} && \text{for } p \leq p' \text{ and } s' = s + d/p' - d/p. \end{aligned}$$

Proposition 3.2 (Besov membership of marginal distributions)

If f is a density function belonging to $B_{spq}(\mathbb{R}^d)$ then each of its marginals belong 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 $\|f_A^{\star l}\|_p \leq \|f_A\|_p$.

With the ℓ^{th} canonical vector of $\mathbb{R}^d$ denoted by e^ℓ and for $t \in \mathbb{R}$, one has,

$$\Delta_t^k f^{\star\ell}(x^\ell) = \sum_{l=0}^{k} (-1)^{l+k} C_k^l f^{\star\ell}(x+t) = \sum_{l=0}^{k} (-1)^{l+k} C_k^l \int_{\mathbb{R}^{d-1}} f(x+te^\ell) dx^{*\ell} = \int_{\mathbb{R}^{d-1}} \Delta_{te^\ell}^k f(x) dx^{*\ell};$$

so that

$$\|\Delta_t^k f^{\star\ell}\|_{L_p(\mathbb{R})}^p = \int_{\mathbb{R}} \left| \int_{\mathbb{R}^{d-1}} \Delta_{te^\ell}^k f(x) dx^{*\ell} \right|^p dx^\ell \leq \int_{\mathbb{R}^d} \left|\Delta_{te^\ell}^k f(x)\right|^p dx \leq \|\Delta_{te^\ell}^k f\|_{L_p(\mathbb{R}^d)}^p,$$

and

$$\omega^k(f^{\star\ell}, \delta)_p = \sup_{|t|\leq\delta} \|\Delta_t^k f^{\star\ell}\|_{Lp(\mathbb{R})} \leq \sup_{|t|\leq\delta} \|\Delta_{te^\ell}^k f\|_{L_p(\mathbb{R}^d)} = \omega_{e^\ell}^k(f,\delta)_p,$$

and

$$\Omega^k(f^{\star\ell}, \delta)_p = \omega^k(f^{\star\ell}, \delta)_p \leq \omega_{e^\ell}^k(f,\delta)_p \leq \sup_{|h|=1, h\in\mathbb{R}^d} \omega_h^k(f,\delta)_p = \Omega_{\mathbb{R}^d}^k(f,\delta)_p.$$

Using the admissible pair $(k, \varrho) = ([s]+1, 0)$, one can see from (3) that $J'_{spq}(f^{\star\ell}) \leq J'_{spq}(f)$.
□

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

Proposition 3.3 (Besov membership of the mixed density)

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

(a) if $f \in L_p(\mathbb{R}^d)$, or if each f^ℓ belongs to $L_p(\mathbb{R})$, then f_A and the product $\prod f_A^{\star\ell}$ belong to $L_p(\mathbb{R}^d)$.

(b) f and f_A have same Besov semi-norm up to a constant.

Hence f and f_A belong to the same (inhomogeneous) Besov space B_{spq}.

proof

For (a), with $p \geq 1$, as in Prop. 3.2 above, one has $\|f_A^{\star\ell}\|_p \leq \|f_A\|_p$. Also, with the determinant of A denoted by $|A|$,

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

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), with a change of variable in the integral one has,

$$\|\Delta_{th} f_A\|_p = |A|^{-1+\frac{1}{p}} \, \|\Delta_{tA^{-1}h} f\|_p \,;$$

so that

$$\omega_h(f_A, \delta)_p = \sup_{|t|\leq\delta, |h|=1} \|\Delta_{th} f_A\|_p = |A|^{-1+\frac{1}{p}} \sup_{|t|\leq\delta|A^{-1}h|, |h|=1} \|\Delta_{th} f\|_p = \omega_l(f, \delta|A^{-1}h|)_p, \quad |h| = 1;$$

and

$$\Omega_{\mathbb{R}^d}(f_A,\delta)_p = |A|^{-1+\frac{1}{p}}\Omega_{\mathbb{R}^d}(f,\delta|A^{-1}h|)_p, \quad |h|=1.$$

Next, with the change of variable $u = t|A^{-1}h|$,

$$\begin{aligned}\int_0^\infty |t^{-\alpha}\Omega(f_A,t)|^q \frac{dt}{t} &= (|A|^{-1+\frac{1}{p}}\,|A^{-1}h|^\alpha)^q \int_0^\infty |u^{-\alpha}\Omega(f,u)|^q \frac{du}{u}, \quad |h|=1\\ &\le (|A|^{-1+\frac{1}{p}}\,\|A^{-1}\|^\alpha)^q \int_0^\infty |u^{-\alpha}\Omega(f,u)|^q \frac{du}{u}\,.\end{aligned}$$

In view of (3), using the admissible pair $(k,\varrho) = ([s]+1,[s])$ yields the desired result when $0<s<1$.

When $1 \le s$, with the same admissible pair $(k,\varrho) = ([s]+1,[s])$, and by recurrence, since $df_A(h) = |A^{-1}|\, df(A^{-1}h)\circ A^{-1}$ one can see in the same way that the modulus of continuity of the (generalized) derivatives of f_A or order k are bounded by those of f.

Note that if A is whitened, in the context of ICA, the norms of f and f_A are equal, at least when $s<1$. □

4. Risk upper bound

Define the experiment $\mathcal{E}^n = (\mathcal{X}^{\otimes n}, \mathcal{A}^{\otimes n}, (X_1,\dots,X_n), P^n_{f_A}, f_A \in B_{spq})$, 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 α_{jk} in the wavelet contrast are estimated as usual by,

$$\hat\alpha_{jk^1\dots,k^d} = \frac{1}{n}\sum_{i=1}^n \varphi_{jk^1}(X_i^1)\dots\varphi_{jk^d}(X_i^d) \text{ and } \hat\alpha_{jk^\ell} = \frac{1}{n}\sum_{i=1}^n \varphi_{jk^\ell}(X_i^\ell),\ \ \ell=1,\dots d. \tag{7}$$

The linear wavelet contrast estimator is given by,

$$\hat C_j(x_1\dots,x_n) = \sum_{k^1\dots,k^d}(\hat\alpha_{jk^1\dots,k^d} - \hat\alpha_{jk^1}\dots\hat\alpha_{jk^d})^2 = \sum_{k\in\mathbb{Z}^d}\hat\delta^2_{jk}, \tag{8}$$

where we define $\hat\delta_{jk}$ as the difference $\hat\alpha_{jk^1\dots,k^d} - \hat\alpha_{jk^1}\dots\hat\alpha_{jk^d}$.

The estimator $\hat\alpha_{jk}$is unbiased under $E^n_{f_A}$, but so is not $\hat\alpha_{jk^1}\dots\hat\alpha_{jk^d}$ unless $A=I$, although it is asymptotically unbiased.

We also make the assumption that both the density and the wavelet are compactly supported so that all sums in k are finite. For simplicity we further suppose the density support to be the hypercube, so that $\sum_k 1 \approx 2^{jd}$.

To bound the risk we use a single lemma whose proof relies on a classical U-statistic lemma, namely the connection between a U-statistic and its associated Von Mises statistic. To fit our purpose the U-statistic lemma first needed to be adapted to kernels that are (generally unsymmetric) products of Φ_{jk} and $\varphi_{jk}\circ\pi^\ell$, and thus depend on the resolution parameter j. This is done in lemmas 7.2 appearing in the Appendix.

Lemma 4.1 (Moments of wavelet coefficients estimators)

Let $\rho, \sigma \geq 0$ be fixed integers; the following relations hold:

$$E^n_{f_A} \hat{\alpha}^\rho_{jk} (\hat{\alpha}_{jk^1} \dots \hat{\alpha}_{jk^d})^\sigma = \alpha^\rho_{jk} (\alpha_{jk^1} \dots \alpha_{jk^d})^\sigma + O(n^{-1}).$$

And in corollary, $E^n_{f_A} \hat{\delta}^\rho_{jk} = \delta^\rho_{jk} + O(n^{-1})$.

proof

To the statistic $V_{nj} = \hat{\alpha}^\rho_{jk} (\hat{\alpha}_{jk^1} \dots \hat{\alpha}_{jk^d})^\sigma$ corresponds a U-statistic U_{nj} with unsymmetric kernel

$$h_j(x_1, \dots, x_{\rho+d\sigma}) = \Phi_{jk}(x_1) \dots \Phi_{jk}(x_\rho) \varphi_{jk^{\ell_1}} \circ \pi^{\ell_1}(x_{\rho+1}) \dots \varphi_{jk^{\ell_{d\sigma}}} \circ \pi^{\ell_{d\sigma}}(x_{\rho+d\sigma}),$$

with π^ℓ the canonical projection on component ℓ, $(\ell_{i\sigma+1}, \dots, \ell_{(i+1)\sigma}) = (1, \dots, d)$, $i = 0 \dots d-1$ and $\Phi(x) = \prod_{\ell=1}^d \varphi \circ \pi^\ell(x)$.

By application of lemma 7.2 in Appendix,

$$|E^n_{f_A} \hat{\alpha}^\rho_{jk} (\hat{\alpha}_{jk^1} \dots \hat{\alpha}_{jk^d})^\sigma - \alpha^\rho (\alpha_{jk^1} \dots \alpha_{jk^d})^\sigma| \leq E^n_{f_A} |V_{nj} - U_{nj}| = O(n^{-1})$$

□

We now express a risk bound for the wavelet contrast estimator. In particular we show that the bias of the estimator is of the order of $C2^{jd}/n$. This is better than the convergence rate of $n^{\frac{-1}{2}}$ for the empirical Hilbert-Schmidt independence criterion (Gretton et al. 2004, theorem 3), except that in our case the resolution parameter j must still be set to some value, especially to cope with the antagonist objectives of reducing the estimator bias and variance.

Proposition 4.4 (Risk upper bound for $\hat{C}_j$)

The quadratic risk $E^n_{f_A} (\hat{C}_j - C_j)^2$ and the bias $E^n_{f_A} \hat{C}_j - C_j$ have convergence rate $2^{jd} O(1/n)$.

In corollary, the variance $E^n_{f_A} \left(\hat{C}_j - E^n_{f_A} \hat{C}_j \right)^2$ has convergence rate $2^{jd} O(1/n)$ and the quadratic risk around zero is $E^n_{f_A} \hat{C}^2_j = C^2_j + 2^{jd} O(1/n)$.

proof

The risk about C_j is written,

$$E^n_{f_A} \left(\hat{C}_j - C_j \right)^2 = E^n_{f_A} \sum_{k,\ell} (\hat{\delta}^2_{jk} - \delta^2_{jk})(\hat{\delta}^2_{j\ell} - \delta^2_{j\ell}) \tag{9}$$

For the squared terms where $k = \ell$, lemma 4.1 yields directly $E^n_{f_A} (\hat{\delta}^2_{jk} - \delta^2_{jk})^2 = O(n^{-1})$, so that the corresponding risk component is bounded by $C2^{jd} n^{-1}$.

For crossed terms where $k \neq \ell$, observe that with a compactly supported Daubechies Wavelet $D2N$, whose support is $[0, 2N-1]$, only a thin band of terms around the diagonal is non zero:

$$\varphi_{jk}\varphi_{j\ell} = 0, \qquad \text{for } |\ell - k| \geq 2N-1.$$

When k is fixed, the cardinal of the set $|\ell - k| \leq 2N-1$ is lower than $(4N)^d$; hence, by Cauchy-Schwarz inequality and lemma 4.1,

$$\begin{aligned} E^n_{f_A} \sum_{k,\ell} (\hat\delta^2_{jk} - \delta^2_{jk})(\hat\delta^2_{j\ell} - \delta^2_{j\ell}) &\leq \sum_k E^n_{f_A}{}^{\frac{1}{2}} (\hat\delta^2_{jk} - \delta^2_{jk})^2 \sum_{|\ell-k|\leq 2N-1} [E^n_{f_A} (\hat\delta^2_{j\ell} - \delta^2_{j\ell})^2]^{\frac{1}{2}} \\ &\leq 2^{jd} C n^{-1/2}\, (4N)^d C n^{-1/2} = C 2^{jd} n^{-1}. \end{aligned}$$

The bias convergence rate is a direct consequence of lemma 4.1, and the two remaining assertions follow from the usual relations, $E^n_{f_A}(\hat C_j - C_j)^2 = E^n_{f_A}(\hat C_j - E^n_{f_A}\hat C_j)^2 + (E^n_{f_A}\hat C_j - C_j)^2$; and $E^n_{f_A}\hat C_j^2 = (E^n_{f_A}\hat C_j)^2 + E^n_{f_A}(\hat C_j - E^n_{f_A}\hat C_j)^2$.
□

We now give a rule for choosing the resolution j minimizing the (about zero) risk upper bound. This rule, obtained as usual through bias-variance balancing, depends on s, the unknown regularity of f, supposed to be a member of some Besov space B_{spq}. The associated convergence rate improves upon the minimax $n^{\frac{-2s}{2s+d}}$ of the underlying density estimations (see Kerkyacharian & Picard, 1992).

Proposition 4.5 (minimizing resolution in the class $B_{s2\infty}$)

Assume that f belongs to $B_{s2\infty}(\mathbb{R}^d)$, and that C_j is based on a r-regular wavelet φ, $r > s$.

The minimizing resolution j is such that $2^j \approx n^{\frac{1}{4s+d}}$ and ensures a quadratic risk converging to zero at rate $n^{-\frac{4s}{4s+d}}$.

proof

By Prop. 3.2 and 3.3 we know that f_A and the product of its marginal distributions belong to the same Besov space than the original f, so that equation (6) becomes

$$|K_\star(A, f) - C_j(f_A)^{\frac{1}{2}}| \leq K 2^{-js}; \tag{10}$$

with K a constant.

Taking power 4 of (10) and using prop. 4.4,

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

with K a placeholder for an unspecified constant, $Q(a, b) = -4a^3 + 6a^2 b - 4ab^2 + b^3$, and R denoting the quadratic risk around zero.

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 $2^{jd} K n^{-1}$.

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

This yields $2^j = O(n^{\frac{1}{4s+d}})$ and convergence rate $n^{\frac{-4s}{4s+d}}$ for the risk relative to zero under independence and also for the risk relative to C_j by substitution in the expression given by Prop. 4.4.
□

Corollary 4.1 (minimizing resolution in the class B_{spq})

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

The minimizing resolution j is such that $2^j \approx n^{\frac{1}{4s'+d}}$, with $s' = s + d/2 - d/p$ if $1 \le p \le 2$ and $s' = s$ if $p > 2$.

This resolution ensures a quadratic risk converging to zero at rate $n^{-\frac{4s'}{4s'+d}}$.

proof

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, with $\{\epsilon_j\} \in \ell_q$,

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

If $p > 2$, since we consider compactly supported densities, with $\{\epsilon_j\} \in \ell_q$,

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

Finally with s' as claimed, equation (6) yields again$|K_\star(A, f) - C_j(f_A)^{\frac{1}{2}}| \le K 2^{-js'}$. □

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

The estimator is computable by means of any Daubechies wavelet, including the Haar wavelet.

For a regular wavelet ($D2N, N > 1$), it is known how to compute the values $\varphi_{jk}(x)$ (and any derivative) at dyadic rational numbers (Nguyen and Strang, 1996); this is the approach we have adopted in this paper.

Alternatively, using the customary 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 appropriate 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 usually gives full satisfaction in density estimation, in the ICA context, without special care, it can lead to an inflation of computational resources, or a possibly inoperative contrast at minimization stage. Indeed, for the Haar contrast to show any variation in response to 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 have then chosen the approach with dyadics for simplicity at the minimization stage and possibly more accurate solutions.

Also for simplicity, in all simulations that follow we have adopted the convention that the whole signal is contained in the hypercube $[0,1]^{\otimes d}$, 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 $(2^j+2N-2)^d$.

Note that relocation in the unit hypercube is not a requirement, but otherwise a sparse array implementation should be used for efficiency.

Sensitivity of the wavelet contrast

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 4.5, 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 mostly 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 passes an end of the observation support. At small j such that $2 \le 2^j \le 2N-1$, 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 the limits.

j	D2 indep	D2 mixed	cpu	j	D4 indep	D4 mixed	cpu
0	0.000E+00	0.000E+00	0.12	0	0.250E+00	0.250E+00	0.21
1	0.184E-06	0.102E-10	0.06	1*	0.239E+00	0.522E+00	0.17
2	0.872E-04	0.199E-04	0.06	2	0.198E-04	0.209E-04	0.17
3	0.585E-03	0.294E-03	0.06	3	0.127E-03	0.159E-03	0.17
4	0.245E-02	0.285E-02	0.06	4	0.635E-03	0.714E-03	0.17
5*	0.926E-02	0.110E-01	0.07	5	0.235E-02	0.282E-02	0.17
6	0.395E-01	0.387E-01	0.07	6	0.988E-02	0.105E-01	0.17
7	0.162E+00	0.162E+00	0.07	7	0.405E-01	0.419E-01	0.17
8	0.651E+00	0.661E+00	0.08	8	0.163E+00	0.165E+00	0.21
9	0.262E+01	0.262E+01	0.12	9	0.653E+00	0.653E+00	0.26
10	0.105E+02	0.105E+02	0.23	10	0.261E+01	0.262E+01	0.39
11	0.419E+02	0.419E+02	0.69	11	0.104E+02	0.105E+02	0.87
12	0.168E+03	0.168E+03	2.48	12	0.419E+02	0.420E+02	2.67

Table 1a. Wavelet contrast values for a D2 and a D4 on a uniform density in dimension 2 under a half degree rotation Amari error $\approx .8$, nobs=100000, L=10,

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. Wavelet contrast values for a D6 and a D8 on a uniform density in dimension 2 under a half degree rotation Amari error $\approx .8$, nobs=100000, L=10,

First, 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, as for the other wavelets, any detection becomes impossible because the ratio $2^{jd}/n$ gets too big, and clearly wanders from the optimal rule of Prop. 4.5.

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 process.

For wavelets D4 and above, the contrast is able to capture the mixing effect especially at low resolution (resolution with largest relative increase 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. Wavelet contrast values on a uniform density, dim=4 , nobs=50000, L=10, Amari error $\approx .5$

Computation uses 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, added to the wavelet contrast computation 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 of n observations on the tensorial space V_j^d and the d margins for a Db(2N) has complexity $O(n(2N-1)^d)$. This is $O(n)$ for a Haar wavelet (2N=2) which boils down to making a histogram. The projection complexity 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 has complexity $O(2^{jd}(2N-1)^d)$. 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 carries with it a lack of sensitivity to small perturbations, which is a problem for empirical gradient evaluation or the detection of a small departure from independence.

For comparison, the Hilbert-Schmidt independence criterion is theoretically computed in $O(n^2)$ (Gretton et al. 2004 section 3), and the Kernel ICA criterion is theoretically computed in $O(n^3 d^3)$. In both cases, using incomplete Choleski decomposition and low-rank approximation of the Gram matrix, the complexity is brought down in practice to $O(nd^2)$ for HSIC and $O(n^2\, log\, n)$ for Kernel ICA(Bach and Jordan 2002 p.19).

6. 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 carry out 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, we can restrict to $SO(d) \subset O(d)$ thus ignoring reflections that are not relevant.

Needed material for minimization on the Stiefel manifold can be found in the paper of Arias et al. (1998). Another very close method uses 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 ∇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, ∇J is computed by first differences; in doing so we cannot keep perturbed W orthogonality, and we actually compute a plain gradient in $\mathbb{R}^{dd}$.

Again, a Haar contrast empirical gradient is tricky to obtain, since a small perturbation in W will likely result in an unchanged histogram at small j, whereas with D4 and above contrasts, 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; `start` columns indicate Amari distance (on the scale 0 to 100) and wavelet contrast on entry; `it` column 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 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 are comparable with the performance of existing ICA algorithms, as presented for instance in the paper of Jordan and Bach (2002) p.30 (average Amari error between 3 and 10 for 2 sources and 1000 observations) or Gretton et al (2004) table 2 (average Amari error between 2 and 6 for 2 sources and 1000 observations).

Finally we give other runs on the example of the uniform density at resolution $j = 2$ under different parameters settings, and relatively fewer number of observations.

obs.	dim	L	Amari start	Amari end	cont. start	cont. end	it.
250	2	10	47.387	38.919	0.279E-01	0.193E-01	2.4
250	2	13	47.387	32.470	0.279E-01	0.170E-01	2.2
250	2	16	47.387	17.915	0.279E-01	0.603E-02	2.3
250	2	19	47.387	19.049	0.279E-01	0.598E-02	2.6
500	2	10	51.097	20.700	0.246E-01	0.106E-01	2.1
500	2	13	51.097	6.644	0.246E-01	0.398E-02	2.2
500	2	16	51.097	21.063	0.246E-01	0.109E-01	2.1
500	2	19	51.097	14.734	0.246E-01	0.839E-02	2.4
1000	2	10	41.064	3.533	0.167E-01	0.186E-02	2.3
1000	2	13	41.064	3.071	0.167E-01	0.190E-02	2.1
1000	2	16	41.064	3.518	0.167E-01	0.194E-02	1.9
1000	3	16	49.607	15.082	0.405E-01	0.127E-01	4.8
5000	3	10	49.575	5.405	0.390E-01	0.399E-02	4.5
5000	3	16	49.575	1.668	0.390E-01	0.960E-03	4.7
5000	4	10	43.004	17.036	0.561E-01	0.190E-01	4.4
5000	5	10	38.400	29.679	0.800E-01	0.559E-01	4.1
5000	5	16	38.400	4.233	0.798E-01	0.700E-02	5.0
5000	6	16	42.529	10.841	0.114E+00	0.278E-01	4.9
5000	7	16	41.128	15.761	0.188E+00	0.573E-01	5.0
5000	8	16	39.883	14.137	0.286E+00	0.743E-01	5.0

Table 5. Average results of 10 runs, j=2, with a D4, truncated at 5 iterations.

One can see that raising the dyadic approximation parameter L tends to improve the minimization when the number of observations is "low" relatively to the number or cells 2^{jd}, but that 500 observations in dimension 2 seems to be a minimum in the current state of the program. In higher dimensions, a higher number of observations is required, and in dimension 6 and above, 5000 is not enough at L=16.

A visual example in dimension 2

In dimension 2, we are exempted from any added complication brought by a gradient descent and Stiefel minimization. After whitening, the inverse of A is an orthogonal matrix, whose membership can be restricted to $SO(2)$, ignoring reflections. So there is only one parameter θ 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), functions of the rotation angle of the matrix W restricted to one period, $[0, \pi/2]$. The minimums are not necessarily at a zero angle, for precisely, mere whitening leaves the signal in a random rotated position (to reproduce the following results run the program `icalette2`).

We see that, provided Amari error and wavelet contrast have coinciding minima, any line search algorithm will find the angle to reverse the mixing effect. We see also in Fig.2 that the Haar wavelet contrast is perfectly suitable to detect independence, so that minimization methods not gradient based might work very well in this case.

On the example of the uniform density (Fig.3) we have an illustration of a non smooth contrast variation typical of a too high resolution j given regularity and number of observations.

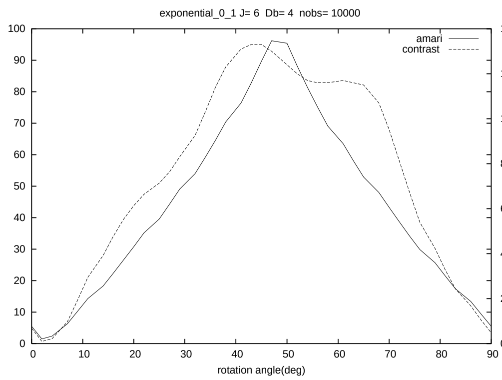


Fig.1. Exponential, D4, j=6, n=10000

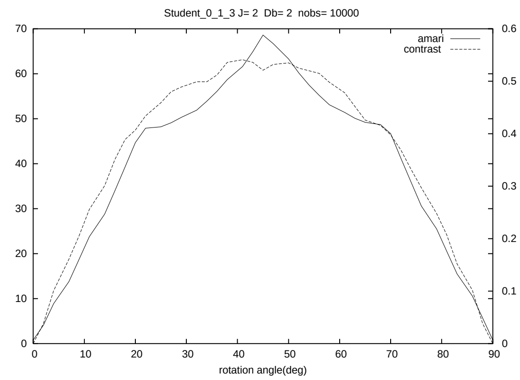


Fig.2. Student, D2, j=2, n=10000

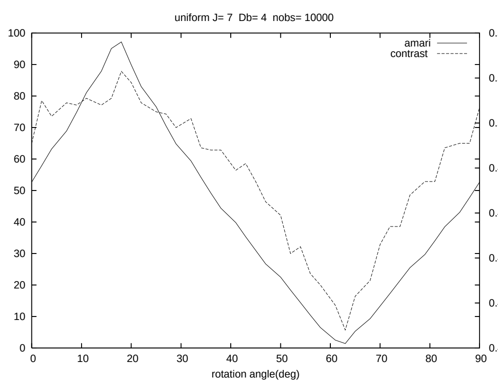


Fig.3. Uniform, D4, j=7, n=10000

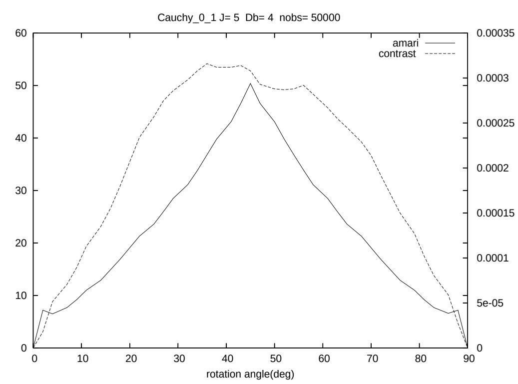


Fig.4. Cauchy, D4, j=5, n=50000

7. Appendix

Lemma 7.2 (Approximation by the j-dependent U-statistic)

Let $(X_1 \dots, X_n)$ be an i.i.d. sample of a random variable on $\mathbb{R}^d$ and let ρ, σ, m be positive integers verifying $\rho + \sigma = m$. Let h_j be a (possibly unsymmetric) kernel function on $\mathbb{R}^{md}$ defined as

$$h_j(x_1, \dots, x_m) = \Phi_{jk}(x_1) \dots \Phi_{jk}(x_\rho) \varphi_{jk^{\ell_1}} \circ \pi^{\ell_1}(x_{\rho+1}) \dots \varphi_{jk^{\ell_\sigma}} \circ \pi^{\ell_\sigma}(x_{\rho+\sigma}),$$

with π^ℓ the canonical projection on component ℓ and $\Phi(x) = \prod_{\ell=1}^d \varphi \circ \pi^\ell(x)$.

Consider the associated U-statistic and Von Mises V statistic,

$$U_{nj} = \frac{(n-m)!}{n!} \sum_{i_1 \neq \dots \neq i_m} h_j(X_{i_1}, \dots, X_{i_m}), \qquad V_{nj} = \frac{1}{n^m} \sum_{i_1, \dots, i_m} h_j(X_{i_1}, \dots, X_{i_m}).$$

The following relation holds:

$$E|U_{nj} - V_{nj}|^r = \left(\frac{2^{j(d\rho+\sigma)}}{n}\right)^{\frac{r}{2}-1} O(n^{-1-\frac{r}{2}}). \tag{11}$$

In corollary, $E|U_{nj} - V_{nj}|^r = O(n^{-r})$ *for* $r = 1, 2$ *and* $E|U_{nj} - V_{nj}|^r = O(n^{-1-\frac{r}{2}})$ *for* $r \geq 2$ *and* $2^{j(d\rho+\sigma)}/n < 1$.

proof

The first lines use the proof of the original lemma (see for example Serfling, 1980), with special care for unsymmetric kernels.

Let W_{nj} be the average of all terms $h_j(X_{i_1}, \dots, X_{i_m})$ with at least one equality $i_a = i_b$ for $a \neq b$ and $1 \leq a, b \leq m$; there are $n^m - A_n^m$ such terms.

Ω denoting the set of n^m unconstrained indexes, one has the relation

$$\begin{aligned}(n^m - A_n^m)(W_{nj} - U_{nj}) &= \sum_{\Omega - \{i_1 \neq \dots \neq i_m\}} h(X_{i_1}, \dots, X_{i_m}) - n^m U_{nj} + \sum_{\{i_1 \neq \dots \neq i_m\}} h(X_{i_1}, \dots, X_{i_m}) \\ &= n^m (V_{nj} - U_{nj})\end{aligned}$$

Hence, using Minkowski inequality and the fact that $(n^m - A_n^m) = O(n^{m-1})$ is positive, one obtains,

$$\begin{aligned}E^n_{f_A}|U_{nj} - V_{nj}|^r &= n^{-mr}(n^m - A_n^m)^r \, E^n_{f_A}|U_{nj} - W_{nj}|^r \\ &\leq n^{-mr}(n^m - A_n^m)^r \left(E^n_{f_A}|U_{nj}|^r + E^n_{f_A}|W_{nj}|^r\right) \\ &\leq O(n^{-r}) \left(E^n_{f_A}|U_{nj}|^r + E^n_{f_A}|W_{nj}|^r\right).\end{aligned}$$

It remains to bound to right parenthesis.

Using Minkowski inequality, one has,

$$\begin{aligned}E^n_{f_A}{}^{\frac{1}{r}}|U_n|^r &\leq [A_n^m]^{-1} \sum_{i_1 \neq \dots \neq i_m} E^n_{f_A}{}^{\frac{1}{r}}|\Phi_{jk}(X_{i_1}) \dots \Phi_{jk}(X_{i_\rho}) \varphi_{jk^{\ell_1}}(X^{\ell_1}_{i_{\rho+1}}) \dots \varphi_{jk^{\ell_{i_\sigma}}}(X^{\ell_{i_\sigma}}_{i_m})|^r \\ &= E^n_{f_A}{}^{\frac{1}{r}}|\varphi_{jk^{\ell_1}}(X_1^{\ell_1})|^r \dots E^n_{f_A}{}^{\frac{1}{r}}|\varphi_{jk^{\ell_\sigma}}(X_1^{\ell_{i_\sigma}})|^r E^n_{f_A}{}^{\frac{\rho}{r}}|\Phi_{jk}(X_1)|^r\end{aligned}$$

Next $E^n_{f_A}|\varphi_{jk^\ell}(X_1^\ell)|^r = 2^{jr/2}\int|\varphi(2^j x - k^\ell)|^r f_A^{\star\ell}(x)dx \leq 2^{j(\frac{r}{2}-1)}\|f_A^{\star\ell}\|_\infty \|\varphi\|_r^r$ and by the same means, $E^n_{f_A}|\Phi_{jk}(X_i)|^r \leq 2^{jd(\frac{r}{2}-1)}\|f_A\|_\infty \|\Phi\|_r^r$

So that,

$$\begin{aligned}E^n_{f_A}|U_n|^r &\leq 2^{j\sigma(\frac{r}{2}-1)}\left(\prod_{i=1,\dots,\sigma}\|f_A^{\star\ell_i}\|_\infty\right)\|\varphi\|_r^{r\sigma} \quad 2^{jd\rho(\frac{r}{2}-1)}\|f_A\|_\infty^\rho \|\Phi\|_r^{r\rho} \\ &= C 2^{j(\sigma+d\rho)(\frac{r}{2}-1)}\end{aligned}$$

Likewise for W_n one obtains terms of the type $E^n_{f_A}\varphi_{jk^{\ell_1}}(X_1^{\ell_1}) \dots \varphi_{jk^{\ell_\kappa}}(X_1^{\ell_\kappa})$ of which $E^n_{f_A}\Phi(X_1)$ is one particular form, which are bounded exactly in the same way, and produce the same power of 2^j. □

I am mostly grateful to my advisor, Dominique Picard, for many suggestions in the writing of this paper.

References

(Achard et al. 2003) Christian Jutten S. Achard, D.T.Pham. A quadratic dependence measure for nonlinear blind sources separation. *Proceeding of ICA 2003 Conference*, pages 263–268, 2003.

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

(Arias et al. 1998) Steven T. Smith Alan Edelman, Tomas Arias. *The geometry of algorithms with orthogonality constraints*. SIAM, 1998.

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

(Bell & Sejnowski, 1995) A. J. Bell. T.J. Sejnowski A non linear information maximization algorithm that performs blind separation. *Advances in neural information processing systems*, 1995.

(Bergh & Löfström, 1976) J. Bergh and J. Löström. *Interpolation spaces*. Springer, Berlin, 1976.

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

(Comon, 1994) P. Comon. Independent component analysis, a new concept ? *Signal processing*, 1994.

(Daubechies, 1992) Ingrid Daubechies. *Ten lectures on wavelets*. SIAM, 1992.

(Devore & Lorentz, 1993) R. Devore, G. Lorentz. *Constructive approximation*. Springer-Verlag, 1993.

(Donoho et al., 1996) G. Kerkyacharian D.L. Donoho, I.M. Johnstone and D. Picard. Density estimation by wavelet thresholding. *Annals of statistics*, 1996.

(Gretton et al. 2003) Alex Smola Arthur Gretton, Ralf Herbrich. The kernel mutual information. Technical report, Max Planck Institute for Biological Cybernetics, April 2003.

(Gretton et al. 2004) A. Gretton, O. Bousquet, A. Smola, B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. Technical report, Max Planck Institute for Biological Cybernetics, October 2004.

(Härdle et al., 1998) Wolfgang Härdle, Gérard Kerkyacharian, Dominique Picard and Alexander Tsybakov. *Wavelets, approximation and statistical applications*. Springer, 1998.

(Hyvarïnen et al. 2001) A. Hyvarïnen, J. Karhunen. E. Oja *Independent component analysis*. Inter Wiley Science, 2001.

(Hyvarinen & Oja, 1997) A. Hyvarinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural computation*, 1997.

(Kerkyacharian & Picard, 1992) Gérard Kerkyacharian Dominique Picard. Density estimation in Besov spaces. *Statistics and Probability Letters*, 13:15–24, 1992.

(Meyer, 1997) Yves Meyer. *Ondelettes et opérateurs*. Hermann, 1997.

(Nguyen Strang, 1996) Truong Nguyen Gilbert Strang. *Wavelets and filter banks.* Wellesley-Cambridge Press, 1996.

(Nikol'skiĭ, 1975) S.M. Nikol'skiĭ. Approximation of functions of several variables and imbedding theorems. *Springer Verlag, 1975.*

(Peetre, 1975) Peetre, J. New Thoughts on Besov Spaces. Dept. Mathematics, Duke Univ, 1975.

(Plumbley, 2004) Mark D. Plumbley. Lie group methods for optimization with orthogonality constraints. *Lecture notes in Computer science*, 3195:1245–1252, 2004.

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

(Rosenblatt, 1975) M. Rosenblatt. A quadratic measure of deviation of two-dimensional density estimates and a test for independence. *Annals of Statistics*, 3:1–14, 1975.

(Rosenthal, 1972) Rosenthal, H. P. On the span in lp of sequences of independent random variables. *Israel J. Math.* 8 273303, 1972.

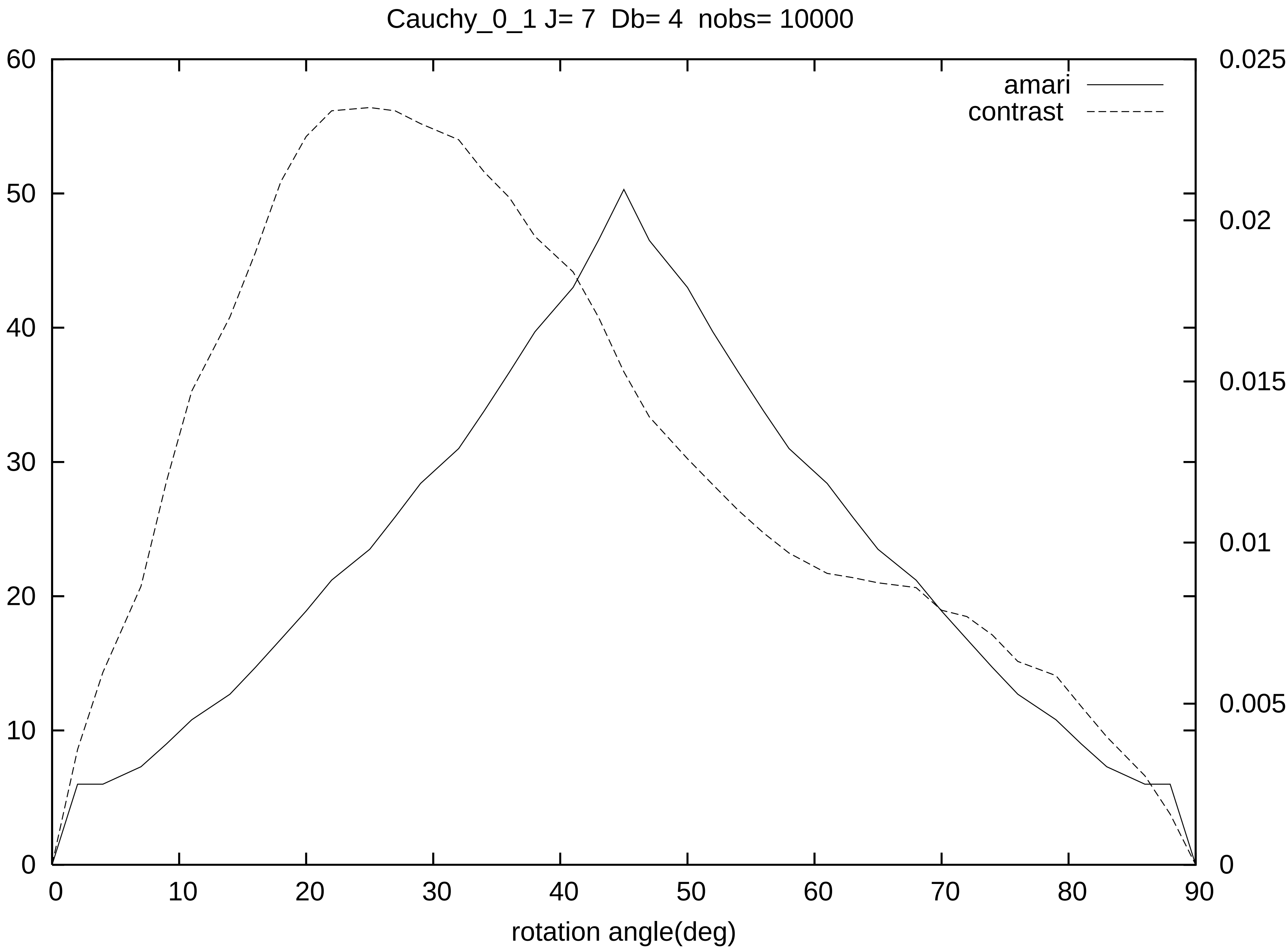
(Serfling, 1980) Robert J. Serfling. *Approximation theorems of mathematical statistics.* Wiley, 1980.

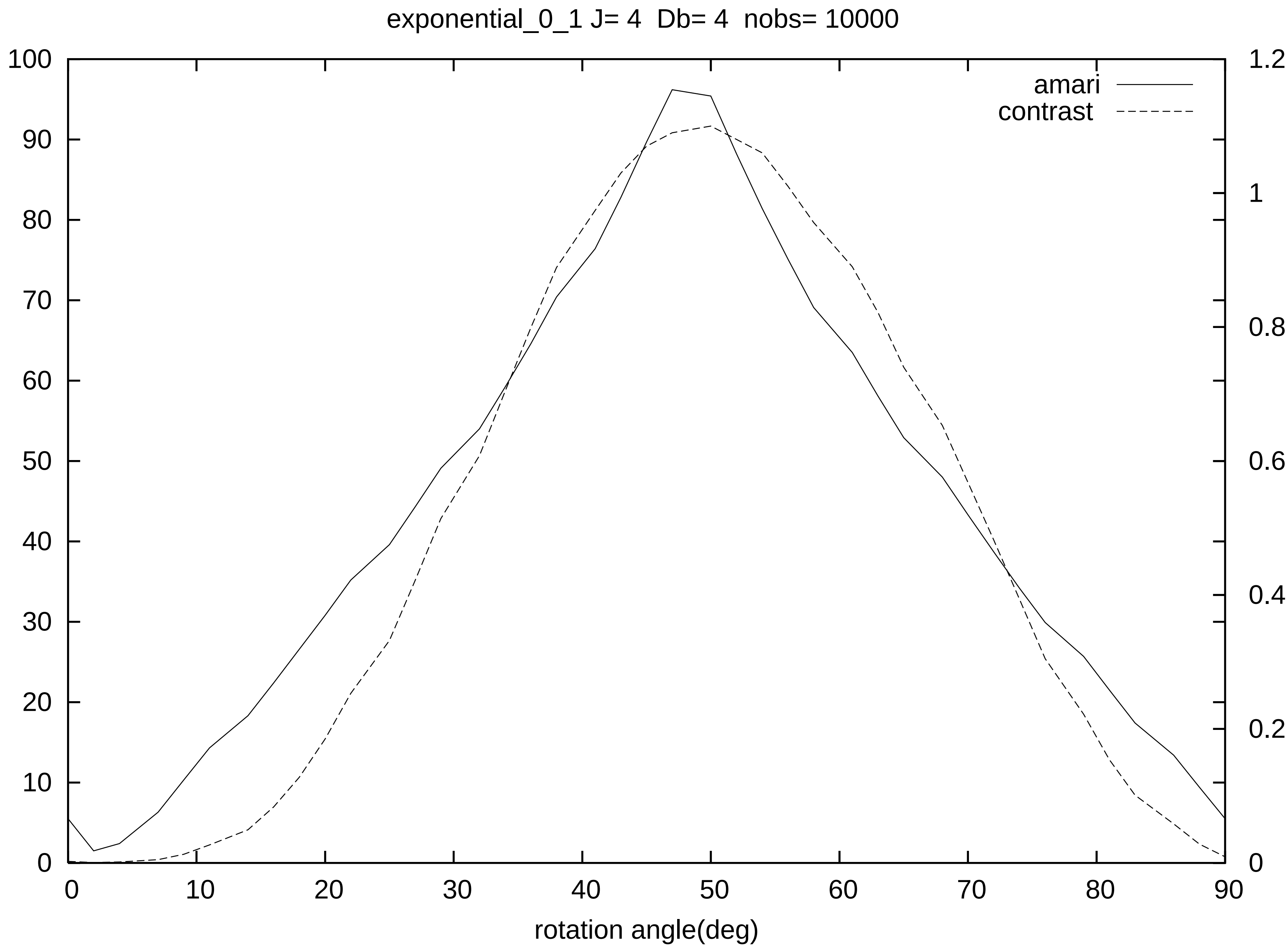
(Sidje, 1998) R. B. Sidje. Expokit. A Software Package for Computing Matrix Exponentials. *ACM Trans. Math. Softw.*, 24(1):130–156, 1998.

(Tsybakov & Samarov, 2004) A. Tsybakov A. Samarov. Nonparametric independent component analysis. *Bernouilli*, 10:565–582, 2004.

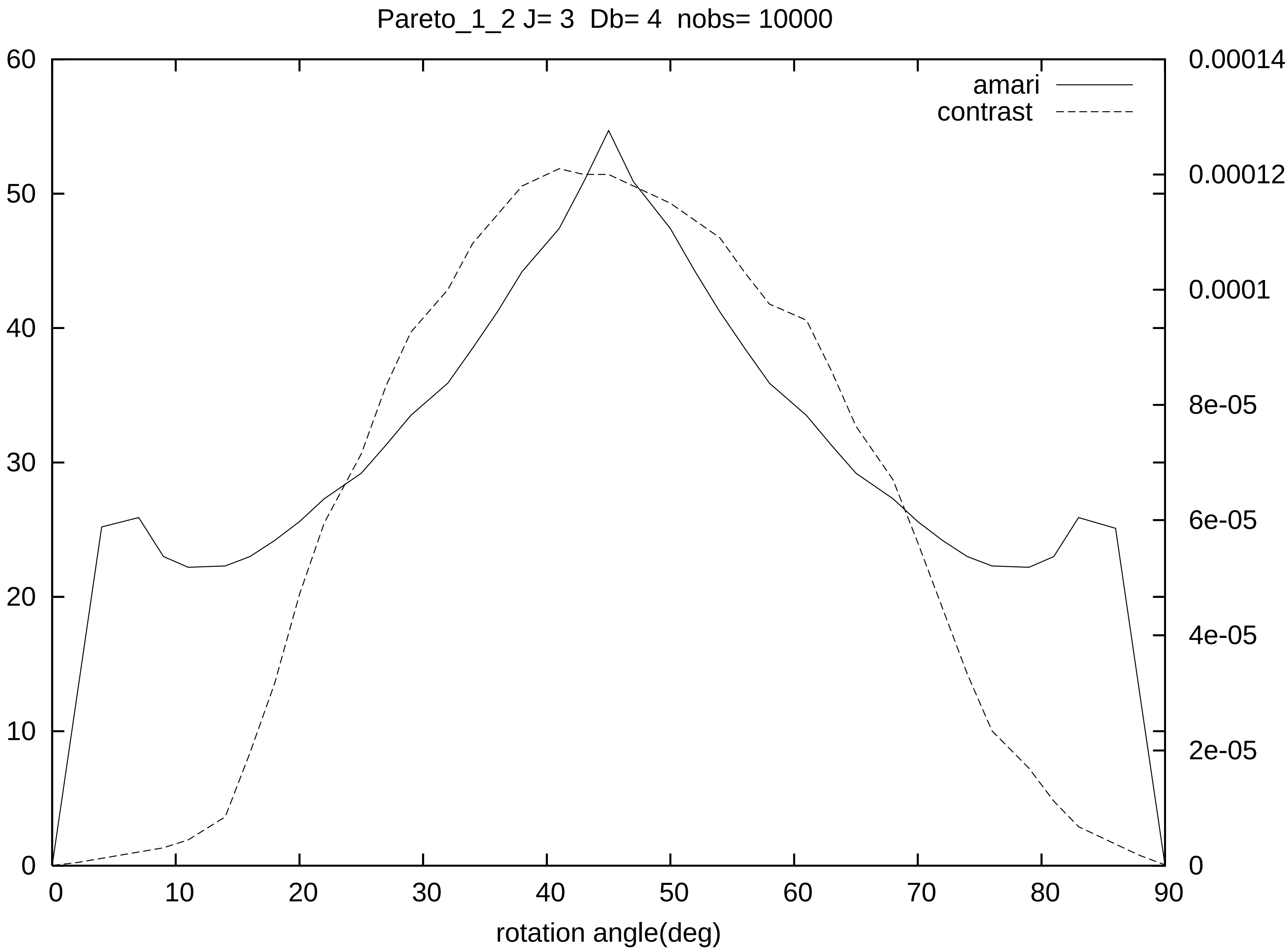
(Triebel, 1992) Triebel, H. Theory of Function Spaces 2. Birkhäuser, Basel, 1992

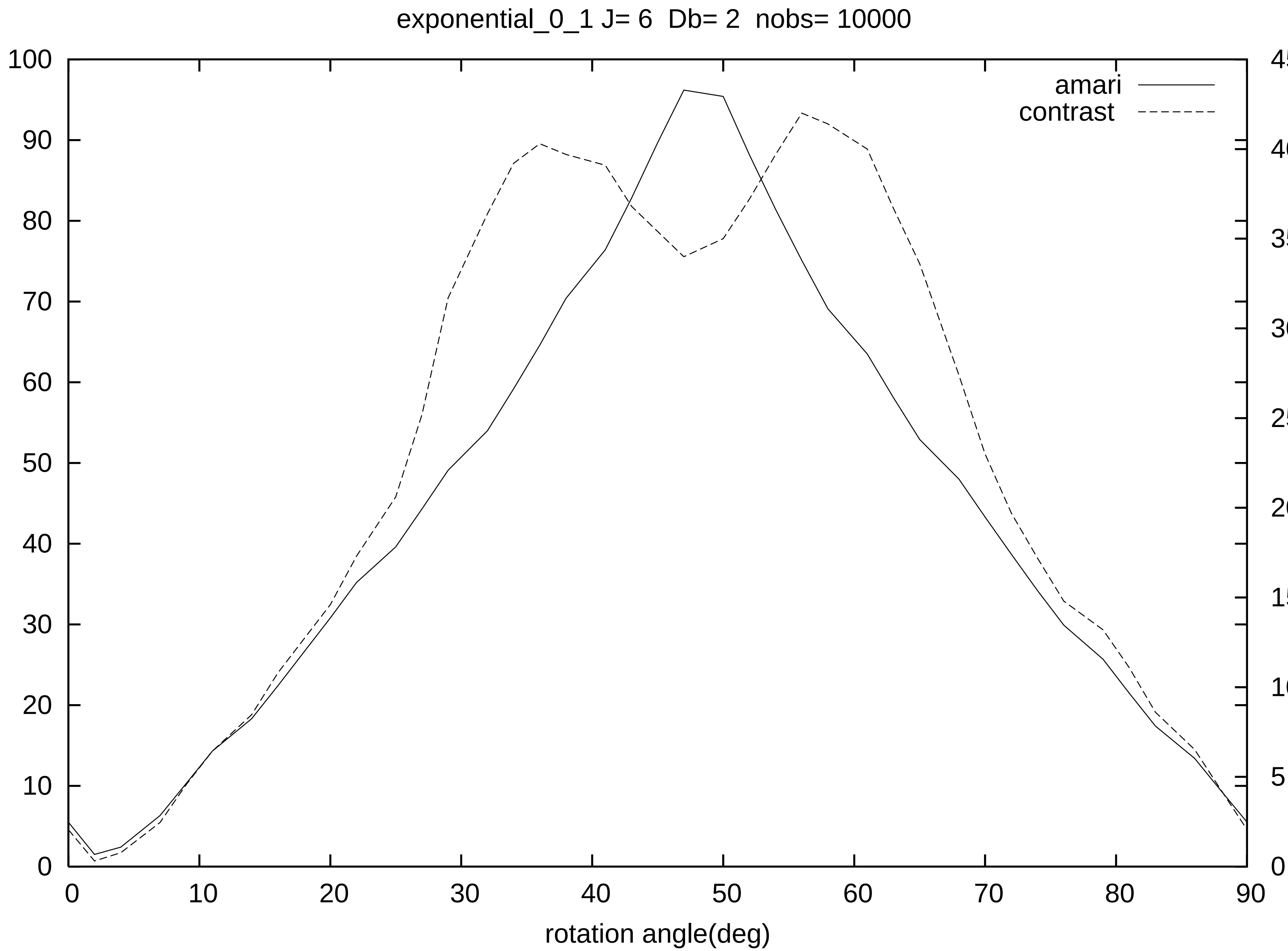
Programs and other runs available at `http://www.proba.jussieu.fr/pageperso/barbedor`


Cauchy_0_1 J= 7 Db= 4 nobs= 10000
amari
contrast
rotation angle(deg)
0
10
20
30
40
50
60
70
80
90
0
10
20
30
40
50
60
0
0.005
0.01
0.015
0.02
0.025

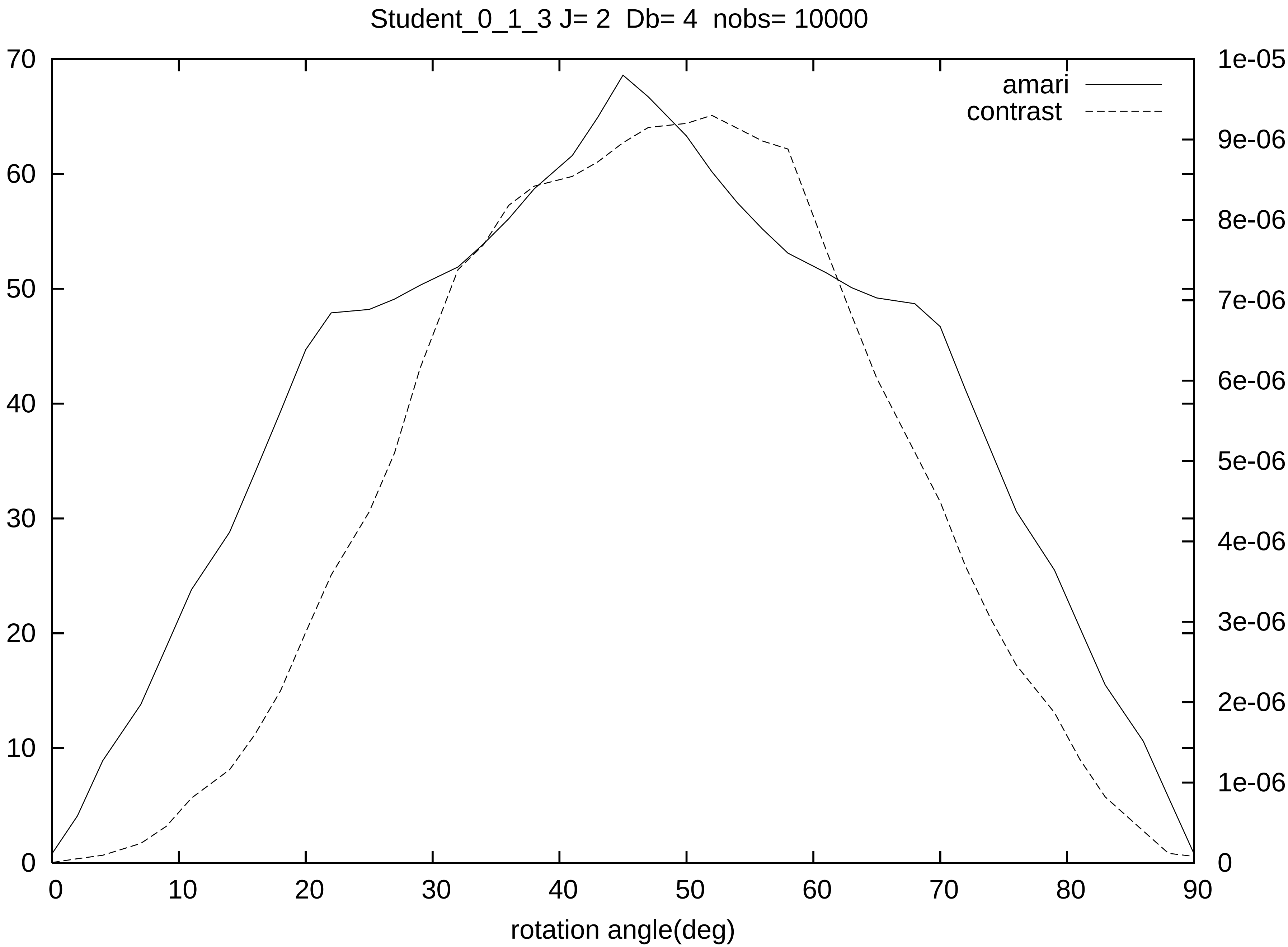


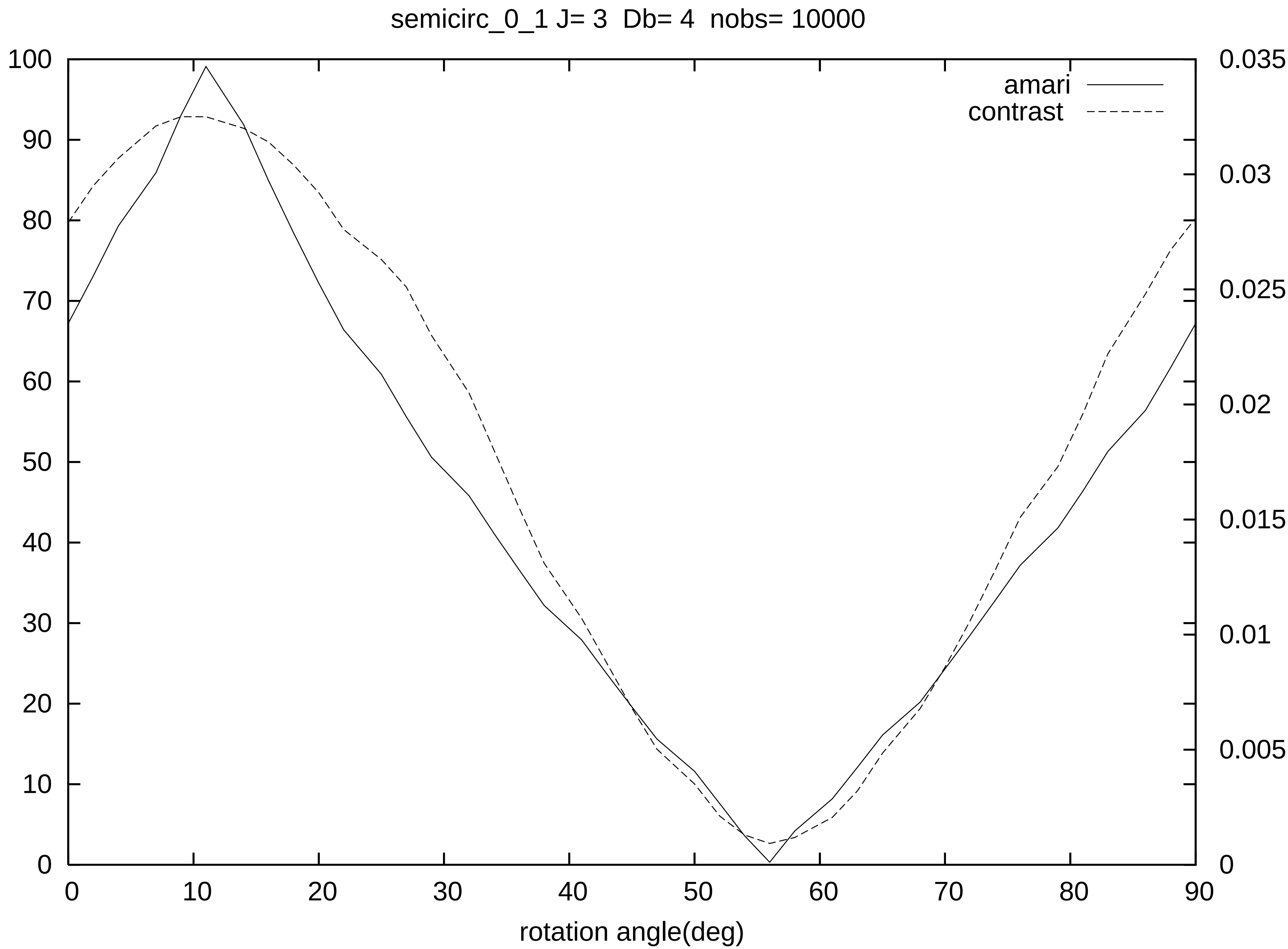
exponential_0_1 J= 4 Db= 4 nobs= 10000
amari
contrast
100
90
80
70
60
50
40
30
20
10
0
1.2
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
70
80
90
rotation angle(deg)


Pareto_1_2 J= 3 Db= 4 nobs= 10000
amari
contrast
rotation angle(deg)
0
10
20
30
40
50
60
70
80
90
0.00014
0.00012
0.0001
8e-05
6e-05
4e-05
2e-05



exponential_0_1 J= 6 Db= 2 nobs= 10000
amari
contrast
0
10
20
30
40
50
60
70
80
90
100
0
5
10
15
20
25
30
35
40
45
0
10
20
30
40
50
60
70
80
90
rotation angle(deg)


Student_0_1_3 J= 2 Db= 4 nobs= 10000
amari
contrast
rotation angle(deg)
0
10
20
30
40
50
60
70
80
90
1e-05
9e-06
8e-06
7e-06
6e-06
5e-06
4e-06
3e-06
2e-06
1e-06



semicirc_0_1 J= 3 Db= 4 nobs= 10000
amari
contrast
rotation angle(deg)
0
10
20
30
40
50
60
70
80
90
100
0.005
0.01
0.015
0.02
0.025
0.03
0.035